Proceedings of the 2021 Winter Simulation Conference
S. Kim, B. Feng, K. Smith, S. Masoud, Z. Zheng, C. Szabo, and M. Loper, eds.

NONPARAMETRIC KULLBACK-LIEBLER DIVERGENCE ESTIMATION USING M-SPACING

Linyun He
Eunhye Song

Department of Industrial and Manufacturing Engineering
Pennsylvania State University
Leonhard Building
University Park, PA 16802, USA

ABSTRACT

Entropy of a random variable with unknown distribution function can be estimated nonparametrically by spacing methods when independent and identically distributed (i.i.d.) observations of the random variable are available. We extend the classical entropy estimator based on sample spacing to define an m-spacing estimator for the Kullback-Liebler (KL) divergence between two i.i.d. observations with unknown distribution functions, which can be applied to measure discrepancy between real-world system output and simulation output as well as between two simulators' outputs. We show that the proposed estimator converges almost surely to the true KL divergence as the numbers of outputs collected from both systems increase under mild conditions and discuss the required choices for m and the simulation output sample size as functions of the real-world sample size. Additionally, we show Central Limit Theorems for the proposed estimator with appropriate scaling.

1 INTRODUCTION

In this paper, we study nonparametric estimation of the Kullback-Liebler (KL) divergence between two continuous-valued random variables whose distribution functions are unknown. Such a problem has relevance in stochastic simulation when the objective is to measure discrepancy between the distribution of the real-world system and simulation outputs when the simulator is built to mimic the system behavior. Similarly, the same measure can be applied to quantify distributional discrepancy between outputs from two different simulators.

Computing the KL divergence requires the probability density functions of two outputs. In the cases of our interest, the distribution functions of the outputs or their parametric forms are typically unknown although their samples may be observed and thus, it is sensible to consider a nonparametric approach. With this motivation, we investigate a sample spacing estimator for the KL divergence constructed from independent and identically distributed (i.i.d.) observations of the two outputs in comparison.

Given a size-n sample of a continuous random variable, the difference between the ith and $(i+m)$th order statistics is referred to as m-spacing for $m \geq 1$. Intuitively, m-spacing provides a measure on how fast the probability distribution changes on the support of the random variable. For fixed m, m-spacing is shorter for the observations near the mode of the distribution than at the tail allowing us to infer the probability density function of the random variable nonparametrically. Thanks to this property, sample spacing has been widely adopted in statistical procedures that require nonparameric density or likelihood estimator, where its application ranges from tests for normality (Vasicek 1976) and uniformity (Dudewicz and van der Meulen 1981; Hall 1986), parameter estimation (Cheng and Amin 1983; Ranneby 1984) and more. A statistic that commonly appears in these work is a spacing estimator for the (differential) entropy of a continuous random variable. First proposed by Tarasenko (1968), several variants of spacing entropy

estimators have been studied. Fundamentally, these estimators take a form of the sample mean of the logarithm of a spacing density estimator evaluated at the data points in the sample. Weak (Vasicek 1976; Hall 1984) and strong (Beirlant and van Zuijlen 1985; van Es 1992) consistency as well as asymptotic normality (van Es 1992) of the spacing entropy estimators are established under various conditions. See Beirlant et al. (1997) for a comprehensive review on nonparametric entropy estimation.

As the focus of the literature moves to estimating the entropy of a high-dimensional random vector, popularity of sample spacing in the entropy (or more generally, information measure) estimation literature seems to have subsided as m-spacing can be defined only for a one-dimensional random variable. Nevertheless, some of more recent high-dimensional estimation approaches can be regarded as extensions of spacing methods. Wang et al. (2006) estimate the KL-divergence of two distributions of d-dimensional random vectors by first estimating their density functions using k-nearest neighbors (k-nn) method in which the volume of the ball including the neighbors is used to measure how fast the probability distribution changes on the support. For $k = 1$, they show their estimator's mean squared error converges to 0 when the sample sizes from both distributions increase to infinity. Wang et al. (2009) extend this work to general k and show strong consistency of their estimator when k grows as a function of the sample size. Moon and Hero (2014b) investigate the convergence rate of the variance and bias of the k-nn method applied to a general f-divergence and devise an ensemble approach to boost the convergence rate. Moon and Hero (2014a) show that the ensemble estimator has asymptotic normality when standardized by its mean and standard deviation. However, such standardization does not provide an explicit form of the asymptotic variance that can be easily computed; Moon and Hero (2014a) adopts bootstrapping to construct a confidence interval of their KL divergence estimator. In all these works, the sample sizes from both distributions are assumed to grow at the same rate.

Considering a simpler case of $d = 1$, our spacing estimator for the KL divergence is an extension of the two-sided $2m$-spacing entropy estimator that appears in Vasicek (1976). We establish strong consistency and central limit theorems (CLTs) for the proposed KL divergence estimator when size-n and size-s i.i.d. samples of real-world and simulation outputs, respectively, are available. In particular, we discuss the requirements for the choices of m and s as increasing functions of n for the asymptotic results to hold. The relationship between n and s provides a guidance on selecting the sample size s for the simulation experiment when n real-world outputs are available, and vice versa. Moreover, our CLTs provide explicit forms of the asymptotic variance, and thus facilitate confidence interval construction.

We note that there are several nonparametric estimators for the KL divergence and other information measures proposed in the literature and only a small subset is reviewed here. We refer the readers to Verdú (2019) for a comprehensive review on recent work.

The remainder of the paper is organized as follows. In Section 2, we provide some background on entropy and its spacing estimator. We discuss the KL divergence and the corresponding spacing estimators in Section 3 assuming the distribution function of the real-world system output is unkonwn, but the simulation output distribution is known. Section 4 extends the KL divergence estimator to the case when both distributions are unknown, and discusses its asymptotic properties. Empirical demonstration is presented in Section 5.

2 BACKGROUND

Throughout the paper, we denote the probability density function (pdf) and cumulative distribution function (cdf) of the continuous-valued real-world output random variable, X, with f and F, respectively. Without loss of generality, X may represent another simulator's output random variable if the objective is to measure the discrepancy between two simulators' output distributions.

For absolutely continuous random variable X with density function f, its differential entropy is

$$H(f) = -\int f(x)\log(f(x))dx = \mathbb{E}[-\log(f(X))]. \tag{1}$$

Note that $\log(\cdot)$ refers to the natural logarithm in this paper while in some work, base 2 logarithm is adopted instead.

Let $X_i \overset{i.i.d.}{\sim} F$, and $X_{(i)}$ denote the ith order statistics such that $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$. An m-spacing estimator for f is

$$f_{n,m}(x) = \frac{m}{n} \frac{1}{X_{(im)} - X_{((i-1)m)}}, \text{ where } x \in \left[X_{((i-1)m)}, X_{(im)}\right). \tag{2}$$

For intuition, consider when $m = 1$. Then, (2) uniformly assigns the probability mass of $1/n$ to the interval, $\left[X_{(i-1)}, X_{(i)}\right)$. For $m > 1$, (2) assigns m/n to each interval of m order statistics. Several variations of m-spacing density estimators have been proposed, but their underlying ideas are similar to that of (2).

Vasicek (1976) adopts a $2m$-spacing density estimator to nonparametrically estimate $H(f)$ by

$$H_{mn} = \frac{1}{n} \sum_{i=1}^{n} \log\left(\frac{n}{2m}\left(X_{(i+m)} - X_{(i-m)}\right)\right), \tag{3}$$

where $X_{(j)} = X_{(n)}$, if $j > n$ and $X_{(j)} = X_{(1)}$, if $j < 1$. Observe that H_{mn} replaces the expectation in (1) with the sample mean while $f(X_{(i)})$ is replaced with a $2m$-spacing density estimator centered at $X_{(i)}$.

To characterize estimation error of H_{mn}, (3) can be written as the sum of three components (Vasicek 1976). Namely, $H_{mn} = -\frac{1}{n}\sum_{i=1}^{n} \log(f(X_i)) + V_{mn} - W_{mn}$, where

$$V_{mn} \triangleq \frac{1}{n} \sum_{i=1}^{n} \log\left(\frac{n}{2m}\left(F\left(X_{(i+m)}\right) - F\left(X_{(i-m)}\right)\right)\right), \text{ and } W_{mn} \triangleq \frac{1}{n} \sum_{i=1}^{n} \log\left(\frac{F\left(X_{(i+m)}\right) - F\left(X_{(i-m)}\right)}{f\left(X_{(i)}\right)\left(X_{(i+m)} - X_{(i-m)}\right)}\right). \tag{4}$$

Observe that the first term in the sum is the Monte Carlo estimate of the entropy given f, which does not depend on m, while V_{mn} corresponds to the error caused by replacing F with the empirical cdf, and W_{mn} can be viewed as the discretization error of the $2m$-spacing density estimator. Thus, consistency and asymptotic normality of H_{mn} can be shown by characterizing each sample mean in the sum.

Note that for $1 \leq i \leq n$, $F\left(X_{(i)}\right) = U_{(i)}$, where $U_{(1)} \leq \cdots \leq U_{(n)}$ is the order statistics from an i.i.d. n-sample $U_1, \ldots, U_n$ of Uniform$(0,1)$ random variables. We additionally define $U_{(0)} = 0$ and $U_{(n+1)} = 1$. From the well-known relationship between uniform order statistics and the Beta distribution, we have $U_{(k)} - U_{(j)} \sim \text{Beta}(k-j, n-(k-j)+1)$ and $\mathbb{E}\left[\log\left(U_{(k)} - U_{(j)}\right)\right] = \psi(k-j) - \psi(n+1)$ for $k > j$, where $\psi(x) = \Gamma'(x)/\Gamma(x)$ is called the digamma function. From these, Vasicek (1976) shows that V_{mn} is independent of F and derives its mean as

$$\mathbb{E}[V_{mn}] = \log(n) - \log(2m) - \psi(n+1) + \left(1 - \frac{2m}{n}\right)\psi(2m) + \frac{2}{n}\sum_{i=1}^{m} \psi(i+m-1), \tag{5}$$

Because $\log(x) - 1/x < \psi(x) < \log(x) - 1/(2x)$ for $x > 0$ (Alzer 1997) and the sum of the last two items in (5) cancels with $\log(2m)$ in the limit, $\mathbb{E}[V_{mn}] \to 0$ as $m \to \infty$ and $n \to \infty$. He further suggests $H'_{mn} \triangleq H_{mn} - \mathbb{E}[V_{mn}]$ as a bias-corrected estimator for $H(f)$, and shows that given f has finite variance, H'_{mn} is weakly consistent when $m = o(n)$ as $m \to \infty$ and $n \to \infty$.

van Es (1992) shows that H'_{mn} is strongly consistent under a different set of assumptions as stated in Assumption 1, which we adopt in the remainder of the paper.

Assumption 1 There exist ρ_f and γ_f such that $0 < \rho_f \leq f(x) \leq \gamma_f < \infty$ for all $x \in \text{supp}(f)$ and X is an absolutely continuous random variable with respect to Lebesgue measure.

For generic density f, $H(f)$ may be arbitrarily large. However, finite variance implies $H(f) < \infty$ and boundedness of f implies $H(f) > -\infty$. Thus Assumption 1 constrains our discussion on finite entropy only. van Es (1992) also shows CLTs for H'_{mn} when scaled appropriately. We restate his results below to later invoke it to show strong consistency and asymptotic normality for our KL divergence estimator.

Lemma 2 (Theorems 2 and 4 in van Es (1992)) Suppose Assumption 1 holds and supp(f) is an interval. Moreover, let $m/\log n \to \infty$ as $n \to \infty$.

(a) If $m = o(n)$, then we have $H'_{mn} \xrightarrow{a.s.} H(f)$.

Furthermore, suppose f is Lipschitz continuous in supp(f).

(b) If f is not the pdf of $U(0,1)$ and $m = o(n^{1/2})$, then $n^{1/2}\left(H'_{mn} - H(f)\right) \xrightarrow{D} \mathcal{N}(0, \mathrm{Var}[\log(f(X))])$.

(c) If f is identical to the pdf of $U(0,1)$ almost everywhere and $m = o(n^{1/3})$, then $(mn)^{1/2} H'_{mn} \xrightarrow{D} \mathcal{N}(0, 1/3)$.

We note that in van Es (1992), Lemma 2 is written for one-sided m-spacing estimator. However, the same results hold for the two-sided variant considered here. One can show that the two estimators are almost surely equivalent under the same set of assumptions made above. van Es (1992) further shows the strong consistency when the support is a finite union of intervals, by replacing $m/\log n \to \infty$ with $m = \mathcal{O}(n^{1-\varepsilon})$ for some $0 < \varepsilon < 1$. This condition is also employed in Beirlant and van Zuijlen (1985), where strong consistency for the two-sided m-spacing estimator is proved, given that f is absolutely continuous and has finite variance, but no constraint on the support.

We close this section by introducing the definition of KL divergence. Let g and G respectively denote the pdf and cdf of the continuous-valued simulation output random variable, Y. The KL divergence of g with respect to f is defined as

$$D(f||g) = \int f(x) \log\left(\frac{f(x)}{g(x)}\right) \mathrm{d}x = \mathbb{E}_f\left[\log\left(\frac{f(X)}{g(X)}\right)\right], \tag{6}$$

where $\mathbb{E}_f$ indicates that the expectation is taken with respect to f. The following two sections discusses estimation of $D(f||g)$ when G is known and unknown, respectively.

3 SPACING ESTIMATOR FOR KL DIVERGENCE WITH KNOWN G

Using the spacing estimator of the density function, H'_{mn} can be extended to estimate $D(f||g)$. Assuming G is known, the corresponding spacing estimator for $D(f||g)$ is

$$D_{mn} = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{2m/n}{G\left(X_{(i+m)}\right) - G\left(X_{(i-m)}\right)}\right) + \mathbb{E}[V_{mn}], \tag{7}$$

where $\mathbb{E}[V_{mn}]$ in (5) is added for bias correction. Essentially, (7) is the sample mean of the log of the estimated likelihood ratio at $X_{(i)}$ where $f\left(X_{(i)}\right)$ is replaced with $\dfrac{2m/n}{X_{(i+m)} - X_{(i-m)}}$ as in (3) and $g\left(X_{(i)}\right)$ is with $\dfrac{G\left(X_{(i+m)}\right) - G\left(X_{(i-m)}\right)}{X_{(i+m)} - X_{(i-m)}}$. Similar to the decomposition of H_{mn}, we can rewrite (7) as

$$D_{mn} = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{f(X_i)}{g(X_i)}\right) - V_{mn} + W_{mn} - Z_{mn} + \mathbb{E}[V_{mn}], \tag{8}$$

where V_{mn} and W_{mn} are defined in (4), and

$$Z_{mn} \triangleq \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{G\left(X_{(i+m)}\right) - G\left(X_{(i-m)}\right)}{g\left(X_{(i)}\right)\left(X_{(i+m)} - X_{(i-m)}\right)}\right). \tag{9}$$

We may interpret Z_{mn} as the discretization error of using the finite difference method to estimate $g\left(X_{(i)}\right)$. Decomposition (8) supports adding $\mathbb{E}[V_{mn}]$ for bias correction in (7).

A quantity similar to (7) is studied by Ekström (1999) in the context of parameter estimation when F has a known distribution family with unknown parameter vector θ_0. As an alternative to maximum likelihood estimation, the maximum spacing estimator for θ_0 is defined as θ that minimizes the estimated KL divergence of $F(x;\theta)$ with respect to true distribution $F(x;\theta_0)$:

$$\frac{1}{n-m+2}\sum_{i=0}^{n-m+1}\log\left(\frac{F\left(X_{(i+m)};\theta_0\right)-F\left(X_{(i)};\theta_0\right)}{F\left(X_{(i+m)};\theta\right)-F\left(X_{(i)};\theta\right)}\right), \tag{10}$$

where $X_{(0)}=-\infty$ and $X_{(n+1)}=\infty$. Compared to (7), observe that (10) uses one-sided m-spacing density estimator. Moreover, (10) assumes both $F(\cdot;\theta)$ and $F(\cdot;\theta_0)$ have known functional forms as the common distribution family is assumed for parameter estimation, whereas (7) assumes only G is known, not F. Ekström (1999) provides several strong consistency results for (10) and related quantities, one of which we restate below as a lemma.

Lemma 3 (Corollary 1 in Ekström (1999)) Suppose Ξ is a nondecreasing bounded function on $[0,1]$ and there exists function ξ such that $\xi(t)=\mathrm{d}\Xi(t)/\mathrm{d}t$ almost everywhere for $t\in(0,1)$ and $\xi(t)\geq\rho_\xi$ for some $\rho_\xi>0$. If $m=o(n)$, then

$$\frac{1}{n-m+2}\sum_{i=0}^{n-m+1}\log\left(\frac{\Xi\left(U_{(i+m)}\right)-\Xi\left(U_{(i)}\right)}{U_{(i+m)}-U_{(i)}}\right)\xrightarrow{a.s.}\int_0^1\log(\xi(t))dt.$$

Although Lemma 3 is written for one-sided m-spacing estimator, the same consistency result holds when $U_{(i)}$ and $n-m+2$ are replaced with $U_{(i-m)}$ and n, respectively, whilst defining $X_{(i)}=X_{(1)}$ for $i<1$ and $X_{(i)}=X_{(n)}$ for $i>n$. We apply this variant of Lemma 3 in the proof of Theorem 5 that states strong consistency of D_{mn}, below. We first state an additional assumption on the density function g.

Assumption 4 There exist ρ_g and γ_g such that $0<\rho_g\leq g(y)\leq\gamma_g<\infty$ for all $y\in\mathrm{supp}(f)$ and Y is an absolutely continuous random variable with respect to Lebesgue measure.

Because the base measure of $D(f||g)$ is f, the boundedness assumption for g is made on the support of f, not of g. The assumption also assures $\mathrm{supp}(f)\subset\mathrm{supp}(g)$, eliminating the case when $D(f||g)=\infty$. Below, we present our first main result on strong consistency of D_{mn}. Unlike Lemma 2 by van Es (1992), note that Theorem 5 does not require $\mathrm{supp}(f)$ to be an interval; it may be a union of intervals.

Theorem 5 Suppose Assumption 1 holds and we have known G whose density function satisfies Assumption 4. If $m/\log n\to\infty$ and $m=o(n)$, then $D_{mn}\xrightarrow{a.s.}D(f||g)$ as $n\to\infty$.

Proof. Let $\Xi=G\circ F^{-1}$. Then, Ξ is nondecreasing in $[0,1]$ and $\xi(t)=\dfrac{g(F^{-1}(t))}{f(F^{-1}(t))}\geq\dfrac{\rho_g}{\gamma_f}>0$ from the assumptions. Let $U_{(i)}=F(X_{(i)})$. Then, $\Xi(U_{(i)})=G(F^{-1}(U_{(i)}))=G(F^{-1}(F(X_{(i)})))=G(X_{(i)})$. Because both F and G are absolutely continuous and thus are differentiable almost everywhere, we apply Lemma 3 to obtain

$$\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{G\left(F^{-1}\left(U_{(i+m)}\right)\right)-G\left(F^{-1}\left(U_{(i-m)}\right)\right)}{U_{(i+m)}-U_{(i-m)}}\right)$$
$$\xrightarrow{a.s.}\int_0^1\log\left(\frac{g(F^{-1}(t))}{f(F^{-1}(t))}\right)\mathrm{d}t=\int_{\mathrm{supp}(f)}\log\left(\frac{g(x)}{f(x)}\right)f(x)\mathrm{d}x=-D(f||g).$$

From (7), observe that

$$D_{mn}=-\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{G\left(X_{(i+m)}\right)-G\left(X_{(i-m)}\right)}{U_{(i+m)}-U_{(i-m)}}\right)-\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{n}{2m}\left(U_{(i+m)}-U_{(i-m)}\right)\right)+\mathbb{E}[V_{mn}], \tag{11}$$

where the first term converges to $D(f||g)$ almost surely as shown above and the remaining part is the bias-corrected entropy estimator, H'_{mn}, for Uniform$(0,1)$ and thus converges to 0 almost surely in the limit. □

To derive CLT results for D_{mn}, we first state extra conditions on smoothness of f and g.

Assumption 6 Suppose that there exist $L_f > 0$ and $L_g > 0$ such that for any x_1 and x_2 in the support of f, $|f(x_1) - f(x_2)| \leq L_f|x_1 - x_2|$ and $|g(x_1) - g(x_2)| \leq L_g|x_1 - x_2|$.

Again, the smoothness condition for g is only needed on the support of f as the base measure of $D(f||g)$ is f.

To pave the way for the CLT results, we restate another lemma from Ekström (1999). The proof of the following lemma can be found in the proof of Theorem 1 in van Es (1992).

Lemma 7 (Lemma 4 in Ekström (1999)) Let $\{m_n\}$ be a sequence of positive integers with $m_n/\log n \to \infty$, then

$$\lim_{n\to\infty} \max_{0\leq j\leq n-m_n+1} \left| \frac{n+1}{m_n} \left(U_{(j+m_n)} - U_{(j)}\right) - 1 \right| = 0, \text{ almost surely.}$$

We show asymptotic normality for D_{mn} when $f \neq g$ and when $f = g$ almost everywhere, respectively, in the following.

Theorem 8 Suppose Assumptions 1, 4, and 6 hold. If the support of f is an interval and $m/\log n \to \infty$ as $n \to \infty$, then we have the following results.

(a) If $f \neq g$ and $m = o\left(n^{1/2}\right)$, then $n^{1/2}(D_{mn} - D(f||g)) \xrightarrow{D} \mathcal{N}(0, \text{Var}[\log(f(X)/g(X))])$.

(b) If $f = g$ almost everywhere and $m = o\left(n^{1/3}\right)$, then $(mn)^{1/2}(D_{mn} - D(f||g)) \xrightarrow{D} \mathcal{N}(0, 1/6)$.

Proof. Because supp(f) is a bounded interval, there exists $M > 0$ such that supp$(f) \subset [-M, M]$. Recall the decomposition of D_{mn} in (8), we examine each error term in this sum starting with W_{mn}. By the mean value theorem, for each $1 \leq i \leq n$, there exists $\tilde{X}_i \in \left(X_{(i-m)}, X_{(i+m)}\right)$ such that $F\left(X_{(i+m)}\right) - F\left(X_{(i-m)}\right) = f\left(\tilde{X}_i\right)\left(X_{(i+m)} - X_{(i-m)}\right)$. Since $n \to \infty$ and $m = o(n)$,

$$\left| \frac{f\left(\tilde{X}_i\right) - f\left(X_{(i)}\right)}{f\left(X_{(i)}\right)} \right| \leq \frac{L_f}{\rho_f} \left|\tilde{X}_i - X_{(i)}\right| \leq \frac{L_f}{\rho_f} \left|X_{(i+m)} - X_{(i-m)}\right| \xrightarrow{a.s.} 0$$

Because $|\log(1+x)| \leq 2|x|$ given $|x| \leq 1/2$, for sufficiently large n, we have

$$\begin{aligned} |W_{mn}| &= \left| \frac{1}{n} \sum_{i=1}^{n} \log\left(\frac{f\left(\tilde{X}_i\right)}{f\left(X_{(i)}\right)} \right) \right| = \left| \frac{1}{n} \sum_{i=1}^{n} \log\left(1 + \frac{f\left(\tilde{X}_i\right) - f\left(X_{(i)}\right)}{f\left(X_{(i)}\right)} \right) \right| \leq \frac{2}{n} \sum_{i=1}^{n} \left| \frac{f\left(\tilde{X}_i\right) - f\left(X_{(i)}\right)}{f\left(X_{(i)}\right)} \right| \\ &\leq \frac{2L_f}{n\rho_f} \sum_{i=1}^{n} \left|X_{(i+m)} - X_{(i-m)}\right| \leq \frac{2L_f}{n\rho_f} 4Mm = \frac{8L_f M}{\rho_f} \frac{m}{n}, \text{ almost surely.} \end{aligned}$$

The last inequality follows from that in $\sum_{i=1}^{n} \left|X_{(i+m)} - X_{(i-m)}\right|$, $\left|X_{(j)} - X_{(j-1)}\right|$ for each $2 \leq j \leq n$ is added at most $2m$ times and that $\left|X_{(n)} - X_{(1)}\right| \leq 2M$. Similarly, we can show $|Z_{mn}| \leq \frac{8L_g M}{\rho_g} \frac{m}{n}$ with probability one. Next we focus on the term V_{mn}. Recall that we define $X_{(j-m)} = X_{(1)}$ for $j - m < 1$ and $X_{(j+m)} = X_{(n)}$ for $j - m > n$. Then, for the first m terms in the sum of V_{mn}, $\frac{1}{n}\sum_{i=1}^{m} \log\left(\frac{n}{2m}\left(F\left(X_{(i+m)}\right) - F\left(X_{(i-m)}\right)\right)\right) = \frac{1}{n}\sum_{i=1}^{m} \log\left(\frac{n}{2m}\left(U_{(i+m)} - U_{(1)}\right)\right)$, and

$$\frac{m}{n} \log\left(\frac{n}{2m}\left(U_{(m+1)} - U_{(1)}\right)\right) \leq \frac{1}{n} \sum_{i=1}^{m} \log\left(\frac{n}{2m}\left(U_{(i+m)} - U_{(1)}\right)\right) \leq \frac{m}{n} \log\left(\frac{n}{2m}\left(U_{(2m)} - U_{(1)}\right)\right).$$

Combined with Lemma 7, we have

$$\frac{m}{n}\log\left(\frac{n}{2m}\left(U_{(2m)}-U_{(1)}\right)\right)=\frac{m}{n}\log\left(\frac{n+1}{2m-1}\left(U_{(2m)}-U_{(1)}\right)\right)+\frac{m}{n}\log\left(\frac{n}{n+1}\frac{2m-1}{2m}\right)=\mathscr{O}\left(\frac{m}{n}\right)$$

almost surely. Similarly, one can show that $\frac{m}{n}\log\left(\frac{n}{2m}\left(U_{(m+1)}-U_{(1)}\right)\right)=\mathscr{O}\left(\frac{m}{n}\right)$. The last m terms in the sum of V_{mn} can also be shown to be $\mathscr{O}\left(\frac{m}{n}\right)$ following the same logic. The remaining part of V_{mn} is

$$\begin{aligned}&\frac{1}{n}\sum_{i=m+1}^{n-m}\log\left(\frac{n}{2m}\left(U_{(i+m)}-U_{(i-m)}\right)\right)\\&\qquad=\frac{(n-2m)}{n}\log\left(\frac{n}{2m}\right)+\frac{1}{n}L_{2m,n}-\frac{1}{n}\log\left(U_{(2m)}\right)-\frac{1}{n}\log\left(1-U_{(n-2m+1)}\right),\end{aligned}$$

where $L_{2m,n}\triangleq\sum_{j=0}^{n-2m+1}\log\left(U_{(j+2m)}-U_{(j)}\right)$, is one of the statistics studied by Cressie (1976). From the discussion on U order statistics in Section 2, it can be seen that $\mathbb{E}[L_{2m,n}]=(n-2m+2)(\psi(2m)-\psi(n+1))$ and $\mathbb{E}[L_{2m,n}]/n+\log(n/2m)\overset{a.s.}{\to}0$ as $n\to\infty$, $m\to\infty$ and $m=o(n)$. Cressie (1976) also shows that the variance of $L_{2m,n}$ is of order $n/(2m)$. Additionally, by Lemma 7,

$$\frac{\sqrt{mn}}{n}\log\left(U_{(2m)}\right)=\sqrt{\frac{m}{n}}\log\left(\frac{n+1}{2m}U_{(2m)}\right)+\sqrt{\frac{m}{n}}\log\left(\frac{2m}{n+1}\right)\overset{a.s.}{\to}0.$$

Similarly, $\frac{1}{n}\log\left(1-U_{(n-2m+1)}\right)\overset{a.s.}{\to}0$. Combining all pieces, we finally have

$$D_{mn}=\frac{1}{n}\sum_{i=1}^{n}\log\left(\frac{f(X_i)}{g(X_i)}\right)-\frac{1}{n}\left(L_{2m,n}-\mathbb{E}[L_{2m,n}]\right)+o\left(\frac{1}{\sqrt{mn}}\right)+\mathscr{O}\left(\frac{m}{n}\right)+\mathbb{E}[V_{mn}]$$

almost surely. The variance of the first four parts are of the order $1/n$, $1/(mn)$, $o(1/(mn))$ and m^2/n^2. Suppose $f\neq g$, then by choosing $m=o\left(n^{1/2}\right)$, the variance of the first term dominates the others. When $f=g$ almost everywhere, the first term is 0. Additionally, W_{mn} and Z_{mn} cancel out each other. Cressie (1976) shows that provided $m=o(n^{1/3})$, $(2m/n)^{1/2}\left(L_{2m,n}-\mathbb{E}[L_{2m,n}]\right)\overset{D}{\to}\mathscr{N}(0,1/3)$, which leads to Part (b) of our theorem. □

We can also extend the result to the case when there are finite discontinuity points in the support of f.

Corollary 9 Suppose Assumptions 1 and 4 hold. If the support of f is a finite union of intervals and Assumptions 6 holds on each interval, then Theorem 8 holds for such f.

Proof. Suppose there are k discontinuity points. We define $\mathscr{A}\subseteq\{1,2,\ldots n\}$ as for any $i\in\mathscr{A}$, $\left(X_{(i-m)},X_{(i+m)}\right)$ overlaps at least one discontinuity point. Therefore the cardinality of $\mathscr{A}$ is at most $2km$. We first consider the case $f\neq g$. From boundedness of f, for the ith term in the sum of W_{mn}, $\left|\log\left(\frac{F\left(X_{(i+m)}\right)-F\left(X_{(i-m)}\right)}{f\left(X_{(i)}\right)\left(X_{(i+m)}-X_{(i-m)}\right)}\right)\right|\leq\left|\log\left(\frac{\gamma_f}{\rho_f}\right)\right|$, for any $1\leq i\leq n$. With a similar analysis on G and g, we can conclude that the contributions to W_{mn} and Z_{mn} from all $i\in\mathscr{A}$ are at most $\mathscr{O}(m/n)$. The analysis on V_{mn} is not affected by the new discontinuity assumption. When $f=g$, W_{mn} cancels Z_{mn}. Therefore, the conclusions still hold. □

4 EXTENSION TO UNKNOWN G

Since we are interested in the case where both F and G are unknown, D_{mn} is not directly applicable to our context. Instead, we consider estimator D^s_{mn}, which replaces G with an empirical cdf. Specifically, suppose

$Y_1, Y_2, \ldots, Y_s$ are i.i.d. observations from G. Then, $G_s(\cdot) = s^{-1}\sum_{j=1}^{s} I(\cdot \le Y_j)$, where $I(\cdot)$ is an indicator function. Replacing G in D_{mn} with G_s, we obtain

$$D^s_{mn} = -\frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{n}{2m}\left(G_s\left(X_{(i+m)}\right) - G_s\left(X_{(i-m)}\right)\right)\right) + \mathbb{E}[V_{mn}]. \tag{12}$$

In the simulation context, a natural question is how to balance the simulation sample size, s, with n and m so that D^s_{mn} exhibits desired statistical properties, which we answer in this section.

Note that we can also write D^s_{mn} as $D^s_{mn} = \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{f(X_i)}{g(X_i)}\right) - V_{mn} + \mathbb{E}[V_{mn}] + W_{mn} - Z_{mn} - S^s_{mn}$, where V_{mn} and W_{mn} are defined as in (4), Z_{mn} is given in (9) and

$$S^s_{mn} \triangleq \frac{1}{n}\sum_{i=1}^{n} \log\left(\frac{G_s\left(X_{(i+m)}\right) - G_s\left(X_{(i-m)}\right)}{G\left(X_{(i+m)}\right) - G\left(X_{(i-m)}\right)}\right). \tag{13}$$

Compared to the representation of D_{mn} in Section3, S^s_{mn} is the additional error due to approximating G with empirical cdf G_s.

Below, we show strong consistency and asymptotic normality of D^s_{mn}.

Theorem 10 Suppose Assumptions 1 and 4 hold. If $m = o(n)$ and $m/\log n \to \infty$ as $n \to \infty$, then we have the following results.

(a) If $s = \omega((n/m)^2)$, then $D^s_{mn} \xrightarrow{p} D(f||g)$.
(b) If $(n/m)^2(\log(\log(s))/s) \to 0$, then $D^s_{mn} \xrightarrow{a.s.} D(f||g)$.

Proof. Notice that $D^s_{mn} = D_{mn} + S^s_{mn}$ and we have $D_{mn} \xrightarrow{a.s.} D(f||g)$ from Theorem 5. In the following, we show that $S^s_{mn} \to 0$ in probability and almost surely with some suitable choices of m, n and s. We have $|S^s_{mn}| \le \frac{1}{n}\sum_{i=1}^{n} |\log(1 + \Delta_{i,1}/\Delta_{i,2})|$, where

$$\Delta_{i,1} \triangleq G_s\left(X_{(i+m)}\right) - G\left(X_{(i+m)}\right) - G_s\left(X_{(i-m)}\right) + G\left(X_{(i-m)}\right) \text{ and } \Delta_{i,2} \triangleq G\left(X_{(i+m)}\right) - G\left(X_{(i-m)}\right).$$

We first analyze $\Delta_{i,2}$. Because F is absolutely continuous, $F^{-1}(F(X)) = X$ with probability one. Therefore, in the almost surely sense, for $m < i \le n-m$, $G\left(X_{(i+m)}\right) - G\left(X_{(i-m)}\right) \ge G\left(X_{(i+m)}\right) - G\left(X_{(i)}\right) = \left(G \circ F^{-1}\right)\left(U_{(i+m)}\right) - \left(G \circ F^{-1}\right)\left(U_{(i)}\right) \ge \frac{\rho_g}{\gamma_f}\left(U_{(i+m)} - U_{(i)}\right)$ from the mean value theorem, where the inequality follows from the boundedness of the densities; for $1 \le i \le m$, $G\left(X_{(i+m)}\right) - G\left(X_{(i-m)}\right) = G\left(X_{(i+m)}\right) - G\left(X_{(1)}\right) \ge G\left(X_{(m+1)}\right) - G\left(X_{(1)}\right) \ge \frac{\rho_g}{\gamma_f}\left(U_{(m+1)} - U_{(1)}\right)$. Performing a similar analysis on $n-m+1 \le i \le n$, we conclude that

$$\min_{1\le i\le n}\left\{G\left(X_{(i+m)}\right) - G\left(X_{(i-m)}\right)\right\} \ge \frac{\rho_g}{\gamma_f}\frac{m}{n+1}\min_{1\le i\le n-m}\left\{\frac{n+1}{m}\left(U_{(i+m)} - U_{(i)}\right)\right\} \tag{14}$$

almost surely. Therefore, we have $\min_{1\le i\le n}\Delta_{i,2} \ge \mathcal{O}(m/n)$ almost surely by applying Lemma 7 to the right-hand side of (14).

For $\Delta_{i,1}$, if there exist a positive sequence $\{a_n\}$ such that $\mathbb{P}\left(\sup_{1\le i\le n}|\Delta_{i,1}| \le a_n\right) \to 1$ and $a_n = o\left(\frac{m}{n}\right)$, then $\sup_{1\le i\le n}|\Delta_{i,1}| / \inf_{1\le i\le n}|\Delta_{i,2}| \to 0$ in probability. Recall that given $|x| \le 1/2$, $|\log(1+x)| \le 2|x|$, we can thus bound S^s_{mn} as

$$|S^s_{mn}| \le \frac{2}{n}\sum_{i=1}^{n}\frac{|\Delta_{i,1}|}{|\Delta_{i,2}|} \le 2\frac{\sup_{1\le i\le n}|\Delta_{i,1}|}{\inf_{1\le i\le n}|\Delta_{i,2}|}$$

for sufficiently large n, and conclude that S^s_{mn} converges to 0 in probability. In the following we give a sufficient condition for such $\{a_n\}$ to exist. By the triangle inequality, $|\Delta_{i,1}| \le |G_s\left(X_{(i+m)}\right) - G\left(X_{(i+m)}\right)| +$

$|G_s\left(X_{(i-m)}\right) - G\left(X_{(i-m)}\right)| \leq 2\sup_x|G_s(x) - G(x)|$, which implies $\sup_{1\leq i\leq n}|\Delta_{i,1}| \leq 2\sup_x|G_s(x) - G(x)|$. Then for any positive sequence $\{a_n\}$, by the Dvorsky-Kiefer-Wolfwitz inequality,

$$\mathbb{P}\left(\sup_{1\leq i\leq n}|\Delta_{i,1}| > a_n\right) \leq \mathbb{P}\left(2\sup_x|G_s(x) - G(x)| > a_n\right) \leq 2\exp\left(-\frac{sa_n^2}{2}\right). \tag{15}$$

Hence, provided that $sa_n^2 \to \infty$, we have $\mathbb{P}\left(\sup_{1\leq i\leq n}|\Delta_{i,1}| \leq a_n\right) \to 1$ as desired. Observe that if m and s are chosen to satisfy $s = \omega((n/m)^2)$, then we can always find $a_n = o(m/n)$ such that $sa_n^2 \to \infty$. Therefore, we have (15)$\to 0$, which concludes the proof of Part (a).

To prove Part (b), we derive an almost sure upper bound for $\sup_{1\leq i\leq n}|\Delta_{i,1}|$. The Glivenko-Cantelli theorem can be strengthened to a law of the iterated logarithm (Van der Vaart 2000):

$$\limsup_{s\to\infty}\sqrt{\frac{s}{2\log(\log(s))}}\|G_s - G\|_\infty \leq \frac{1}{2} \quad \text{almost surely.}$$

Thus, choosing m and s such that$(n/m)^2(\log(\log(s))/s) \to 0$ as $n \to \infty$ and combining the two inequalities, $\sup_{1\leq i\leq n}|\Delta_{i,1}| \leq 2\sup_x|G_s(x) - G(x)|$ and $\inf_{1\leq i\leq n}\Delta_{i,2} \geq \mathcal{O}(m/n)$ almost surely, we have that $\sup_{1\leq i\leq n}|\Delta_{i,1}|\,/\,\inf_{1\leq i\leq n}|\Delta_{i,2}| \to 0$ with probability one. We proceed as

$$\begin{aligned}|S_{mn}^s| &\leq \frac{1}{n}\sum_{i=1}^n\left|\log\left(1+\frac{\Delta_{i,1}}{\Delta_{i,2}}\right)\right| \leq \frac{2}{n}\sum_{i=1}^n\left|\frac{\Delta_{i,1}}{\Delta_{i,2}}\right| \leq \frac{2\sup_{1\leq i\leq n}|\Delta_{i,1}|}{\inf_{1\leq i\leq n}|\Delta_{i,2}|} \leq \frac{4\rho_g}{\gamma_f}\frac{n}{m}\|G_s - G\|_\infty \\ &= \sqrt{\frac{2\log(\log(s))}{s}}\frac{4\rho_g}{\gamma_f}\frac{n}{m}\sqrt{\frac{s}{2\log(\log(s))}}\|G_s - G\|_\infty \leq \frac{4\rho_g}{\gamma_f}\sqrt{\frac{2\log(\log(s))}{s}}\frac{n}{m}, \text{ almost surely.}\end{aligned}$$

□

Notice that strong consistency (Part (b)) requires slightly larger s than weak consistency (Part (a)). Theorem 10 allows s to grow slower than n; for instance, $m = n^{3/4}$ and $s = n^{3/4}$ satisfy sample size requirements for both weak and strong consistency. Despite the savings in the simulation effort, this choice makes m large and thus introduces large discretization error of f and g, as we have shown that both $|W_{mn}|$ and $|Z_{mn}|$ are controlled by rate of $\mathcal{O}(m/n)$.

Lastly, CLT results for D_{mn}^s are derived in the following theorem.

Theorem 11 Suppose Assumption 1, 4, and 6 hold. If the support of f is an interval and $m/\log n \to \infty$ as $n \to \infty$, then we have the following results.

(a) If $f \neq g$, $m = o\left(n^{1/2}\right)$ and $s = \omega(n^3/m^2)$, then $n^{1/2}(D_{mn}^s - D(f||g)) \xrightarrow{D} \mathcal{N}(0, \text{Var}[\log(f(X)/g(X))])$.

(b) If $f = g$ almost everywhere, $m = o\left(n^{1/3}\right)$ and $s = \omega(n^3/m)$, $(mn)^{1/2}(D_{mn}^s - D(f||g)) \xrightarrow{D} \mathcal{N}(0, 1/6)$.

Proof. Suppose $f \neq g$ and $m = o\left(n^{1/2}\right)$. Following the same steps as in the proof of Theorem 8, we only need to show that $n^{1/2}S_{mn}^s \xrightarrow{p} 0$ under some suitable choices of m, n and s. Similar to Part (a) of Theorem 10, if $s = \omega(n^3/m^2)$, then we can always find $\{a_n\}$ such that $a_n = o(m/n^{3/2})$, $sa_n^2 \to \infty$. Therefore, we have $\mathbb{P}\{n^{1/2}|S_{mn}^s| \leq \frac{n^{3/2}}{m}a_n\} \to 1$ and $\frac{n^{3/2}}{m}a_n \to 0$, which concludes Part (a). Part (b) can be shown in a similar fashion given $m = o\left(n^{1/3}\right)$. □

Notice that Part (a) of Theorem 11 requires $s = \omega(n^2)$. In contrast with Theorem 10, m/n needs to decrease at a faster rate in order to make $n^{-1/2}\{\sum_{i=1}^n \log(f(X_i)/g(X_i)) - D(f||g)\}$ the dominant term of D_{mn}^s. As a consequence, s needs to increase faster to compensate for smaller m. For Part (b), notice that s is required to grow even faster in n. Notice that in both Parts (a) and (b), the explicit expressions for the asymptotic variances are provided.

Similar to Corollary 9, we can extend the CLTs result to the case where the support of f is a finite union of intervals.

Corollary 12 Suppose Assumptions 1 and 4 hold. If the support of f is a finite union of intervals and Assumptions 6 holds on each of them, then Theorem 10 holds for such f.

5 EMPIRICAL STUDIES

In this section, we examine the empirical performance of the KL divergence estimator, D^s_{mn}. To meet Assumptions 1 and 4, we first choose interval $[a,b]$ to be the support of both f and g and consider the truncated versions of f and g so that both the densities are bounded away from zero and infinity within the interval. The truncated version of f can be computed as $\tilde{f}(x) = f(x)/\int_a^b f(x)\mathrm{d}x$ if $x \in [a,b]$ and $\tilde{f}(x) = 0$ if $x \notin [a,b]$. Similarly, $\tilde{g}$ can be computed from g.

From the experiments, we observed that $D^s_{mn} = \infty$ for some instances. This is due to the fact that multiple realizations of X can lie within a single interval $[Y_{(j)}, Y_{(j+1)}]$ with nonzero probability. When this causes $G_s(X_{(i+m)} - G_s(X_{i-m})) = 0$ for some i, we have $D^s_{mn} = \infty$. To avoid this situation, we modify G_s via linear interpolation. Namely, we define $Y_{(0)} = a$ and $Y_{(s+1)} = b$ and let

$$G_s(z) = \frac{j}{s+1} + \frac{z - Y_{(j)}}{Y_{(j+1)} - Y_{(j)}}, \text{ if } z \in [Y_{(j)}, Y_{(j+1)}).$$

In the following, we compare $D\left(\tilde{f}||\tilde{g}\right)$ with D^s_{mn} for several choices of f and g. For all cases in Tables 1–3, f and g are truncated to $\tilde{f}$ and $\tilde{g}$, respectively, in $[a,b] = [0,20]$. The true value of $D\left(\tilde{f}||\tilde{g}\right)$ is computed via numerical integration. For all Gamma distributions discussed below, we adopt the shape-scale parameterization. In all experiments, we chose $m = n^{\frac{1}{2}}\log(\log(n))$ and $s = n$ and 100 macro runs were made to provide the average D^s_{mn} and its standard error.

Table 1 shows the case when f and g are the pdfs of $\mathcal{N}(5,1^2)$ and $\Gamma(10,0.5)$, respectively. Although $D(f||g) = \infty$ since $g(x) = 0$ for $x \leq 0$, the KL divergence after truncation, $D\left(\tilde{f}||\tilde{g}\right)$, is finite. The parameter choices for this example makes g look similar to f, which results in relatively small $D\left(\tilde{f}||\tilde{g}\right) = 0.1592$. In Table 1, we compare D^s_{mn} and D_{mn} with $D\left(\tilde{f}||\tilde{g}\right)$. Additionally, W_{mn}, Z_{mn} and S^s_{mn}, are displayed to understand the source of estimation error in D_{mn} and D^s_{mn}. All values in Table 1 are averaged over 100 macro runs and the standard errors are presented in the parentheses. As n and s grow, D^s_{mn} is as good as D_{mn} on average with a slightly larger standard error. All error terms are relatively small, among which S^s_{mn} shrinks to zero fastest. For our choice of m, the discretization errors W_{mn} and Z_{mn} are small; $\tilde{f}$ and $\tilde{g}$ have similar shapes, making most X and Y samples generated within the same region, and the errors due to empirical cdf surrogates become negligible since few points are evaluated at the right tail of G_s.

Table 2 shows the results when f and g are pdfs of $\mathcal{N}(5,1^2)$ and $\Gamma(2.5,2)$, respectively. Notice that both f and g have the same mean, however, compared to the first experiment, g is more right-skewed as the shape parameter is smaller. This change is reflected in increased $D\left(\tilde{f}||\tilde{g}\right) = 0.7153$. Similar observations can be made as in the first experiment; D_{mn} only outperforms D^s_{mn} in standard error when n and s are big.

Table 3 shows the case when we swap the choices for f and g from those in the second experiment. This change results in much increased $D\left(\tilde{f}||\tilde{g}\right) = 3.3209$. Observe that both D_{mn} and D^s_{mn} have significant

Table 1: Comparison of estimated KL divergence with 100 macro-replications with $f : \mathcal{N}(5,1^2)$ and $g : \Gamma(10,0.5)$ before truncation and benchmark $D\left(\tilde{f}||\tilde{g}\right) = 0.159$.

$n = s$	m	D_{mn}	D^s_{mn}	W_{mn}	Z_{mn}	S^s_{mn}
1000	61	0.194 (0.001)	0.197 (0.002)	-0.045 (0.001)	-0.039 (0.001)	-0.003 (0.002)
5000	151	0.174 (0.001)	0.177 (0.001)	-0.033 (0.001)	-0.028 (0.001)	-0.003 (0.001)
10000	222	0.154 (0.000)	0.156 (0.001)	-0.027 (0.000)	-0.023 (0.000)	-0.002 (0.000)
100000	772	0.157 (0.000)	0.157 (0.000)	-0.012 (0.000)	-0.010 (0.000)	-0.000 (0.000)

Table 2: Comparison of estimated KL divergence with 100 macro-replications with $f : \mathcal{N}(5,1^2)$ and $g : \Gamma(2.5,2)$ before truncation and benchmark $D\left(\tilde{f}||\tilde{g}\right) = 0.7153$.

$n = s$	m	D_{mn}	D^s_{mn}	W_{mn}	Z_{mn}	S^s_{mn}
1000	61	0.720 (0.002)	0.730 (0.004)	-0.046 (0.001)	-0.010 (0.000)	-0.010 (0.004)
5000	151	0.709 (0.001)	0.711 (0.002)	-0.033 (0.001)	-0.007 (0.000)	-0.002 (0.002)
10000	222	0.694 (0.001)	0.694 (0.001)	-0.027 (0.000)	-0.006 (0.000)	0.000 (0.001)
100000	772	0.707 (0.000)	0.708 (0.001)	-0.013 (0.000)	-0.004 (0.000)	-0.001 (0.000)

Table 3: Comparison of estimated KL divergence with 100 macro-replications with $f : \Gamma(2.5,2)$ and $g : \mathcal{N}(5,1^2)$ before truncation and benchmark $D\left(\tilde{f}||\tilde{g}\right) = 3.3209$.

$n = s$	m	D_{mn}	D^s_{mn}	W_{mn}	Z_{mn}	S^s_{mn}	inf
1000	61	1.850 (0.015)	1.458 (0.012)	-0.027 (0.001)	1.414 (0.017)	0.392 (0.013)	0
5000	151	2.386 (0.008)	1.780 (0.009)	-0.015 (0.000)	0.914 (0.006)	0.605 (0.010)	0
10000	222	2.568 (0.006)	1.887 (0.007)	-0.011 (0.000)	0.732 (0.003)	0.681 (0.008)	4
100000	772	Inf (NaN)	2.141 (0.004)	-0.003 (0.000)	-Inf (NaN)	Inf (NaN)	100

biases. Compared to normal distribution, Gamma distribution, which serves as the base measure $\tilde{f}$ in this experiment, has a heavier tail. Therefore, we have a relatively larger number of observations of X at the right tail than that of Y and the function value of $\tilde{G}$ at the tail are not estimated very well, which results in large biases.

While in all macro runs, D^s_{mn} were finite, some macro runs returned $D^s_{mn} = Z_{mn} = S^s_{mn} = \infty$. The last column of Table 3 counts the number of such instances out of the 100 macro-runs. For $n = 10000$, we calculate the averages and standard errors for D^s_{mn}, Z_{mn}, and S^s_{mn} based on the 96 finite instances. These infinite values are due to sknewness of f. As explained earlier, relatively larger number of X are observed at the tail of f and the difference, $G(X_{(i+m)}) - G(X_{(i-m)})$, is indistinguishable from 0 given the computer's numerical precision, which becomes infinity after the logarithmic operation.

ACKNOWLEDGMENTS

This work is partially supported by National Science Foundation Grant DMS-1854659.

REFERENCES

Alzer, H. 1997. "On some inequalities for the gamma and psi functions". *Mathematics of computation* 66(217):373–389.

Beirlant, J., E. Dudewicz, L. Gyor, and E. Meulen. 1997, 01. "Nonparametric Entropy Estimation: An Overview". *International Journal of Mathematical and Statistical Sciences* 6.

Beirlant, J., and M. van Zuijlen. 1985. "The empirical distribution function and strong laws for functions of order statistics of uniform spacings". *Journal of Multivariate Analysis* 16(3):300–317.

Cheng, R. C. H., and N. A. K. Amin. 1983. "Estimating Parameters in Continuous Univariate Distributions with a Shifted Origin". *Journal of the Royal Statistical Society. Series B (Methodological)* 45(3):394–403.

Cressie, N. 1976. "On the logarithms of high-order spacings". *Biometrika* 63(2):343–355.

Dudewicz, E. J., and E. C. van der Meulen. 1981. "Entropy-Based Tests of Uniformity". *Journal of the American Statistical Association* 76(376):967–974.

Ekström, M. 1999. "Strong Limit Theorems for Sums of Logarithms of High Order Spacings". *Statistics* 33(2):153–169.

Hall, P. 1984. "Limit theorems for sums of general functions of *m*-spacings". *Mathematical Proceedings of the Cambridge Philosophical Society* 96(3):517–532.

Hall, P. 1986. "On powerful distributional tests based on sample spacings". *Journal of Multivariate Analysis* 19(2):201–224.

Moon, K., and A. Hero. 2014a. "Multivariate f-divergence Estimation With Confidence". In *Advances in Neural Information Processing Systems*, edited by Z. Ghahramani, M. Welling, C. Cortes, N. Lawrence, and K. Q. Weinberger, Volume 27: Curran Associates, Inc.

Moon, K. R., and A. O. Hero. 2014b. "Ensemble estimation of multivariate f-divergence". In *2014 IEEE International Symposium on Information Theory*, 356–360.

Ranneby, B. 1984. "The Maximum Spacing Method. An Estimation Method Related to the Maximum Likelihood Method". *Scandinavian Journal of Statistics* 11(2):93–112.

Tarasenko, F. 1968. "On the evaluation of an unknown probability density function, the direct estimation of the entropy from independent observations of a continuous random variable, and the distribution-free entropy test of goodness-of-fit". *Proceedings of the IEEE* 56(11):2052–2053.

Van der Vaart, A. W. 2000. *Asymptotic statistics*, Volume 3. Cambridge university press.

van Es, B. 1992. "Estimating Functionals Related to a Density by a Class of Statistics Based on Spacings". *Scandinavian Journal of Statistics* 19(1):61–72.

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

Verdú, S. 2019. "Empirical Estimation of Information Measures: A Literature Guide". *Entropy* 21(8).

Wang, Q., S. R. Kulkarni, and S. Verdú. 2006. "A Nearest-Neighbor Approach to Estimating Divergence between Continuous Random Vectors". In *2006 IEEE International Symposium on Information Theory*, 242–246.

Wang, Q., S. R. Kulkarni, and S. Verdú. 2009. "Divergence Estimation for Multidimensional Densities Via k-Nearest-Neighbor Distances". *IEEE Transactions on Information Theory* 55(5):2392–2405.

AUTHOR BIOGRAPHIES

LINYUN HE is a Ph.D. student in the Department of Industrial and Manufacturing Engineering at the Penn State University. His research interests include simulation optimization, stochastic optimization, non-parametric methods and high-dimensional statistics. His email address is ljh5602@psu.edu. His website is https://dongfengguzhu.github.io.

EUNHYE SONG is the Harold and Inge Marcus Early Career assistant professor in Industrial and Manufacturing Engineering at Penn State. She earned her Ph.D. in Industrial Engineering and Management Sciences at Northwestern University. Her research interests include simulation model risk analysis, robust simulation optimization, and large-scale discrete simulation optimization. Her email address is eus358@psu.edu. Her website is https://eunhyesong.info.