Nearest Neighbor Density Functional Estimation From Inverse Laplace Transform

IEEE Transactions on Information Theory, vol. 68, no. 6, pp. 3511–3551, June 2022

Jongha (Jon) Ryu

UCSD $\rightarrow$ MIT

Joint work with Shouvik Ganguly, Young-Han Kim, Yung-Kyun Noh, Daniel D. Lee

KIAS
August 10, 2022

Outline

1. Introduction
2. The Proposed Estimators
3. Theoretical Guarantees and Proofs
   - Asymptotic L_2-consistency
   - L_2-Convergence Rates

Introduction

Problem Setting (1)

- A distribution P over $\mathcal{X} = \mathbb{R}^d$ with density p

- Q. How to characterize a property of a distribution by a single number?

- A. mean, variance, entropy, ...

- An one-density functional: for some $f : \mathbb{R}_+ \to \mathbb{R}$,

$$T_f(p) \triangleq \mathbb{E}_{\mathbf{X}\sim p}[f(p(\mathbf{X}))] = \int f(p(\mathbf{x}))p(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

- **Estimation**: given $\mathbf{X}_{1:m} \sim p$, how to estimate $T_f(p)$?

Problem Setting (2)

- Two distributions P, Q over $\mathcal{X} = \mathbb{R}^d$ with density p, q

  - Q. How to characterize a *dissimilarity* of the distributions?

  - A. KL divergence, f-divergences, integral probability metrics, Wasserstein distance, maximum mean discrepancy, ...

- A two-density functional: for some $f\colon \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$,

$$T_f(p, q) \triangleq \mathbb{E}_{\mathbf{X}\sim p}[f(p(\mathbf{X}), q(\mathbf{X}))] = \int f(p(\mathbf{x}), q(\mathbf{x}))p(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

- **Estimation**: given $\mathbf{X}_{1:m} \sim p$ and $\mathbf{Y}_{1:n} \sim q$, how to estimate $T_f(p, q)$?

- This talk will focus on the one-density case

Motivation

- Wish to construct an L_2-consistent estimator $\hat{T}_f(\mathbf{X}_{1:m})$ of $T_f(p)$, which satisfies

$$\lim_{m\to\infty} \mathbb{E}_{\mathbf{X}_{1:m}\sim p}[(\hat{T}_f(\mathbf{X}_{1:m}) - T_f(p))^2] = 0$$

- A naive, plug-in solution: given a density estimator $\hat{p}(\mathbf{x})$,

$$T_f(p) \approx \tilde{T}_f(p) \triangleq \frac{1}{m}\sum_{i=1}^{m} f(\hat{p}(\mathbf{X}_i))$$

- One can plug-in a k-nearest-neighbors (k-NN) density estimator, but it is NOT consistent for fixed k

- **This paper**: Construct a class of L_2-consistent estimators based on k-NNs

Using Nearest-Neighbors

- Classification, regression: "your neighbors can tell about you"
- Density (functional) estimation: "how far your neighbors tell how crowded you are at"
- Samples $\mathbf{X}_{1:m} \triangleq \{\mathbf{X}_1, \ldots, \mathbf{X}_m\} \sim$ i.i.d. p
- Given a query point $\mathbf{x}$,

$$\mathbf{X}_{(k)}(\mathbf{x}) = \mathbf{X}_{(k)}(\mathbf{x}; \mathbf{X}_{1:m}) \triangleq \text{(the } k\text{-th nearest neighbor)}$$

$$r_k(\mathbf{x}) = r_k(\mathbf{x}; \mathbf{X}_{1:m}) \triangleq \text{(the distance from } \mathbf{x} \text{ to the } k\text{-th nearest neighbor)}$$

- Intuition:

$$p(\mathbf{x}) \times \text{(volume of the } k\text{-NN ball at } \mathbf{x}) \approx \frac{k}{m}$$

- The standard k-NN density estimator:

$$\hat{p}_{km}(\mathbf{x}) \triangleq \frac{k}{m \times \text{(volume of the } k\text{-NN ball at } \mathbf{x})} = \frac{k}{m v_d r_k^d(\mathbf{x})}$$

- Let $v_d \triangleq$ (volume of the unit ball in $\mathbb{R}^d$)

A Plug-in Approach

- Recall

$$\hat{p}_{km}(\mathbf{x}) = \frac{k}{m v_d r_k^d(\mathbf{x})}$$

- **Fact**: $\hat{p}_{km}(\mathbf{x}) \to p(\mathbf{x})$ (weakly consistent) as $m \to \infty$ if $k \to \infty$ with $k = o(m)$
- **Example**: differential entropy ($f(p) = \ln \frac{1}{p}$)

$$h(p) \triangleq \int p(\mathbf{x}) \log \frac{1}{p(\mathbf{x})} \, \mathrm{d}\mathbf{x}$$

- Let's build a plug-in estimator with $\hat{p}_{km}(\mathbf{x})$:

$$\tilde{h}_k(\mathbf{X}_{1:m}) \triangleq \frac{1}{m} \sum_{i=1}^{m} \log \frac{1}{\hat{p}_{km}(\mathbf{X}_i)}$$

- For fixed $k \in \mathbb{N}$, it is NOT consistent!

Kozachenko–Leonenko Estimator

- We need to correct its bias...
- The (generalized) Kozachenko–Leonenko estimator [Kozachenko and Leonenko, 1987, Singh et al., 2003, Goria et al., 2005]:

$$\begin{aligned}\hat{T}_{\mathsf{KL}}^{(k)}(\mathbf{X}_{1:m}) &= \tilde{T}_f(\hat{p}_{km}) + \ln k - \Psi(k) \\ &= \frac{1}{m}\sum_{i=1}^{m} \ln \frac{1}{\hat{p}_{km}(\mathbf{X}_i)} + \ln k - \Psi(k),\end{aligned} \tag{1}$$

  where $\Psi(x) \triangleq \Gamma'(x)/\Gamma(x)$ denotes the digamma function [Korn and Korn, 2000]
- **Fact 1**: $\hat{T}_{\mathsf{KL}}^{(k)}(\mathbf{X}_{1:m})$ is L_2-consistent for any fixed $k \geq 1$ [Tsybakov and van der Meulen, 1996, Goria et al., 2005, Gao et al., 2018]
- **Fact 2**: $\hat{T}_{\mathsf{KL}}^{(k=1)}(\mathbf{X}_{1:m})$ is minimax-rate-optimal for a certain class of densities [Jiao et al., 2018]
- **Q.** Given a general f, how can we build a L_2-consistent estimator based on fixed-k-NNs?

A Brief History of Bias-Corrected Plug-in Estimators

- In a similar spirit, L_2-consistent fixed-k or fixed-(k,l) plug-in estimators with proper additive or multiplicative bias correction were proposed and analyzed for KL divergence [Wang et al., 2009], Rényi entropies [Leonenko et al., 2008], Rényi divergences [Póczos and Schneider, 2011], and several other divergences of a specific polynomial form [Póczos et al., 2012]:

$$\tilde{T}_f^{\mathsf{aff}}(\hat{p}) = a_k \tilde{T}_f(\hat{p}) + b_k, \tag{2}$$

$$\tilde{T}_f^{\mathsf{aff}}(\hat{p},\hat{q}) = a_{kl} \tilde{T}_f(\hat{p},\hat{q}) + b_{kl}, \tag{3}$$

  where (a_k, b_k) and (a_{kl}, b_{kl}) determine functional-specific bias correction
- Singh and Póczos [2016] analyzed a bias-corrected estimatorof the following form

$$\tilde{T}_{b\circ f}(\hat{p}) = \frac{1}{m}\sum_{i=1}^{m} b_{km}(f(\hat{p}_{km}(\mathbf{X}_i))) \tag{4}$$

  and established L_2-consistency for fixed k with convergence rate **if** there exists a bias-correcting function b_{km} that satisfies a strict condition depending on p

The Proposed Estimators

Our General Recipe

- Given f and $k \geq 1$, define

$$\hat{T}_f^{(k)}(\mathbf{X}_{1:m}) \triangleq \frac{1}{m}\sum_{i=1}^{m} \phi_k(U_{km}(\mathbf{X}_i)), \tag{5}$$

where we denote a *normalized* volume of the k-NN ball at $\mathbf{x}$ as

$$U_{km}(\mathbf{x}) \triangleq U_k(\mathbf{x}; \mathbf{X}_{1:m}) \triangleq m v_d r_k^d(\mathbf{x})$$

and **choose** a function ϕ_k so that

$$\lim_{m\to\infty} \mathbb{E}[\hat{T}_f^{(k)}(\mathbf{X}_{1:m})] = T_f(p) \quad \text{(asymptotic unbiasedness)}$$

An Useful Asymptotic Property

- A *normalized* volume of the k-NN ball at $\mathbf{x}$:

$$U_{km}(\mathbf{x}) \triangleq U_k(\mathbf{x}; \mathbf{X}_{1:m}) \triangleq m v_d r_k^d(\mathbf{x})$$

- A Gamma random variable $U \sim \mathsf{G}(\alpha, \beta)$ with shape parameter $\alpha > 0$ and rate parameter $\beta > 0$ is defined by its density

$$f_{\alpha,\beta}(u) \triangleq \frac{\beta^\alpha}{\Gamma(\alpha)} u^{\alpha-1} e^{-\beta u}, \quad u > 0$$

Proposition ⋆

For any $k \in \mathbb{N}$, for p-almost every $\mathbf{x}$,

$$U_{km}(\mathbf{x}) \xrightarrow{d} U_{k\infty}(\mathbf{x}) \text{ as } m \to \infty,$$

where $U_{k\infty}(\mathbf{x}) \sim \mathsf{G}(k, p(\mathbf{x}))$

Distribution of the k-NN Distance

Lemma 2.1

The cdf of $r_{km}(\mathbf{x})$ is

$$F_{r_{km}(\mathbf{x})}(r) = \Pr\{B_{m,\mathsf{P}(\mathbb{B}(\mathbf{x},r))} \geq k\}$$

Proof.

$$\begin{aligned}
F_{r_{km}(\mathbf{x})}(r) &= \Pr\{r_{km}(\mathbf{x}) \leq r\} \\
&= \Pr\Big\{|\{i \in [m] \colon \mathbf{X}_i \in \mathbb{B}(\mathbf{x}, r)\}| \geq k\Big\} \\
&= \Pr\{B_{m,\mathsf{P}(\mathbb{B}(\mathbf{x},r))} \geq k\} \qquad \square
\end{aligned}$$

Proof of Proposition $\star$

- Fix $\mathbf{x} \in \mathbb{R}^d$ and $u > 0$
- Since $F_{U_{km}(\mathbf{x})}(u) = F_{r_{km}(\mathbf{x})}(\varrho(\frac{u}{m}))$, we have $F_{U_{km}(\mathbf{x})}(u) = \Pr\{B_{m,P_m} \ge k\}$ from Lemma 2.1, where $P_m \triangleq \mathsf{P}(\mathbb{B}(\mathbf{x}, \varrho(\frac{u}{m})))$
- By the Lebesgue differentiation theorem (see, e.g., Rudin [1987]), for a.e. $\mathbf{x}$,

$$\lim_{m\to\infty} mP_m = \lim_{m\to\infty} u\frac{\mathsf{P}(\mathbb{B}(\mathbf{x}, \varrho(\frac{u}{m})))}{\mathsf{Vol}(\mathbb{B}(\mathbf{x}, \varrho(\frac{u}{m})))} = up(\mathbf{x})$$

- Therefore, for each $i = 0, \ldots, k-1$, we have

$$\binom{m}{i} P_m^i (1-P_m)^{m-i} = \frac{i!}{m^i}\binom{m}{i}(1-P_m)^{m-i}\frac{(mP_m)^i}{i!} \xrightarrow{m\to\infty} e^{-up(\mathbf{x})}\frac{(up(\mathbf{x}))^i}{i!},$$

since $\lim_{m\to\infty} \frac{i!}{m^i}\binom{m}{i} = 1$ and $\lim_{m\to\infty}(1-P_m)^{m-i} = e^{-up(\mathbf{x})}$

- This leads us to concludes that

$$\lim_{m\to\infty} \Pr\{U_{km}(\mathbf{x}) > u\} = \sum_{i=0}^{k-1} e^{-up(\mathbf{x})}\frac{up(\mathbf{x})^i}{i!} = \Pr\{U_{k\infty}(\mathbf{x}) > u\} \qquad \square$$

- Observe

$$\mathbb{E}_{\mathbf{X}_{1:m}}[\hat{T}_f^{(k)}(\mathbf{X}_{1:m})] = \mathbb{E}_{\mathbf{X}_{1:m}}\Big[\frac{1}{m}\sum_{i=1}^{m}\phi_k(U_{km}(\mathbf{X}_i))\Big]$$

$$= \mathbb{E}_{\mathbf{X}_m}[\phi_k(U_{km}(\mathbf{X}_m))] = \mathbb{E}_{\mathbf{X}}[\phi_k(U_{k,m-1}(\mathbf{X}))] \qquad (*)$$

- Since $U_{k,m-1}(\mathbf{x}) \xrightarrow{d} U_{k\infty}(\mathbf{x}) \sim \mathsf{G}(k, p(\mathbf{x}))$ by Proposition $\star$, we **expect**

$$\lim_{m\to\infty}\mathbb{E}_{\mathbf{X}_{1:m}}[\hat{T}_f^{(k)}(\mathbf{X}_{1:m})] \overset{(*)}{=} \lim_{m\to\infty}\mathbb{E}[\phi_k(U_{k,m-1}(\mathbf{X}))]$$

$$\overset{(?)}{=} \mathbb{E}[\phi_k(U_{k\infty}(\mathbf{X}))]$$

- Hence, the desired unbiasedness might be attained if we choose ϕ_k such that

$$\mathbb{E}[\phi_k(U_{k\infty}(\mathbf{X}))] = T_f(p)$$

$$\Leftrightarrow \int \mathbb{E}[\phi_k(U_{k\infty}(\mathbf{x}))]p(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int f(p(\mathbf{x}))p(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

$$\Leftarrow \mathbb{E}[\phi_k(U)] = f(p) \quad \text{for } U \sim \mathsf{G}(k,p)$$

The Estimator Function via Inverse Laplace Transform

- Given f and $k \geq 1$, we choose ϕ_k such that for every $p > 0$, if $U \sim \mathsf{G}(k,p)$, then

$$\begin{aligned} f(p) &= \mathbb{E}[\phi_k(U)] \\ &= \int_0^\infty \phi_k(u) \frac{p^k}{\Gamma(k)} u^{k-1} e^{-pu} \, \mathrm{d}u \\ &= \frac{p^k}{\Gamma(k)} \mathcal{L}\{u^{k-1}\phi_k(u)\}(p), \end{aligned}$$

where $\mathcal{L}\{\cdot\}$ represents the *one-sided Laplace transform*, defined as

$$\mathcal{L}\{g(u)\}(p) \triangleq \int_0^\infty g(\tilde{u}) e^{-p\tilde{u}} \, \mathrm{d}\tilde{u}$$

- Rearranging the terms leads to defining the *estimator function* ϕ_k for f with parameter k:

$$\phi_k(u) \triangleq \frac{\Gamma(k)}{u^{k-1}} \mathcal{L}^{-1}\left\{ \frac{f(p)}{p^k} \right\}(u)$$

The Proposed Estimator

- Given f and k, define

$$\hat{T}_f^{(k)}(\mathbf{X}_{1:m}) \triangleq \frac{1}{m}\sum_{i=1}^{m} \phi_k(U_{km}(\mathbf{X}_i)),$$

  where

$$\phi_k(u) \triangleq \frac{\Gamma(k)}{u^{k-1}} \mathcal{L}^{-1}\left\{\frac{f(p)}{p^k}\right\}(u),$$

  if the inverse Laplace transform exists

- This estimator **unifies** almost all existing bias-corrected estimators, and is **new** for several other density functionals
- This is different from the existing bias-correction approaches such as [Singh and Póczos, 2016] and more widely applicable

The Proposed Estimator: Examples

Table: Examples of functionals of one density and their estimator functions $\phi_k(u)$. The last column presents a pair of exponents (a_k, b_k) of the polynomial envelope of the estimator function $\phi_k(u)$. The constant ϵ, if any, can be chosen as an arbitrarily small positive number. For the first three examples, $k > -a_k$ is required to guarantee the existence of the corresponding inverse Laplace transform.

Name	$T_f(p) = \mathbb{E}_p[f(p)]$	$\phi_k(u) = \frac{\Gamma(k)}{u^{k-1}} \mathcal{L}^{-1}\Big\{\frac{f(p)}{p^k}\Big\}(u)$	(a_k, b_k)
Differential entropy	$\mathbb{E}\Big[\ln \frac{1}{p}\Big]$	$\ln u - \Psi(k)$	$(-\epsilon, \epsilon)$
α-entropy ($\alpha \geq 0$)	$\mathbb{E}[p^{\alpha-1}]$	$\frac{\Gamma(k)}{\Gamma(k-\alpha+1)}\Big(\frac{1}{u}\Big)^{\alpha-1}$	$(1-\alpha, 1-\alpha)$
Logarithmic α-entropy ($\alpha > 0$)	$\mathbb{E}\Big[p^{\alpha-1} \ln \frac{1}{p}\Big]$	$\frac{\Gamma(k)}{\Gamma(k-\alpha+1)} u^{-\alpha+1}(\ln u - \Psi(k-\alpha+1))$	$(1-\alpha-\epsilon, 1-\alpha+\epsilon)$
Exponential (α,β)-entropy ($\alpha > 0, \beta \geq 0$)	$\mathbb{E}[p^{\alpha-1} e^{-\beta p}]$	$\frac{\Gamma(k)}{\Gamma(k-\alpha+1)} \frac{(u-\beta)^{k-\alpha}}{u^{k-1}} \mathbb{1}_{[\beta,\infty)}(u)$	$(0, 1-\alpha)$ for $k \geq \alpha$

The Proposed Estimator with Two Densities

- Recall $\mathbf{X}_{1:m} \sim p$ and $\mathbf{Y}_{1:n} \sim q$
- Given f and k, l, define

$$\hat{T}_f^{(k,l)}(\mathbf{X}_{1:m}, \mathbf{Y}_{1:n}) \triangleq \frac{1}{m} \sum_{i=1}^{m} \phi_{kl}(U_{km}(\mathbf{X}_i), V_{ln}(\mathbf{X}_i)),$$

where

$$\phi_{kl}(u, v) \triangleq \frac{\Gamma(k)\Gamma(l)}{u^{k-1} v^{l-1}} \mathcal{L}^{-1}\left\{ \frac{f(p, q)}{p^k q^l} \right\}(u, v).$$

if the inverse Laplace transform exists

The Proposed Estimator with Two Densities: Examples

Table: Examples of functionals of two densities and their estimator functions $\phi_{kl}(u,v)$.

Name	$T_f(p,q)=\mathbb{E}_p[f(p,q)]$	$\phi_{kl}(u,v)=\frac{\Gamma(k)\Gamma(l)}{u^{k-1}v^{l-1}}\mathcal{L}^{-1}\left\{\frac{f(p,q)}{p^k q^l}\right\}(u,v)$	(a_{kl},b_{kl}); $(\tilde{a}_{kl},\tilde{b}_{kl})$
KL divergence	$\mathbb{E}\left[\ln\frac{p}{q}\right]$	$\ln\frac{v}{u}+\Psi(k)-\Psi(l)$	$(-\epsilon,\epsilon)$; $(-\epsilon,\epsilon)$
α-divergence ($\alpha>0$)	$\mathbb{E}\left[\left(\frac{p}{q}\right)^{\alpha-1}\right]$	$\frac{\Gamma(k)\Gamma(l)}{\Gamma(k-\alpha+1)\Gamma(l+\alpha-1)}\left(\frac{v}{u}\right)^{\alpha-1}$	$(1-\alpha,1-\alpha)$; $(\alpha-1,\alpha-1)$
Logarithmic α-divergence ($\alpha>0$)	$\mathbb{E}\left[\left(\frac{p}{q}\right)^{\alpha-1}\ln\frac{p}{q}\right]$	$\frac{\Gamma(k)\Gamma(l)}{\Gamma(k-\alpha+1)\Gamma(l+\alpha-1)}\left(\frac{v}{u}\right)^{\alpha-1}\times$ $\left(\ln\frac{v}{u}+\Psi(k-\alpha+1)-\Psi(l+\alpha-1)\right)$	$(1-\alpha-\epsilon,1-\alpha+\epsilon)$; $(\alpha-1-\epsilon,\alpha-1+\epsilon)$
Le Cam distance	$\mathbb{E}\left[\frac{(p-q)^2}{2p(p+q)}\right]$	$2\binom{k+l-2}{k-1}^{-1}\left\{\sum_{j=0}^{l-1}\binom{k+l-2}{k-1+j}\left(-\frac{u}{v}\right)^j-\right.$ $\left.\left(-\frac{u}{v}\right)^{l-1}\left(1-\frac{v}{u}\right)^{k+l-2}\mathbb{1}_{[v,\infty)}(u)\right\}$	$(-k+1,l-1)$; $(-l+1,k-1)$
Entropy difference (Q $\ll$ P)	$\mathbb{E}\left[\ln\frac{1}{p}-\frac{q}{p}\ln\frac{1}{q}\right]$	$\frac{(l-1)}{k}\frac{u}{v}(\Psi(l-1)-\ln v)-(\Psi(k)-\ln u)$	$(-\epsilon,1)$; $(-1-\epsilon,-1+\epsilon)$
Reverse KL divergence (Q $\ll$ P)	$\mathbb{E}\left[\frac{q}{p}\ln\frac{q}{p}\right]$	$\frac{l-1}{k}\frac{u}{v}\left(\ln\frac{u}{v}+\Psi(l-1)-\Psi(k+1)\right)$	$(1-\epsilon,1+\epsilon)$; $(-1-\epsilon,-1+\epsilon)$
Jensen–Shannon divergence (Q $\ll$ P)	$\mathbb{E}\left[\frac{1}{2}\ln\frac{2p}{p+q}+\frac{q}{2p}\ln\frac{2q}{p+q}\right]$	(omitted; see paper)	$(-k+1,l-1)$; $(-l+1,k-1)$

Theoretical Guarantees and Proofs

Polynomial Envelope

- Wish to analyze the estimator in a unified manner for general functionals
- **Idea**: abstract tail behaviors of the estimator function $\phi_k(u)$ (i.e., how $\phi_k(u)$ varies when $u \downarrow 0$ and $u \uparrow \infty$) by a pair of constants $(a_k, b_k) \in \mathbb{R}^2$ such that

$$|\phi_k(u)| \lesssim \psi_{a_k,b_k}(u),$$

where we define a piecewise polynomial function $\psi_{a,b} \colon \mathbb{R}_+ \to \mathbb{R}$ for $a, b \in \mathbb{R}$ as

$$\psi_{a,b}(u) \triangleq \begin{cases} u^a & \text{if } 0 < u \le 1, \\ u^b & \text{if } u > 1 \end{cases} \tag{6}$$

- As a gets larger, $\psi_{a,b}(u)$ decays faster as $u \downarrow 0$
  $\Rightarrow$ a quantifies the amount of contribution of low density values *through* $\phi_k(u)$
- As b gets smaller, $\psi_{a,b}(u)$ decays faster as $u \uparrow \infty$
  $\Rightarrow$ b quantifies the amount of contribution of high density values *through* $\phi_k(u)$
- We will establish stronger statements for functionals with larger a and smaller b

Examples

Example 3.1 (Differential entropy [Kozachenko and Leonenko, 1987])

For $f(p) = \ln(1/p)$ and any $k \geq 1$, we can compute

$$\phi_k(u) = \ln u - \Psi(k).$$

As a bound on the estimator function $\phi_k(u)$, we consider

$$|\phi_k(u)| \lesssim |\ln u| + 1 \lesssim \psi_{-\epsilon,\epsilon}(u)$$

for any arbitrarily small $\epsilon > 0$ throughout the paper[a]

[a] A finer analysis without relying on the polynomial bound $\psi_{-\epsilon,\epsilon}(u)$ may lead to a marginal improvement in the resulting performance guarantee [Gao et al., 2018, Bulinski and Dimitrov, 2019a,b].

Example 3.2 (α-entropy [Leonenko et al., 2008])

- For $f(p) = p^{\alpha-1}$ $(\alpha \geq 0)$, we refer to the density functional $T_f(p) = \int p^\alpha(\mathbf{x})\,\mathrm{d}\mathbf{x}$ as the *α-entropy*
- In the literature, this functional appears in Rényi [1961] entropy $h_\alpha(p) = (\ln T_f(p))/(1-\alpha)$ and Harvda and Charvat [1967] or Tsallis [1988] entropy $\tilde{h}_\alpha(p) = (1 - T_f(p))/(\alpha - 1)$
- For any $k \in \mathbb{N}$ such that $k > \alpha - 1$, we can compute

$$\phi_k(u) = \frac{\Gamma(k)}{\Gamma(k-\alpha+1)}\left(\frac{1}{u}\right)^{\alpha-1},$$

which allows the tight polynomial bound

$$|\phi_k(u)| \lesssim \psi_{1-\alpha,1-\alpha}(u)$$

Asymptotic L_2-consistency

Local Extremal Operators

- The standard simplifying assumptions: there exist $c > 0$ and $C > 0$ such that

$$c \le p(\mathbf{x}) \le C \text{ for any } \mathbf{x} \in \mathsf{supp}(p)$$

- Instead, we consider weaker conditions than the boundedness assumptions, adopting conditions from [Bulinski and Dimitrov, 2019a,b].
- For each $r > 0$, define the local extremal operators on $\mathbb{R}^d$ for a density p by

$$\text{(local maximal operator)} \quad M_r p(\mathbf{x}) \triangleq \sup_{r' \in (0,r]} \frac{\mathsf{P}(\mathbb{B}(\mathbf{x}, r'))}{\mathsf{Vol}(\mathbb{B}(\mathbf{x}, r'))},$$

$$\text{(local minimal operator)} \quad m_r p(\mathbf{x}) \triangleq \inf_{r' \in (0,r]} \frac{\mathsf{P}(\mathbb{B}(\mathbf{x}, r'))}{\mathsf{Vol}(\mathbb{B}(\mathbf{x}, r'))}$$

  - $m_r p(\mathbf{x}) \le p(\mathbf{x}) \le M_r p(\mathbf{x})$
  - By the Lebesgue differentiation theorem, $M_r p(\mathbf{x}) \downarrow p(\mathbf{x})$ and $m_r p(\mathbf{x}) \uparrow p(\mathbf{x})$ as $r \downarrow 0$, for p-a.e. $\mathbf{x}$
  - For each $r > 0$, $\mathbf{x} \mapsto M_r p(\mathbf{x})$ and $\mathbf{x} \mapsto m_r p(\mathbf{x})$ are lower- and upper-semicontinuous, respectively, and so are Borel measurable [Bulinski and Dimitrov, 2019a,b]

Functionals Based on Local Extremal Operators

- Given a non-decreasing function $\xi\colon \mathbb{R}_+ \to \mathbb{R}_+$, for densities p and $\tilde{p}$, define

$$\text{(upper bound on } p) \quad W(p,\tilde{p};\vartheta,r) \triangleq \int p(\mathbf{x})(M_r\tilde{p}(\mathbf{x}))^{\vartheta}\,\mathrm{d}\mathbf{x},$$

$$\text{(lower bound on } p) \quad w(p,\tilde{p};\xi,\vartheta,r) \triangleq \int p(\mathbf{x})\xi((m_r\tilde{p}(\mathbf{x}))^{-\vartheta})\,\mathrm{d}\mathbf{x},$$

$$\text{(bounded support)} \quad R(p,\tilde{p};\xi,\vartheta,r) \triangleq \iint_{\rho(\mathbf{x},\mathbf{y})>r} p(\mathbf{x})\tilde{p}(\mathbf{y})\xi(\upsilon^{\vartheta}(\rho(\mathbf{x},\mathbf{y})))\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y}$$

  for each $\vartheta > 0$ and $r > 0$
- **Note**: $R(p,\tilde{p};\xi,\vartheta,r) \to 0$ as $r \to \infty$
  - As the tails of p and $\tilde{p}$ decay faster, so does $R(p,\tilde{p};\xi,\vartheta,r)$
  - In particular, if p and $\tilde{p}$ have bounded support, then $R(p,\tilde{p};\xi,\vartheta,r) = 0$ for $r \gg 1$
- **Note**: W, w, and R become larger as ϑ increases

Regularity Conditions

- Given $k \in \mathbb{N}$ and $(a, b) \in \mathbb{R}^2$, consider the following conditions

($\mathbf{U}_{p\tilde{p}}$; k, a) Either $a \geq 0$, or if $a < 0$, then there exists $r > 0$ such that $W(p, \tilde{p}; k, r) < \infty$

($\mathbf{L}_{p\tilde{p}}$; ξ, b) Either $b \leq 0$, or if $b > 0$, then there exists $r > 0$ such that $w(p, \tilde{p}; \xi, b, r) < \infty$ and

$$\limsup_{m\to\infty} \xi(m^b) R\Big(p, \tilde{p}; \xi, b, \varrho\Big(\frac{\kappa_m}{m}\Big)\Big) < \infty \tag{7}$$

for some κ_m such that $\kappa_m/m \to \infty$ and $(\ln \kappa_m)/m \to 0$ as $m \to \infty$

- **Recall**: the polynomial tail exponents a and b of $\phi_k(u)$ quantify the amount of contribution of high and low density values to the estimator, resp.
- Hence, $a \leftrightarrow W$ that captures the upper boundedness of the density; while $b \leftrightarrow w$ and R that quantify the lower boundedness
- **Note**: as a gets larger, k gets smaller, and b gets smaller, conditions **($\mathbf{L}_{pp}$; ξ, b)** and **($\mathbf{U}_{pp}$; k, a)** become weaker, thus encompassing a larger class of densities.

L_2-consistency

- Let Ξ be the class of non-decreasing functions $\xi\colon \mathbb{R}_+ \to \mathbb{R}_+$ such that
  1. $\xi(t)/t \to \infty$ as $t \to \infty$;
  2. $\xi(t_1 t_2) \le \xi(t_1)\xi(t_2)$ for any $x, y > t_0$ for some $t_0 \in \mathbb{R}_+$;
  3. $\omega(\xi) \triangleq \inf\{\eta > 1\colon \xi(t)/t^\eta \to 0 \text{ as } t \to \infty\} < \infty$
  - **Examples**: $\xi_1(t) = (t \ln t) \vee 0 \in \Xi$ with $t_0 = e$ and $\omega(\xi_1) = 1$;
    $\xi_2(t) = t^\alpha \in \Xi$ for $\alpha > 1$ with $t_0 = 0$ and $\omega(\xi_2) = \alpha$
- Bias–variance decomposition of mean-squared error (MSE):

$$\begin{aligned}\mathbb{E}[(\hat{T}_f(\mathbf{X}_{1:m}) - T_f(p))^2] &= (\mathbb{E}[\hat{T}_f(\mathbf{X}_{1:m})] - T_f(p))^2 + \mathrm{Var}(\hat{T}_f(\mathbf{X}_{1:m})) \\ &= (\text{bias})^2 + (\text{variance})\end{aligned}$$

- Analyzing the bias is often involved, and controlling the variance is relatively easier

L_2-consistency (Cont'd)

Theorem 3.3 (Vanishing bias)

For $T_f(\cdot)$, if ϕ_k is continuous and p satisfies **(U_{pp}; k, a)** *and* **(L_{pp}; ξ, b)** *with some function $\xi \in \Xi$, then the estimator (5) with fixed $k > -\omega(\xi)a$ is asymptotically unbiased*

Theorem 3.4 (Vanishing variance)

For $T_f(\cdot)$, if p satisfies **(U_{pp}; k, a)** *and* **(L_{pp}; ξ, b)** *with $\xi(t) = t^2$, the variance of the estimator (5) with fixed $k > -2a$ converges to zero as $m \to \infty$*

Corollary 3.5 (L_2-consistency)

For $T_f(\cdot)$, if ϕ_k is continuous and p satisfies **(U_{pp}; k, a)** *and* **(L_{pp}; ξ, b)** *with $\xi(t) = t^2$, then the estimator (5) with fixed $k > -2a$ is L_2-consistent*

Examples

Example 3.6 (Differential entropy; Example 3.1 contd.)

- **Recall**: for any $k \in \mathbb{N}$, $|\phi_k(u)| \lesssim \psi_{-\epsilon,\epsilon}(u)$ for arbitrarily small $\epsilon > 0$
- By Corollary 3.5, the estimator (5) is L_2-consistent if p satisfies that **($\mathbf{U}_{pp}$; $k, -\epsilon$)** and **($\mathbf{L}_{pp}$; ξ, ϵ)** with $\xi(t) = t^2$ for some $\epsilon > 0$
- We note that the condition (7) in **($\mathbf{L}_{pp}$; ξ, ϵ)** can be relaxed to a milder condition in which there exist some $\delta, R > 0$ such that

$$\iint_{\rho(\mathbf{x},\mathbf{y})>R} p(\mathbf{x})p(\mathbf{y}) |\ln v(\rho(\mathbf{x},\mathbf{y}))|^{\delta} \,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y} < \infty$$

  by performing a similar analysis based on the upper bound $|\phi_k(u)| \lesssim |\ln u| + 1$
- This recovers a similar result reported in [Bulinski and Dimitrov, 2019b]

Examples

Example 3.7 (α-entropy; Example 3.2 contd.)

- Recall that for any $k \in \mathbb{N}$, $|\phi_k(u)| \lesssim \psi_{1-\alpha,1-\alpha}(u)$
- For $\alpha > 1$, since $b = 1 - \alpha < 0$, the estimator with fixed $k > 2(\alpha - 1)$ is L_2-consistent if p satisfies **(U$_{pp}$; k, a)**, which slightly generalizes the upper-boundedness condition and the requirement $k > 2\alpha - 1$ assumed in [Leonenko et al., 2008]
- For $\alpha < 1$, since $a = 1 - \alpha > 0$, the estimator with fixed $k \geq 1$ is L_2-consistent if p satisfies **(L$_{pp}$; ξ, b)** with $\xi(t) = t^2$, for examples, if p is bounded away from zero and supported over a hyperrectangle (Leonenko and Pronzato [2010] reported the L_2-consistency of the estimator for densities satisfying alternative conditions when $\alpha < 1$)

Proof of Theorem 3.3 (Vanishing Bias)

- Since ϕ_k is continuous, from Proposition $\star$, we have $\phi_k(U_{k,m-1}(\mathbf{X}_m)) \xrightarrow{d} \phi_k(U_{k\infty}(\mathbf{X}))$ as $m \to \infty$ by the continuous mapping theorem, where $U_{k\infty}(\mathbf{x})$ is a $\mathsf{G}(k, p(\mathbf{x}))$ random variable, independent of $\mathbf{X} \sim p$ for P-a.e. $\mathbf{x}$
- **Recall**: a collection of random variables $(X_i)_{i\in I}$ is said to be *uniformly integrable (U.I.)* if for any $\epsilon > 0$, there exists $K \geq 0$ such that $\sup_{i\in I} \mathbb{E}[X_i \mathbb{1}_{[K,\infty)}(X_i)] \leq \epsilon$
- **Proposition**: if $(X_n)_{n\in\mathbb{N}}$ is U.I. and $X_n \xrightarrow{d} X_\infty$, then

$$\lim_{n\to\infty} \mathbb{E}[X_n] = \mathbb{E}[X_\infty]$$

- Hence, if the sequence of random variables $(\phi_k(U_{k,m-1}(\mathbf{X}_m)))_{m\geq 1}$ is **U.I.**, the asymptotic unbiasedness readily follows:

$$\lim_{m\to\infty} \mathbb{E}[\hat{T}_f^{(k)}(\mathbf{X}_{1:m})] = \lim_{m\to\infty} \mathbb{E}[\phi_k(U_{k,m-1}(\mathbf{X}_m))]$$
$$\stackrel{\text{(U.I.?)}}{=} \mathbb{E}[\phi_k(U_{k\infty}(\mathbf{X}))] = T_f(p)$$

Proof of Theorem 3.3 (Vanishing Bias) (Cont'd)

- To show the uniform integrability of $(\phi_k(U_{k,m-1}(\mathbf{X}_m)))_{m\geq 1}$, we invoke:

Lemma 3.8 (De la Vallée Poussin theorem [Borkar, 1995, Theorem 1.3.4])

A collection of random variables $(X_i)_{i\in I}$ is uniformly integrable
$\Leftrightarrow \exists$ *a non-decreasing function* $\xi\colon \mathbb{R}_+ \to \mathbb{R}_+$ *such that*

1. $\sup_{i\in I} \mathbb{E}[\xi(|X_i|)] < \infty$; *and*
2. $\lim_{t\to\infty} \frac{\xi(t)}{t} = \infty$

- (This is why we introduced the class of functions Ξ)
- The second condition is satisfied since $\xi \in \Xi$ by assumption
- Only need to check the first condition
- We will plug-in $X_i \leftarrow \phi_k(U_{k,m-1}(\mathbf{X}_m))$

Proof of Theorem 3.3 (Vanishing Bias) (Cont'd)

- Observe that we have

$$\mathbb{E}[\xi(|\phi_k(U_{k,m-1}(\mathbf{X}_m))|)] = \int p(\mathbf{x})\mathbb{E}[\xi(|\phi_k(U_{k,m-1}(\mathbf{x}))|)]\,\mathrm{d}\mathbf{x}$$
$$\lesssim \int p(\mathbf{x})\mathbb{E}[\xi(\psi_{a,b}(U_{km}(\mathbf{x})))]\,\mathrm{d}\mathbf{x} \qquad \text{(polynomial envelope)}$$

- Since $\xi \in \Xi$, we have $-\int_0^1 u^k\,\mathrm{d}\xi(u^{a\wedge 0}) < \infty$ for $k > -\omega(\xi)a$ and $\int_0^\infty e^{-t}\xi(t^{b\vee 0})\,\mathrm{d}t < \infty$, and thus we can apply Lemma 3.9 (next slide), which yields

$$\limsup_{m\to\infty} \mathbb{E}[\xi(|\phi_k(U_{k,m-1}(\mathbf{X}_m))|)] \lesssim \limsup_{m\to\infty} \int p(\mathbf{x})\mathbb{E}[\xi(\psi_{a,b}(U_{km}(\mathbf{x})))]\,\mathrm{d}\mathbf{x} < \infty$$

- This ensures the uniform integrability of $(\phi_k(U_{k,m-1}(\mathbf{X}_m)))_{m\geq 1}$ by the de la Vallée Poussin theorem, and thus concludes the proof □

A Technical Lemma

Lemma 3.9

Assume that $-\int_0^1 u^k \,\mathrm{d}\xi(u^{a\wedge 0}) < \infty$ *and* $\int_0^\infty e^{-t}\xi(t^{b\vee 0})\,\mathrm{d}t < \infty$.
If the density p *satisfies* **($\mathbf{U}_{pp}$; k, a)** *and* **($\mathbf{L}_{pp}$; ξ, b)**, *we have*

$$\limsup_{m\to\infty} \int p(\mathbf{x})\mathbb{E}[\xi(\psi_{a,b}(U_{km}(\mathbf{x})))]\,\mathrm{d}\mathbf{x} < \infty$$

- The proof is rather involved
- **Idea**: Break the inner integral over $(0,\infty)$ over four intervals $(0,1),(1,\nu_m),(\nu_m,\kappa_m),(\kappa_m,\infty)$, and analyze each term by bounding the cumulative density function of $U_{km}(\mathbf{x})$

A Generic Lemma for Bounding Variance

Lemma 3.10 ([Singh and Póczos, 2016])

For a given function $\phi\colon \mathbb{R}_+ \to \mathbb{R}$, *let* $\zeta_k(\mathbf{x}|\mathbf{x}_{1:m}) \triangleq \phi(r_k(\mathbf{x}|\mathbf{x}_{1:m}))$ *for any points* $\mathbf{x}, \mathbf{x}_{1:m}$ *in the* d*-dimensional Euclidean space* $(\mathbb{R}^d, \|\cdot\|)$. *Let*

$$\Phi(\mathbf{x}_{1:m}) = \frac{1}{m}\sum_{i=1}^{m} \zeta_k(\mathbf{x}_i|\mathbf{x}_{1:m}^{\sim i}). \tag{8}$$

If the samples $\mathbf{X}_{1:m}$ *are i.i.d., then*

$$\begin{aligned}\mathrm{Var}(\Phi(\mathbf{X}_{1:m})) \le \frac{2(1+k\gamma_d)}{m}\{&(2k+1)\mathbb{E}[\zeta_k^2(\mathbf{X}_m|\mathbf{X}_{1:m-1})]\\ &+ 2k\mathbb{E}[\zeta_{k+1}^2(\mathbf{X}_m|\mathbf{X}_{1:m-1})]\},\end{aligned}$$

where $\gamma_d \in \mathbb{N}$ *is a constant which depends only on* d

Proof Techniques for the Variance Lemma

Lemma 3.11 (Efron–Stein inequality [Efron and Stein, 1981, Steele, 1986])

Let $X_1, \ldots, X_n$ be independent random variables, and let $g(X_{1:n}) = g(X_1, \ldots, X_n)$ be a square-integrable function of $X_1, \ldots, X_n$.
Then if $X_1', \ldots, X_n'$ are independent copies of $X_1, \ldots, X_n$, we have

$$\mathrm{Var}(g(X_{1:n})) \le \frac{1}{2} \sum_{i=1}^{n} \mathbb{E}\big[|g(X_{1:n}) - g(X_{1:i-1} X_i' X_{i+1:n})|^2\big]$$

Lemma 3.12 ([Biau and Devroye, 2015, Lemma 20.6])

In $(\mathbb{R}^d, \|\cdot\|)$, there exists a constant $\gamma_d > 0$ which depends only on d such that for any $m \in \mathbb{N}$ and for any distinct points $\mathbf{x}, \mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^d$,

$$\sum_{i=1}^{m} \mathbb{1}_{N_k(\mathbf{x}_i | \mathbf{x}_{1:m}^{\sim i}, \mathbf{x})}(\mathbf{x}) \le k\gamma_d$$

Proof of Theorem 3.4 (Vanishing Variance)

- By Lemma 3.10 for the Euclidean space $(\mathbb{R}^d, \|\cdot\|)$, we have

$$\begin{aligned}\mathrm{Var}(\hat{T}_f^{(k)}) \le \frac{2(1+k\gamma_d)}{m}\{&(2k+1)\mathbb{E}[\phi_k^2(U_{k,m-1}(\mathbf{X}_m))]\\ &+ 2k\mathbb{E}[\phi_k^2(U_{k+1,m-1}(\mathbf{X}_m))]\},\end{aligned}$$

  where γ_d is a constant which only depends on d; see Lemma 3.10
- Since $\xi(t) = t^2$ and $k > -2a$ imply that $-\int_0^1 u^k \,\mathrm{d}\xi(u^{a\wedge 0}) < \infty$ and $\int_0^\infty e^{-t}\xi(t^{b\vee 0})\,\mathrm{d}t < \infty$, we can apply Lemma 3.9, which ensures for $k' \in \{k, k+1\}$ that

$$\limsup_{m\to\infty} \mathbb{E}[\phi_k^2(U_{k',m-1}(\mathbf{X}_m))] < \infty$$

- It establishes $\mathrm{Var}(\hat{T}_f^{(k)}) = O(m^{-1})$ for m sufficiently large $\square$

L_2-Convergence Rates

Boundedness Conditions

Upper bound

($\mathbf{U}_p$) there exists $0 < C_p < \infty$ such that $p(\mathbf{x}) \le C_p$ almost everywhere (a.e.)

Lower bound

($\mathbf{L1}_p$) there exists $c_p > 0$ such that $p(\mathbf{x}) \ge c_p$ for $\mathbf{x} \in \mathsf{supp}(p)$;

($\mathbf{L2}_p$) the support of p is bounded;

($\mathbf{L3}_p$) there exists $r > 0$ such that

$$\eta_p \triangleq \inf_{\mathbf{x} \in \mathsf{supp}(p)} \inf_{r' \in (0,r]} \frac{\mathsf{Vol}(\mathbb{B}(\mathbf{x}, r') \cap \mathsf{supp}(p))}{\mathsf{Vol}(\mathbb{B}(\mathbf{x}, r'))} > 0$$

Boundedness Conditions

Remark 3.1

- The upper-boundedness condition **($\mathbf{U}_p$)** implies **($\mathbf{U}_{pp}; k, a$)**, since $M_r p(\mathbf{x}) \le C_p < \infty$ for every $\mathbf{x} \in \mathbb{R}^d$ and any $r > 0$
- Also, the lower-boundedness conditions **($\mathbf{L1}_p$)**, **($\mathbf{L2}_p$)**, and **($\mathbf{L3}_p$)** imply **($\mathbf{L}_{pp}; \xi, b$)** for any nonnegative function ξ, since for $b > 0$ we have

$$\begin{aligned} w(p, p; \xi, b, r) &= \int p(\mathbf{x}) \xi((m_r p(\mathbf{x}))^{-b}) \, \mathrm{d}\mathbf{x} \\ &\le \int p(\mathbf{x}) \xi((\eta_p c_p)^{-b}) \, \mathrm{d}\mathbf{x} = \xi((\eta_p c_p)^{-b}) < \infty \end{aligned}$$

for some $r > 0$ by **($\mathbf{L1}_p$)**, **($\mathbf{L2}_p$)**, and **($\mathbf{L3}_p$)**, and

$$R(p, p; \xi, b, \varrho(\kappa_m / m))) = 0$$

for m sufficiently larger than an absolute constant, by **($\mathbf{L2}_p$)**

Variance Rate

Theorem 3.13 (Variance rate)

For $T_f(\cdot)$, if p satisfies ***(U_p)***, ***($L1_p$)***, ***($L2_p$)***, *and* ***($L3_p$)***, *then the estimator (5) with fixed $k > -2a$ satisfies*

$$\mathrm{Var}(\hat{T}_f^{(k)}) = O(m^{-1}) \tag{9}$$

Proof of Theorem 3.13 (Variance Rate)

- **Recall**: By Lemma 3.10 for the Euclidean space $(\mathbb{R}^d, \|\cdot\|)$, we have

$$\mathrm{Var}(\hat{T}_f^{(k)}) \le \frac{2(1+k\gamma_d)}{m}\{(2k+1)\mathbb{E}[\phi_k^2(U_{k,m-1}(\mathbf{X}_m))] + 2k\mathbb{E}[\phi_k^2(U_{k+1,m-1}(\mathbf{X}_m))]\},$$

  where γ_d is a constant which only depends on d; see Lemma 3.10
- Since the boundedness conditions **(U$_p$)**, **(L1$_p$)**, **(L2$_p$)**, and **(L3$_p$)** imply stronger conditions than **(U$_{pp}$; k, a)** and **(L$_{pp}$; ξ, b)** (see Remark 3.1), we can prove:

Lemma 3.14

Assume that $-\int_0^1 u^k \,\mathrm{d}\xi(u^{a\wedge 0}) < \infty$ *and* $\int_0^\infty e^{-t}\xi(t^{b\vee 0})\,\mathrm{d}t < \infty$.
If the density p *satisfies* ***(U$_p$)***, ***(L1$_p$)***, ***(L2$_p$)***, *and* ***(L3$_p$)***, *we have*

$$\sup_{m\ge 1}\int p(\mathbf{x})\mathbb{E}[\xi(\psi_{a,b}(U_{km}(\mathbf{x})))]\,\mathrm{d}\mathbf{x} < \infty$$

- Hence, the variance rate directly follows by setting $\xi(t) = t^2$ $\square$

Smoothness Conditions

Definition 3.15

For $\sigma > 0$, a function $g\colon \mathbb{R}^d \to \mathbb{R}$ is said to be *σ-Hölder continuous over an open subset* $\Omega \subseteq \mathbb{R}^d$ if g is continuously differentiable over Ω up to order $\kappa \triangleq \lceil \sigma \rceil - 1$ and

$$L(g;\Omega) \triangleq \sup_{\substack{\mathbf{r}\in\mathbb{Z}_+^d \\ |\mathbf{r}|=\kappa}} \sup_{\substack{\mathbf{y},\mathbf{z}\in\Omega \\ \mathbf{y}\neq\mathbf{z}}} \frac{|\partial^{\mathbf{r}} g(\mathbf{y}) - \partial^{\mathbf{r}} g(\mathbf{z})|}{\|\mathbf{y}-\mathbf{z}\|^{\beta}} < \infty, \tag{10}$$

where $\beta \triangleq \sigma - \kappa$. Here we use a multi-index notation (see, e.g., [Folland, 2013, Ch. 8]), that is, $|\mathbf{r}| \triangleq r_1 + \cdots + r_d$ for $\mathbf{r} \in \mathbb{Z}_+^d$ and $\partial^{\mathbf{r}} g(\mathbf{x}) \triangleq \frac{\partial^{\kappa} g(\mathbf{x})}{\partial x_1^{r_1} \cdots \partial x_d^{r_d}}$

Smoothness Conditions (Cont'd)

- Due to the lower-boundedness condition **(L1$_p$)**, the density is NOT smooth on the boundary of the support
- Hence, we assume a smoothness condition on the underlying density only over the interior of its support and impose a separate regularity condition on the boundary:

Smoothness

(S$_p$) The density p is σ_p-Hölder continuous over the interior of $\mathrm{supp}(p)$ for $\sigma_p \in (0, 2]$;

(B$_p$) the boundary of $\mathrm{supp}(p)$ has finite $(d-1)$-dimensional Hausdorff measure [Folland, 2013]

Bias Rate

Theorem 3.16 (Bias rate)

For $T_f(\cdot)$, if p satisfies the conditions ***(U$_p$)***, ***(L1$_p$)***, ***(L2$_p$)***, ***(L3$_p$)***, ***(S$_p$)***, *and* ***(B$_p$)***, *then the estimator* (5) *with fixed $k > -a$ satisfies*

$$\left|\mathbb{E}[\hat{T}_f^{(k)}] - T_f(p)\right| = \tilde{O}(m^{-\lambda(\sigma_p, a, k)}), \tag{11}$$

$$\textit{where} \quad \lambda(\sigma, a, k) = \begin{cases} \frac{1}{d}(\sigma \wedge 1)(\frac{k+a}{k-1}) & \textit{if } a \leq -\frac{\sigma}{d} - 1, \\ \frac{1}{d}(\sigma \wedge \frac{k+a}{k-1}) & \textit{if } -\frac{\sigma}{d} - 1 < a \leq -1, \\ \frac{1}{d}(\sigma \wedge 1) & \textit{if } a > -1 \end{cases} \tag{12}$$

Remark 3.2

- The rate exponent λ increases as the lower-tail-polynomial exponent a increases, or equivalently, the estimator function $\phi_k(u)$ converges to 0 faster as $u \downarrow 0$
- If a is independent of k (which is true for most cases), the rate exponent λ becomes larger with larger k

Proof of Theorem 3.16 (Bias Rate)

- First note that $U_{km}(\mathbf{X}_1), \ldots, U_{km}(\mathbf{X}_m)$ are identically distributed, and $U_{km}(\mathbf{X}_m) = U_{k,m-1}(\mathbf{X}_m)$. Hence, we can write

$$\begin{aligned}
\mathbb{E}[\hat{T}_f^{(k)}] &= \mathbb{E}[\phi_k(U_{k,m-1}(\mathbf{X}_m))] \\
&= \int \mathbb{E}[\phi_k(U_{k,m-1}(\mathbf{X}_m))|\mathbf{X}_m = \mathbf{x}]p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\
&= \int \mathbb{E}[\phi_k(U_{k,m-1}(\mathbf{x}))]p(\mathbf{x})\,\mathrm{d}\mathbf{x}, \qquad (13)
\end{aligned}$$

  where the last equality holds since $\mathbf{X}_m$ and $\mathbf{X}_{1:m-1}$ are independent. Recall from Proposition $\star$ that $U_{km}(\mathbf{x}) \xrightarrow{d} U_{k\infty}(\mathbf{x}) \sim \mathsf{G}(k, p(\mathbf{x}))$ for p-a.e. $\mathbf{x}$

- Thus, by the construction of $\phi_k(u)$, we can express the density functional as

$$T_f(p) = \int f(p(\mathbf{x}))p(\mathbf{x})\,\mathrm{d}\mathbf{x} = \int \mathbb{E}[\phi_k(U_{k\infty}(\mathbf{x}))]p(\mathbf{x})\,\mathrm{d}\mathbf{x}$$

Proof of Theorem 3.16 (Bias Rate) (Cont'd)

- Applying the triangle inequality, we first have

$$\begin{aligned}\left|\mathbb{E}[\hat{T}_f^{(k)}] - T_f(p)\right| &\le \int p(\mathbf{x}) |\mathbb{E}[\phi_k(U_{k,m-1}(\mathbf{x})) - \phi_k(U_{k\infty}(\mathbf{x}))]| \,\mathrm{d}\mathbf{x} \\ &= \int p(\mathbf{x}) \left| \int_0^\infty \phi_k(u)(\rho_{U_{k,m-1}(\mathbf{x})}(u) - \rho_{U_{k\infty}(\mathbf{x})}(u)) \,\mathrm{d}u \right| \mathrm{d}\mathbf{x} \qquad (14)\end{aligned}$$

- Pick any $0 \le \tau_m \le 1 \le \nu_m < \infty$, which are to be determined later as functions of k, a, d, and σ_p
- Break the inner integral and apply the polynomial bound $|\phi_k(u)| \lesssim \psi_{a,b}(u)$ with the triangle inequality to obtain

$$\left|\mathbb{E}[\hat{T}_f^{(k)}] - T_f(p)\right| \lesssim I_{\mathsf{out},1} + I_{\mathsf{in},1} + I_{\mathsf{in},2} + I_{\mathsf{out},2}, \qquad (15)$$

Proof of Theorem 3.16 (Bias Rate) (Cont'd)

- where

$$I_{\text{out},1} \triangleq \mathbb{E}_p[I_{\text{out},1}(\mathbf{X})] = \mathbb{E}_p\Big[\int_0^{\tau_m} \psi_{a,b}(u)(\rho_{U_{k,m-1}(\mathbf{X})}(u) + \rho_{U_{k\infty}(\mathbf{X})}(u))\,\mathrm{d}u\Big],$$

$$I_{\text{in},1} \triangleq \mathbb{E}_p[I_{\text{in},1}(\mathbf{X})] = \mathbb{E}_p\Big[\int_{\tau_m}^{1} \psi_{a,b}(u)|\rho_{U_{k,m-1}(\mathbf{X})}(u) - \rho_{U_{k\infty}(\mathbf{X})}(u)|\,\mathrm{d}u\Big],$$

$$I_{\text{in},2} \triangleq \mathbb{E}_p[I_{\text{in},2}(\mathbf{X})] = \mathbb{E}_p\Big[\int_{1}^{\nu_m} \psi_{a,b}(u)|\rho_{U_{k,m-1}(\mathbf{X})}(u) - \rho_{U_{k\infty}(\mathbf{X})}(u)|\,\mathrm{d}u\Big], \quad \text{and}$$

$$I_{\text{out},2} \triangleq \mathbb{E}_p[I_{\text{out},2}(\mathbf{X})] = \mathbb{E}_p\Big[\int_{\nu_m}^{\infty} \psi_{a,b}(u)(\rho_{U_{k,m-1}(\mathbf{X})}(u) + \rho_{U_{k\infty}(\mathbf{X})}(u))\,\mathrm{d}u\Big]$$

- The *inner bias* terms $I_{\text{in},1}$ and $I_{\text{in},2}$ can be controlled under the conditions **(U$_p$)**, **(S$_p$)**, and **(B$_p$)**
- The *outer bias* terms $I_{\text{out},1}$ and $I_{\text{out},2}$ can be bounded under the conditions **(U$_p$)**, **(L1$_p$)**, **(L2$_p$)**, and **(L3$_p$)**
- After putting the bounds together, a proper choice of the break points (τ_m, ν_m) concludes the proof □

Technical Lemmas for the Proof

- The following lemma establishes a rate of convergence of a Poisson binomial random variable $B_{m,Q/m} \sim \mathrm{Binom}(m, Q/m)$ to a Poisson random variable $P_Q \sim \mathrm{Poisson}(Q)$ in distribution

Lemma 3.17 (Generalization of [Gao et al., 2018, Lemma 5])

For any $Q, k = o(\sqrt{m})$ as $m \to \infty$, there exists a constant $C_0 > 0$ such that for m sufficiently large

$$\left|\Pr\{B_{m,\frac{Q}{m}} = k\} - \Pr\{P_Q = k\}\right| \le C_0 \frac{Q^k e^{-Q}}{k!} \frac{(k^2 + Q^2)}{m}.$$

Technical Lemmas for the Proof (Cont'd)

Lemma 3.18 (Generalization of [Gao et al., 2018, Lemma 4])

If a density p is σ_p-Hölder continuous with constant $L > 0$ over $\mathbb{B}(\mathbf{x}, R)$ for $\mathbf{x} \in \mathbb{R}^d$ and some $\sigma_p \in [0, 2]$, we have for any $0 < r < R$,

$$\left| \frac{\mathsf{P}(\mathbb{B}(\mathbf{x}, r))}{\mathsf{Vol}(\mathbb{B}(\mathbf{x}, r))} - p(\mathbf{x}) \right| \le \frac{d}{\sigma_p + d} L r^{\sigma_p},$$

$$\left| \frac{\mathrm{d}\,\mathsf{P}(\mathbb{B}(\mathbf{x}, r))}{\mathrm{d}\mathsf{Vol}(\mathbb{B}(\mathbf{x}, r))} - p(\mathbf{x}) \right| \le L r^{\sigma_p}.$$

- The convergence speed of $U_{km}(\mathbf{x}) \xrightarrow{d} U_{k\infty}(\mathbf{x})$ can be quantified in terms of a gap between the densities using this lemma and the order of smoothness σ_p of p
- However, $O(r^{\sigma_p})$ in Lemma 3.18 cannot be improved further beyond $O(r^2)$ [Han et al., 2020]
- In general, nonnegative kernel-based methods cannot exploit $\sigma_p > 2$

On the Proof of Theorem 3.16 (Bias Rate)

Remark 3.3

- The key step in this analysis is the decomposition in (15), which is based on the construction of the estimator from its asymptotic unbiasedness
- By considering only the polynomial tail behavior of each estimator function and using (15), our analysis can deal with a general functional in a simple, unified manner
- The rest of the bias analysis, that is, bounding the four bias terms, closely follows and naturally extends that of [Gao et al., 2018] for a truncated version of the Kozachenko–Leonenko estimator of differential entropy

MSE Rate

Corollary 3.19

Under the same assumptions in Theorem 3.16, then the estimator (5) with fixed $k > -2a$ satisfies

$$\mathbb{E}\Big[\big(\hat{T}_f^{(k)} - T_f(p)\big)^2\Big] = \tilde{O}(m^{-2\lambda(\sigma_p, a, k)} + m^{-1}) \tag{16}$$

Remark 3.4

- For $d \geq 2$, the bias bound always dominates the variance bound so that the MSE is bounded as $\tilde{O}(m^{-2\lambda})$
- For $d = 1$, the variance bound may dominate the bias bound, depending on σ_p and a

Example 3.20 (Differential entropy; Example 3.1 contd.)

- Recall from Example 3.1 that $|\phi_k(u)| \lesssim \psi_{-\epsilon,\epsilon}(u)$ for any arbitrarily small $\epsilon > 0$. Suppose that p satisfies the conditions **(U$_p$)**, **(L1$_p$)**, **(L2$_p$)**, **(L3$_p$)**, **(S$_p$)**, and **(B$_p$)**, in Theorem 3.16 with some $\sigma_p \in (0, 2]$
- Then we have the bias exponent $\lambda = \sigma_p/d$ as in the third case of (12) and the variance exponent of 1 from (9)
- Consequently, by Corollary 3.19 the MSE of our estimator is bounded as $\tilde{O}(m^{-2(\sigma_p \wedge 1)/d} + m^{-1})$. This result recovers the same MSE rate of a truncated Kozachenko–Leonenko estimator in [Gao et al., 2018] for $\sigma_p = 2$
- We remark that Gao et al. [2018] reported a lower bound $\Omega(m^{-\frac{16}{d+8}} + m^{-1})$ for estimating differential entropy under $\sigma = 2$, and indeed the convergence rate is **not** minimax optimal!

Example 3.21 (α-entropy; Example 3.2 contd.)

- Recall from Example 3.2 that $|\phi_k(u)| \lesssim \psi_{1-\alpha,1-\alpha}(u)$ for any $k \in \mathbb{N}$ such that $k > \alpha - 1$
- Hence, for densities satisfying the conditions **(U$_p$)**, **(L1$_p$)**, **(L2$_p$)**, **(L3$_p$)**, **(S$_p$)**, and **(B$_p$)**, the MSE of our estimator (5) with fixed $k > 2(\alpha - 1)$ is bounded as (16) with the bias rate exponent

$$\lambda(\sigma_p, a, k) = \begin{cases} \frac{1}{d}(\sigma_p \wedge 1) & \text{if } \alpha < 2, \\ \frac{1}{d}(\sigma_p \wedge \frac{k+1-\alpha}{k-1}) & \text{if } 2 \leq \alpha < 2 + \frac{\sigma_p}{d}, \\ \frac{1}{d}(\sigma_p \wedge 1)(\frac{k+1-\alpha}{k-1}) & \text{if } \alpha \geq 2 + \frac{\sigma_p}{d} \end{cases} \tag{17}$$

- Note that similar convergence rates can be established for the logarithmic α-entropy and the exponential (α, β)-entropy

On the Rate Suboptimality

Remark 3.5

- An estimator of a given density functional is said to be *minimax optimal* if its MSE for the worst-case density is no larger than that of any other estimator
- In general, the established convergence rates in MSE in this paper are **not** minimax optimal [Singh and Póczos, 2014a,b, Krishnamurthy et al., 2014, Kandasamy et al., 2015] due to the suboptimal bias rates; see, e.g., Example 3.20
- For the special case of differential entropy, we note that Jiao et al. [2018] established an asymptotic minimax optimality of the Kozachenko–Leonenko estimator for smooth densities of order $\sigma \in (0, 2]$ *over a torus* (no boundary condition), matching the lower bound of [Han et al., 2020] up to a polylogarithmic factor

More Technical Results

- Convergence rates for smooth densities of unbounded support
- Functionals of two densities
- Adaptive choices of k: Using $k = \Theta((\ln m)^{1.01})$ may improve the rates

Concluding Remarks

- The established convergence rates are not minimax optimal; see Remark 3.5

Q. Can we extend the analysis of [Jiao et al., 2018] and establish a minimax optimality of the estimator under the torus condition?

- As noted earlier, the proposed estimators **cannot** adapt to a higher order of smoothness $\sigma > 2$, due to the inherent limitation of positive-valued kernels

- One possible solution to both problems is the ensemble approach [Sricharan et al., 2013, Moon and Hero, 2014] that takes a weighted average of multiple estimators based on the asymptotic bias expansion of each density functional estimator

Q. Is an ensemble version of the estimators minimax-rate optimal?

See Berrett and Samworth [2019] for a weighted version of the proposed divergence functional estimator from this paper with local minimax optimality

References I

Thomas B Berrett and Richard J Samworth. Efficient two-sample functional estimation and the super-oracle phenomenon. *arXiv preprint arXiv:1904.09347*, 2019.

Gérard Biau and Luc Devroye. *Lectures on the Nearest Neighbor Method*. Springer International Publishing, 2015.

Vivek S Borkar. *Probability theory: an advanced course*. Springer Science & Business Media, 1995.

Alexander Bulinski and Denis Dimitrov. Statistical estimation of the shannon entropy. *Acta Mathematica Sinica, English Series*, 35(1):17–46, 2019a.

Alexander Bulinski and Denis Dimitrov. Statistical estimation of the Kullback–Leibler divergence. *arXiv preprint arXiv:1907.00196*, 2019b.

B Efron and C Stein. The Jackknife Estimate of Variance. *Ann. Statist.*, 9(3):586–596, 1981.

Gerald B Folland. *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons, 2013.

References II

Weihao Gao, Sewoong Oh, and Pramod Viswanath. Demystifying fixed k-nearest neighbor information estimators. *IEEE Trans. Inf. Theory*, 64(8):5629–5661, August 2018.

M. N. Goria, N. N. Leonenko, V. V. Mergel, and P. L.Novi Inverardi. A new class of random vector entropy estimators and its applications in testing statistical hypotheses. *J. Nonparametr. Statist.*, 17(3):277–297, 2005.

Yanjun Han, Jiantao Jiao, Tsachy Weissman, and Yihong Wu. Optimal rates of entropy estimation over Lipschitz balls. *Ann. Statist.*, 48(6):3228–3250, 2020.

J Harvda and F Charvat. Quantification method of classification processes. concept of structural α-entropy. *Kybernetika (Prague)*, 3:30–35, 1967.

Jiantao Jiao, Weihao Gao, and Yanjun Han. The Nearest Neighbor Information Estimator is Adaptively Near Minimax Rate-Optimal. In *Adv. Neural Inf. Proc. Syst.*, volume 31, December 2018.

References III

Kirthevasan Kandasamy, Akshay Krishnamurthy, Barnabas Poczos, Larry A Wasserman, and James M Robins. Nonparametric von Mises estimators for entropies, divergences and mutual informations. In *Adv. Neural Inf. Proc. Syst.*, volume 28, pages 397–405, 2015.

Granino Arthur Korn and Theresa M Korn. *Mathematical Handbook for Scientists and Engineers: Definitions, Theorems, and Formulas for Reference and Review*. Courier Corporation, 2000.

L F Kozachenko and Nikolai N Leonenko. Sample estimate of the entropy of a random vector. *Probl. Inf. Transm.*, 23(2):9–16, 1987. (Russian).

Akshay Krishnamurthy, Kirthevasan Kandasamy, Barnabás Póczos, and Larry Wasserman. Nonparametric estimation of Rényi divergence and friends. In *Proc. Int. Conf. Mach. Learn.*, pages 919–927, 2014.

References IV

Nikolai Leonenko, Luc Pronzato, and Vippal Savani. A class of Rényi information estimators for multidimensional densities. *Ann. Statist.*, 36(5):2153–2182, October 2008. Corrected in Leonenko, N. and Prozanto, L. (2010). Correction: A class of Rényi information estimators for multidimensional densities. *Ann. Statist.* **38** 3837–3838.

Nikolai N Leonenko and Luc Pronzato. Correction: A class of Rényi information estimators for multidimensional densities. *Ann. Statist.*, 38(6):3837–3838, 2010.

Kevin R Moon and Alfred O Hero. Ensemble estimation of multivariate f-divergence. In *Proc. IEEE Int. Symp. Inf. Theory*, pages 356–360. IEEE, 2014.

Barnabás Póczos and Jeff G Schneider. On the Estimation of alpha-Divergences. *Int. Conf. Artif. Int. Statist.*, pages 609–617, 2011.

Barnabás Póczos, Liang Xiong, Dougal J Sutherland, and Jeff Schneider. Nonparametric kernel estimators for image classification. In *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.*, pages 2989–2996. IEEE, June 2012.

Alfréd Rényi. On measures of entropy and information. In *Proc. 4th Berkeley Sympos. Math. Statist. Probab.*, volume 1, pages 547–761. Univ. California Press, Berkeley, 1961.

References V

Walter Rudin. *Real and Complex Analysis*. McGraw-Hill Education, 1987.

Harshinder Singh, Neeraj Misra, Vladimir Hnizdo, Adam Fedorowicz, and Eugene Demchuk. Nearest neighbor estimates of entropy. *Am. J. Math. Manag. Sci.*, 23(3-4): 301–321, 2003.

Shashank Singh and Barnabás Póczos. Generalized exponential concentration inequality for rényi divergence estimation. In *Proc. Int. Conf. Mach. Learn.*, pages 333–341. PMLR, 2014a.

Shashank Singh and Barnabás Póczos. Exponential concentration of a density functional estimator. In *Adv. Neural Inf. Proc. Syst.*, volume 27, pages 3032–3040, 2014b.

Shashank Singh and Barnabás Póczos. Finite-sample analysis of fixed-k nearest neighbor density functional estimators. In *Adv. Neural Inf. Proc. Syst.*, volume 29, pages 1217–1225. Curran Associates, Inc., 2016.

Kumar Sricharan, Dennis Wei, and Alfred O Hero. Ensemble estimators for multivariate entropy estimation. *IEEE Trans. Inf. Theory*, 59(7):4374–4388, 2013.

References VI

J. Michael Steele. An Efron–Stein Inequality for Nonsymmetric Statistics. *Ann. Statist.*, 14(2):753–758, June 1986.

Constantino Tsallis. Possible generalization of boltzmann-gibbs statistics. *J. of Statist. Phys.*, 52(1-2):479–487, 1988.

A. B. Tsybakov and E. C. van der Meulen. Root-n Consistent Estimators of Entropy for Densities with Unbounded Support. *Scand. Statist. Theory Appl.*, 1996.

Qing Wang, Sanjeev R Kulkarni, and Sergio Verdú. Divergence estimation for multidimensional densities via k-nearest-neighbor distances. *IEEE Trans. Inf. Theory*, 55 (5):2392–2405, 2009.