

Notes about statistical estimation of Shannon entropy

YU SUN^{1,2} 

¹ Systemic Risk Centre, London School of Economics and Political Science, London WC2A 2AE, UK

² Department of Economics and Management, University of Trento, 38 122 Trento, Italy

Abstract. We improve the method of Bulinski and Dimitrov (2019) to prove the L^2 -consistency of the Kozachenko-Leonenko estimator for (differential) Shannon entropy.

AMS subject classifications: 60F25, 62G05, 62G20, 62H12

Keywords: Shannon differential entropy; nearest neighbor graphs; unbiasedness; consistency; Kozachenko-Leonenko estimate; L^2 -consistency

Received March 25, 2025; accepted September 1, 2025

1. Introduction

Shannon entropy is a fundamental concept in information theory that has numerous applications in fields like statistics, probability, and combinatorics. Let $X \in \mathbb{R}^d$ be a random vector defined on a probability space (Ω, \mathcal{F}, P) . Suppose that the joint distribution of X has a density $f(x)$ with respect to the Lebesgue measure dx , with the support $\mathcal{S} = \mathcal{S}(f) = \{x \in \mathbb{R}^d : f(x) > 0\}$. Consider $h(X) = -\log f(X)$, which can be thought of as the (random) information content of X (or as log-likelihood). The average value of information content of X is known as Shannon (or Boltzmann-Gibbs) entropy,

$$H(X) := \mathcal{E}[h(X)] = - \int_{\mathcal{S}} f(x) \log f(x) dx. \quad (1)$$

In [2], Kozachenko and Leonenko construct a popular nonparametric (differential) entropy estimator, the so-called Kozachenko-Leonenko (K-L) estimator H_N , based on the nearest neighbor (NN) graph. Let X_1, \dots, X_N be i.i.d. random vectors having the same law as the random vector $X \in \mathbb{R}^d$. For each $i = 1, \dots, N$ and $N \geq 2$, set $\rho_i = \min \{\rho(X_i, X_j), \forall i \in \{1 \dots N\}, \forall j \in \{1, \dots, N\} \setminus \{i\}\}$, where $\rho(x, y) = \|x - y\|$ denotes the *Euclidean distance* between $x, y \in \mathbb{R}^d$. In other words, ρ_i is the distance from X_i to its nearest neighbor (NN) in the sample $\{X_1, \dots, X_N\} \setminus \{X_i\}$, and $\{\rho_i, i = 1, \dots, N\}$ defines the NN random graph. Recall the NN in [2], the estimator of H defined in equation (1) is provided by the formula

$$H_N = \frac{1}{N} \sum_{i=1}^N \zeta_i(N),$$

$$\text{with } \zeta_i(N) = \log [\rho_i^d V_d e^\gamma (N-1)], \quad i = 1, \dots, N,$$

where $V_d := \pi^{(d/2)} / \Gamma(\frac{d}{2} + 1)$ and $\gamma = -\int_0^\infty e^{-t} \log t dt \approx 0.5772$ are the volume of a unit ball in \mathbb{R}^d and the Euler-Mascheroni constant, respectively.

2. Main results

In [1], Bulinski and Dimitrov propose using an analogue of the Hardy-Littlewood maximal functions for an investigation of unbiasedness and consistency of H_N . They provide an interesting and detailed proof method; however, some errors and shortcomings emerged. In this paper, we correct the errors, refine the proof method, and propose a more rigorous and general proof framework. This framework can be applied

Email address: y.u.sun@outlook.com

<https://www.mathos.unios.hr/mc>

© 2026 School of Applied Mathematics and Informatics, University of Osijek

not only to the entropy proof of the NN but also to the entropy proof of k -th nearest neighbor (k -NN) graphs, as well as to VarEntropy and high-order statistics of entropy estimators. Before stating the main results, some more definitions are needed. Let $B(x, r) = \{y \in \mathbb{R}^d : \rho(x, y) < r\}$ be a ball of a radius $r > 0$ with a center $x \in \mathbb{R}^d$. Clearly, $|B(x, r)| = \mu(B(x, r)) = r^d V_d$. Let $G(t)$ be a monotonically increasing function on $[0, \infty)$:

$$G(t) = \begin{cases} 0, & 0 \leq t < 1, \\ t \log t, & t \geq 1. \end{cases}$$

Recall the following functionals:

$$I_f(x, r) = \frac{\int_{B(x, r)} f(y) dy}{r^d V_d}, \quad M_f(x, R) = \sup_{r \in (0, R]} I_f(x, r) \quad \text{and} \quad m_f(x, R) = \inf_{r \in (0, R]} I_f(x, r).$$

It is known that the function $I_f(x, r)$ is continuous in $(x, r) \in \mathbb{R}^d \times (0, \infty)$, while for each $R > 0$, the functions $m_f(\cdot, R)$ and $M_f(\cdot, R)$ are upper semicontinuous and lower semicontinuous, respectively. Hence, these non-negative functions are Borel measurable. Clearly, for each $x \in \mathbb{R}^d$, $m_f(\cdot, R)$ is nonincreasing and $M_f(\cdot, R)$ is nondecreasing. Note in passing that substituting $\sup_{r \in (0, R]}$ by $\sup_{r \in (0, \infty)}$ in the definition of $M_f(x, R)$ leads to the celebrated Hardy-Littlewood maximal function $M_f(x)$ that is widely used in harmonic analysis. The main results in [1] are

Assumption 1. For a continuous density f in \mathbb{R}^d , given positive $\varepsilon_0, \varepsilon_1, \varepsilon_2, R_1, R_2, \alpha = 1, 2$, it holds that

$$\begin{aligned} K_f(\varepsilon_0, \alpha) &:= \int_{\mathbb{R}^d} \left(\int_{\mathbb{R}^d \setminus \{x\}} G(|\log^\alpha \rho(x, y)|) f(y) dy \right)^{1+\varepsilon_0} f(x) dx < \infty, \\ Q_f(\varepsilon_1, R_1) &:= \int_{\mathbb{R}^d} M_f^{\varepsilon_1}(x, R_1) f(x) dx < \infty, \\ T_f(\varepsilon_2, R_2) &:= \int_{\mathbb{R}^d} m_f^{-\varepsilon_2}(x, R_2) f(x) dx < \infty. \end{aligned}$$

Theorem 1. Under Assumption 1 and $\alpha = 1$, the estimator H_N is asymptotically unbiased, i.e.

$$\mathbb{E}(H_N) \rightarrow H, \quad N \rightarrow \infty.$$

Theorem 2. Under Assumption 1 and $\alpha = 2$, the estimator H_N is L^2 -consistent, i.e.

$$\mathbb{E}(H_N - H)^2 \rightarrow 0, \quad N \rightarrow \infty.$$

The proofs of these two theorems, while drawing inspiration from the classical approach in [2], innovatively incorporate the Hardy-Littlewood maximal function, widely used in harmonic analysis. However, the proof of Theorem 2 lacks rigor and contains errors. We have made the following corrections and refinements to address these issues.

The proof of *Theorem 2* (i.e. the L^2 -consistency of H_N) is provided in *Section 4 Proof of Theorem 2.8* in [1]. Note that the L^2 -consistency of H_N is guaranteed by $\text{Var}(H_N) \rightarrow 0$, as $N \rightarrow \infty$, and

$$\text{Var}(H_N) = \frac{1}{N} \text{Var}[\zeta_1(N)] + \frac{2}{N^2} \sum_{1 \leq i < j \leq N} \text{Cov}[\zeta_i(N), \zeta_j(N)].$$

The proof is composed of two steps. Step 1 and Step 2 are established to prove the asymptotic zero of variance and covariance terms separately, i.e. $\frac{1}{N} \text{Var}[\zeta_1(N)] \rightarrow 0$ and $\frac{1}{N^2} \sum_{1 \leq i < j \leq N} \text{Cov}[\zeta_i(N), \zeta_j(N)] \rightarrow 0$, as $N \rightarrow \infty$.

In Step 1, Bulinski and Dimitrov adopt the setting of $N = 2$ and replace the cumulative distribution functions $p_{N,x}(u)$ and $F_{N,x}(u)$ with $p_{2,x}(w)$ and $F_{2,x}(w)$ to facilitate the proof. In particular, they change the variable from u to $w = u/(N-1)$ and thus, the radius $r_N(u)$ defined in [1, Eq. (3.7)], i.e.

$r_N(u) := \left(\frac{u}{V_d \tilde{\gamma} (N-1)} \right)^{1/d}$, becomes $r_N(u) = r_N(w(N-1)) = r_N\left(\frac{w}{V_d \tilde{\gamma}}\right)^{1/d} = r_2(w)$. Hence, the accumulative distribution functions are changed accordingly (see, for instance, [1, Eq. (4.4)]).

In Step 2, Bulinski and Dimitrov specify that $N \geq 3$ is required (see, for instance, the settings of [1, Eq. (4.6)]). However, they apply the results of the Step 1 in [1, pp. 31-34], which was based on the settings of $N = 2$, e.g. $r_N(u) = r_2(w)$. In particular, in [1, p. 37], it is written: “*The same reasoning as was used at Step 1 of the proof of Theorem 2.8 leads.....*”.

Note that we require $N \geq 2$, as we use the Euclidean distance of two different points $\rho(x, y)$ to define the nearest neighbor. In other words, we need at least two points to define a ball $B(x, r) := \{y \in \mathbb{R}^d : \rho(x, y) < r\}$. Consider a ball centered at x with radius r . We can adopt the setting of $N = 2$ for Step 1, where we only consider the variance of the single random variable y or $\zeta_1(N)$. But we need at least two points y and z to calculate the covariance in Step 2, i.e. we need $N \geq 3$.

We improved the proof of Step 2 by amending the settings of $N = 2$ to adopt the requirement of $N \geq 3$. To avoid confusion, we use $J_2^y(N, x)$ to replace $J_2(N, x)$ for the case when $N \geq 3$. We define $\tilde{u} = 2u/(N-1)$ and the change of variables leads to the following remarkable results:

$$(i) \quad u \in [\sqrt{N-1}, \infty) \iff \tilde{u} = \frac{2u}{N-1} \in \left[\frac{2}{\sqrt{N-1}}, \infty \right), \quad N \geq 3,$$

$$(ii) \quad r_N(u) = \left[\frac{u}{(N-1)V_d \tilde{\gamma}} \right]^{1/d} = \left[\frac{\tilde{u}}{2V_d \tilde{\gamma}} \right]^{1/d} = r_3(\tilde{u}),$$

$$(iii) \quad \mathbf{P}_{N,x}(u) = \int_{B(x, r_N(u))} f(\xi) d\xi = \int_{B(x, r_3(\tilde{u}))} f(\xi) d\xi = \mathbf{P}_{3,x}(\tilde{u}),$$

$$(iv) \quad F_{N,x}^y(u) = F_{3,x}^y(\tilde{u}).$$

Recall from the last equation in [1, p. 32] that the integral $\int_{[e, \infty]}$ is split into $\int_{[e, \sqrt{N-1}]} + \int_{(\sqrt{N-1}, \infty)}$:

$$I_2^y(N, x) = J_1^y(N, x) + J_2^y(N, x),$$

with

$$\begin{aligned} J_1^y(N, x) &= \int_{[e, \sqrt{N-1}]} [1 - F_{N,x}^y(u)] \frac{\log u}{u} \left[\log(\log u) + \frac{1}{2} \right] du, \\ J_2^y(N, x) &= \int_{(\sqrt{N-1}, \infty)} [1 - F_{N,x}^y(u)] \frac{\log u}{u} \left[\log(\log u) + \frac{1}{2} \right] du. \end{aligned} \quad (2)$$

We substitute $\tilde{u} = \frac{2u}{N-1}$ into [1, Eq. (4.4)] and split the integrals as follows:

$$\begin{aligned} J_2^y(N, x) &\leq \frac{(3\tilde{\gamma})^\varepsilon}{m_f^\varepsilon(x, R_2)(N-1)^{\varepsilon/2}} \left(\int_{\left(\frac{2}{\sqrt{N-1}}, e^{1+\Delta}\right]} + \int_{(e^{1+\Delta}, \infty)} \right) \\ &\quad \frac{\log \frac{(N-1)\tilde{u}}{2} \left[\log \log \frac{(N-1)\tilde{u}}{2} + \frac{1}{2} \right]}{\tilde{u}} [1 - \mathbf{P}_{3,x}(\tilde{u})] d\tilde{u} \end{aligned} \quad (3)$$

Recall from (iv) that we have

$$\begin{aligned} 1 - F_{3,x}^y(\tilde{u}) &= \mathbf{1} \left[\rho(x, y) > r_3(\tilde{u}) \right] \left[1 - \mathbf{P}_{3,x}(\tilde{u}) \right] \\ &= \begin{cases} 1 - \mathbf{P}_{3,x}(\tilde{u}), & \rho(x, y) > r_3(\tilde{u}), \\ 0, & \rho(y, x) \leq r_3(\tilde{u}). \end{cases} \end{aligned}$$

Therefore, when $\rho(y, x) \leq r_3(\tilde{u})$, we have $F_{3,x}^y(\tilde{u}) = F_{N,x}^y(u) = 1$. From equation (2), we obtain that $J_2^y(N, x)$ is equal to zero. Obviously, the boundedness of $J_2^y(N, x)$ is proved in this case. Let us focus on the case $\rho(x, y) > r_3(\tilde{u})$, and then we have

$$F_{3,x}^y(\tilde{u}) = P_{3,x}(\tilde{u}). \quad (4)$$

From equations (3) to (4), we obtain a refined $J_2^y(N, x)$, which can replace $J_2(N, x)$ in [1, Eq. (4.4)]. Analogously, we can amend the related formulas in Step 2 by replacing w , $P_{2,x}(w)$, $F_{2,x}(w)$ with $\tilde{u}/2$, $P_{3,x}(\tilde{u})$, $F_{3,x}(\tilde{u})$. After this amendment, the main results of Bulinski and Dimitrov remain unchanged. This work provides a rigorous correction to the proof of Theorem 2 in the original article. The refined methodology offers researchers a more precise framework for applying similar proof techniques to related problems. Notably, this improved approach has already demonstrated its utility through successful application and extension in the recent work in [3].

Acknowledgements

Yu Sun would like to thank for support of PRIN 2022 ‘‘Prediction and causal inference on the tail index for policy decisions’’ - CUP E53D23006380006.

References

- [1] A. BULINSKI and D. DIMITROV, [Statistical estimation of the Shannon entropy](#), *Acta Math. Sin. (Engl. Ser.)* **35** (2019), no. 1, 17–46.
- [2] L. F. KOZACHENKO and N. N. LEONENKO, Sample estimate of the entropy of a random vector, *Probl. Inf. Transm.* **23** (1987), no. 1, 95–101.
- [3] N. LEONENKO, Y. SUN, and E. TAUFER, Varentropy estimation via nearest neighbor graphs, preprint. Available at <https://arxiv.org/abs/2402.09374>.