Nearest Neighbor Density Functional Estimation from Inverse Laplace Transform

J. Jon Ryu*, *Student Member, IEEE,* Shouvik Ganguly*, *Member, IEEE,* Young-Han Kim, *Fellow, IEEE,* Yung-Kyun Noh, *Member, IEEE* and Daniel D. Lee *Fellow, IEEE*

***Abstract*—A new approach to L_2-consistent estimation of a general density functional using k-nearest neighbor distances is proposed, where the functional under consideration is in the form of the expectation of some function f of the densities at each point. The estimator is designed to be asymptotically unbiased, using the convergence of the normalized volume of a k-nearest neighbor ball to a Gamma distribution in the large-sample limit, and naturally involves the inverse Laplace transform of a scaled version of the function f. Some instantiations of the proposed estimator recover existing k-nearest neighbor based estimators of Shannon and Rényi entropies and Kullback–Leibler and Rényi divergences, and discover new consistent estimators for many other functionals such as logarithmic entropies and divergences. The L_2-consistency of the proposed estimator is established for a broad class of densities for general functionals, and the convergence rate in mean squared error is established as a function of the sample size for smooth, bounded densities.**

***Index Terms*—Density functional estimation, information measure, nearest neighbor, inverse Laplace transform.**

I. Introduction

THIS paper studies the problem of estimating an entropy functional of the form

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

where $f\colon \mathbb{R}_+ \to \mathbb{R}$ is a given function and p is a probability density over $\mathbb{R}^d$. Table I lists examples of f and the corresponding functional T_f. The goal is to estimate $T_f(p)$ based on independent and identically distributed (i.i.d.) samples $\mathbf{X}_{1:m} = (\mathbf{X}_1, \ldots, \mathbf{X}_m)$ from p by forming an estimator $\hat{T}_f^m(\mathbf{X}_{1:m})$ that converges to $T_f(p)$ in L_2 as the sample size m grows to infinity, that is,

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

More generally, let $f\colon \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ and consider a divergence functional

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

of a pair of probability densities p and q over $\mathbb{R}^d$. Table II lists examples of f and the corresponding T_f. In this case, the main problem is to construct an estimator $\hat{T}_f^{m,n}(\mathbf{X}_{1:m}, \mathbf{Y}_{1:n})$ based on i.i.d. samples $\mathbf{X}_{1:m}$ from p and $\mathbf{Y}_{1:n}$ from q, independent of each other, such that

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

Consistent estimation of such quantities, such as Shannon's differential entropy ($f = \ln(1/p)$), (exponentiated) Rényi α-entropies ($f = p^{\alpha-1}$), Kullback–Leibler (KL) divergence ($f = \ln(p/q)$), Hellinger distance ($f = \sqrt{q/p}$), (exponentiated) Rényi α-divergences ($f = p^{\alpha-1}q^{-\alpha}$), and Jensen–Shannon divergence (see Table II), is a problem of considerable practical interest, having wide-ranging applications in parameter estimation [1, 2], goodness-of-fit testing [3, 4, 5], quantization [6], independent component analysis [7, 8, 9], texture classification [10, 11], design of experiments [12, 13], pattern recognition [14, 15, 16, 17], clustering and feature selection [16, 18, 19, 20], and statistical inference [21]. In addition, divergence estimates can be used as measures of distance between two distributions and thus can generalize distance-based algorithms for metric spaces to the space of probability distributions; see, for example, [22, 23] and the references therein.

One of the most basic and prominent nonparametric approaches is the k-nearest neighbor (k-NN) based method, which is appealing since its hyperparameter tuning is relatively simple and is computationally efficient, especially when k is held fixed, independent of the sample sizes m and n. In this paper, we propose a new, universal design principle of a L_2-consistent k-NN based estimator for a wide class of the density functionals $T_f(p)$ and $T_f(p,q)$ based on the inverse Laplace transform, which generalizes many existing estimators which have been developed and analyzed separately. Based on the proposed mathematical framework, we establish the

Manuscript received July 31, 2020; revised January 14, 2021; accepted January 26, 2022. This work was supported in part by the National Science Foundation under Grant CCF-1911238 and the National Research Foundation of Korea Grant (NRF/MSIT2017R1E1A1A03070945) and Hanyang University (HY-2019).The associate editor coordinating the review of this article and approving it for publication was O. Johnson. (Corresponding author: J. Jon Ryu.)

*J. J. Ryu and S. Ganguly contributed equally to this work.

J. J. Ryu and Y.-H. Kim are with the Department of Electrical and Computer Engineering, University of California, San Diego, La Jolla, CA 92093 USA (e-mail: jongharyu@ucsd.edu; yhk@ucsd.edu).

S. Ganguly was with the Department of Computer Engineering, University of California, San Diego, La Jolla, CA 92093 USA. He is now affiliated at XCOM Labs, San Diego, CA 92121 USA (e-mail: sganguly@xcom-labs.com).

Y.-H. Kim is with the Department of Electrical and Computer Engineering, University of California at San Diego, La Jolla, CA 92093 USA and Gauss Labs Inc, Seoul, South Korea (e-mail: yhk@ucsd.edu).

Y.-K. Noh is with Department of Computer Science, Hanyang University, Seoul 04763, Republic of Korea and School of Computational Sciences, Korea Institute for Advanced Study, Seoul 02455, Republic of Korea (e-mail: nohyung@hanyang.ac.kr).

D. D. Lee is with Cornell Tech, New York, NY 10044 USA and Global AI Center for Samsung Research (e-mail: ddl46@cornell.edu).

TABLE I

EXAMPLES OF FUNCTIONALS OF ONE DENSITY AND THEIR ESTIMATOR FUNCTIONS $\phi_k(u)$. A REFERENCE IS GIVEN WHENEVER AN ESTIMATOR ALREADY EXISTS IN THE LITERATURE. 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. HERE, $\Psi(\alpha)$ DENOTES THE DIGAMMA FUNCTION [24]; SEE ALSO EXAMPLE III.1.

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 [25, 26, 5] (Examples III.1, III.5, III.7, III.9, V.1)	$\mathbb{E}\Big[\ln \frac{1}{p}\Big]$	$\ln u - \Psi(k)$	$(-\epsilon, \epsilon)$
α-entropy [27] ($\alpha \geq 0$) (Examples III.2, III.6, III.8, III.10, V.2)	$\mathbb{E}[p^{\alpha-1}]$	$\frac{\Gamma(k)}{\Gamma(k-\alpha+1)}\Big(\frac{1}{u}\Big)^{\alpha-1}$	$(1-\alpha, 1-\alpha)$
Logarithmic α-entropy ($\alpha > 0$) (Example III.3)	$\mathbb{E}\Big[p^{\alpha-1} \ln \frac{1}{p}\Big]$	$\frac{\Gamma(k)}{\Gamma(k-\alpha+1)} u^{-\alpha+1}(\ln u - \Psi(k-\alpha+1))$	$(1-\alpha-\epsilon, 1-\alpha+\epsilon)$
Exponential (α,β)-entropy ($\alpha > 0, \beta \geq 0$) (Example III.4)	$\mathbb{E}[p^{\alpha-1}e^{-\beta p}]$	$\frac{\Gamma(k)}{\Gamma(k-\alpha+1)} \frac{(u-\beta)^{k-\alpha}}{u^{k-1}} \mathbb{1}_{[\beta,\infty)}(u)$	$(0, 1-\alpha)$ for $k \geq \alpha$

consistency and the rate of convergence in MSE of the density functional estimator under fairly general regularity conditions, by extending and simplifying the existing analyses of the KL estimator by Bulinski and Dimitrov [28, 29] and Gao et al. [30].

A. The proposed single-density functional estimators

Suppose that a metric $\rho\colon \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}_+$ is associated with the d-dimensional space $\mathbb{R}^d$. Given samples $\mathbf{X}_{1:m}$ and a point $\mathbf{x} \in \mathbb{R}^d$, we denote the k-NN distance of $\mathbf{x}$ from the samples by $r_{km}(\mathbf{x}) := r_k(\mathbf{x}|\mathbf{X}_{1:m})$ for $k \leq m$. Here, $r_k(\mathbf{x}|A)$ denotes the k-NN distance of $\mathbf{x}$ from a set $A \subseteq \mathbb{R}^d$, where the distance tie is broken arbitrarily. The key statistic in this paper is a normalized volume

$$U_{km}(\mathbf{x}) := U_k(\mathbf{x}|\mathbf{X}_{1:m}) := m\,\lambda_{\mathrm{Leb}}(\mathbb{B}(\mathbf{x}, r_k(\mathbf{x}|\mathbf{X}_{1:m}))) \tag{I.1}$$

of the k-NN ball centered at $\mathbf{x}$ with respect to $\mathbf{X}_{1:m}$. Here and henceforth, λ_{Leb} denotes the Lebesgue measure over $\mathbb{R}^d$, $\mathbb{B}(\mathbf{x}, r) := \{\mathbf{y} \in \mathbb{R}^d\colon \rho(\mathbf{x},\mathbf{y}) < r\}$ denotes the open ball of radius $r > 0$ centered at $\mathbf{x} \in \mathbb{R}^d$, and $\overline{\mathbb{B}}(\mathbf{x}, r)$ denotes the closure of $\mathbb{B}(\mathbf{x}, r)$. When the k-NN distance r_k is evaluated at one of the samples $\mathbf{x} = \mathbf{X}_i$ $(1 \leq i \leq m)$, we define it as $r_k(\mathbf{X}_i|\mathbf{X}_{1:i-1}\mathbf{X}_{i+1:m})$ to exclude the trivial zero distance. Consequently, we use the convention

$$U_{km}(\mathbf{X}_i) := (m-1)\,\lambda_{\mathrm{Leb}}(\mathbb{B}(\mathbf{x}, r_k(\mathbf{x}|\mathbf{X}_{1:i-1}\mathbf{X}_{i+1:m}))).$$

Note that under this convention, we have

$$U_{km}(\mathbf{X}_m) = U_{k,m-1}(\mathbf{X}_m). \tag{I.2}$$

Let $\mathsf{G}(\alpha,\beta)$ denote the Gamma distribution with shape parameter $\alpha > 0$ and rate parameter $\beta > 0$, whose density is

$$\frac{\beta^\alpha}{\Gamma(\alpha)} u^{\alpha-1} e^{-\beta u}, \quad u \geq 0.$$

Here $\Gamma(\alpha) := \int_0^\infty x^{\alpha-1}e^{-x}\,\mathrm{d}x$ denotes the Gamma function. The following fact on the asymptotic distribution of $U_{km}(\mathbf{x})$ is well known [26, 5, 27]. The proof is presented in Appendix B-B for completeness.

Proposition I.1. *Suppose that $k \geq 1$ is a fixed integer, and let $\mathbf{X}_{1:m}$ be i.i.d. samples drawn from p on $\mathbb{R}^d$. Then, for almost every $\mathbf{x}$, $U_{km}(\mathbf{x})$ converges to a $\mathsf{G}(k, p(\mathbf{x}))$ random variable in distribution as m goes to infinity.*

This general convergence result is the cornerstone of the design of our estimator. To be more specific, for functionals of one density p, consider an estimator of the form

$$\hat{T}_f^{(k)}(\mathbf{X}_{1:m}) = \frac{1}{m}\sum_{i=1}^m \phi_k(U_{km}(\mathbf{X}_i)) \tag{I.3}$$

that depends on the samples only through the k-NN distance evaluated at each of them. As a necessary condition for the L_2-consistency of this estimator, the function ϕ_k should be chosen such that

$$\lim_{m\to\infty} \mathbb{E}[\hat{T}_f^{(k)}] = T_f(p),$$

that is, the estimator is asymptotically unbiased. On the one hand, since $\mathbf{X}_{1:m}$ are identically distributed, we have, from (I.2) and (I.3), that $\mathbb{E}[\hat{T}_f^{(k)}] = \mathbb{E}[\phi_k(U_{k,m-1}(\mathbf{X}_m))]$, and thus the desired asymptotic unbiasedness for a *fixed* k can be expressed equivalently as

$$\lim_{m\to\infty} \mathbb{E}[\phi_k(U_{k,m-1}(\mathbf{X}_m))] = T_f(p) = \int p(\mathbf{x}) f(p(\mathbf{x}))\,\mathrm{d}\mathbf{x}. \tag{I.4}$$

On the other hand, from Proposition I.1, we expect that under certain regularity conditions,

$$\begin{aligned} \lim_{m\to\infty} \mathbb{E}[\phi_k(U_{k,m-1}(\mathbf{X}_m))] &= \mathbb{E}[\phi_k(U_{k\infty}(\mathbf{X}))] \qquad \text{(I.5)} \\ &= \int p(\mathbf{x})\mathbb{E}[\phi_k(U_{k\infty}(\mathbf{x}))]\,\mathrm{d}\mathbf{x}, \end{aligned}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3151231, IEEE Transactions on Information Theory

TABLE II

EXAMPLES OF FUNCTIONALS OF TWO DENSITIES AND THEIR ESTIMATOR FUNCTIONS $\phi_{kl}(u,v)$. THE ABSOLUTE CONTINUITY $\mathcal{P} \ll \mathcal{Q}$ IS ASSUMED IMPLICITLY UNLESS STATED OTHERWISE. A REFERENCE IS GIVEN WHENEVER AN ESTIMATOR ALREADY EXISTS IN THE LITERATURE. THE LAST COLUMN PRESENTS PAIRS OF EXPONENTS (a_{kl}, b_{kl}) AND $(\tilde{a}_{kl}, \tilde{b}_{kl})$ OF THE POLYNOMIAL ENVELOPES OF THE ESTIMATOR FUNCTION $\phi_{kl}(u,v)$ IN u AND v, RESPECTIVELY. THE CONSTANT ϵ, IF ANY, CAN BE CHOSEN AS AN ARBITRARILY SMALL POSITIVE NUMBER. FOR EACH CASE, $k > -a_{kl}$ AND $l > -\tilde{a}_{kl}$ IS REQUIRED TO GUARANTEE THE EXISTENCE OF THE CORRESPONDING INVERSE LAPLACE TRANSFORM.

Name	$T_f(p,q) = \mathbb{E}_p[f(p,q)]$	$\phi_{kl}(u,v) = \frac{\Gamma(k)\Gamma(l)}{u^{k-1}v^{l-1}}\mathcal{L}^{-1}\Big\{\frac{f(p,q)}{p^k q^l}\Big\}(u,v)$	(a_{kl}, b_{kl}); $(\tilde{a}_{kl}, \tilde{b}_{kl})$
KL divergence [31] (Examples IV.1, IV.6, IV.8, V.3, E.1)	$\mathbb{E}\Big[\ln \frac{p}{q}\Big]$	$\ln \frac{v}{u} + \Psi(k) - \Psi(l)$	$(-\epsilon, \epsilon)$; $(-\epsilon, \epsilon)$
α-divergence [32] ($\alpha > 0$) (Examples IV.2, IV.7, IV.9, V.4, E.2)	$\mathbb{E}\Big[\Big(\frac{p}{q}\Big)^{\alpha-1}\Big]$	$\frac{\Gamma(k)\Gamma(l)}{\Gamma(k-\alpha+1)\Gamma(l+\alpha-1)}\Big(\frac{v}{u}\Big)^{\alpha-1}$	$(1-\alpha, 1-\alpha)$; $(\alpha-1, \alpha-1)$
Logarithmic α-divergence ($\alpha > 0$) (Examples IV.3, E.3)	$\mathbb{E}\Big[\Big(\frac{p}{q}\Big)^{\alpha-1}\ln \frac{p}{q}\Big]$	$\frac{\Gamma(k)\Gamma(l)}{\Gamma(k-\alpha+1)\Gamma(l+\alpha-1)}\Big(\frac{v}{u}\Big)^{\alpha-1}\times$ $\Big(\ln\frac{v}{u} + \Psi(k-\alpha+1) - \Psi(l+\alpha-1)\Big)$	$(1-\alpha-\epsilon, 1-\alpha+\epsilon)$; $(\alpha-1-\epsilon, \alpha-1+\epsilon)$
Le Cam distance (Examples IV.4, IV.10, E.4)	$\mathbb{E}\Big[\frac{(p-q)^2}{2p(p+q)}\Big]$	$2\binom{k+l-2}{k-1}^{-1}\Big\{\sum_{j=0}^{l-1}\binom{k+l-2}{k-1+j}\Big(-\frac{u}{v}\Big)^j -$ $\Big(-\frac{u}{v}\Big)^{l-1}\Big(1-\frac{v}{u}\Big)^{k+l-2}\mathbb{1}_{[v,\infty)}(u)\Big\}$	$(-k+1, l-1)$; $(-l+1, k-1)$
Entropy difference ($\mathcal{Q} \ll \mathcal{P}$) (Example E.5)	$\mathbb{E}\Big[\ln \frac{1}{p} - \frac{q}{p}\ln\frac{1}{q}\Big]$	$\frac{(l-1)}{k}\frac{u}{v}(\Psi(l-1) - \ln v) - (\Psi(k) - \ln u)$	$(-\epsilon, 1)$; $(-1-\epsilon, -1+\epsilon)$
Reverse KL divergence ($\mathcal{Q} \ll \mathcal{P}$) (Example E.6)	$\mathbb{E}\Big[\frac{q}{p}\ln\frac{q}{p}\Big]$	$\frac{l-1}{k}\frac{u}{v}\Big(\ln\frac{u}{v} + \Psi(l-1) - \Psi(k+1)\Big)$	$(1-\epsilon, 1+\epsilon)$; $(-1-\epsilon, -1+\epsilon)$
Jensen–Shannon divergence ($\mathcal{Q} \ll \mathcal{P}$) (Examples IV.5, IV.11, E.7)	$\mathbb{E}\Big[\frac{1}{2}\ln\frac{2p}{p+q} + \frac{q}{2p}\ln\frac{2q}{p+q}\Big]$	See Example IV.5.	$(-k+1, l-1)$; $(-l+1, k-1)$

where $U_{k\infty}(\mathbf{x})$ is a $\mathsf{G}(k, p(\mathbf{x}))$ random variable, independent of $\mathbf{X} \sim p$ for every $\mathbf{x}$. We choose $\phi_k(u)$ so as to equate the integrands in (I.4) and (I.5), i.e., for every $p > 0$, if $U \sim \mathsf{G}(k,p)$, then

$$\begin{aligned} f(p) &= \mathbb{E}[\phi_k(U)] \\ &= \int_0^\infty \phi_k(u)\frac{p^k}{\Gamma(k)}u^{k-1}e^{-up}\,\mathrm{d}u \\ &= \frac{p^k}{\Gamma(k)}\mathcal{L}\{u^{k-1}\phi_k(u)\}(p), \end{aligned} \tag{I.6}$$

where $\mathcal{L}\{\cdot\}$ represents the *one-sided Laplace transform* (see, e.g., [24, Ch. 29]), defined as

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

Rearranging the terms in (I.6), we obtain the key equation of this paper via inverse Laplace transform

$$\phi_k(u) = \frac{\Gamma(k)}{u^{k-1}}\mathcal{L}^{-1}\Big\{\frac{f(p)}{p^k}\Big\}(u), \tag{I.7}$$

which we refer to as the *estimator function* ϕ_k for f with parameter k. In general, inverse Laplace transform $\mathcal{L}^{-1}\{\cdot\}(\cdot)$ can be obtained by the *Bromwich integral*, which is the contour integral

$$\mathcal{L}^{-1}\{f(p)\}(u) = \frac{1}{2\pi i}\lim_{T\to\infty}\int_{\gamma-iT}^{\gamma+iT} e^{pu}f(p)\,\mathrm{d}p,$$

where γ is chosen so that all singularities of $f(p)$ lie to the left of the vertical line $\mathrm{Re}(p) = \gamma$ in the complex plane and that $f(p)$ is bounded on the line (see, e.g., [33, Ch. 2]). For most cases of our interest (see Tables I and II), however, inverse Laplace transforms can be computed using known transforms of elementary functions [24], along with several properties of Laplace transform, such as linearity, time-scaling, and convolution. The reader is referred to Table III in Appendix E for a list of elementary Laplace transforms. Note, for example, that by the linearity of the inverse Laplace transform, if ϕ_k is the estimator function for f, then the estimator function for $af + b$ is $a\phi_k + b$ for any $a, b \in \mathbb{R}$. Concrete examples of estimator functions for different choices of f are presented in Table I. See Appendix E for detailed derivation of these examples.

The main contributions of this paper, for single-density functionals, are as follows: By establishing the asymptotic unbiasedness condition in (I.4) and (I.5) of the proposed estimator (I.3), the necessity of which was first observed in the Ph.D. thesis of one of the authors [34, Ch. 5], and by establishing that the variance of the estimator also vanishes asymptotically, we show that the proposed estimator is L_2-consistent under mild regularity conditions on densities. The general statement (Corollary III.3) capture the hardness of estimating a given functional based on k-NN statistics as a polynomial tail behavior of its corresponding inverse Laplace transform. For smooth, bounded densities, we also establish the polynomial convergence rate in mean-squared error (MSE) by carefully bounding nonasymptotic error terms. Informally, under certain regularity conditions, we establish that

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

where σ_p is the order of smoothness of the underlying distribution p, a quantifies how much the functional T_f is affected by *high* densities (see (III.2)), and $\lambda(\sigma, a, k)$ is the bias rate exponent defined in (III.6); see Section III-B and Corollary III.7 for details. For example, when the densities are sufficiently smooth, i.e., $\sigma_p \geq 1$, the rate exponent becomes $\lambda \approx 1/d$ for k sufficiently large, implying the approximate MSE rate of $\tilde{O}(m^{-1/\max\{d,2\}})$.

B. The proposed double-density functional estimators

For functionals of two densities, we naturally extend the same idea to the Laplace transform in two dimensional spaces. For $g : \mathbb{R}_+^2 \to \mathbb{R}$, we use (u, v) and (p, q) to denote "time domain" and "frequency domain" variables, respectively, and define

$$\mathcal{L}\{g(u,v)\}(p,q) := \int_0^\infty \int_0^\infty g(\tilde{u},\tilde{v})e^{-p\tilde{u}}e^{-q\tilde{v}}\,\mathrm{d}\tilde{u}\,\mathrm{d}\tilde{v}.$$

Note we keep dummy variables such as u and v in $\mathcal{L}\{g(u,v)\}(p,q)$ explicit, so as to avoid any confusion on which function is being transformed. We define the *estimator function* ϕ_{kl} for f with parameters (k, l), computed through the two-dimensional inverse Laplace transform, as

$$\phi_{kl}(u,v) = \frac{\Gamma(k)\Gamma(l)}{u^{k-1}v^{l-1}}\mathcal{L}^{-1}\Big\{\frac{f(p,q)}{p^k q^l}\Big\}(u,v). \tag{I.8}$$

When $T_f(p, q)$ is in the form of divergence, i.e., $f(p, q)$ is a function of p/q, the corresponding estimator function $\phi_{kl}(u, v)$ is also a function of u/v; see Proposition E.1 in Appendix E. Concrete examples of estimator functions for different choices of f are presented in Table II. See Appendix E for detailed derivations of these examples. Given two sets of samples $\mathbf{X}_{1:m}$ from p and $\mathbf{Y}_{1:n}$ from q, we further define

$$V_{ln}(\mathbf{x}) := V_l(\mathbf{x}|\mathbf{Y}_{1:n}) := n\,\lambda_{\mathrm{Leb}}(\mathbb{B}(\mathbf{x}, r_l(\mathbf{x}|\mathbf{Y}_{1:n}))).$$

We then propose a (k, l)-NN estimator of the form

$$\hat{T}_f(\mathbf{X}_{1:m}, \mathbf{Y}_{1:n}) = \frac{1}{m}\sum_{i=1}^m \phi_{kl}(U_{km}(\mathbf{X}_i), V_{ln}(\mathbf{X}_i)). \tag{I.9}$$

As in the single-density case, we establish the L_2-consistency and MSE convergence rate of our estimator (I.9) under respective regularity conditions.

Throughout the paper, we assume the Euclidean distance, i.e., $\rho(\mathbf{x}, \mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|$, but the results will continue to hold for the p-norm ($p \geq 1$) with minor modifications; see Section VII for related remarks.

Notation. We use $\varrho_d(v) := (v/\upsilon_d)^{1/d}$ to denote the radius of a d-dimensional ball of a volume v and $\upsilon_d(r) := \varrho_d^{-1}(r) = \lambda_{\mathrm{Leb}}(\mathbb{B}(0, r))$ to denote the volume of ball of radius r. We further use $\upsilon_d := \upsilon_d(1) = 2^d\Gamma(1 + \frac{1}{2})^d\Gamma(1 + \frac{d}{2})^{-1}$ to denote the volume of the unit ball $\mathbb{B}(0, 1)$. We denote the density of a random variable U as $\rho_U(u)$. We use the calligraphic letters $\mathcal{P}$ and $\mathcal{Q}$ to denote the probability measures corresponding to the density p and q, respectively, and denote the support of a density p as

$$\mathrm{supp}(p) := \{\mathbf{x} \in \mathbb{R}^d \colon\ \mathcal{P}(\mathbb{B}(\mathbf{x}, r)) > 0,\ \forall r > 0\}.$$

We use $\mathcal{P} \ll \mathcal{Q}$ to denote the absolute continuity of $\mathcal{P}$ with respect to $\mathcal{Q}$. For nonnegative functions $A(x)$ and $B(x)$ of $x \in \mathcal{X}$, we write $A(x) \lesssim_\alpha B(x)$ if there exists $C(\alpha) > 0$, depending only on some parameter α, such that $A(x) \leq C(\alpha)B(x)$ for all $x \in \mathcal{X}$. We use the standard Bachmann–Landau notation O and Θ (see, e.g., [35]) throughout the paper, and write $f(n) = \tilde{O}(g(n))$ to represent the polylogarithmic order $f(n) = O(g(n)(\ln g(n))^k)$ for some $k \in \mathbb{R}$. We use the shorthand notation $a \wedge b = \min\{a, b\}$ and $a \vee b = \max\{a, b\}$. Finally, $\mathbb{1}_A$ stands for the indicator function of a set A.

Organization: The rest of the paper is organized as follows. Section II discusses the relevant literature and positions our contributions in that context. We analyze the proposed estimator for functionals of one density (cf. (I.3) and (I.7)) in Section III and of two densities (cf. (I.9) and (I.8)) in Section IV. We discuss the convergence rate of the estimators with adaptive choices of k and l in Section V. We present in Section VI numerical results to demonstrate the proposed estimator for a few synthetic examples. Section VII concludes the paper.

II. Related Work

One of the most straightforward estimators of the density functional $T_f(p) = \mathbb{E}_{\mathbf{X}\sim p}[f(p(\mathbf{X}))]$ is the "plug-in" estimator that first forms a density estimate $\mathbf{x} \mapsto \hat{p}(\mathbf{x})$ from the samples $\mathbf{X}_{1:m}$, such as the standard k-NN density estimate

$$\hat{p}_{km}(\mathbf{x}) = \hat{p}(\mathbf{x}) = \frac{k/m}{\lambda_{\mathrm{Leb}}(\mathbb{B}(\mathbf{x}, r_{km}(\mathbf{x})))}, \tag{II.1}$$

then plugs it in as

$$\tilde{T}_f(\hat{p}) = \frac{1}{m}\sum_{i=1}^m f(\hat{p}(\mathbf{X}_i)). \tag{II.2}$$

Building on the consistency of the k-NN density estimate $\hat{p}_{km}$ when k increases sublinearly with m [36, 37], one can establish the consistency and finite-sample analysis of the plug-in estimator when $k \to \infty$ [38, 39, 40, 41]. For estimating the double-density functional $T_f(p, q) = \mathbb{E}_{\mathbf{X}\sim p}[f(p(\mathbf{X}), q(\mathbf{X}))]$, Berrett and Samworth [42] recently proposed a weighted version of the plug-in (k, l)-NN estimators of the form

$$\tilde{T}_f(\hat{p}, \hat{q}) = \frac{1}{m}\sum_{i=1}^m f(\hat{p}(\mathbf{X}_i), \hat{q}(\mathbf{X}_i)), \tag{II.3}$$

with the k-NN density estimate $\hat{p}_{km}$ and the l-NN density estimate $\hat{q}_{kn}$ based on the samples $\mathbf{X}_{1:m}$ from p and $\mathbf{Y}_{1:n}$ from q, respectively. They proved its efficiency by establishing a tight local asymptotic minimax lower bound and established a corresponding central limit theorem, given that k and l of the weighted-averaged plug-in estimators grow to infinity.

For a *fixed* k, however, an appropriate "bias correction" is necessary for the plug-in estimator in (II.2) to be asymptotically unbiased, since the fixed-k-NN density estimator in (II.1) is not consistent for a finite k. A fixed-k plug-in estimator with bias correction was first studied by Kozachenko and Leonenko [25], who applied 1-NN distances to estimate differential entropies of densities on $\mathbb{R}^d$ based on an idea of Dobrushin [43], and established the L_2-consistency of their estimator.

Subsequently, Singh et al. [26] and Goria et al. [5] generalized the 1-NN Kozachenko–Leonenko estimator to $k \ge 1$ as

$$\begin{aligned}\hat{T}^{(k)}_{\mathsf{KL}}(\mathbf{X}_{1:m}) &= \tilde{T}_f(\hat{p}_{km}) + \ln k - \Psi(k) \qquad \text{(II.4)}\\ &= \frac{1}{m}\sum_{i=1}^{m} \ln \frac{1}{\hat{p}_{km}(\mathbf{X}_i)} + \ln k - \Psi(k),\end{aligned}$$

where $\Psi(x) := \Gamma'(x)/\Gamma(x)$ denotes the digamma function [24]. As the canonical fixed-k density functional estimator, the Kozachenko–Leonenko estimator has been investigated extensively in the literature. Beyond the L_2-consistency, Tsybakov and van der Meulen [44] first established $\sqrt{m}$-consistency, i.e., the L_2-convergence rate of $O(m^{-1})$, of a truncated version of the 1-NN Kozachenko–Leonenko estimator in $\mathbb{R}$, which was extended by Gao et al. [30] to $k \ge 1$ and $d \ge 1$. Some recent developments include a central limit theorem [45], results on large-k behavior [46], and minimax optimality [47, 48].

Along the same line, L_2-consistent fixed-k or fixed-(k, l) plug-in estimators with proper additive or multiplicative bias correction were proposed[1] for KL divergence (Wang et al. [31]), Rényi entropies (Leonenko et al. [27]), Rényi divergences (Póczos and Schneider [32]), and several other divergences of a specific polynomial form (Póczos et al. [51]). These plug-in estimators can be expressed in general as

$$\tilde{T}^{\mathsf{aff}}_f(\hat{p}) = a_k \tilde{T}_f(\hat{p}) + b_k, \qquad \text{(II.5)}$$

or

$$\tilde{T}^{\mathsf{aff}}_f(\hat{p}, \hat{q}) = a_{kl} \tilde{T}_f(\hat{p}, \hat{q}) + b_{kl}, \qquad \text{(II.6)}$$

where $\hat{p}$ is the fixed-k-NN density estimator from $\mathbf{X}_{1:m}$ in (II.1), $\hat{q}$ is the fixed-l-NN density estimate similarly obtained from $\mathbf{Y}_{1:n}$, and (a_k, b_k) and (a_{kl}, b_{kl}) determine functional-specific bias correction, respectively. Many density functionals beyond the special examples mentioned earlier, however, do not allow such affine bias correction. For example, a plug-in estimator for the logarithmic α-entropy in Table I cannot be made unbiased, even asymptotically, by any affine bias correction.

A more general approach to correcting bias of the fixed-k plug-in estimator was proposed by Singh and Póczos [52] as

$$\tilde{T}_{b\circ f}(\hat{p}) = \frac{1}{m}\sum_{i=1}^{m} b_{km}(f(\hat{p}_{km}(\mathbf{X}_i))), \qquad \text{(II.7)}$$

which obviously subsumes affine bias correction. This estimator was shown to be L_2-consistent for a fixed k with definite convergence rate if there exists a bias-correcting function b_{km} that satisfies

$$\mathbb{E}[b_{km}(f(\hat{p}_{km}(\mathbf{x})))] = \mathbb{E}[f(\bar{p}_{km}(\mathbf{x}))] \qquad \text{(II.8)}$$

for every m and any underlying density p, and for $\mathcal{P}$-a.e. $\mathbf{x}$, where

$$\bar{p}_{km}(\mathbf{x}) = \frac{\mathcal{P}(\mathbb{B}(\mathbf{x}, r_{km}(\mathbf{x})))}{\lambda_{\text{Leb}}(\mathbb{B}(\mathbf{x}, r_{km}(\mathbf{x})))}$$

is the average density over the k-NN ball $\mathbb{B}(\mathbf{x}, r_{km}(\mathbf{x}))$. Despite the general form of this estimator, however, the existence of b_{km} satisfying the stringent condition of equality in (II.8) for every m could be established only for differential entropy (and only for KL divergence in case of functionals of two densities).

In contrast to the existing literature, our estimator

$$\begin{aligned}\hat{T}^{(k)}_f(\mathbf{X}_{1:m}) &= \frac{1}{m}\sum_{i=1}^{m} \phi_k(U_{km}(\mathbf{X}_i)) \qquad \text{(II.9)}\\ &= \frac{1}{m}\sum_{i=1}^{m} \phi_k\Bigl(\frac{k}{\hat{p}_{km}(\mathbf{X}_i)}\Bigr)\end{aligned}$$

bypasses the whole bias correction issue of the plug-in approach by specifying the estimator function ϕ_k directly via the inverse Laplace transform (I.7). Here, we identified that $U_{km}(\mathbf{x}) = k/\hat{p}_{km}(\mathbf{x})$ by the respective definitions in (I.1) and (II.1). Our approach naturally unifies all existing estimators of the form (II.5) or (II.6), and finds new estimators for logarithmic entropies and divergences that cannot be obtained even in the most general bias-corrected form (II.7) of the traditional plug-in estimator (II.2). For example, our estimator for the logarithmic α-entropy ($f(p) = p^{\alpha-1}\ln(1/p)$) is characterized by the estimator function

$$\begin{aligned}\phi_k(u) &= \phi_k\Bigl(\frac{k}{p}\Bigr) \qquad \text{(II.10)}\\ &= \frac{\Gamma(k)}{\Gamma(k-\alpha+1)} k^{-\alpha+1} p^{\alpha-1}\Bigl(\ln\frac{k}{p} - \Psi(k-\alpha+1)\Bigr),\end{aligned}$$

which cannot be expressed as a function $b_{km}(f(p))$ for some b_{km}.

We comment on how analysis techniques of the proposed estimators are related to those in the literature. Through the design of our estimator functions (I.7) and (I.8) via inverse Laplace transform, we can naturally extend and simplify existing analyses for differential entropy and KL divergence by Bulinski and Dimitrov [28, 29], and establish the asymptotic unbiasedness of our estimators (I.3) and (I.9) for a general functional. By adapting the nonasymptotic analysis for differential entropy in Gao et al. [30], we can also establish the bias convergence rate of the estimator for a general functional, but without truncation. For variance analysis, we deviate from the aforementioned work [28, 29, 30] for simplicity and deploy a technique for the Euclidean space used by Singh and Póczos [52]; see also [37, Ch. 7]. Note, however, that the established variance results of our estimator continue to hold under the p-norm; see Remark III.2. Our consistency analysis (unbiasedness and vanishing variance) strengthens and simplifies many existing ones including those for Rényi entropies [27], Rényi divergences [32], and divergences of polynomial form [51]. The convergence rates for the functionals in Tables I and II are established in this paper for the first time, except the Kozachenko–Leonenko estimator [44, 30, 52, 48] and the KL divergence estimator [52].

In a different direction of investigation, kernel density estimator (KDE)-based approaches have been widely studied in the literature for estimation of smooth density functionals, which also include many of the examples presented in

[1] As pointed out in [49], there are slight errors in the original analyses in [25, 5, 31, 27] when invoking asymptotic theory to establish L_2-consistency. Correct proofs were given later in [28, 29, 50].

Sections IV and VI as special cases. Birge and Massart [53] established a minimax optimal rate $O(m^{-\frac{8\sigma}{d+4\sigma}}+m^{-1})$ on convergence rates in MSE of estimators of certain integral functionals involving the density and its derivatives under Hölder smoothness of order σ (Definition III.1) on the density and demonstrated that the parametric rate $O(1/m)$ is achievable if the density is sufficiently smooth, say, $\sigma \geq d/4$. For estimating polynomial divergence functionals, Krishnamurthy et al. [54] proposed plug-in estimators corrected through estimating higher-order terms in the von Mises expansions, which may require computationally demanding numerical integration, and established a minimax lower bound $\Omega(m^{-\frac{8\sigma}{4\sigma+d}}+m^{-1})$ under Hölder smoothness of order $\sigma > 0$. Kandasamy et al. [55] generalized this approach to more general functionals and mutual information and established similar rates. In another line of work, extending the boundary-corrected plug-in estimator for mutual information of [56], Singh and Póczos [57, 58] established the MSE rate $O(m^{-\frac{2\sigma}{\sigma+d}}+m^{-1})$ for a kernel-based plug-in estimator of a class of density functionals under certain regularity conditions; we remark that this approach commonly requires a prior knowledge on the support.

Convergence of k-NN distance-based estimators of density functionals can be improved by using the so-called "ensemble method", where a convex combination of estimators with different k values is used. Moon et al. [59] studied the ensemble method for estimation of the mutual information between two continuous random variables, and demonstrated that under certain broad regularity conditions on the density, the optimal convex combination, which can be computed by solving a convex optimization problem, yields the *parametric* MSE rate $O(1/m)$ provided that the density is sufficiently smooth. In a similar spirit, Moon et al. [60], Noshad et al. [61], and Wisler et al. [62] obtained the MSE rate $O(1/m)$ for estimating the KL divergence, f-divergences, and a wider class of density functionals including f-divergences, respectively, using the ensemble method. Analyzing the ensemble version of the proposed estimators is beyond the scope of this paper.

We finally remark that Nguyen et al. [63] studied the estimation of f-divergences through minimization of empirical risk, by formulating the problem as a convex program. They established convergence rates when the likelihood ratio between the two distributions belongs to a reproducing kernel Hilbert space. It seems, however, quite nontrivial to compare these assumptions with those on smoothness used in the present work.

III. Functionals of One Density

Recall that we define the estimator function $\phi_k\colon \mathbb{R}_+ \to \mathbb{R}$ for a given $f\colon \mathbb{R}_+ \to \mathbb{R}$, with parameter $k \in \mathbb{N}$ as

$$\phi_k(u) = \frac{\Gamma(k)}{u^{k-1}}\mathcal{L}^{-1}\Big\{\frac{f(p)}{p^k}\Big\}(u), \tag{I.7}$$

whenever the inverse Laplace transform exists, and then define the estimator as

$$\hat{T}_f^{(k)}(\mathbf{X}_{1:m}) = \frac{1}{m}\sum_{i=1}^{m}\phi_k(U_{km}(\mathbf{X}_i)). \tag{I.3}$$

Remark III.1. One can check that, for all the examples in Table I,

$$\lim_{k\to\infty}\phi_k\Big(\frac{k}{p}\Big) = f(p) \tag{III.1}$$

for each $p > 0$. In light of (II.9), this observation heuristically indicates that our estimator becomes closer to the plug-in estimator (II.2) as we use larger, fixed k. This observation is consistent with our intuition that we do not need any bias correction for the plug-in estimator with very large k, since the plugged-in k-NN density estimate (II.1) becomes consistent as $k \to \infty$ in the sample limit [36].

To analyze the proposed estimator for general functionals $T_f(p)$ in a unified manner, we abstract polynomial tail behaviors of each estimator function $\phi_k(u)$ as $u \downarrow 0$ 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

$$\psi_{a,b}(u) := \begin{cases} u^a & \text{if } 0 < u \leq 1, \\ u^b & \text{if } u > 1. \end{cases} \tag{III.2}$$

Note that as a gets larger and b gets smaller, the piecewise polynomial function $\psi_{a,b}(u)$ decays faster as $u \downarrow 0$ and as $u \uparrow \infty$, respectively. Therefore, a and b quantify the amount of contribution of low and high density values to the estimator function $\phi_k(u)$, respectively. Consistent with the observation that such extreme density values typically make the density functional estimation problem harder, we will establish stronger statements for functionals with larger a and smaller b. Below we present the estimator functions for a few representative functionals.

Example III.1 (Differential entropy [25]). For $f(p) = \ln(1/p)$ and any $k \geq 1$, we can compute, as detailed in Example E.1 in Appendix E,

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

Note that we can write $\Psi(k) = H_{k-1} - \gamma$ for $k \in \mathbb{N}$, where $H_k = \sum_{i=1}^k (1/i)$ denotes the k-th harmonic number and $\gamma := \lim_{k\to\infty}(H_k - \ln k)$ denotes the Euler–Mascheroni constant [24]. 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 finer analysis without relying on the polynomial bound $\psi_{-\epsilon,\epsilon}(u)$ may lead to a marginal improvement in the resulting performance guarantee [30, 28, 29], but we do not pursue that in this paper.

Example III.2 (α-entropy [27]). 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 [64] entropy $h_\alpha(p) = (\ln T_f(p))/(1-\alpha)$ and Harvda and Charvat [65] or Tsallis [66] 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, as verified in Example E.2 in Appendix E,

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

which allows the tight polynomial bound

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

Example III.3 (Logarithmic α-entropy). For $f(p) = p^{\alpha-1}\ln(1/p)$ ($\alpha > 0$), we refer to the density functional $T_f(p) = \int p^\alpha(\mathbf{x})\ln(1/p(\mathbf{x}))\,\mathrm{d}\mathbf{x}$ as the *logarithmic α-entropy*. For any $k \in \mathbb{N}$ such that $k > \alpha - 1$, we can compute, as verified in Example E.3 in Appendix E,

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

and we consider

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

for any arbitrarily small $\epsilon > 0$ as its polynomial bound.

Example III.4 (Exponential (α,β)-entropy). For $f(p) = p^{\alpha-1}e^{-\beta p}$ ($\alpha > 0$, $\beta \geq 0$), we refer to the density functional $T_f(p) = \int p^\alpha(\mathbf{x})e^{-\beta p(\mathbf{x})}\,\mathrm{d}\mathbf{x}$ as the *exponential (α,β)-entropy*. For any $k \in \mathbb{N}$ such that $k > \alpha - 1$, we can compute

$$\phi_k(u) = \frac{\Gamma(k)}{\Gamma(k-\alpha+1)} \frac{(u-\beta)^{k-\alpha}}{u^{k-1}} \mathbb{1}_{[\beta,\infty)}(u)$$

using time shifting property of Laplace transform from the estimator function expression of the α-entropy. The estimator function ϕ_k can be bounded as

$$|\phi_k(u)| \lesssim \psi_{0,1-\alpha}(u)$$

for $k \geq \alpha$ and cannot be bounded by a piecewise polynomial function if $k < \alpha$.

In our subsequent analysis, regularity conditions for the consistency and convergence rate of the proposed estimator depend on k and f via the lower tail exponent a and the upper tail exponent b. By (II.9), extreme values of $\hat{p}_{km}$ are amplified more via ϕ_k as a decreases and and b increases. Hence, intuitively, when a is large and b is small, the regularity conditions are milder and the estimator converges faster.

A. Consistency

Focusing solely on the asymptotic behavior of our estimator, we can establish the L_2-consistency for general functionals under mild assumptions on densities. To state the results rigorously, we first define certain technical conditions. For future use in Section IV-A for functionals of two densities, we state the conditions in terms of two densities p and $\tilde{p}$ such that $\mathcal{P} \ll \tilde{\mathcal{P}}$. Later, we identify $\tilde{p}$ as the density p for samples $\mathbf{X}_{1:m}$ or the density q for samples $\mathbf{Y}_{1:n}$.

For the sake of easy analysis of density functional estimators, the standard simplifying assumptions are global upper- and lower-boundedness on the underlying density p, i.e., there exist $c > 0$ and $C > 0$ such that $c \leq p(\mathbf{x}) \leq C$ for any $\mathbf{x} \in \operatorname{supp}(p)$; note that the boundedness of the support follows from the lower boundedness of the density. In what follows, to establish the asymptotic consistency of the proposed estimators for a larger class of densities, we will consider weaker conditions than the boundedness assumptions, similar to those in [28, 29].

For each $r > 0$, we define the local maximal operator M_r on $\mathbb{R}^d$ for a density p by

$$M_r p(\mathbf{x}) := \sup_{r' \in (0,r]} \frac{\mathcal{P}(\mathbb{B}(\mathbf{x}, r'))}{\lambda_{\mathrm{Leb}}(\mathbb{B}(\mathbf{x}, r'))}.$$

Similarly, for each $r > 0$, we define the local minimal operator m_r on $\mathbb{R}^d$ for a density p by

$$m_r p(\mathbf{x}) := \inf_{r' \in (0,r]} \frac{\mathcal{P}(\mathbb{B}(\mathbf{x}, r'))}{\lambda_{\mathrm{Leb}}(\mathbb{B}(\mathbf{x}, r'))}.$$

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 [28, 29]. In particular, $M_r p(\mathbf{x})$ and $m_r p(\mathbf{x})$ are pointwise upper and lower bounds, respectively, on the density p.

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

$$W(p,\tilde{p};\vartheta,r) := \int p(\mathbf{x})(M_r\tilde{p}(\mathbf{x}))^\vartheta\,\mathrm{d}\mathbf{x},$$
$$w(p,\tilde{p};\xi,\vartheta,r) := \int p(\mathbf{x})\xi((m_r\tilde{p}(\mathbf{x}))^{-\vartheta})\,\mathrm{d}\mathbf{x},$$

and

$$R(p,\tilde{p};\xi,\vartheta,r) := \iint_{\rho(\mathbf{x},\mathbf{y})>r} p(\mathbf{x})\tilde{p}(\mathbf{y})\xi(v^\vartheta(\rho(\mathbf{x},\mathbf{y})))\,\mathrm{d}\mathbf{x}\,\mathrm{d}\mathbf{y}$$

for each $\vartheta > 0$ and $r > 0$. Here we define these quantities with possibly different densities p and $\tilde{p}$ for the future use with double-density functionals; for single-density functionals, the readers can simply assume $p = \tilde{p}$. In place of the upper- and lower- boundedness assumptions on the density $\tilde{p}$, we will impose the finiteness of the expected values $W(p,\tilde{p};\vartheta,r)$ and $w(p,\tilde{p};\xi,\vartheta,r)$, respectively. Further, $R(p,\tilde{p};\xi,\vartheta,r)$ roughly quantifies how fast p and $\tilde{p}$ decay to zero in there tails. Observe that $R(p,\tilde{p};\xi,\vartheta,r) \to 0$ as $r \to \infty$. Intuitively, as the tails of p and $\tilde{p}$ decay faster, the speed of convergence of $R(p,\tilde{p};\xi,\vartheta,r)$ will be faster. In particular, if both p and $\tilde{p}$ have bounded support, then $R(p,\tilde{p};\xi,\vartheta,r) = 0$ for r sufficiently large. Note further that W, w, and R become larger as ϑ increases.

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

$$\limsup_{m\to\infty} \xi(m^b)R\Big(p,\tilde{p};\xi,b,\varrho\Big(\frac{\kappa_m}{m}\Big)\Big) < \infty \quad \text{(III.3)}$$

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

Recall that the polynomial tail exponents a and b of the the k-NN estimator function (II.9) of a given density functional quantify the amount of contribution of high and low density values to the estimator, respectively. Hence, a is coupled with W that captures the upper boundedness of the density, while b is pertinent to w and R that quantify the lower boundedness. We note that as a gets larger, k gets smaller, and b gets smaller, conditions ($\mathbf{L}_{p\tilde{p}}$; ξ, b) and ($\mathbf{U}_{p\tilde{p}}$; k, a) become weaker, thus encompassing a larger class of densities.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3151231, IEEE Transactions on Information Theory

Let Ξ be the class of non-decreasing functions $\xi\colon \mathbb{R}_+ \to \mathbb{R}_+$ such that $\xi(t)/t \to \infty$ as $t \to \infty$, that $\xi(t_1 t_2) \le \xi(t_1)\xi(t_2)$ for any $x, y > t_0$ for some $t_0 \in \mathbb{R}_+$, and that $\omega(\xi) := \inf\{\eta > 1\colon \xi(t)/t^{\eta} \to 0 \text{ as } t \to \infty\} < \infty$. For example, $\xi_1(t) = (t \ln t) \vee 0 \in \Xi$ with $t_0 = e$ and $\omega(\xi_1) = 1$, and $\xi_2(t) = t^{\alpha} \in \Xi$ for $\alpha > 1$ with $t_0 = 0$ and $\omega(\xi_2) = \alpha$.

We are now ready to state the L_2-consistency results. We show separately that the bias and variance converge to zero under certain regularity conditions. Note that all estimator functions presented in Table I are continuous. Throughout, we consider a fixed $(a, b) \in \mathbb{R}^2$ for a target functional $T_f(\cdot)$ that satisfies $|\phi_k(u)| \lesssim \psi_{a,b}(u)$, provided that the estimator function $\phi_k(u)$ exists for $k > -a$.

Theorem III.1 (Vanishing bias)**.** *For a target functional $T_f(\cdot)$, if the estimator function ϕ_k is continuous and the underlying density p satisfies* $(\mathbf{U}_{pp};\ k, a)$ *and* $(\mathbf{L}_{pp};\ \xi, b)$ *with some function $\xi \in \Xi$, then the estimator* (I.3) *with fixed $k > -\omega(\xi)a$ is asymptotically unbiased.*

Theorem III.2 (Vanishing variance)**.** *For a target functional $T_f(\cdot)$, if the underlying density p satisfies* $(\mathbf{U}_{pp};\ k, a)$ *and* $(\mathbf{L}_{pp};\ \xi, b)$ *with $\xi(t) = t^2$, the variance of the estimator* (I.3) *with fixed $k > -2a$ converges to zero as $m \to \infty$.*

Combining Theorems III.1 and III.2, the L_2-consistency readily follows as a corollary.

Corollary III.3 (Consistency)**.** *For a target functional $T_f(\cdot)$, if the estimator function ϕ_k is continuous and the underlying density p satisfies* $(\mathbf{U}_{pp};\ k, a)$ *and* $(\mathbf{L}_{pp};\ \xi, b)$ *with $\xi(t) = t^2$, then the estimator* (I.3) *with fixed $k > -2a$ is L_2-consistent.*

In the following examples, we illustrate how Corollary III.3 can be instantiated for a few representative functionals.

Example III.5 (Differential entropy; Example III.1 contd.). Recall that for any $k \in \mathbb{N}$, $|\phi_k(u)| \lesssim \psi_{-\epsilon,\epsilon}(u)$ for arbitrarily small $\epsilon > 0$. By Corollary III.3, the estimator (I.3) is L_2-consistent if the underlying density p satisfies that $(\mathbf{U}_{pp};\ k, -\epsilon)$ and $(\mathbf{L}_{pp};\ \xi, \epsilon)$ with $\xi(t) = t^2$ for some $\epsilon > 0$. We note that the condition (III.3) in $(\mathbf{L}_{pp};\ \xi, \epsilon)$ can be relaxed to a milder condition in which there exist some $\delta, R > 0$ such that

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

by performing a similar analysis based on the upper bound $|\phi_k(u)| \lesssim |\ln u| + 1$, i.e., without invoking the polynomial bound $\psi_{-\epsilon,\epsilon}(u)$ for an arbitrarily small $\epsilon > 0$. This recovers a similar result reported in [29].

Example III.6 (α-entropy; Example III.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 $(\mathbf{U}_{pp};\ k, a)$, which slightly generalizes the upper-boundedness condition and the requirement $k > 2\alpha - 1$ assumed in Leonenko et al. [27]. For $\alpha < 1$, since $a = 1 - \alpha > 0$, the estimator with fixed $k \ge 1$ is L_2-consistent if p satisfies $(\mathbf{L}_{pp};\ \xi, b)$ with $\xi(t) = t^2$, for examples, if p is bounded away from zero and supported over a hyperrectangle. We remark that Leonenko and Pronzato [50] reported the L_2-consistency of the estimator for densities satisfying alternate conditions when $\alpha < 1$.

1) *Proof of Theorem III.1 (vanishing bias):* If the estimator function ϕ_k is continuous, by the continuous mapping theorem and Proposition I.1, we have the convergence of the statistic $\phi_k(U_{k,m-1}(\mathbf{X}_m))$ to $\phi_k(U_{k\infty}(\mathbf{X}))$ in distribution as $m \to \infty$, where $U_{k\infty}(\mathbf{x})$ is a $\mathsf{G}(k, p(\mathbf{x}))$ random variable, independent of $\mathbf{X} \sim p$ for $\mathcal{P}$-a.e. $\mathbf{x}$. Hence, if the sequence of random variables $(\phi_k(U_{k,m-1}(\mathbf{X}_m)))_{m\ge1}$ is uniformly integrable, we readily establish the asymptotic unbiasedness:

$$\begin{aligned} \lim_{m\to\infty} \mathbb{E}[\hat{T}_f^{(k)}(\mathbf{X}_{1:m})] &= \lim_{m\to\infty} \mathbb{E}[\hat{p}_k(U_{k,m-1}(\mathbf{X}_m))] \\ &= \mathbb{E}[\phi_k(U_{k\infty}(\mathbf{X}))] = T_f(p). \end{aligned}$$

To show the uniform integrability of $(\phi_k(U_{k,m-1}(\mathbf{X}_m)))_{m\ge1}$, we invoke the following lemma.

Lemma III.4 (De la Vallée Poussin theorem [67, Theorem 1.3.4])**.** *A collection of random variables $(X_i)_{i\in I}$ is uniformly integrable if and only if there exists a non-decreasing function $\xi\colon \mathbb{R}_+ \to \mathbb{R}_+$ such that $\sup_{i\in I} \mathbb{E}[\xi(|X_i|)] < \infty$ and $\xi(t)/t \to \infty$ as $t \to \infty$.*

Observe that we have

$$\begin{aligned} &\mathbb{E}[\xi(|\phi_k(U_{k,m-1}(\mathbf{X}_m))|)] \\ &= \int p(\mathbf{x}) \mathbb{E}[\xi(|\phi_k(U_{k,m-1}(\mathbf{x}))|)] \, \mathrm{d}\mathbf{x} \\ &\lesssim \int p(\mathbf{x}) \mathbb{E}[\xi(\psi_{a,b}(U_{k,m-1}(\mathbf{x})))] \, \mathrm{d}\mathbf{x} \\ &= \int p(\mathbf{x}) \int_0^{\infty} \xi(\psi_{a,b}(u)) \, \mathrm{d}F_{km}(u|\mathbf{x}) \, \mathrm{d}\mathbf{x}. \end{aligned}$$

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 B.17 in Appendix B-D, which yields

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

This ensures the uniform integrability of $(\phi_k(U_{k,m-1}(\mathbf{X}_m)))_{m\ge1}$ by the de la Vallée Poussin theorem (Lemma III.4), and thus concludes the proof. □

2) *Proof of Theorem III.2 (vanishing variance):* By Lemma B.24 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 B.24. 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 B.17, 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. □

Remark III.2. The variance analysis relies on the Efron–Stein inequality (Lemma B.25) and a covering lemma (Lemma B.26) that only applies to the Euclidean space; see Appendix B-F.

An idea for the generic variance bound (Lemma B.24) first appeared in Singh and Póczos [52] as a generalization of a technique for analyzing the 1-NN Kozachenko–Leonenko estimator by Biau and Devroye [37, Ch. 7], and has been employed in the literature to bound the variance of k-NN based estimators; see, e.g., Moon et al. [59]. We note that one can attain the same rate (up to polylogarithmic factors) under the p-norm, by instead adapting the analysis in Gao et al. [30]. As it demands a rather involved argument to bound a covariance term, however, we present a simpler approach in this paper.

B. Convergence rates for smooth, bounded densities

So far, we have established the L_2-consistency of the proposed estimator for general functionals under mild assumptions on densities. Under rather stronger assumptions such as smoothness and boundedness, we can actually establish the convergence rate of the proposed estimator in MSE. Specifically, we consider certain regularity conditions adapted from [30].

First, we assume that

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

Further, we impose a few conditions related to lower-boundedness of the density, that is,

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

(L2$_p$) the support of p is bounded, and

(L3$_p$) there exists $r > 0$ such that

$$\eta_p := \inf_{\mathbf{x}\in\operatorname{supp}(p)} \inf_{r'\in(0,r]} \frac{\lambda_{\mathrm{Leb}}(\mathbb{B}(\mathbf{x},r')\cap\operatorname{supp}(p))}{\lambda_{\mathrm{Leb}}(\mathbb{B}(\mathbf{x},r'))} > 0.$$

The last condition **(L3$_p$)** is called the (η_p, r)-regularity of $\operatorname{supp}(\mu)$ in the literature [68].

Remark III.3. The upper-boundedness condition **(U$_p$)** implies the condition **(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 conditions **(L1$_p$)**, **(L2$_p$)**, and **(L3$_p$)** on lower-boundedness of p imply the condition **(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 **(L1$_p$)** and **(L3$_p$)**, and $R(p,p;\xi,b,\varrho(\kappa_m/m))) = 0$ for m sufficiently large by the boundedness of the support of p from **(L2$_p$)**.

We recall the following notion of Hölder continuity for smoothness of the density p, which is assumed commonly in nonparametric statistics; see, e.g., [53, 54, 52, 47, 48].

Definition III.1. 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 $\kappa := \lceil \sigma \rceil - 1$ and

$$L(g;\Omega) := \sup_{\substack{\mathbf{r}\in\mathbb{Z}_+^d \\ |\mathbf{r}|=\kappa}} \sup_{\substack{\mathbf{y},\mathbf{z}\in\Omega \\ \mathbf{y}\neq\mathbf{z}}} \frac{|\partial^{\mathbf{r}} g(\mathbf{y}) - \partial^{\mathbf{r}} g(\mathbf{z})|}{\|\mathbf{y}-\mathbf{z}\|^{\beta}} < \infty, \tag{III.4}$$

where $\beta := \sigma - \kappa$. Here we use a multi-index notation (see, e.g., [69, Ch. 8]), that is, $|\mathbf{r}| := r_1 + \cdots + r_d$ for $\mathbf{r} \in \mathbb{Z}_+^d$ and $\partial^{\mathbf{r}} g(\mathbf{x}) := \partial^{\kappa} g(\mathbf{x})/(\partial x_1^{r_1} \cdots \partial x_d^{r_d})$.

Since the density is not smooth on the boundary of the support due to the lower-boundedness condition **(L1$_p$)**, 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:

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

(B$_p$) the boundary of $\operatorname{supp}(p)$ has finite $(d-1)$-dimensional Hausdorff measure [69].

Truncated versions of well-known distributions such as exponential, Gaussian, and Cauchy distributions, as well as distributions with bounded support, such as uniform distribution and beta distributions with parameters $\alpha, \beta \ge 1$, satisfy these conditions with $\sigma_p = 2$, and the truncated Laplace distribution satisfies the conditions with $\sigma_p = 1$; see Appendix F for details on these examples. For densities of *unbounded* support, we provide a separate treatment using a variant of our estimator; see Section III-C.

Equipped with these regularity conditions, we upper bound the MSE of our estimator by considering its bias and variance separately.

Theorem III.5 (Bias rate)**.** *For a target functional $T_f(\cdot)$, if the underlying density p satisfies the conditions* ***(U$_p$)***, ***(L1$_p$)***, ***(L2$_p$)***, ***(L3$_p$)***, ***(S$_p$)***, *and* ***(B$_p$)***, *then the estimator* (I.3) *with fixed $k > -a$ satisfies*

$$\left|\mathbb{E}[\hat{T}_f^{(k)}] - T_f(p)\right| = \tilde{O}(m^{-\lambda(\sigma_p,a,k)}) \tag{III.5}$$

as $m \to \infty$, where

$$\lambda(\sigma,a,k) = \begin{cases} \frac{1}{d}(\sigma\wedge 1)(\frac{k+a}{k-1}) & \text{if } a \le -\frac{\sigma}{d} - 1, \\ \frac{1}{d}(\sigma \wedge \frac{k+a}{k-1}) & \text{if } -\frac{\sigma}{d} - 1 < a \le -1, \\ \frac{1}{d}(\sigma\wedge 1) & \text{if } a > -1. \end{cases} \tag{III.6}$$

Remark III.4. Since $k > -a$ is required to apply Theorem III.5, when $a \le -1$ (for example, the 2-entropy), our estimator is well-defined and λ in (III.6) is positive only for $k > 1$. Conversely, our bias bound holds for 1-NN estimators of any functional $T_f(p)$ with estimator function $\phi_1(u)$ of lower tail exponent $a > 1$, the examples of which include differential entropy, the α-entropy with $\alpha < 2$, the logarithmic α-entropy with $\alpha < 2$, and exponential (α, β)-entropy with $\alpha \le 1$ in Table I.

Remark III.5. 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, the rate exponent λ becomes larger with larger k. In Section V, we show that a properly growing k in sample size can guarantee the largest rate exponent in (III.6). Note, however, that if a decreases as k increases, which is the case for some exceptional cases (Examples IV.10 and IV.11), the rate exponent could become slower with larger k. This is in contrast to the large-k requirement for *plug-in* estimators, to guarantee the underlying k-NN density estimate to be consistent. We remind that our estimator is designed to be asymptotically unbiased for every fixed k, without appealing to the consistency of the k-NN density estimator, and it thus does not contradict the behavior of plug-in estimators.

Remark III.6. The upper tail exponent b appears only in the exponent of polylogarithmic factors $O(\mathrm{poly}\ln(m))$ in the rate, and thus is hidden by $\tilde{O}$ in (III.6). At a finer scale, the rate increases as b decreases; see the proof of Theorem B.23 and Lemmas B.21 and B.23 in Appendix C-A.

The variance of the estimator can be bounded without the smoothness conditions.

Theorem III.6 (Variance rate)**.** *For a target functional $T_f(\cdot)$, if the underlying density p satisfies **($\mathbf{U}_p$)**, **($\mathbf{L1}_p$)**, **($\mathbf{L2}_p$)**, and **($\mathbf{L3}_p$)**, then the estimator* (I.3) *with fixed $k > -2a$ satisfies*

$$\mathrm{Var}(\hat{T}_f^{(k)}) = O(m^{-1}). \tag{III.7}$$

Combining Theorem III.5 on bias and Theorem III.6 on variance, we can obtain the convergence rate in MSE and establish the L_2-consistency of the estimator.

Corollary III.7 (Convergence rate)**.** *Under the same assumptions in Theorem III.5, then the estimator* (I.3) *with fixed $k > -2a$ satisfies*

$$\mathbb{E}\big[\big(\hat{T}_f^{(k)} - T_f(p)\big)^2\big] = \tilde{O}(m^{-2\lambda(\sigma_p,a,k)} + m^{-1}). \tag{III.8}$$

Remark III.7. 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.

Remark III.8. We note that the bias rate of the proposed estimators under Hölder smoothness of order $\sigma > 0$ is at most $O(m^{-(\sigma\wedge 1)/d})$; it may be improved to $O(m^{-(\sigma\wedge 2)/d})$ if the *boundary bias* is ignored, as remarked in [30], but it still suffers the curse of dimensionality. As pointed out in Jiao et al. [48], it is an inherent problem with any *positive*-kernel-based estimator that a higher smoothness $\sigma > 2$ cannot be exploited in density functional estimation [70, Chapter 1]. In particular, the key component in our analysis is Lemma B.6 from [48], which cannot be improved for $\sigma > 2$. See [47] for an extensive deliberation on this issue and see [45, 38, 39, 59, 60, 61, 62] for a solution based on the jackknife idea for some density functionals. Providing a remedy to the limitation of the proposed estimators is left as an open problem.

Remark III.9. 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, including the rates for divergence functional estimators in Corollaries IV.6, are not minimax optimal [57, 58, 54, 55] due to the suboptimal bias rates; see, e.g., Example III.7. Since our main focus is on providing unified consistency and convergent rate analyses of the proposed generic estimators, we leave proving minimax optimality under proper regularity conditions with or without modifications of the proposed estimators as important future directions. For the special case of differential entropy, we note that Jiao et al. [48] established an asymptotic minimax optimality of the Kozachenko–Leonenko estimator [48] for for smooth densities of order $\sigma \in (0, 2]$ *over a torus* (no boundary condition), matching the lower bound of [47] up to a polylogarithmic factor.

Example III.7 (Differential entropy; Example III.1 contd.). Recall from Example III.1 that $|\phi_k(u)| \lesssim \psi_{-\epsilon,\epsilon}(u)$ for any arbitrarily small $\epsilon > 0$. Suppose that the underlying density p satisfies the conditions **($\mathbf{U}_p$)**, **($\mathbf{L1}_p$)**, **($\mathbf{L2}_p$)**, **($\mathbf{L3}_p$)**, **($\mathbf{S}_p$)**, and **($\mathbf{B}_p$)**, in Theorem III.5 with some $\sigma_p \in (0, 2]$. Then we have the bias exponent $\lambda = \sigma_p/d$ as in the third case of (III.6) and the variance exponent of 1 from (III.7). Consequently, by Corollary III.7 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 [30] for $\sigma_p = 2$. We remark that Gao et al. [30] reported a lower bound $\Omega(m^{-\frac{16}{d+8}} + m^{-1})$ for estimating differential entropy under $\sigma = 2$ and hence, the convergence rate is not minimax optimal.

Example III.8 (α-entropy; Example III.2 contd.). Recall from Example III.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 **($\mathbf{U}_p$)**, **($\mathbf{L1}_p$)**, **($\mathbf{L2}_p$)**, **($\mathbf{L3}_p$)**, **($\mathbf{S}_p$)**, and **($\mathbf{B}_p$)**, the MSE of our estimator (I.3) with fixed $k > 2(\alpha - 1)$ is bounded as (III.8) with the bias rate exponent

$$\lambda(\sigma_p, a, k) = \begin{cases} \frac{1}{d}(\sigma_p \wedge 1) & \text{if } \alpha < 2, \\ \frac{1}{d}(\sigma_p \wedge \frac{k+1-\alpha}{k-1}) & \text{if } 2 \leq \alpha < 2 + \frac{\sigma_p}{d}, \\ \frac{1}{d}(\sigma_p \wedge 1)(\frac{k+1-\alpha}{k-1}) & \text{if } \alpha \geq 2 + \frac{\sigma_p}{d}. \end{cases} \tag{III.9}$$

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

1) Proof of Theorem III.5 (bias rate): First note that $U_{km}(\mathbf{X}_1), \ldots, U_{km}(\mathbf{X}_m)$ are identically distributed, and $U_{km}(\mathbf{X}_m) = U_{k,m-1}(\mathbf{X}_m)$ by definition; see (I.2). Hence, we can write

$$\begin{aligned} \mathbb{E}[\hat{T}_f^{(k)}] &= \mathbb{E}[\phi_k(U_{k,m-1}(\mathbf{X}_m))] \\ &= \int \mathbb{E}[\phi_k(U_{k,m-1}(\mathbf{X}_m))|\mathbf{X}_m = \mathbf{x}]p(\mathbf{x})\,\mathrm{d}\mathbf{x} \\ &= \int \mathbb{E}[\phi_k(U_{k,m-1}(\mathbf{x}))]p(\mathbf{x})\,\mathrm{d}\mathbf{x}, \end{aligned} \tag{III.10}$$

where the last equality holds since $\mathbf{X}_m$ and $\mathbf{X}_{1:m-1}$ are independent. Recall from Proposition I.1 that $U_{km}(\mathbf{x})$ converges to a $\mathsf{G}(k, p(\mathbf{x}))$ random variable $U_{k\infty}(\mathbf{x})$ for $\mathcal{P}$-a.e. $\mathbf{x}$. Thus, by the construction (I.6) of the estimator function $\phi_k(u)$, we can express the density functional as

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

Applying the triangle inequality, we first have

$$\begin{aligned} &\big|\mathbb{E}[\hat{T}_f^{(k)}] - T_f(p)\big| \\ &\leq \int p(\mathbf{x})|\mathbb{E}[\phi_k(U_{k,m-1}(\mathbf{x})) - \phi_k(U_{k\infty}(\mathbf{x}))]|\,\mathrm{d}\mathbf{x} \\ &= \int p(\mathbf{x})\Big|\int_0^\infty \phi_k(u)(\rho_{U_{k,m-1}(\mathbf{x})}(u) - \rho_{U_{k\infty}(\mathbf{x})}(u))\,\mathrm{d}u\Big|\,\mathrm{d}\mathbf{x}. \end{aligned} \tag{III.11}$$

For some real numbers τ_m and ν_m such that $0 \leq \tau_m \leq 1 \leq \nu_m < \infty$, which are to be determined later as functions of k, a, d, and σ_p, we break the inner integral and apply the polynomial bound $|\phi_k(u)| \lesssim \psi_{a,b}(u)$ with the triangle inequality to obtain

$$\left|\mathbb{E}[\hat{T}_f^{(k)}] - T_f(p)\right| \lesssim I_{\text{out},1} + I_{\text{in},1} + I_{\text{in},2} + I_{\text{out},2}, \tag{III.12}$$

where

$$\begin{aligned} I_{\text{out},1} &:= \mathbb{E}_p[I_{\text{out},1}(\mathbf{X})] \\ &= \mathbb{E}_p\Big[\int_0^{\tau_m} \psi_{a,b}(u)(\rho_{U_{k,m-1}(\mathbf{X})}(u) + \rho_{U_{k\infty}(\mathbf{X})}(u))\,\mathrm{d}u\Big], \\ I_{\text{in},1} &:= \mathbb{E}_p[I_{\text{in},1}(\mathbf{X})] \\ &= \mathbb{E}_p\Big[\int_{\tau_m}^{1} \psi_{a,b}(u)|\rho_{U_{k,m-1}(\mathbf{X})}(u) - \rho_{U_{k\infty}(\mathbf{X})}(u)|\,\mathrm{d}u\Big], \\ I_{\text{in},2} &:= \mathbb{E}_p[I_{\text{in},2}(\mathbf{X})] \\ &= \mathbb{E}_p\Big[\int_{1}^{\nu_m} \psi_{a,b}(u)|\rho_{U_{k,m-1}(\mathbf{X})}(u) - \rho_{U_{k\infty}(\mathbf{X})}(u)|\,\mathrm{d}u\Big], \end{aligned}$$

and

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

The *inner bias* terms $I_{\text{in},1}$ and $I_{\text{in},2}$ can be bounded by Lemma B.21 under the conditions **(U$_p$)**, **(S$_p$)**, and **(B$_p$)**, and the *outer bias* terms $I_{\text{out},1}$ and $I_{\text{out},2}$ can be bounded by Lemma B.23 under the conditions **(U$_p$)**, **(L1$_p$)**, **(L2$_p$)**, and **(L3$_p$)**. After putting the bounds from Lemmas B.21 and B.23 together, a proper choice of the break points (τ_m, ν_m) concludes the proof; see Appendix C-A for the details. □

Remark III.10. The key step in this analysis is the decomposition in (III.12), which is based on the construction of the estimator (I.6) from its asymptotic unbiasedness. Moreover, by considering only the polynomial tail behavior of each estimator function and using (III.12), 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 [30] for a truncated version of the Kozachenko–Leonenko estimator of differential entropy.

2) Proof of Theorem III.6 (variance rate): Since the boundedness conditions **(U$_p$)**, **(L1$_p$)**, **(L2$_p$)**, and **(L3$_p$)** imply **(U$_{pp}$;** k, a**)** and **(L$_{pp}$;** ξ, b**)** (see Remark III.3), the variance rate directly follows from the proof of Theorem III.2 in Section III-A2. □

C. Convergence rates for smooth densities of unbounded support

Theorem III.5 establishes the bias rate of the proposed estimator for smooth, bounded densities that inherently assume nonsmooth boundary. In this section, we establish convergence rate of a truncated version of the estimator for densities of unbounded support.

For functionals of one density, we define a truncated version of the estimator (I.3) as

$$\bar{T}_f^{(k)}(\mathbf{X}_{1:m}) := \frac{1}{m}\sum_{i=1}^{m} \bar{\phi}_k(U_{km}(\mathbf{X}_i); \tau_m, \nu_m), \tag{III.13}$$

where we define the truncated estimator function

$$\bar{\phi}_k(u; \tau, \nu) := \phi_k(u)\mathbb{1}_{(\tau,\nu)}(u)$$

and the *lower and upper truncation points* $\tau_m, \nu_m \in \mathbb{R}_+$ are hyperparameters such that $0 \leq \tau_m \leq 1 \leq \nu_m < \infty$ that are to be determined based on the function f, the dimension d, the number of nearest neighbors k, and/or the smoothness order of the underlying density p.

We assume the following condition on the tail behavior of the underlying density, which is more general than **(L1$_p$)**:

(L1$'_p$) There exist $\theta > 0$ and $D_0 > 0$ such that $\int p(\mathbf{x})e^{-\beta p(\mathbf{x})}\,\mathrm{d}\mathbf{x} \leq D_0\beta^{-\theta}$ for all $\beta > 1$.

This tail condition with $\theta = 1$ was originally considered by Tsybakov and van der Meulen [44] for their analysis in $\mathbb{R}$. As pointed out in [44], densities with strictly sub-exponential tails, such as Gaussian distributions, satisfy **(L1$'_p$)** with $\theta = 1$. It can also be shown that densities with polynomially decaying tails satisfy condition **(L1$'_p$)** for some $0 < \theta < 1$.

We additionally introduce the following functional-dependent condition on the behavior of the estimator function for small density values:

(L4$_p$) There exists $\delta > 0$ such that $\int p(\mathbf{x})(p(\mathbf{x}))^{-(1+\delta)b}\,\mathrm{d}\mathbf{x} < \infty$.

Finally, as we consider densities with unbounded support, we assume that

(S$'_p$) the density p is σ_p-Hölder continuous over $\mathbb{R}^d$ for $\sigma_p \in (0, 2]$,

in place of **(S$_p$)**.

Exclusively for the following proposition, we additionally assume that $\phi_k(u)$ satisfies $|\phi_k(u)| \lesssim \psi_{a,b}(u)$, $\phi_k(u)$ is differentiable at any $u > 0$, and $|\phi'_k(u)| \lesssim \psi_{a-1,b-1}(u)$, which hold for all the examples in Table I.

Proposition III.8 (Bias rate for smooth densities of unbounded support)**.** *For a target functional $T_f(\cdot)$, if the underlying density p satisfies the conditions **(U$_p$)**, **(L1$'_p$)**, **(L4$_p$)**, and **(S$'_p$)**, then the truncated estimator* (III.13) *with $-a < k < -b+\theta+1$ and truncation points*

$$\tau_m = \begin{cases} \Theta(m^{-\frac{\sigma_p}{d}\frac{1}{k-\frac{\sigma_p}{d}-1}}) & \textit{if } a \leq -\frac{\sigma_p}{d} - 1, \\ O(m^{-\frac{\sigma_p}{d}\frac{1}{k+a}}) & \textit{o.w.} \end{cases} \tag{III.14}$$

and

$$\begin{aligned} &\nu_m \\ &= \begin{cases} \Theta(m^{(\frac{\sigma_p}{d}\wedge 1)\frac{1}{\theta-k-b+1}}) & \textit{if } k \leq -b-1, b \leq -\frac{\sigma_p}{d} - 1, \\ \Theta(m^{\frac{\sigma_p}{d}\frac{1}{\theta-k+\frac{\sigma_p}{d}+2}}) & \textit{if } k \leq -b-1, b > -\frac{\sigma_p}{d} - 1, \\ \Theta(m^{\frac{1}{\theta+2}}) & \textit{if } k > -b-1, b \leq -\frac{\sigma_p}{d} - 1, \\ \Theta(m^{(\frac{\sigma_p}{d}\wedge 1)\frac{1}{\theta+2}}) & \textit{if } k > -b-1, b > -\frac{\sigma_p}{d} - 1, \end{cases} \end{aligned} \tag{III.15}$$

with $\nu_m = o(\sqrt{m})$ as $m \to \infty$ satisfies

$$\left|\mathbb{E}\left[\overline{T}_f^{(k)}\right] - T_f(p)\right| = O\left(m^{-\lambda_\tau \wedge \lambda_\nu}\right),$$

where

$$\lambda_\tau = \begin{cases} \frac{\sigma_p}{d} \frac{k+a}{k-\frac{\sigma_p}{d}-1} & \textit{if } a \le -\frac{\sigma_p}{d} - 1, \\ \frac{\sigma_p}{d} & \textit{o.w.}, \end{cases} \tag{III.16}$$

and

$$\lambda_\nu \tag{III.17}$$

$$= \begin{cases} \frac{\sigma_p}{d} \wedge 1 & \textit{if } k \le -b-1, b \le -\frac{\sigma_p}{d} - 1, \\ (\frac{\sigma_p}{d}(1 - \frac{b+\frac{\sigma_p}{d}+1}{\theta-k+\frac{\sigma_p}{d}+2})) \wedge 1 & \textit{if } k \le -b-1, b > -\frac{\sigma_p}{d} - 1, \\ \frac{\sigma_p}{d} \wedge (1 - \frac{k+b+1}{\theta+2}) & \textit{if } k > -b-1, b \le -\frac{\sigma_p}{d} - 1, \\ (\frac{\sigma_p}{d} \wedge 1)(1 - \frac{k+b+1}{\theta+2}) & \textit{if } k > -b-1, b > -\frac{\sigma_p}{d} - 1. \end{cases}$$

We can establish the variance rate with truncation under only the upper-boundedness condition, without explicitly imposing the condition $k > -2a$ as required in Theorem III.6.

Proposition III.9 (Variance rate of truncated estimator)**.** *For a target functional $T_f(\cdot)$, if the underlying density p satisfies* **(U$_p$)***, then the estimator* (III.13) *with $k > -a$ satisfies*

$$\mathrm{Var}\left(\overline{T}_f^{(k)}\right) = O\left(\frac{k^2}{m}\left(k^{-k}\tau_m^{(k+2a)\wedge 0} + \nu_m^{2b\vee 0}\right)\right). \tag{III.18}$$

Combining Propositions III.8 and III.9, we can obtain a corresponding consistency result as in Corollary III.7, the formal statement of which is omitted.

At face value, Proposition III.8 enlarges considerably the class of densities under the purview of our analyses. On the flip side, however, it requires the underlying density to be smooth over the whole of $\mathbb{R}^d$ and this rules out, for example, the uniform distribution, which is covered by **(L1$_p$)**. Thus, Proposition III.8 and Theorem III.5 complement each other.

The stringent requirement $k < -b+\theta+1$ in Proposition III.8 is due to a bias term $O(\nu_m^{b+k-1-\theta})$ that appears in the analysis; a smaller k, which is, of course, still larger than $-a$, gives a tighter bound on this term, whereas a larger k is desired to reduce the bias due to the lower truncation. Proposition III.8 thus cannot guarantee the L_2-consistency of the estimator when k grows as $m \to \infty$, as the condition $k < \theta - b + 1$ is violated.

Example III.9 (Differential entropy; Example III.1 contd.). For estimating differential entropy, recall that $|\phi_k(u)| \lesssim \psi_{-\epsilon,\epsilon}(u)$ for arbitrarily small $\epsilon > 0$. Consider densities that satisfy the conditions **(U$_p$)**, **(L1$'_p$)**, **(L4$_p$)**, and **(S$'_p$)** for some $0 < \theta \le 1$. Since Proposition III.8 requires $k < \theta + 1 - \epsilon$, we need to choose $k = 1$ to guarantee the L_2-consistency of our estimator. We obtain a bias bound $O(m^{-\frac{\theta-\epsilon}{\theta+2}(\frac{\sigma_p}{d}\wedge 1)})$, a variance bound $O(m^{-(1-\delta)})$ for arbitrarily small $\delta > 0$ from Proposition III.9, and thus the MSE rate $O(m^{-\frac{2(\theta-\epsilon)}{\theta+2}(\frac{\sigma_p}{d}\wedge 1)})$. In particular, for one-dimensional densities with $\sigma_p \ge 1$ and $\theta = 1$, we obtain the MSE rate $O(m^{-\frac{2(1-\epsilon)}{3}})$. Note that this rate is slightly worse than $O(m^{-1})$, as obtained by Tsybakov and van der Meulen [44, Section 2, pp. 77–78] under different regularity conditions with a faster growing upper truncation point $\nu_m = \Theta(\sqrt{m})$.

Example III.10 (α-entropy; Example III.2 contd.). Consider estimating the α-entropy ($\alpha \ne 1$) of densities that satisfy the conditions **(U$_p$)**, **(L1$'_p$)**, **(L4$_p$)**, and **(S$'_p$)** with some $\theta > 0$. Since $|\phi_k(u)| \lesssim \psi_{1-\alpha,1-\alpha}(u)$, we need to use $k \in (\alpha-1, \alpha+\theta)$ for our estimator to apply Proposition III.8. By setting the truncation points as

$$(\tau_m, \nu_m)$$

$$= \begin{cases} (O(m^{-\frac{\sigma_p}{d}\frac{1}{k-\alpha+1}}), \Theta(m^{(\frac{\sigma_p}{d}\wedge 1)\frac{1}{\theta+2}})), & \text{if } \alpha < \frac{\sigma_p}{d} + 2, \\ (\Theta(m^{-\frac{\sigma_p}{d}\frac{1}{k-\frac{\sigma_p}{d}-1}}), \Theta(m^{\frac{1}{\theta+2}})), & \text{if } \alpha \ge \frac{\sigma_p}{d} + 2, \end{cases}$$

our estimator achieves the bias rate $O(m^{-(\lambda_\tau \wedge \lambda_\nu)})$, where

$$(\lambda_\tau, \lambda_\nu) = \begin{cases} (\frac{\sigma_p}{d}, (\frac{\sigma_p}{d} \wedge 1)\frac{\theta+\alpha-k}{\theta+2}) & \text{if } \alpha < \frac{\sigma_p}{d} + 2, \\ (\frac{\sigma_p}{d}\frac{k-\alpha+1}{k-\frac{\sigma_p}{d}-1}, \frac{\sigma_p}{d} \wedge \frac{\theta+\alpha-k}{\theta+2}) & \text{if } \alpha \ge \frac{\sigma_p}{d} + 2. \end{cases}$$

From Proposition III.9, we can bound the variance of our estimator as $O(m^{-\lambda_{\mathrm{v}}})$, where

$$\lambda_{\mathrm{v}} = \begin{cases} 1 - (\frac{\sigma_p}{d} \wedge 1)\frac{2(1-\alpha)\vee 0}{\theta+2} & \text{if } \alpha < \frac{\sigma_p}{d} + 2, \\ 1 - \frac{\sigma_p}{d}\frac{(2\alpha-k-2)\vee 0}{k-\frac{\sigma_p}{d}-1} & \text{if } \alpha \ge \frac{\sigma_p}{d} + 2, \end{cases}$$

and thus we establish the MSE rate $O(m^{-2(\lambda_\tau\wedge\lambda_\nu)} + m^{-\lambda_{\mathrm{v}}})$.

Remark III.11. We remark in passing on the consistency of the truncated estimator (without convergence rate analysis). With lower truncation point τ_m such that $\tau_m^{k+2a} = o(m)$, the conditions $k > -2a$ can be relaxed to $k > -a$ in Corollary III.3. Moreover, a very mild upper truncation of speed $\nu_m = e^{o(m)}$ can relax the condition **(L$_{pp}$)** assumed in the consistency results to a milder one, i.e.,

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

with $\tilde{p} = p$.

IV. Functionals of Two Densities

We now consider estimating a functional $T_f(p, q)$ of two densities p and q. Henceforth, we assume that $\mathcal{P} \ll \mathcal{Q}$. Recall that for fixed $k, l \in \mathbb{N}$ and a given $f\colon \mathbb{R}_+^2 \to \mathbb{R}$, we define the *estimator function* $\phi_{kl}\colon \mathbb{R}_+^2 \to \mathbb{R}$ of f with parameters k, l as

$$\phi_{kl}(u, v) = \frac{\Gamma(k)\Gamma(l)}{u^{k-1}v^{l-1}}\mathcal{L}^{-1}\left\{\frac{f(p,q)}{p^k q^l}\right\}(u, v), \tag{I.8}$$

whenever the inverse Laplace transform exists, and then define the estimator as

$$\hat{T}_f^{(k)}(\mathbf{X}_{1:m}, \mathbf{Y}_{1:n}) = \frac{1}{m}\sum_{i=1}^m \phi_{kl}(U_{km}(\mathbf{X}_i), V_{ln}(\mathbf{Y}_i)). \tag{I.9}$$

Here we define

$$V_{ln}(\mathbf{x}) := U_l(\mathbf{x}|\mathbf{Y}_{1:n}) = n\,\lambda_{\mathrm{Leb}}(\mathbb{B}(\mathbf{x}, r_l(\mathbf{x}|\mathbf{Y}_{1:n}))).$$

Remark IV.1. Similar to the observation made in Remark III.1, an analogous limiting behavior

$$\lim_{k,l\to\infty} \phi_{kl}\left(\frac{k}{p}, \frac{l}{q}\right) = f(p, q)$$

can be verified for all the examples in Table II except Le Cam distance and Jensen–Shannon divergence.

As for the single-density case, a polynomial tail behavior of the estimator function $\phi_{kl}(u,v)$ affects the convergence rate of each instantiated estimator. We describe a tail behavior of $\phi_{kl}(u,v)$ by a quadruple $(a_{kl}, b_{kl}, \tilde{a}_{kl}, \tilde{b}_{kl}) \in \mathbb{R}^4$ such that $|\phi_{kl}(u,v)| \lesssim \psi_{a_{kl},b_{kl}}(u)\psi_{\tilde{a}_{kl},\tilde{b}_{kl}}(v)$. This characterization allows us to handle the convergence of $U_{km}(\mathbf{x})$ and $V_{ln}(\mathbf{x})$ separately so that we can extend the analysis for the single-density case in a straightforward manner. Note that for all the examples presented in Table II, $(a_{kl}, b_{kl}, \tilde{a}_{kl}, \tilde{b}_{kl})$ can be found as constants independent of k and l, except Le Cam distance and Jensen–Shannon divergence. Also note that all the estimator functions $\phi_{kl}(u,v)$ presented in Table II are continuous.

Example IV.1 (KL divergence [31]). For $f(p,q) = \ln(p/q)$, we can compute, as shown in Example E.1 in Appendix E,

$$\phi_{kl}(u,v) = \ln\frac{v}{u} + H_{k-1} - H_{l-1}.$$

As a bound on the estimator function $\phi_{kl}(u,v)$, we consider

$$\begin{aligned}|\phi_{kl}(u,v)| &\lesssim 1 + |\ln u| + |\ln v|\\ &\lesssim (1+|\ln u|)(1+|\ln v|) \lesssim \psi_{-\epsilon,\epsilon}(u)\psi_{-\epsilon,\epsilon}(v)\end{aligned}$$

for any arbitrarily small $\epsilon > 0$.

Example IV.2 (Polynomial functional [32, 51]). For $f(p,q) = p^{\alpha-1}q^{\beta}$ $(\alpha > 0, \beta > 1-\alpha)$ and any $k,l \in \mathbb{N}$ such that $k > \alpha - 1$ and $l > \beta$, we can compute, as shown in Example E.2 in Appendix E,

$$\phi_{kl}(u,v) = \frac{\Gamma(k)\Gamma(l)}{\Gamma(k-\alpha+1)\Gamma(l-\beta)}u^{1-\alpha}v^{-\beta},$$

which allows the tight polynomial bound

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

This class of polynomial functionals includes many important functionals. For the special instance of $\beta = 1-\alpha$, we refer to the density functional $T_f(p,q) = \int p^{\alpha}(\mathbf{x})q^{1-\alpha}(\mathbf{x})\,\mathrm{d}\mathbf{x}$ as the *α-divergence*, which appears in the literature in a few different forms; see, e.g., Rényi [64] and Cichocki et al. [71].

Example IV.3 (Logarithmic α-divergence). For $f(p,q) = (p/q)^{\alpha-1}\ln(p/q)$ $(\alpha > 0)$, we refer to the density functional $T_f(p,q) = \int p^{\alpha}(\mathbf{x})q^{1-\alpha}(\mathbf{x})\ln(p(\mathbf{x})/q(\mathbf{x}))\,\mathrm{d}\mathbf{x}$ as the *logarithmic α-divergence*. For any $k,l \in \mathbb{N}$ such that $k > \alpha - 1$ and $l > 1-\alpha$, we can compute, as shown in Example E.3 in Appendix E,

$$\begin{aligned}\phi_{kl}(u,v) = &\frac{\Gamma(k)\Gamma(l)}{\Gamma(k-\alpha+1)\Gamma(l+\alpha-1)}\\ &\times u^{-\alpha+1}\Big(\ln\frac{v}{u} + \Psi(k-\alpha+1) - \Psi(l+\alpha-1)\Big).\end{aligned}$$

As a bound on the estimator function $\phi_{kl}(u,v)$, we consider

$$\begin{aligned}|\phi_{kl}(u,v)| &\lesssim u^{-\alpha+1}v^{\alpha-1}(1+|\ln u| + |\ln v|)\\ &\lesssim u^{-\alpha+1}v^{\alpha-1}(1+|\ln u|)(1+|\ln v|)\\ &\lesssim \psi_{1-\alpha-\epsilon,1-\alpha+\epsilon}(u)\psi_{\alpha-1-\epsilon,\alpha-1+\epsilon}(v)\end{aligned}$$

for any arbitrarily small $\epsilon > 0$.

Example IV.4 (Le Cam distance). For $f(p,q) = (p-q)^2/(2p(p+q))$, the corresponding divergence functional

$$\begin{aligned}D_{\mathsf{LC}}(p,q) &= \frac{1}{2}\int\frac{(p(\mathbf{x})-q(\mathbf{x}))^2}{p(\mathbf{x})+q(\mathbf{x})}\,\mathrm{d}\mathbf{x}\\ &= 1 - \int\frac{2p(\mathbf{x})q(\mathbf{x})}{p(\mathbf{x})+q(\mathbf{x})}\,\mathrm{d}\mathbf{x}\end{aligned}$$

is called Le Cam distance [72, p. 47] in the literature [73]. We note in passing that this functional has a connection to the nearest neighborhood binary classification rule: it is well known that the asymptotic error of the nearest neighborhood binary classification for equiprobable classes is given as $\frac{1}{2}(1 - T_f(p,q))$ [74]. For any $k,l \in \mathbb{N}$, we can compute, as shown in Example E.4 in Appendix E,

$$\begin{aligned}\phi_{kl}(u,v) = 2&\binom{k+l-2}{k-1}^{-1}\Big(-\frac{u}{v}\Big)^{l-1}\times\\ &\Big\{\sum_{i=0}^{l-1}\binom{k+l-2}{i}\Big(-\frac{v}{u}\Big)^{i}\\ &\quad - \Big(1-\frac{v}{u}\Big)^{k+l-2}\mathbb{1}_{[v,\infty)}(u)\Big\} - 1.\end{aligned}$$

As a bound on the estimator function $\phi_{kl}(u,v)$, we have

$$|\phi_{kl}(u,v)| \lesssim \psi_{-k+1,l-1}(u)\psi_{-l+1,k-1}(v).$$

Example IV.5 (Jensen–Shannon divergence). When $\mathcal{Q} \ll \mathcal{P}$, we can write Jensen–Shannon divergence as

$$D_{\mathsf{JS}}(p,q) = \frac{1}{2}\Big(D\Big(p\;\Big\|\;\frac{p+q}{2}\Big) + D\Big(q\;\Big\|\;\frac{p+q}{2}\Big)\Big) = T_f(p,q)$$

for

$$f(p,q) = \frac{1}{2}\Big(\frac{q}{p}+1\Big)\ln\frac{2}{(q/p)+1} + \frac{q}{2p}\ln\frac{q}{p},$$

where $D(p\,\|\,q)$ denotes the KL divergence between p and q. For any $k \geq 1$ and $l \geq 2$, we can compute, as shown in Example E.7 in Appendix E,

$$\begin{aligned}\phi_{kl}(u,v) = \frac{1}{2}\Big\{&\ln 2 + \frac{l-1}{k}\frac{u}{v}\Big(\Psi(l-1) - \Psi(k+1) + \ln 2\frac{u}{v}\Big)\\ &+ B_{kl}(u,v) + \frac{l-1}{k}\frac{u}{v}B_{k+1,l-1}(u,v)\Big\},\end{aligned}$$

where $B_{kl}(u,v)$ is defined in (IV.1). As a polynomial bound, we have

$$|\phi_{kl}(u,v)| \lesssim \psi_{-k+1,l-1}(u)\psi_{-l+1,k-1}(v).$$

A. *Consistency*

As in Section III-A, we can establish the L_2-consistency of the estimator of functionals of two densities under mild regularity conditions. Throughout, we consider a fixed $(a,b,\tilde{a},\tilde{b}) \in \mathbb{R}^4$ for a target functional $T_f(\cdot,\cdot)$ whose estimator function ϕ_{kl} satisfies $|\phi_{kl}(u,v)| \lesssim \psi_{a,b}(u)\psi_{\tilde{a},\tilde{b}}(v)$, provided that the estimator function ϕ_{kl} exists for $k > -a$ and $l > -\tilde{a}$.

Theorem IV.1 (Vanishing bias). *For a target functional $T_f(\cdot,\cdot)$, if the estimator function $\phi_{kl}(u,v)$ is continuous and the underlying densities p and q satisfy **($\mathbf{U}_{pp}$;** k,a**)**, **($\mathbf{L}_{pp}$;** ξ^2,b**)**, **($\mathbf{U}_{pq}$;** $l,\tilde{a}$**)**, and **($\mathbf{L}_{pq}$;** $\xi^2,\tilde{b}$**)** for some function $\xi \in \Xi$,*

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3151231, IEEE Transactions on Information Theory

$$B_{kl}(u,v) = \begin{cases} \binom{k+l-2}{k-1}^{-1} \sum_{j=1}^{l-1} \binom{k+l-2}{k-1+j} \frac{(-u/v)^j}{j} & \text{if } \frac{u}{v} < 1, \\ -\ln\frac{u}{v} + \binom{k+l-2}{k-1}^{-1} \Big\{ -\sum_{j=-k+1}^{-1} \binom{k+l-2}{k-1+j} \frac{(-u/v)^j}{j} + \sum_{\substack{j=-k+1 \\ j\neq 0}}^{l-1} \binom{k+l-2}{k-1+j} \frac{(-1)^j}{j} \Big\} & \text{if } \frac{u}{v} \ge 1. \end{cases} \tag{IV.1}$$

then the estimator (I.9) *with $k > -2\omega(\xi)a$ and $l > -2\omega(\xi)\tilde{a}$ is asymptotically unbiased as $m, n \to \infty$.*

Theorem IV.2 (Vanishing variance)**.** *For a target functional $T_f(\cdot,\cdot)$, if the underlying densities p and q satisfy* ***(U$_{pp}$; k, a)***, ***(L$_{pp}$; ξ^2, b)***, ***(U$_{pq}$; $l, \tilde{a}$)***, *and* ***(L$_{pq}$; $\xi^2, \tilde{b}$)*** *with $\xi(t) = t^2$, then then the variance of the estimator* (I.9) *with fixed $k > -4a$ and fixed $l > -4\tilde{a}$ converges to zero as $m, n \to \infty$.*

Corollary IV.3 (Consistency)**.** *For a target functional $T_f(\cdot,\cdot)$, if the estimator function $\phi_{kl}(u,v)$ is continuous and the underlying densities p and q satisfy* ***(U$_{pp}$; k, a)***, ***(L$_{pp}$; ξ^2, b)***, ***(U$_{pq}$; $l, \tilde{a}$)***, *and* ***(L$_{pq}$; $\xi^2, \tilde{b}$)*** *with $\xi(t) = t^2$, then the estimator* (I.9) *with fixed $k > -4a$ and fixed $l > -4\tilde{a}$ is L_2-consistent.*

In the following examples, we illustrate how Corollary IV.3 can be instantiated for a few representative functionals.

Example IV.6 (KL divergence; Example IV.1 contd.). Recall that for estimating differential entropy, $|\phi_{kl}(u,v)| \lesssim \psi_{-\epsilon,\epsilon}(u)\psi_{-\epsilon,\epsilon}(v)$ for arbitrarily small $\epsilon > 0$ and for any $k, l \in \mathbb{N}$. By Corollary IV.3, the estimator (I.9) with fixed $k \ge 1$ and $l \ge 1$ is L_2-consistent if the underlying densities p and q satisfy **(U$_{pp}$; $k, -\epsilon$)**, **(L$_{pp}$; ξ^2, ϵ)**, **(U$_{pq}$; $l, -\epsilon$)**, and **(L$_{pq}$; ξ^2, ϵ)** with $\xi(t) = t^2$. As discussed in Example III.5, a finer analysis recovers a similar consistency result established in [29].

The proofs of the main results (Theorems IV.1, IV.2, IV.4, and IV.5) in this section follow with minor extensions to those of the single-density case, and are deferred to Appendix C.

Example IV.7 (α-divergence; Example IV.2 contd.). Recall that for estimating the α-divergence ($\alpha \neq 1$), we have $|\phi_{kl}(u,v)| \lesssim \psi_{1-\alpha,1-\alpha}(u)\psi_{\alpha-1,\alpha-1}(v)$ for any $k, l \in \mathbb{N}$ such that $k > \alpha - 1$ and $l > 1 - \alpha$. For $\alpha > 1$, since $b = 1 - \alpha < 0$ and $\tilde{a} = \alpha - 1 > 0$, the estimator with fixed $k > 4(\alpha - 1)$ and $l \ge 1$ is L_2-consistent if the underlying densities p and q satisfy that **(U$_{pp}$; $k, 1-\alpha$)** and **(L$_{pq}$; $\xi^2, \alpha-1$)** with $\xi(t) = t^2$. For $\alpha < 1$, since $a = 1 - \alpha > 0$ and $\tilde{b} = \alpha - 1 < 0$, the estimator with $k \ge 1$ and $l > 4(1 - \alpha)$ is L_2-consistent if the underlying densities p and q satisfy that **(L$_{pp}$; ξ^2, b)** and **(U$_{pq}$; $l, \tilde{a}$)** with $\xi(t) = t^2$. This consistency result covers a strictly larger class of densities than an earlier result by Póczos and Schneider [32], whereby the L_2-consistency of the estimator with $l = k$ is established under rather stronger assumptions such as boundedness and uniform continuity of densities. Moreover, Propositions IV.3 and IV.3 strengthen the L_2-consistency result established in Póczos et al. [51] for a polynomial functional (see Example IV.2), which subsumes α-divergence.

B. Convergence rates for smooth, bounded densities

Theorem IV.4 (Bias rate)**.** *For a target functional $T_f(\cdot,\cdot)$, if the underlying density p satisfies* ***(U$_p$)***, ***(L1$_p$)***, ***(L2$_p$)***, ***(L3$_p$)***, ***(S$_p$)***, *and* ***(B$_p$)***, *and q satisfies* ***(U$_q$)***, ***(L1$_q$)***, ***(L2$_q$)***, ***(L3$_q$)***, ***(S$_q$)***, *and* ***(B$_q$)***, *then the estimator* (I.9) *with fixed $k > -a$ and $l > -\tilde{a}$ satisfies*

$$\big|\mathbb{E}[\hat{T}_f^{(k,l)}] - T_f(p,q)\big| = \tilde{O}\big(m^{-\lambda(\sigma_p,a,k)} + n^{-\lambda(\sigma_q,\tilde{a},l)}\big),$$

as $m, n \to \infty$, where the rate exponent function $\lambda(\sigma, a, k)$ is as defined in (III.6).

Theorem IV.5 (Variance rate)**.** *For a target functional $T_f(\cdot,\cdot)$, if the underlying density p satisfies* ***(U$_p$)***, ***(L1$_p$)***, ***(L2$_p$)***, *and* ***(L3$_p$)***, *and q satisfies* ***(U$_q$)***, ***(L1$_q$)***, ***(L2$_q$)***, *and* ***(L3$_q$)***, *then the estimator* (I.9) *with fixed $k > -2a$ and fixed $l > -2\tilde{a}$ satisfies*

$$\mathrm{Var}\big(\hat{T}_f^{(k,l)}\big) = O(m^{-1}). \tag{IV.2}$$

Combining Theorems IV.4 and Theorem IV.5, we obtain the convergence rate in MSE and conclude the L_2-consistency of the estimator.

Corollary IV.6 (Convergence rate)**.** *Under the same assumptions in Theorem IV.4, then the estimator* (I.9) *with fixed $k > -2a$ and fixed $l > -2\tilde{a}$ satisfies*

$$\begin{aligned} &\mathbb{E}\big[\big(\hat{T}_f^{(k,l)} - T_f(p,q)\big)^2\big] \\ &= \tilde{O}\big(m^{-2\lambda(\sigma_p,a,k)} + n^{-2\lambda(\sigma_q,\tilde{a},l)} + m^{-1}\big) \end{aligned} \tag{IV.3}$$

and thus is L_2-consistent.

Remark IV.2. Similar to the single-density case, if $d \ge 2$, the bias bound dominates the variance bound.

Example IV.8 (KL divergence; Example IV.1 contd.). For estimating KL divergence, recall that $|\phi_{kl}(u,v)| \lesssim \psi_{-\epsilon,\epsilon}(u)\psi_{-\epsilon,\epsilon}(v)$ for any arbitrarily small $\epsilon > 0$. It can be shown, using Theorems IV.4 and IV.5, that for estimating the (forward) KL or reverse KL divergences between any two densities p and q such that $\mathcal{P} \ll \mathcal{Q}$, each of which is either the uniform distribution, or one of the truncated Gaussian, Cauchy, Laplace, or exponential distributions, we obtain a bias bound of $\tilde{O}(m^{-1/d})$ and a variance bound of $O(m^{-1})$, and therefore, the MSE rate of $\tilde{O}(m^{-2/d} + n^{-2/d} + m^{-1})$ as established in Corollary IV.6.

Example IV.9 (α-divergence; Example IV.2 contd.). For estimating the α-divergence ($\alpha > 0$), recall that $|\phi_{kl}(u,v)| \lesssim \psi_{1-\alpha,1-\alpha}(u)\psi_{\alpha-1,\alpha-1}(v)$ for any $k, l \in \mathbb{N}$ such that $k > \alpha - 1$ and $l > 1 - \alpha$. Hence, if p satisfies **(U$_p$)**, **(L1$_p$)**, **(L2$_p$)**, **(L3$_p$)**, **(S$_p$)**, and **(B$_p$)**, and q satisfies **(U$_q$)**, **(L1$_q$)**, **(L2$_q$)**, **(L3$_q$)**, **(S$_q$)**, and **(B$_q$)**, then the MSE of the estimator (I.9) with $k > 2(\alpha - 1)$

and $l > 2(1-\alpha)$ is bounded as (IV.3) with the bias rate exponents

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

and

$$\lambda(\sigma_q, \tilde{a}, l) = \frac{1}{d}(\sigma_q \wedge 1).$$

This result also holds for the logarithmic α-divergence.

C. Le Cam distance and Jensen–Shannon divergence: Performance guarantee with truncation

The statements in the previous section do not apply to the estimators for Le Cam distance (Example IV.4) and Jensen–Shannon divergence (Example IV.5). The difficulty arises from the fact that the estimator function ϕ_{kl} for these divergences have lower-polynomial-tail exponents $(a, \tilde{a}) = (-k+1, -l+1)$ which become smaller with larger k and l. Therefore, while the bias guarantees (Theorems IV.1 and IV.4) are still applicable, we cannot control the variance of the estimator using Theorems IV.2 or IV.5, as $(a, \tilde{a}) = (-k+1, -l+1)$ does not meet the requirements $\{k > -4a,\ l > -4\tilde{a}\}$ or $\{k > -2a,\ l > -2\tilde{a}\}$.

To handle the variance of the estimator for these exceptional cases, we consider a truncated version of the estimator (I.9). For functionals of two densities, we define the truncated estimator as

$$\begin{aligned} &\overline{T}_f^{(k,l)}(\mathbf{X}_{1:m}, \mathbf{Y}_{1:n}) \\ &:= \frac{1}{m}\sum_{i=1}^{m} \bar{\phi}_{kl}(U_{km}(\mathbf{X}_i), V_{ln}(\mathbf{X}_i); \tau_m, \nu_m, \tilde{\tau}_n, \tilde{\nu}_n), \end{aligned} \tag{IV.4}$$

where we define the truncated estimator function

$$\bar{\phi}_{kl}(u, v; \tau, \nu, \tilde{\tau}, \tilde{\nu}) := \phi_{kl}(u, v)\mathbb{1}_{(\tau,\nu)}(u)\mathbb{1}_{(\tilde{\tau},\tilde{\nu})}(v)$$

and the *truncation points* $\tau_m, \nu_m, \tilde{\tau}_n, \tilde{\nu}_n$ are hyperparameters such that $0 \leq \tau_m \leq 1 \leq \nu_m \leq \infty$ and $0 \leq \tilde{\tau}_n \leq 1 \leq \tilde{\nu}_n \leq \infty$. As noted earlier, we do not require the upper-truncation points in contrast to Section III-C and thus only consider a *lower-truncated estimator* with $\nu_m = \infty$ and $\tilde{\nu}_n = \infty$ in this section.

We can first establish the consistency of the lower-truncated estimator.

Proposition IV.7 (Consistency)**.** *For a target functional $T_f(\cdot,\cdot)$, if the estimator function $\phi_{kl}(u, v)$ is continuous and the underlying densities p and q satisfy **(U$_{pp}$;** k, a**)**, **(L$'_{pp}$;** ξ^2, b**)**, **(U$_{pq}$;** $l, \tilde{a}$**)**, and **(L$'_{pq}$;** $\xi^2, \tilde{b}$**)** with $\xi(t) = t^2$, then the lower-truncated estimator* (IV.4) *with fixed $k > -a$ and $l > -\tilde{a}$ and with lower-truncation points such that $\tau_m^{(k+4a)\wedge 0}\tilde{\tau}_n^{(l+4\tilde{a})\wedge 0} = o(m)$ is L_2-consistent.*

We can also establish convergence rate of the truncated estimator IV.4 for functionals of two densities. Define a lower truncation point function as

$$\begin{aligned} &\tau(m, \sigma, a, k) \\ &= \begin{cases} \Theta\big(m^{-\frac{\sigma\wedge 1}{d(k-1)}}\big) & \text{if } a \leq -\frac{\sigma}{d} - 1, \\ \Theta\big(m^{-\frac{1}{d(k-1)}}\big) & \text{if } -\frac{\sigma}{d} - 1 < a \leq -1, \\ O\big(m^{-\frac{1}{d(a+1)}}\big) & \text{if } a > -1. \end{cases} \end{aligned} \tag{IV.5}$$

Proposition IV.8 (Convergence rate)**.** *For a target functional $T_f(\cdot,\cdot)$, if the underlying density p satisfies the conditions **(U$_p$)**, **(L1$_p$)**, **(S$_p$)**, and **(B$_p$)**, and q satisfies the conditions **(U$_q$)**, **(L1$_q$)**, **(S$_q$)**, and **(B$_q$)**, the truncated estimator* (IV.4) *with fixed $k > -a$ and $l > -\tilde{a}$ satisfies*

$$\begin{aligned} &\mathbb{E}\big[\big(\overline{T}_f^{(k,l)} - T_f(p,q)\big)^2\big] \\ &= \tilde{O}\big(m^{-2\lambda(\sigma_p,a,k)} + n^{-2\lambda(\sigma_q,\tilde{a},l)} + m^{-1}\tau_m^{(2a+k)\wedge 0}\tilde{\tau}_n^{(2\tilde{a}+l)\wedge 0}\big), \end{aligned}$$

as $m, n \to \infty$, and thus is L_2-consistent.

Example IV.10 (Le Cam distance; Example IV.4 contd.). For estimating $T_f(p,q)$ with $f(p,q) = q/(p+q)$, recall that $|\phi_{kl}(u,v)| \lesssim \psi_{-k+1,l-1}(u)\psi_{-l+1,k-1}(v)$ for any $k \geq 1$ and $l \geq 1$. For densities p and q satisfying conditions in Proposition IV.7, the lower-truncated estimator (IV.4) for Le Cam distance is L_2-consistent. In particular, the estimator with $k = l = 1$ is consistent even without lower truncation, since $\tau_m^{(k+4a)\wedge 0}\tilde{\tau}_n^{(l+4\tilde{a})\wedge 0} = \tau_m^0\tilde{\tau}_n^0 = 0$ with $\tau_m = \tilde{\tau}_n = 0$ and $k = l = 1$. If the underlying densities p and q satisfy the conditions in Proposition IV.8, then the lower-truncated estimator with fixed $k \geq 1$ and $l \geq 1$ and truncation points $\tau_m = \tau(m, \sigma_p, -k+1, k)$, and $\tilde{\tau}_n = \tau(n, \sigma_q, -l+1, l)$ satisfies

$$\begin{aligned} &\mathbb{E}\big[\big(\hat{T}_f^{(k,l)} - T_f(p,q)\big)^2\big] \\ &= \tilde{O}\big(m^{-2\lambda_k(\sigma_p)} + n^{-2\lambda_l(\sigma_q)} + m^{-1}\tau_m^{(-k+2)\wedge 0}\tilde{\tau}_n^{(-l+2)\wedge 0}\big), \end{aligned} \tag{IV.6}$$

as $m, n \to \infty$, where $\lambda_p = \lambda_k(\sigma_p)$ and $\lambda_q = \lambda_l(\sigma_q)$, where

$$\lambda_k(\sigma) := \lambda(\sigma, -k+1, k) = \begin{cases} \frac{1}{d}(\sigma \wedge 1) & \text{if } k = 1, \\ \frac{1}{d}(\sigma \wedge \frac{1}{k-1}) & \text{if } 2 \leq k < 2 + \frac{\sigma}{d}, \\ \frac{1}{d}\frac{\sigma\wedge 1}{k-1}, & \text{if } k > 2 + \frac{\sigma}{d}. \end{cases}$$

Based on this rate-exponent expression and the additional factor of $\tau_m^{(2a+k)\wedge 0}\tilde{\tau}_n^{(2\tilde{a}+l)\wedge 0}$ in the variance rate which only worsens the rate with larger k and $l4$, one would expect that the convergence becomes only slower as k and/or l become large, and thus, the fastest rate achieved is $\tilde{O}(m^{-\frac{2}{d}(\sigma_p\wedge 1)} + n^{-\frac{2}{d}(\sigma_q\wedge 1)} + m^{-1})$, when $k = 1$ and $l = 1$ with lower truncation points $\tau_m = 0$ and $\tilde{\tau}_n = \Theta(n^{-\frac{1}{d}})$. This is in contrast with Remark III.5, where we observed faster convergence with larger values of k when a does not decrease in k. We note that the experiments with synthetic data in Section VI show that the estimator performs well even for large values of k and l, suggesting that the detrimental effect of the lower tail exponents might be removed with a tighter analysis.

Example IV.11 (Jensen–Shannon divergence; Example IV.5 contd.). For estimating Jensen–Shannon divergence, recall that $|\phi_{kl}(u,v)| \lesssim \psi_{-k+1,l-1}(u)\psi_{-l+1,k-1}(v)$ for any $k \geq 1$ and

$l \geq 2$. For densities p and q satisfying conditions in Proposition IV.7, the lower-truncated estimator (IV.4) for Jensen–Shannon divergence is L_2-consistent. Also, we do not require the lower-truncation τ_m for $k = 1$, by the same argument in the previous example. If the underlying densities p and q satisfy the conditions in Proposition IV.8 and additionally $\mathcal{Q} \ll \mathcal{P}$, then the estimator (I.9) with fixed $k \geq 1$ and $l \geq 2$ and the same truncation points in Example IV.10 satisfies (IV.6). The established rate seems to get only slower as k and/or l become large, and thus achieves its fastest rate $\tilde{O}(m^{-\frac{2}{d}(\sigma_p \wedge 1)} + n^{-\frac{2}{d}(\sigma_q \wedge 1)} + m^{-1})$ when $k = 1$ and $l = 2$ with lower truncation points $\tau_m = 0$ and $\tilde{\tau}_n = \Theta(n^{-\frac{1}{d}})$. Note, however, this conclusion might not hold in practice; see Example IV.10.

V. Adaptive Choices of k and l

In Section III, we established the convergence rate of the proposed estimator (I.3) for fixed k. Since $\mathbb{E}[\phi_k(U_{k\infty}(\mathbf{x}))] = f(p(\mathbf{x}))$ for each valid $k \in \mathbb{N}$ by design, we can choose any valid k without violating the asymptotic unbiasedness. In Remark III.5, we observed that a larger *fixed* k in general leads to a larger rate exponent in (III.6), and thus, a faster convergence rate. This prompts the question of whether increasing $k \to \infty$ along with m improves the convergence rate upon fixed k. The following proposition answers this in the affirmative. The proof is deferred to Appendix D-B.

Proposition V.1 (Convergence rate and L_2-consistency with increasing k)**.** *For a target functional $T_f(\cdot)$, if the underlying density p satisfies* ***(U$_p$)**, **(L1$_p$)**, **(L2$_p$)**, **(L3$_p$)**, **(S$_p$)**, and **(B$_p$)**, then the estimator* (I.3) *with $k = \Theta((\ln m)^{1.1})$ satisfies*

$$\left|\mathbb{E}[\hat{T}_f^{(k)}] - T_f(p)\right| = \tilde{O}\left(m^{-\frac{\sigma_p \wedge 1}{d}}\right) \tag{V.1}$$

as $m \to \infty$. Furthermore, the estimator (I.3) *satisfies*

$$\mathbb{E}\left[\left(\hat{T}_f^{(k)} - T_f(p)\right)^2\right] = \tilde{O}\left(m^{-\frac{2(\sigma_p \wedge 1)}{d}} + m^{-1}\right) \tag{V.2}$$

and thus is L_2-consistent.

Remark V.1. As expected heuristically, the bias rate exponent $(\sigma_p \wedge 1)/d$ in (V.1) equals the limit of the finite-k rate exponent in (III.6) as $k \to \infty$.

Remark V.2. There is no consensus on the optimal choice of k for functional estimation in the literature. For example, Singh and Póczos [52] analyzed $k = O(1)$, whereas Berrett et al. [46] suggested $k = O((\ln m)^5)$ for asymptotic efficiency of the estimator, a slightly faster choice than the previous theorem, for differential entropy. Pérez-Cruz [75] discussed some relevant empirical results on the choice of k.

Remark V.3. While our main focus in this paper is to establish consistency and convergence rates for the proposed estimators with fixed k (and l), we point out that a tighter analysis on the dependence on k may lead to a better asymptotic convergence rate. Note that the analysis of Kozachenko–Leonenko estimator by Berrett et al. [46] allows polynomial growth of k in the sample size. The loose dependence on k in our analysis can be traced back to Lemma B.4, which quantifies the gap between densities of the normalized volume of k-NN ball $U_{km}(\mathbf{x})$ and its limiting Poisson random variable $U_{k\infty}(\mathbf{x})$. To tighten the bound, one needs to sharpen Lemma B.5 on the speed of convergence of a Poisson binomial random variable to a Poisson random variable.

Example V.1 (Differential entropy; Example III.7 contd.). Applying Proposition V.1 on differential entropy with $k = \Theta((\ln m)^{1.05})$, we obtain the MSE rate (V.2). This rate is the same as the fixed-k case in Example III.7.

Example V.2 (α-entropy; Example III.8 contd.). Applying Proposition V.1 on α-entropy with $k = \Theta((\ln m)^{1.05})$, we obtain the bias rate exponent $(\sigma_p \wedge 1)/d$, which is greater than or equal to that in Example III.8 with k fixed.

Similarly to the single-density case, we can establish the convergence rate when k and l vary polylogarithmically with m and n, provided that m and n grow to infinity in the same speed, i.e., $m \asymp n$. The following proposition can be proved by extending the proof of Proposition V.1 to the double-density case as in the proofs of Theorems IV.4 and IV.5, and thus is omitted.

Proposition V.2 (Convergence rate and L_2-consistency with increasing k and l)**.** *For a target functional $T_f(\cdot,\cdot)$, if the underlying densities p and q satisfy the conditions* ***(U$_p$)**, **(L1$_p$)**, **(L2$_p$)**, **(L3$_p$)**, **(S$_p$)**, **(B$_p$)**, **(U$_q$)**, **(L1$_q$)**, **(L2$_q$)**, **(L3$_q$)**, **(S$_q$)**, and **(B$_q$)**, then the estimator* (I.9) *with $k = \Theta((\ln m)^{1.1})$ and $l = \Theta((\ln n)^{1.1})$ satisfies*

$$\left|\mathbb{E}[\hat{T}_f^{(k,l)}] - T_f(p,q)\right| = \tilde{O}(m^{-\frac{\sigma_p \wedge 1}{d}} + n^{-\frac{\sigma_q \wedge 1}{d}}),$$

as $m, n \to \infty$ with $m \asymp n$. Furthermore, the estimator (I.9) *satisfies*

$$\begin{aligned}&\mathbb{E}\left[(\hat{T}_f^{(k,l)} - T_f(p,q))^2\right] \\ &\quad = \tilde{O}(m^{-\frac{2(\sigma_p \wedge 1)}{d}} + n^{-\frac{2(\sigma_q \wedge 1)}{d}} + m^{-1}),\end{aligned} \tag{V.3}$$

and thus is L_2-consistent, provided that $m \asymp n$.

Remark V.4. For $d \geq 2$, if k and l increase as in Proposition V.2, the bias bound always dominates the variance bound so that the MSE is bounded as $O(m^{-1})$. For $d = 1$, the variance bound may dominate the bias bound depending on σ_p, σ_q, d, and/or the choices of k and l.

Example V.3 (KL divergence; Example IV.8 contd.). Letting k and l increase as $k = \Theta((\ln m)^{1.05}))$ and $l = \Theta((\ln n)^{1.05}))$, we obtain the MSE rate (V.3) for estimating KL divergence. As a complementary asymptotic result, Wang et al. [31] showed that the (k,l)-NN KL divergence estimator with $k = k_m$ and $l = l_n$ such that $k_m/m \to 0$ and $k_m/(\ln m) \to \infty$ as $m \to \infty$ and $l_n/n \to 0$ and $l_n/(\ln n) \to \infty$ as $n \to \infty$ converges to the true KL divergence almost surely for uniformly continuous densities bounded from below on their support.

Example V.4 (α-divergence; Example IV.9 contd.). Letting k and l increase as $k = \Theta((\ln m)^{1.05}))$ and $l = \Theta((\ln n)^{1.05}))$, the MSE of our estimator is bounded as (V.3).

VI. Numerical Results

The performance of the proposed estimators (I.3) and (I.9) for several density functionals were simulated over 500 runs for sample sizes ranging from 100 till 25600.[2] For each dimension d from 1 through 5, we considered the uniform density $\mathsf{Unif}([0,1]^d)$, the Gaussian density $\mathsf{N}(0,I_d)$ restricted to $\|\mathbf{x}\| \le 3$, and the Gaussian density $\mathsf{N}(0,I_d)$ as the density p. For double-density functionals, we considered $\mathsf{Unif}([0,2]^d)$, $\mathsf{N}(0,4I_d)$ restricted to $\mathbb{B}(0,3)$, and $\mathsf{N}(0,4I_d)$ as the density q.[3] Note that all the functionals considered in these simulations can be expressed in closed form up to incomplete gamma function, except the exponential entropies, Le Cam distance, and Jensen–Shannon divergences for Gaussian densities. We estimated the latter using Monte Carlo approximation. Polynomial rates of convergence were observed for all cases, and in each case, the exponent was calculated by ordinary least-squares linear regression between the logarithms of the sample sizes and the MSE. We considered $k \in \{1,2,3,4,5,10,15\}$ and, for double-density functional estimators, $l = k$ for simplicity.

Figure 1 presents the convergence of the estimator for differential entropy, α-entropies for $\alpha \in \{0.5,1.5\}$, logarithmic 2-entropy, and exponential $(2.5,1)$-entropy for 3-dimensional densities. The simulation results show that smaller k yields faster convergence while incurring larger variance, which suggests the use of a moderate size of k in practice. Figure 2 summarizes the empirical exponents of the estimator for each functional and density. A simple upper bound $(2/d) \wedge 1$ on the theoretical exponents established in Corollary III.7 is also plotted for comparison; see also Examples III.7 and III.8. Empirical convergence rates are consistently better than theoretical bounds for the truncated densities.

Corresponding simulation results for a few representative double-density functionals (KL divergence, α-divergence, logarithmic α-divergence, Le Cam distance, and Jensen–Shannon divergence) are presented in Figures 3 and 4. These simulations indicate that the requirement $k > -4a$ and $l > -4\tilde{a}$ in Theorem IV.2 may be relaxed to the milder condition $k > -2a$ and $l > -2\tilde{a}$. For example, the estimator with $k = l = 4$ for logarithmic 2-divergence ($k = 3 \le -4(1-2) = 4$ and $l = 3 \le -4(1-2)$) still exhibit consistency in Figure 3. As presented in the last two rows in Figures 3 and 4, simulations also indicate that our estimator is consistent in practice for the exceptional examples of Le Cam distance and Jensen–Shannon divergence even without truncation. For estimating Le Cam distance, we observed that using too large values for k or l lead to bad convergence behavior for small dimensions; see, e.g., the case of $k = l = 15$ for $d = 1$ at the second column of the fourth row in Figure 4.

VII. Concluding Remarks

In this paper, we developed a systematic approach to designing k-NN based consistent estimators for a variety of functionals, starting from the fundamental requirement of asymptotic unbiasedness and utilizing the limiting behavior of the k-NN statistics (Proposition I.1). The proposed estimators rediscovered and unified several existing k-NN based estimators for Shannon entropy, KL divergence, α-entropies and α-divergences, and polynomial functionals, which have been sporadically studied and individually analyzed in the literature. It demystified the need of the known, but rather ad-hoc "bias corrections" for some functionals, providing an alternative, principled recipe to identify L_2-consistent estimators. Our list of examples is not exhaustive; other density functionals in the same form may exist or may be discovered in future, and our recipe will furnish consistent k-NN estimators for the same, with nonasymptotic performance predicted by our current analysis.

We remark that the established convergence rates are not minimax optimal; see Remark III.9. As further noted in Remark III.8, 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 [39, 40] that takes a weighted average of multiple estimators based on the asymptotic bias expansion of each density functional estimator. Studying the ensemble version of the estimators is beyond our scope and left as a future direction; see [42] for a weighted version of the proposed divergence functional estimator with local minimax optimality.

Throughout the paper, we assumed the Euclidean distance $\rho(\mathbf{x},\mathbf{y}) = \|\mathbf{x}-\mathbf{y}\|$. We conclude the paper with specifying technical issues one needs to address in order to extend the results of this paper to a general metric measure space $(\mathcal{X},\rho,\mu)$, where $(\mathcal{X},\rho)$ is a complete separable metric space and μ is a locally finite measure on the Borel σ-algebra of $\mathcal{X}$ (see, e.g., Sturm [76]). Consider a μ-absolutely continuous probability measure $\mathcal{P}$ with density p. In general, the weak convergence property in Proposition I.1 for asymptotic unbiasedness (Theorems III.1 and IV.1) requires the Lebesgue differentiation theorem to hold in the metric measure space $(\mathcal{X},\rho,\mu)$, i.e., we need

$$\lim_{r\to 0} \frac{\mathcal{P}(\mathbb{B}(x,r))}{\mu(\mathbb{B}(x,r))} = p(x)$$

for μ-a.e. $x \in \mathcal{X}$. Further, for the bias rate analysis to work, we need to extend Lemma B.6, which states that if p is locally σ-Hölder smooth on $\mathbb{B}(x,R)$, then for $r < R$,

$$\left|\frac{\mathcal{P}(\mathbb{B}(x,r))}{\mu(\mathbb{B}(x,r))} - p(x)\right| \lesssim r^\sigma \text{ and } \left|\frac{\mathrm{d}\,\mathcal{P}(\mathbb{B}(x,r))}{\mathrm{d}\mu(\mathbb{B}(x,r))} - p(x)\right| \lesssim r^\sigma .$$

If there exists a nonsmooth boundary, we then further need Lemma B.22 to hold in the metric measure space. For the variance analysis to hold under p-norm and other norms, we can apply and extend the analysis in [30] as pointed out earlier in Remark III.2.

Appendix A
Notation

In what follows, let $\mathrm{P}_U(u) = \Pr\{U \le u\}$ and $\rho_U(u) = \mathrm{d}\mathrm{P}_U(u)/\,\mathrm{d}u$ denote the cumulative distribution function (cdf)

[2] The code is available at https://github.com/jongharyu/knn-functional-estimation.

[3] As an exception for the experiment with the Jensen–Shannon divergence estimator, instead of $\mathsf{Unif}([0,1]^d)$ and $\mathsf{Unif}([0,2]^d)$, we used piecewise constant densities p and q supported on $[0,1]^d$, which are defined as follows:

$$p(\mathbf{x}) = \begin{cases} 3/2 & \text{if } 0 \le x_1 \le 1/2, \\ 1/2 & \text{if } 1/2 < x \le 1, \end{cases} \quad \text{and} \quad q(\mathbf{x}) = \begin{cases} 1/2 & \text{if } 0 \le x_1 \le 1/2, \\ 3/2 & \text{if } 1/2 < x \le 1. \end{cases}$$

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3151231, IEEE Transactions on Information Theory

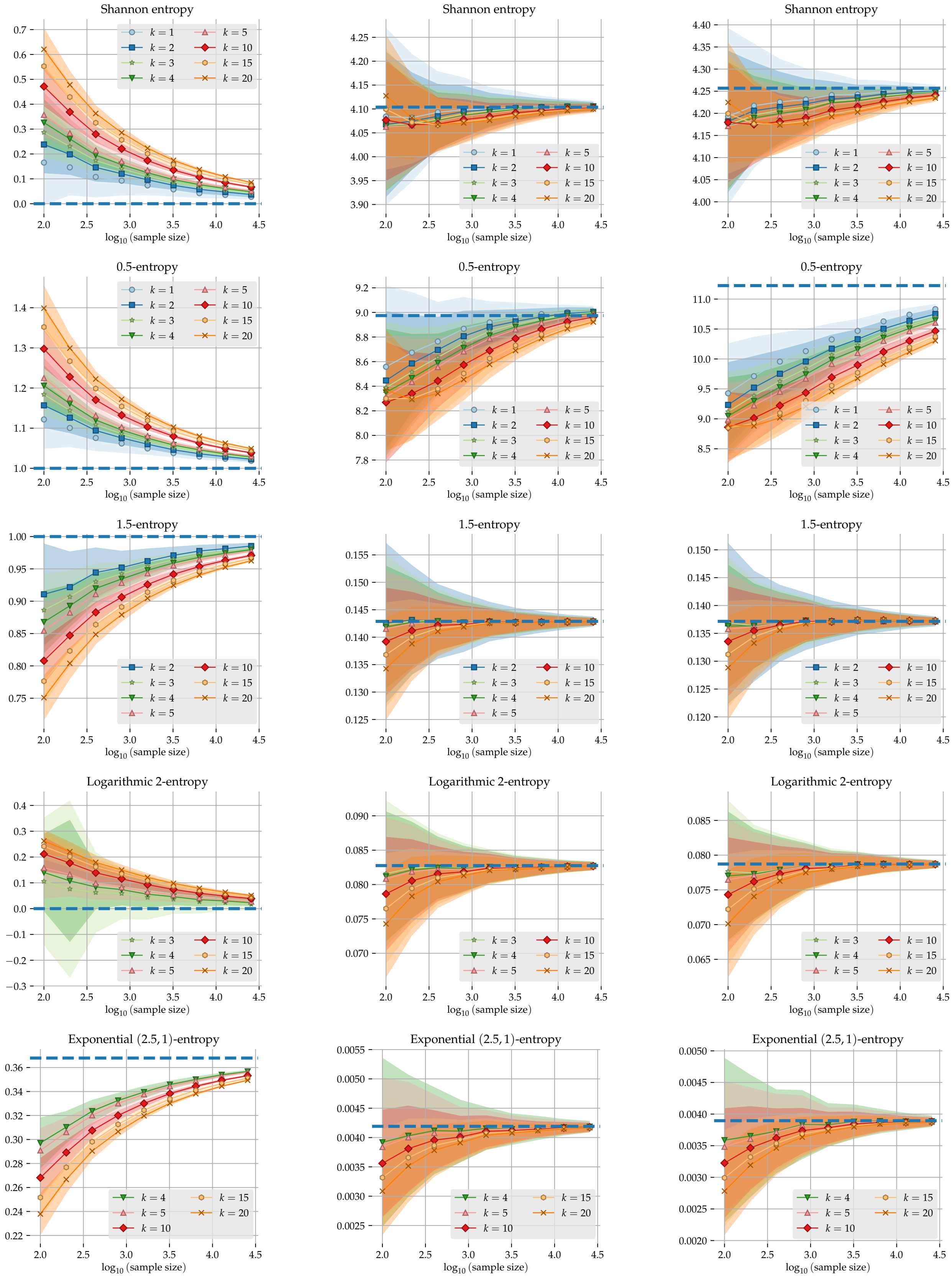

Fig. 1. Convergence of the single-density functional estimator for differential entropy, α-entropies $\alpha \in \{0.5, 1.5\}$, logarithmic 2-entropy, and exponential $(2.5, 1)$-entropy for 3-dimensional densities. The first, second, and third columns present simulation results with $\mathsf{Unif}([0,1]^3)$, $\mathsf{N}(0, I_3)$ restricted to $\|\mathbf{x}\| \leq 3$, and $\mathsf{N}(0, I_3)$, respectively. The true functional values are indicated as dashed lines and one sample standard deviations of the estimates are indicated as shaded area.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3151231, IEEE Transactions on Information Theory

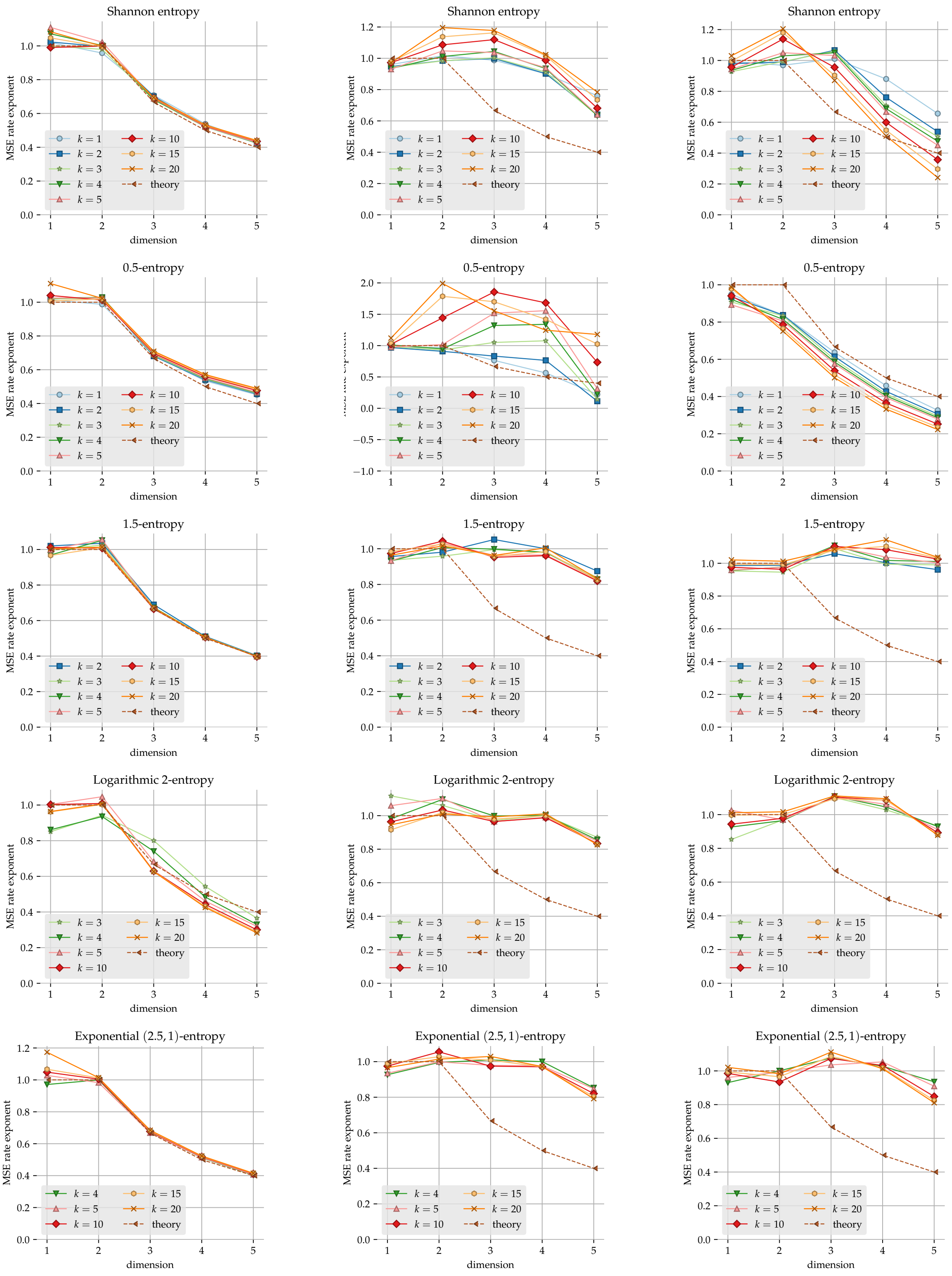

Fig. 2. Simulated MSE rate exponents of the single-density functional estimator for differential entropy, α-entropies for $\alpha \in \{0.5, 1.5\}$, logarithmic 2-entropy, and exponential $(2.5, 1)$-entropy. The first, second, and third columns present simulation results with $\mathsf{Unif}([0,1]^d)$, $\mathsf{N}(0, I_d)$ restricted to $\|\mathbf{x}\| \leq 3$, and $\mathsf{N}(0, I_d)$, respectively, for $d \in \{1, 2, 3, 4, 5\}$.

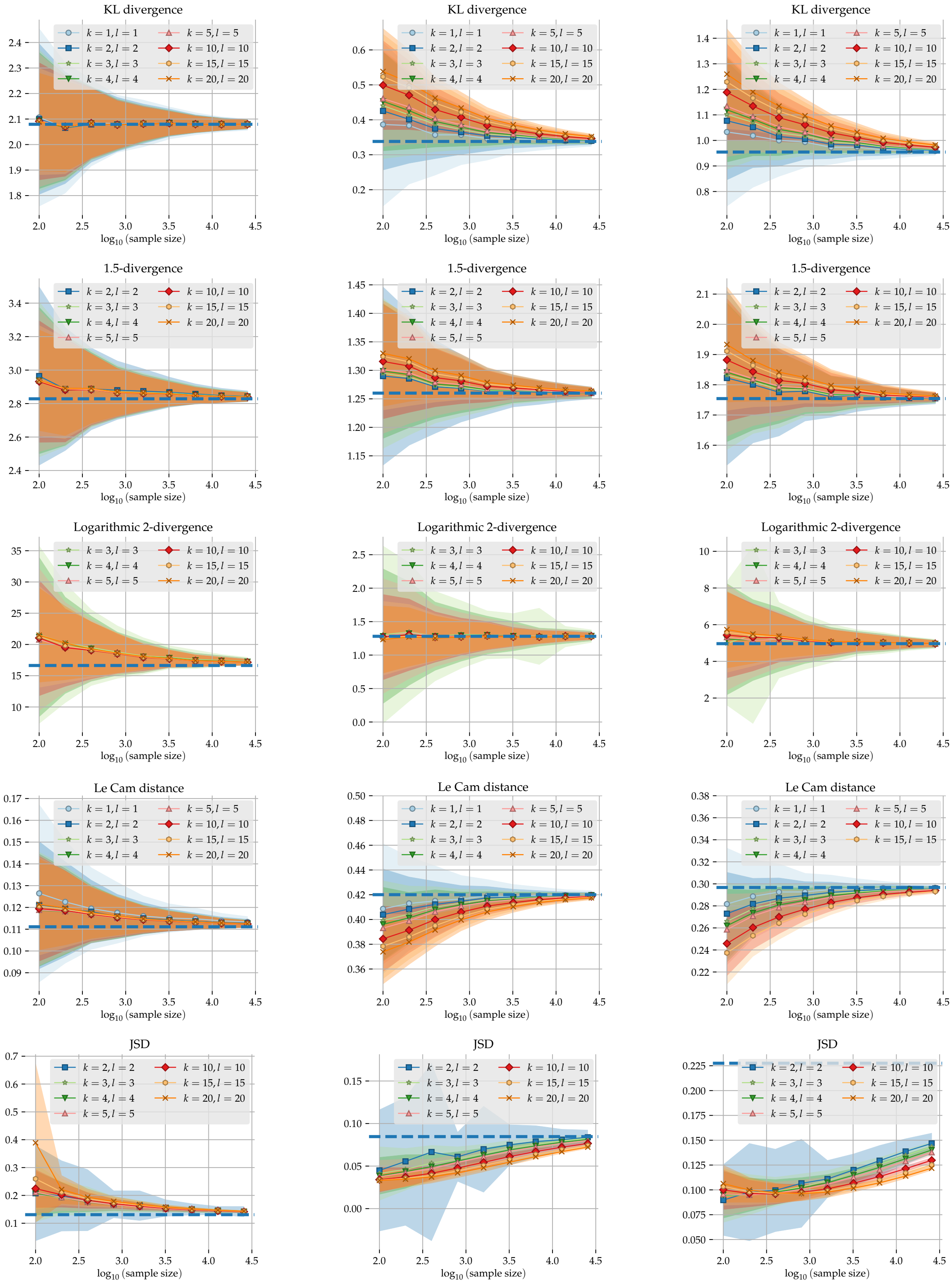

Fig. 3. Convergence of the double-density functional estimator for KL divergence, 1.5-divergence, and logarithmic 2-divergence for 3-dimensional densities. The first, second, and third columns present simulation results for the densities p and q considered as $\mathsf{Unif}([0,1]^3)$ and $\mathsf{Unif}([0,2]^3)$, $\mathsf{N}(0,I_3)$ restricted to $\|\mathbf{x}\| \le 3$ and $\mathsf{N}(0,4I_3)$ restricted to $\|\mathbf{x}\| \le 3$, and $\mathsf{N}(0,I_3)$ and $\mathsf{N}(0,4I_3)$, respectively. The true functional values are indicated as dashed lines and one sample standard deviations of the estimates are indicated as shaded area. LCD and JSD are abbreviations for Le Cam distance and Jensen–Shannon divergence, respectively.

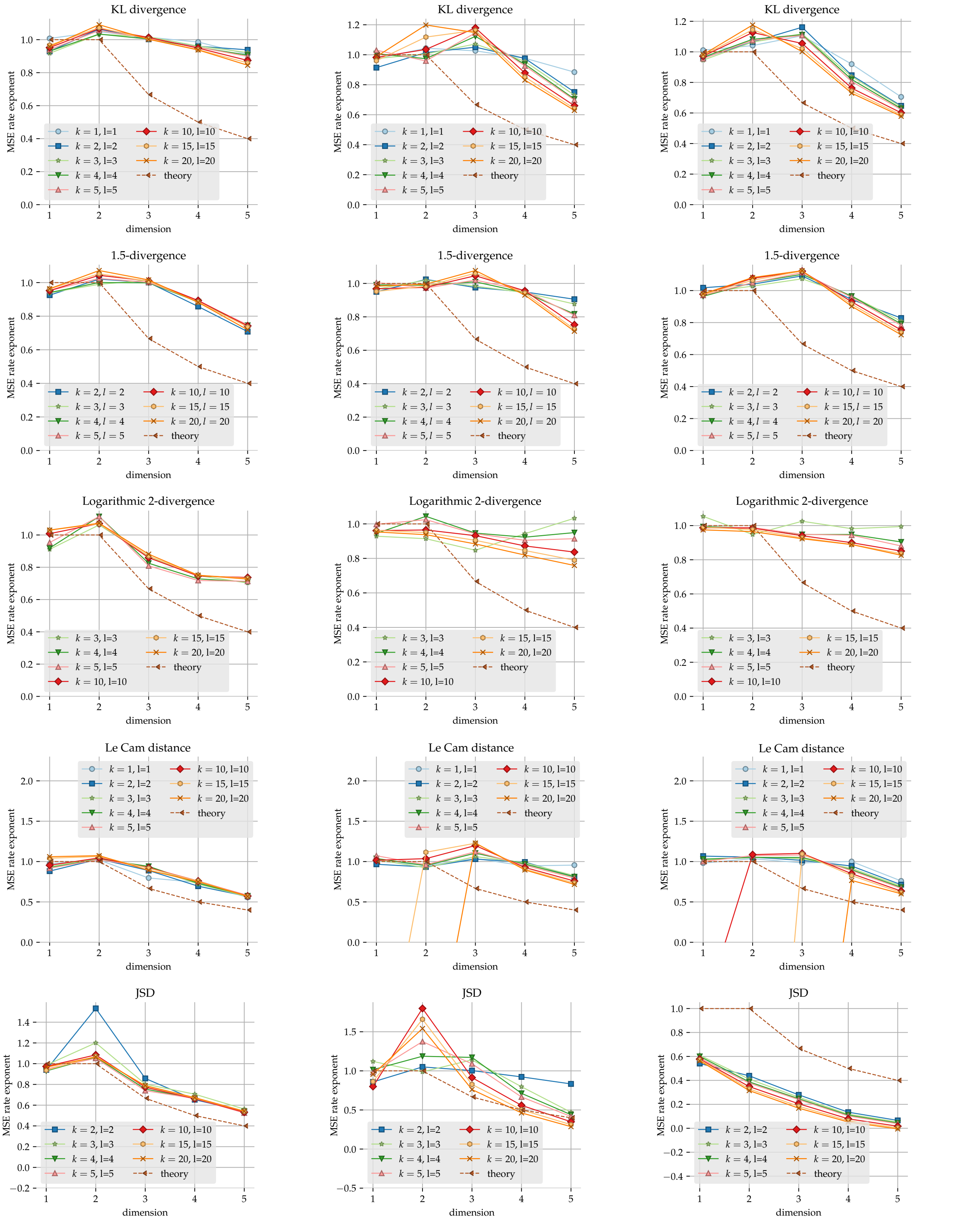

Fig. 4. Simulated MSE rate exponents of the double-density functional estimator for KL divergence, 1.5-divergence, and logarithmic 2-divergence. The first, second, and third columns present simulation results for the densities p and q considered as $\mathsf{Unif}([0,1]^3)$ and $\mathsf{Unif}([0,2]^3)$, $\mathsf{N}(0, I_3)$ restricted to $\|\mathbf{x}\| \leq 3$ and $\mathsf{N}(0, 4I_3)$ restricted to $\|\mathbf{x}\| \leq 3$, and $\mathsf{N}(0, I_3)$ and $\mathsf{N}(0, 4I_3)$, respectively, for $d \in \{1, 2, 3, 4, 5\}$. LCD and JSD are abbreviations for Le Cam distance and Jensen–Shannon divergence, respectively.

and the density of a random variable U, respectively. We use $B_{n,P}$ to denote a binomial random variable with parameters n and P. We use P_q to denote a Poisson random variable with rate $q > 0$. We use $X_{\alpha,\beta}$ to denote a beta random variable with parameters $\alpha, \beta > 0$ for $\alpha, \beta > 0$, whose density is

$$\frac{t^{\alpha-1}(1-t)^{\beta-1}}{\mathsf{B}(\alpha,\beta)}, \quad 0 \le t \le 1.$$

Here $\mathsf{B}(\alpha,\beta) := \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}\,\mathrm{d}t$ denotes the beta function. Finally, we use H^{d-1} to denote the $(d-1)$-dimensional Hausdorff measure.

Appendix B
Technical lemmas

A. Auxiliary lemmas

Lemma B.1. *Assume that* $\mathcal{P}$ *and* $\tilde{\mathcal{P}}$ *have densities* p *and* $\tilde{p}$*, respectively, with respect to the Lebesgue measure* λ_{Leb}*. If* $\mathcal{P} \ll \tilde{\mathcal{P}}$*, then* $\mathcal{P}(\{\mathbf{x}\colon m_r\tilde{p}(\mathbf{x}) > 0\}) = 1$ *for any* $r > 0$*.*

Proof. Let $r > 0$ be fixed. We first observe that $\mathcal{P}(\mathrm{supp}(\tilde{p})) = 1$, since

$$\begin{aligned} 1 - \mathcal{P}(\mathrm{supp}(\tilde{p})) &= \int p(\mathbf{x})(1 - \mathbb{1}_{\mathrm{supp}(\tilde{p})}(\mathbf{x}))\,\mathrm{d}\mathbf{x} \\ &= \int p(\mathbf{x})\mathbb{1}_{\{\exists\delta>0 \text{ s.t.} \tilde{\mathcal{P}}(\mathbb{B}(\mathbf{x},\delta))=0\}}\,\mathrm{d}\mathbf{x} \\ &\overset{(a)}{\le} \int p(\mathbf{x})\mathbb{1}_{\{\exists\delta>0 \text{ s.t. } \mathcal{P}(\mathbb{B}(\mathbf{x},\delta))=0\}}\,\mathrm{d}\mathbf{x}, \\ &\overset{(b)}{=} 0. \end{aligned}$$

Here, (a) follows from the absolute continuity $\mathcal{P} \ll \tilde{\mathcal{P}}$, and (b) follows since $p(\mathbf{x}) = 0$ for a.e. $\mathbf{x}$ over the set $\{\mathbf{x}\colon \exists\delta > 0 \text{ s.t.} \mathcal{P}(\mathbb{B}(\mathbf{x},\delta)) = 0\}$, by the Lebesgue differentiation theorem.

Now, define $A_\delta\tilde{p}(\mathbf{x}) = \tilde{\mathcal{P}}(\mathbb{B}(\mathbf{x},\delta))/\lambda_{\mathrm{Leb}}(\mathbb{B}(\mathbf{x},\delta))$ for each $\delta > 0$ and $\mathbf{x} \in \mathbb{R}^d$. On the one hand, we have

$$\lim_{\delta\to 0} A_\delta\tilde{p}(\mathbf{x}) = \tilde{p}(\mathbf{x})$$

for λ_{Leb}-a.e. $\mathbf{x}$ by the Lebesgue differentiation theorem. On the other hand, for $\mathbf{x} \in T \cap \mathrm{supp}(\tilde{p})$ where $T := \{\mathbf{x}\colon m_r\tilde{p}(\mathbf{x}) = 0\}$, we have $A_\delta\tilde{p}(\mathbf{x}) > 0$ for every $\delta > 0$ and

$$0 = m_r\tilde{p}(\mathbf{x}) = \inf_{0<\delta\le r} A_\delta\tilde{p}(\mathbf{x})$$

for any $r > 0$. Hence, we must have

$$\tilde{p}(\mathbf{x}) = \lim_{\delta\to 0} A_\delta\tilde{p}(\mathbf{x}) = 0$$

for λ_{Leb}-a.e. $\mathbf{x} \in T \cap \mathrm{supp}(\tilde{p})$, which implies that $\tilde{\mathcal{P}}(T \cap \mathrm{supp}(\tilde{p})) = 0$, and thus $\mathcal{P}(T \cap \mathrm{supp}(\tilde{p})) = 0$ since $\mathcal{P} \ll \tilde{\mathcal{P}}$. This, together with $\mathcal{P}(\mathrm{supp}(\tilde{p})) = 1$, establishes that $\mathcal{P}(T) = 0$. □

Lemma B.2. *For the* lower incomplete gamma function $\gamma(s,x) := \int_0^x t^{s-1}e^{-t}\,\mathrm{d}t$ *and the* upper incomplete gamma function $\Gamma(s,x) := \int_x^\infty t^{s-1}e^{-t}\,\mathrm{d}t$*, we have*

$$\gamma(s,x) \le \Gamma(s) \wedge \frac{x^s}{s}, \qquad \forall s > 0, x > 0, \tag{B.1}$$

$$\Gamma(s,x) \le \Gamma(s)x^{s-1}e^{-x+1}, \qquad \forall s \ge 1, x \ge 1. \tag{B.2}$$

Proof. As $\Gamma(s,x)/\Gamma(s)$ is decreasing in s for fixed $x \ge 1$, we have that for $s \ge 1$,

$$\begin{aligned} \frac{\Gamma(s,x)}{\Gamma(s)} \le \frac{\Gamma(\lfloor s\rfloor, x)}{\Gamma(\lfloor s\rfloor)} &= e^{-x}\sum_{k=0}^{\lfloor s\rfloor-1}\frac{x^k}{k!} \\ &\le e^{-x}x^{\lfloor s\rfloor-1}\sum_{k=0}^{\infty}\frac{1}{k!} \le e^{-x+1}x^{s-1}. \end{aligned}$$

The second inequality follows since, for any $x > 0$, letting $t = xe^{-u}$, we have

$$\begin{aligned} \gamma(s,x) = \int_0^x t^{s-1}e^{-t}\,\mathrm{d}t &= x^s\int_0^\infty e^{-(su+xe^{-u})}\,\mathrm{d}u \\ &\le x^s\int_0^\infty e^{-su}\,\mathrm{d}u = \frac{x^s}{s}. \end{aligned}$$

□

B. Convergence of distribution of k-NN statistics

We first state a basic statistical property of k-NN statistics.

Lemma B.3 (Distribution of k-NN distance)**.** *The cdf of* $r_{km}(\mathbf{x})$ *is*

$$\mathrm{P}_{r_{km}(\mathbf{x})}(r) = \Pr\{B_{m,\mathcal{P}(\mathbb{B}(\mathbf{x},r))} \ge k\} = \mathrm{P}_{X_{k,m-k+1}}(\mathcal{P}(\mathbb{B}(\mathbf{x},r))).$$

Proof. Consider

$$\begin{aligned} \mathrm{P}_{r_{km}(\mathbf{x})}(r) &= \Pr\{r_{km}(\mathbf{x}) \le r\} \\ &= \Pr\{\rho(\mathbf{x}, \mathbf{X}_{(k)}(\mathbf{x})) \le r\} \\ &= \Pr\big\{|\{i \in [m]\colon \mathbf{X}_i \in \mathbb{B}(\mathbf{x},r)\}| \ge k\big\} \\ &= \Pr\{B_{m,\mathcal{P}(\mathbb{B}(\mathbf{x},r))} \ge k\} \\ &= \mathrm{P}_{X_{k,m-k+1}}(\mathcal{P}(\mathbb{B}(\mathbf{x},r))). \end{aligned}$$

The last equality follows from the identity

$$\Pr\{B_{m,P} \ge k\} = \mathrm{P}_{X_{k,m-k+1}}(P).$$

□

Using this fact, Proposition I.1, which claims the weak convergence of the k-NN statistics $U_{km}(\mathbf{x})$ to a Gamma random variable, readily follows.

Proof of Proposition I.1. Fix $\mathbf{x} \in \mathbb{R}^d$ and $u > 0$, and let $P_m := \mathcal{P}(\mathbb{B}(\mathbf{x}, \varrho(\frac{u}{m})))$. Since $\mathrm{P}_{U_{km}(\mathbf{x})}(u) = \mathrm{P}_{r_{km}(\mathbf{x})}(\varrho(\frac{u}{m}))$, we have

$$\mathrm{P}_{U_{km}(\mathbf{x})}(u) = \Pr\{B_{m,P_m} \ge k\}$$

from Lemma B.3. By the Lebesgue differentiation theorem (see, e.g., [77]), for λ_{Leb}-a.e. $\mathbf{x}$,

$$\lim_{m\to\infty} mP_m = \lim_{m\to\infty} u\frac{\mathcal{P}(\mathbb{B}(\mathbf{x},\varrho(\frac{u}{m})))}{\lambda_{\mathrm{Leb}}(\mathbb{B}(\mathbf{x},\varrho(\frac{u}{m})))} = up(\mathbf{x}).$$

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

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

since

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

This leads us to concludes that

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

where $U_{k\infty}(\mathbf{x})$ is a $\mathsf{G}(k,p(\mathbf{x}))$ random variable. □

Moreover, if the density p is locally smooth, then one can establish a polynomial convergence rate of the density of $U_{km}(\mathbf{x})$ to $U_{k\infty}(\mathbf{x})$ as follows.

Lemma B.4 (Generalization of [30, Lemma 2])**.** *Suppose that $\nu_m=o(\sqrt{m})$ and $k=k_m=o(\sqrt{m})$ as $m\to\infty$. For $\mathbf{x}\in$ $\mathrm{supp}(p)$, if $p(\mathbf{x})\le C_p<\infty$ and p is σ_p-Hölder continuous ($\sigma_p\in[0,2]$) over $\mathbb{B}(\mathbf{x},\varrho(\frac{u}{m}))$ with Hölder constant L, we have*

$$\begin{aligned}&\big|\rho_{U_{km}(\mathbf{x})}(u)-\rho_{U_{k\infty}(\mathbf{x})}(u)\big|\\&\lesssim_{\sigma_p,L,C_p,d}(1+u)\Big(\frac{u}{m}\Big)^{\frac{\sigma_p}{d}}+k^{-k}\frac{(k^2+u^2)u^{k-1}e^{-up(\mathbf{x})}}{m}\end{aligned}$$

for $u\in[0,\nu_m]$ and m sufficiently large.

We first state two technical lemmas required to prove Lemma B.4, whose proofs are omitted here; we refer the interested readers to [30]. The first lemma in the following establishes a rate of convergence of a Poisson binomial random variable $B_{m,Q/m}\sim\mathsf{Bin}(m,Q/m)$ to a Poisson random variable $P_Q\sim\mathsf{Poi}(Q)$ in distribution.

Lemma B.5 (Generalization of [30, 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*

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

The second lemma establishes the speed of convergence of $\mathcal{P}(\mathbb{B}(\mathbf{x},r))/\lambda_{\mathrm{Leb}}(\mathbb{B}(\mathbf{x},r))$ and $\mathrm{d}\,\mathcal{P}(\mathbb{B}(\mathbf{x},r))/\mathrm{d}\,\lambda_{\mathrm{Leb}}(\mathbb{B}(\mathbf{x},r))$ to $p(\mathbf{x})$ as $r\to0$, when p is locally smooth at $\mathbf{x}$.

Lemma B.6 (Generalization of [30, 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 $0<r<R$,*

$$\begin{aligned}\Big|\frac{\mathcal{P}(\mathbb{B}(\mathbf{x},r))}{\lambda(\mathbb{B}(\mathbf{x},r))}-p(\mathbf{x})\Big|&\le\frac{d}{\sigma_p+d}Lr^{\sigma_p},\\\Big|\frac{\mathrm{d}\,\mathcal{P}(\mathbb{B}(\mathbf{x},r))}{\mathrm{d}\lambda(\mathbb{B}(\mathbf{x},r))}-p(\mathbf{x})\Big|&\le Lr^{\sigma_p}.\end{aligned}$$

The proof of the first inequality can be found in [48] and the second inequality can be proved by a similar argument.

Remark B.1. If g is bounded above over $\mathbb{B}(\mathbf{x},R)$, then g is σ_p-Hölder continuous over $\mathbb{B}(\mathbf{x},R)$ with $\sigma_p=0$. The convergence of $U_{km}(\mathbf{x})$ to a $\mathsf{G}(k,p(\mathbf{x}))$ random variable as $m\to\infty$ can be quantified in terms of a gap between the densities using this lemma and the order of smoothness σ_p of the underlying density p; however, the bounds in Lemma B.6 cannot be improved further beyond $O(r^2)$. It is consistent with the observation that the higher-order smoothness beyond 2 cannot be exploited with k-NN methods [44, 47].

Now we are ready to present the proof of Lemma B.4.

Proof of Lemma B.4. First note that the density of the k-th NN statistics $r_{km}(\mathbf{x})$ is

$$\begin{aligned}\rho_{r_{km}(\mathbf{x})}(r)&=m\Pr\{B_{m-1,\mathcal{P}(\mathbb{B}(\mathbf{x},r))}=k-1\}\frac{\mathrm{d}\,\mathcal{P}(\mathbb{B}(\mathbf{x},r))}{\mathrm{d}r}\\&=g_{km}(\mathcal{P}(\mathbb{B}(\mathbf{x},r)))\frac{\mathrm{d}\,\mathcal{P}(\mathbb{B}(\mathbf{x},r))}{\mathrm{d}r}\end{aligned}$$

from Lemma B.3 in Appendix B-B. Here we define

$$g_{km}(P):=m\Pr\{B_{m-1,P}=k-1\}$$

for $p\in[0,1]$, which is the density of the k-th order statistic from among m random samples drawn from the uniform distribution over $[0,1]$. It is easy to check that $g_{km}(P)\le m$ and $g_{km}'(P)\le 2m(m-1)\le 2m^2$ for any $P\in[0,1]$. Recall that $P_m(u|\mathbf{x}):=\mathcal{P}(\mathbb{B}(\mathbf{x},\varrho(\frac{u}{m})))$. The density of $U_{km}(\mathbf{x})$ can then be written as

$$\begin{aligned}\rho_{U_{km}(\mathbf{x})}(u)&=\rho_{r_{km}(\mathbf{x})}\big(\varrho(\frac{u}{m})\big)\frac{\mathrm{d}\,\varrho(\frac{u}{m})}{\mathrm{d}u}\\&=g_{km}(P_m(u|\mathbf{x}))\frac{\mathrm{d}P_m(u|\mathbf{x})}{\mathrm{d}u}.\end{aligned}$$

We define an intermediate density approximation

$$\rho_{km}(u):=g_{km}\Big(\frac{up(\mathbf{x})}{m}\Big)\frac{p(\mathbf{x})}{m}$$

for $u\le m/C_p$, and bound the density gap by

$$\begin{aligned}&\big|\rho_{U_{km}(\mathbf{x})}(u)-\rho_{U_{k\infty}(\mathbf{x})}(u)\big|\\&\le\big|\rho_{U_{km}(\mathbf{x})}(u)-\rho_{km}(u)\big|+\big|\rho_{km}(u)-\rho_{U_{k\infty}(\mathbf{x})}(u)\big|.\end{aligned}$$

We bound each term on the right hand side.

For the first term, consider

$$\begin{aligned}&|\rho_{U_{km}(\mathbf{x})}(u)-\rho_{km}(u)|\\&\le g_{km}(P_m(u|\mathbf{x}))\Big|\frac{\mathrm{d}P_m(u|\mathbf{x})}{\mathrm{d}u}-\frac{p(\mathbf{x})}{m}\Big|\\&\quad+\Big|g_{km}(P_m(u|\mathbf{x}))-g_{km}\Big(\frac{up(\mathbf{x})}{m}\Big)\Big|\frac{p(\mathbf{x})}{m}\\&\le g_{km}(P_m(u|\mathbf{x}))\Big|\frac{\mathrm{d}P_m(u|\mathbf{x})}{\mathrm{d}u}-\frac{p(\mathbf{x})}{m}\Big|\\&\quad+\max_{p\in(0,1)}|g_{km}'(p)|\Big|P_m(u|\mathbf{x})-\frac{up(\mathbf{x})}{m}\Big|\frac{p(\mathbf{x})}{m}\\&\le m\Big|\frac{dP_m(u|\mathbf{x})}{\mathrm{d}u}-\frac{p(\mathbf{x})}{m}\Big|+2m^2\Big|P_m(u|\mathbf{x})-\frac{up(\mathbf{x})}{m}\Big|\frac{p(\mathbf{x})}{m}\\&=\Big|\frac{\mathrm{d}\,\mathcal{P}(\mathbb{B}(\mathbf{x},\varrho(\frac{u}{m})))}{\mathrm{d}\lambda(\mathbb{B}(\mathbf{x},\varrho(\frac{u}{m}))}-p(\mathbf{x})\Big|+2up(\mathbf{x})\Big|\frac{\mathcal{P}(\mathbb{B}(\mathbf{x},\varrho(\frac{u}{m})))}{\lambda_{\mathrm{Leb}}(\mathbb{B}(\mathbf{x},\varrho(\frac{u}{m}))}-p(\mathbf{x})\Big|\\&\le\Big(1+2C_p\frac{d}{\sigma_p+d}u\Big)L\,\varrho^{\sigma_p}\big(\frac{u}{m}\big)\\&\lesssim_{\sigma_p,L,C_p,d}(1+u)\Big(\frac{u}{m}\Big)^{\frac{\sigma_p}{d}}.\end{aligned}$$

The second last inequality follows from Lemma B.6. Note that this term is independent of k.

The second term can be bounded using Lemma B.5. For m sufficiently large, we have

$$\begin{aligned}
&\big|\rho_{km}(u)-\rho_{U_{k\infty}(\mathbf{x})}(u)\big|\\
&=\frac{k}{u}\big|\Pr\{B_{m,up(\mathbf{x})/m}=k\}-\Pr\{P_{up(\mathbf{x})}=k\}\big|\\
&\le\frac{k}{u}C_0\frac{(up(\mathbf{x}))^k e^{-up(\mathbf{x})}}{k!}\frac{k^2+u^2p^2(\mathbf{x})}{m}\\
&=\frac{C_0}{\Gamma(k)}\frac{(k^2+(up(\mathbf{x}))^2)(up(\mathbf{x}))^k e^{-up(\mathbf{x})}}{mu}\\
&\lesssim_{C_0,C_p} k^{-k}\frac{(k^2+u^2)u^{k-1}e^{-up(\mathbf{x})}}{m},
\end{aligned}$$

which holds uniformly for all $u, k = o(\sqrt{m})$ as $m \to \infty$. Here we use the Stirling approximation $C_p^k/k! \sim (eC_p)^k/k^{k+\frac{1}{2}}$. □

Remark B.2. This proof closely follows the one in [30], while keeping track of the explicit dependence on the constants C_0, C_p and k.

The following lemma quantifies the convergence of the cdf of $U_{km}(\mathbf{x})$ to the cdf of $U_{k\infty}(\mathbf{x})$ when the underlying density p is smooth.

Lemma B.7 (Generalization of [30, Lemma 3])**.** *Suppose that $\nu_m = o(\sqrt{m})$ and $k = k_m = o(\sqrt{m})$ as $m \to \infty$. For $\mathbf{x} \in \operatorname{supp}(p)$, if $p(\mathbf{x}) \le C_p < \infty$ and p is σ_p-Hölder continuous ($\sigma_p \in [0, 2]$) over $\mathbb{B}(\mathbf{x}, \varrho(u/m))$ with Hölder constant L, we have*

$$\begin{aligned}
&\big|\mathrm{P}_{U_{km}(\mathbf{x})}(u)-\mathrm{P}_{U_{k\infty}(\mathbf{x})}(u)\big|\\
&\lesssim_{\sigma_p,L,C_p,d} ku\Big(\frac{u}{m}\Big)^{\frac{\sigma_p}{d}}+\frac{(k^2+u^2)u^{k-1}e^{-up(\mathbf{x})}}{m},
\end{aligned}\tag{B.3}$$

for $u \in [1/(p(\mathbf{x})), \nu_m]$ for m sufficiently large.

Proof. First, note that

$$\mathrm{P}_{U_{k\infty}(\mathbf{x})}(u)=1-\sum_{j=0}^{k-1}\Pr\{P_{up(\mathbf{x})}=j\}$$

and

$$\mathrm{P}_{U_{km}(\mathbf{x})}(u)=1-\sum_{j=0}^{k-1}\Pr\{B_{m,P_m(u|\mathbf{x})}=j\},$$

from Lemma B.3 in Appendix B-B. By triangle inequality, we have

$$\begin{aligned}
&\big|\mathrm{P}_{U_{km}(\mathbf{x})}(u)-\mathrm{P}_{U_{k\infty}(\mathbf{x})}(u)\big|\\
&\le\sum_{j=0}^{k-1}\big|\Pr\{P_{up(\mathbf{x})}=j\}-\Pr\{B_{m,P_m(u|\mathbf{x})}=j\}\big|\\
&\le\sum_{j=0}^{k-1}\Big\{\big|\Pr\{P_{up(\mathbf{x})}=j\}-\Pr\{B_{m,\frac{up(\mathbf{x})}{m}}=j\}\big|\\
&\qquad+\big|\Pr\{B_{m,\frac{up(\mathbf{x})}{m}}=j\}-\Pr\{B_{m,P_m(u|\mathbf{x})}=j\}\big|\Big\}.
\end{aligned}$$

For the first term, using Lemma B.5, we obtain

$$\begin{aligned}
&\big|\Pr\{P_{up(\mathbf{x})}=j\}-\Pr\{B_{m,\frac{up(\mathbf{x})}{m}}=j\}\big|\\
&\le C_0\frac{(up(\mathbf{x}))^j e^{-up(\mathbf{x})}}{j!}\frac{j^2+(up(\mathbf{x}))^2}{m},
\end{aligned}$$

for each $j = 0, \ldots, k-1$, which implies that

$$\begin{aligned}
&\sum_{j=0}^{k-1}\big|\Pr\{P_{up(\mathbf{x})}=j\}-\Pr\{B_{m,\frac{up(\mathbf{x})}{m}}=j\}\big|\\
&\le C_0\frac{k^2+(up(\mathbf{x}))^2}{m}e^{-up(\mathbf{x})}\sum_{j=0}^{k-1}\frac{(up(\mathbf{x}))^j}{j!}\\
&=C_0\frac{k^2+(up(\mathbf{x}))^2}{m}\frac{\Gamma(k,up(\mathbf{x}))}{\Gamma(k)}\\
&\le C_0\frac{k^2+(up(\mathbf{x}))^2}{m}(up(\mathbf{x}))^{k-1}e^{-up(\mathbf{x})+1},
\end{aligned}$$

where the last inequality follows from Lemma B.2.

For the second term, we have

$$\begin{aligned}
&\big|\Pr\{B_{m,\frac{up(\mathbf{x})}{m}}=j\}-\Pr\{B_{m,P_m(u|\mathbf{x})}=j\}\big|\\
&\le 2m\Big|P_m(u|\mathbf{x})-\frac{up(\mathbf{x})}{m}\Big|\\
&=2u\Big|\frac{\mathcal{P}(\mathbb{B}(\mathbf{x},\varrho(\frac{u}{m})))}{\lambda_{\mathrm{Leb}}(\mathbb{B}(\mathbf{x},\varrho(\frac{u}{m})))}-p(\mathbf{x})\Big|\\
&\le 2u\frac{d}{\sigma_p+d}L\,\varrho^{\sigma_p}\big(\frac{u}{m}\big),
\end{aligned}$$

for each $j = 0, \ldots, k-1$, from Lemma B.6.

Putting the bounds together and using the triangle inequality, we have that for $k, u = o(\sqrt{m})$

$$\begin{aligned}
&\big|\mathrm{P}_{U_{km}(\mathbf{x})}(u)-\mathrm{P}_{U_{k\infty}(\mathbf{x})}(u)\big|\\
&\le\frac{2kud}{\sigma_p+d}L\,\varrho^{\sigma_p}\big(\frac{u}{m}\big)+C_0\frac{k^2+(up(\mathbf{x}))^2}{m}(up(\mathbf{x}))^{k-1}e^{-up(\mathbf{x})+1}\\
&\lesssim_{\sigma_p,d,L,C_0,C_p} ku\Big(\frac{u}{m}\Big)^{\frac{\sigma_p}{d}}+\frac{(k^2+u^2)u^{k-1}e^{-up(\mathbf{x})}}{m},
\end{aligned}$$

which concludes the proof. □

C. Bounds on distribution of k-NN statistics

We now present several bounds on

$$\begin{aligned}
F_{km}(u|\mathbf{x}) &:= \Pr\{U_{km}(\mathbf{x})\le u\}\\
&=\Pr\{r_{km}(\mathbf{x})\le\varrho(\frac{u}{m})\}\\
&=\Pr\{B_{m,P_m(u|\mathbf{x})}\ge k\},
\end{aligned}$$

which is the cdf of $U_{km}(\mathbf{x})$. Here and henceforth, for $\mathbf{x} \in \mathbb{R}^d$ and $u \ge 0$, we define

$$P_m(u|\mathbf{x}) := \mathcal{P}\big(\mathbb{B}\big(\mathbf{x},\varrho(\frac{u}{m})\big)\big)=\frac{u}{m}\frac{\mathcal{P}(\mathbb{B}(\mathbf{x},\varrho(\frac{u}{m})))}{\lambda(\mathbb{B}(\mathbf{x},\varrho(\frac{u}{m})))}.$$

Note that by the definitions of $m_r p(\mathbf{x})$ and $M_r p(\mathbf{x})$, we have

$$u' m_r p(\mathbf{x}) \le m P_m(u'|\mathbf{x}) \le m \wedge (u' M_r p(\mathbf{x}))$$

for $r = \varrho(\frac{u}{m})$ and for any $0 < u' \le u$.

The following lemma presents an upper bound on the cdf $F_{km}(u|\mathbf{x})$.

Lemma B.8 (Generalization of [29, Eq. (3.19)])**.** *For any $\mathbf{x} \in \mathbb{R}^d$ and $u > 0$, we have*

$$F_{km}(u|\mathbf{x}) \le \frac{(mP_m(u|\mathbf{x}))^k}{k!}.\tag{B.4}$$

Proof. Since $F_{km}(u|\mathbf{x}) = \mathrm{P}_{T_{k,m-k+1}}(P_m(u|\mathbf{x}))$ from Lemma B.3, we have

$$\begin{aligned}F_{km}(u|\mathbf{x}) &= \int_0^{P_m(u|\mathbf{x})} \frac{t^{k-1}(1-t)^{m-k}}{\mathsf{B}(k,m-k+1)}\,\mathrm{d}t\\ &\le \frac{P_m^k(u|\mathbf{x})}{k\,\mathsf{B}(k,m-k+1)}\\ &= \binom{m}{k} P_m^k(u|\mathbf{x})\\ &\le \frac{(mP_m(u|\mathbf{x}))^k}{k!},\end{aligned}$$

which concludes the proof. □

We present two upper bounds on the complementary cdf $1 - F_{km}(u|\mathbf{x})$.

Lemma B.9 ([29, Eq. (3.23)]). *For any $\mathbf{x} \in \mathbb{R}^d$, $0 < D < 1$, and $u \ge 0$, we have*

$$1 - F_{km}(u|\mathbf{x}) \le (1-D)^{-k+1} e^{-DmP_m(u|\mathbf{x})}. \tag{B.5}$$

In particular, if $mP_m(u|\mathbf{x}) > k$, we have

$$1 - F_{km}(u|\mathbf{x}) \le \Bigl(\frac{emP_m(u|\mathbf{x})}{k}\Bigr)^k e^{-mP_m(u|\mathbf{x})}. \tag{B.6}$$

Proof. Since we can write $1-F_{km}(u|\mathbf{x}) = \Pr\{B_{m,P_m(u|\mathbf{x})} < k\}$ from Lemma B.3, the bound follows immediately from a Chernoff bound on a binomial random variable. For any $\lambda > 0$,

$$\begin{aligned}\Pr\{B_{m,P} < k\} &\le e^{\lambda k}\mathbb{E}[e^{-\lambda B_{m,P}}]\\ &= e^{\lambda k}(1-P+Pe^{-\lambda})^m\\ &\le e^{\lambda k}e^{-mP(1-e^{-\lambda})},\end{aligned}$$

and this proves (B.5) if we set $D := 1-e^{-\lambda} \in (0,1)$. If $mP > k$, we then can minimize the right hand side by plugging in $\lambda = \ln\frac{mp}{k}$, which obtains

$$\Pr\{B_{m,P} < k\} \le \Bigl(\frac{emP}{k}\Bigr)^k e^{-mP}.$$ □

Lemma B.10 ([29, Eq. (3.32)]). *For any $\mathbf{x} \in \mathbb{R}^d$, $\delta > 0$, $m \ge (1+1/\delta)(k-1)$, and $u \ge 0$, we have*

$$1 - F_{km}(u|\mathbf{x}) \le (1+\delta)(1-P_m(u|\mathbf{x})). \tag{B.7}$$

Proof. Consider

$$\begin{aligned}&1-F_{km}(u|\mathbf{x})\\ &= \sum_{j=0}^{k-1}\binom{m}{j}P_m^j(u|\mathbf{x})(1-P_m(u|\mathbf{x}))^{m-j}\\ &= (1-P_m(u|\mathbf{x}))\\ &\qquad\times\sum_{j=0}^{k-1}\frac{m}{m-j}\binom{m-1}{j}P_m^j(u|\mathbf{x})(1-P_m(u|\mathbf{x}))^{m-j-1}.\end{aligned}$$

For any fixed $\delta > 0$, if $m \ge (1+\delta^{-1})(k-1)$, then

$$\frac{m}{m-j} \le \frac{m}{m-k+1} \le 1+\delta$$

for $j = 0, \dots, k-1$. Therefore, we have

$$1-F_{km}(u|\mathbf{x}) \le (1+\delta)(1-P_m(u|\mathbf{x})).$$ □

Lemma B.11. *If $p(\mathbf{z}) \le C_p$ for $\mathbf{z} \in \overline{\mathbb{B}}(\mathbf{x},r)$, we have*

$$\rho_{U_{km}(\mathbf{x})}(u) \le \frac{C_p^k u^{k-1}}{\Gamma(k)}.$$

We first prove the following lemma. Let us denote the sphere centered at $\mathbf{x} \in \mathbb{R}^d$ of radius $r > 0$ by $\mathbb{S}(\mathbf{x},r) := \{\mathbf{y}\colon \rho(\mathbf{x},\mathbf{y}) = r\}$. Note that the the Hausdorff measure $H^{d-1}(\mathbb{S}(\mathbf{x},r))$ of the sphere is $dv_d r^{d-1}$.

Lemma B.12. *If $p(\mathbf{z}) \le C_p$ for $\mathbf{z} \in \mathbb{S}(\mathbf{x},r)$, we have*

$$\frac{\mathrm{d}\,\mathcal{P}(\mathbb{B}(\mathbf{x},r))}{\mathrm{d}\,\lambda_{\mathrm{Leb}}(\mathbb{B}(\mathbf{x},r))} \le C_p.$$

Proof of Lemma B.12. It is easy to see that $p(\mathbf{x}) \le M_r p(\mathbf{x})$ for any $r > 0$ by contradiction. From the coarea formula [78], we have

$$\begin{aligned}\frac{\mathrm{d}\,\mathcal{P}(\mathbb{B}(\mathbf{x},r))}{\mathrm{d}r} &= \frac{\mathrm{d}}{\mathrm{d}r}\int_{\mathbb{B}(\mathbf{x},r)} p(\mathbf{y})\,\mathrm{d}\mathbf{y}\\ &= \int_{\mathbb{S}(\mathbf{x},r)} p(\mathbf{y})H^{d-1}(\mathrm{d}\mathbf{y})\\ &\le C_p(dv_d r^{d-1})\end{aligned}$$

since $p(\mathbf{x}) \le C_p$ for $\mathbf{x} \in \mathbb{S}(\mathbf{x},r)$. Therefore, we have

$$\frac{\mathrm{d}\,\mathcal{P}(\mathbb{B}(\mathbf{x},r))}{\mathrm{d}\,\lambda_{\mathrm{Leb}}(\mathbb{B}(\mathbf{x},r))} = \frac{\frac{\mathrm{d}}{\mathrm{d}r}\mathcal{P}(\mathbb{B}(\mathbf{x},r))}{\frac{\mathrm{d}}{\mathrm{d}r}\lambda_{\mathrm{Leb}}(\mathbb{B}(\mathbf{x},r))} \le C_p.$$ □

Proof of Lemma B.11. Now, from Lemma B.3 and Lemma B.12, if $p(\mathbf{y}) \le C_p$ for $\mathbf{y} \in \mathbb{B}(\mathbf{x},r)$, then

$$\begin{aligned}\rho_{r_{km}(\mathbf{x})}(r) &= \rho_{X_{k,m-k+1}}(\mathcal{P}(\mathbb{B}(\mathbf{x},r)))\frac{\mathrm{d}\,\mathcal{P}(\mathbb{B}(\mathbf{x},r))}{\mathrm{d}r}\\ &\le \frac{m^k}{\Gamma(k)}\mathcal{P}^{k-1}(\mathbb{B}(\mathbf{x},r))\frac{\mathrm{d}\,\mathcal{P}(\mathbb{B}(\mathbf{x},r))}{\mathrm{d}r}\\ &\le \frac{(C_p m)^k}{\Gamma(k)}\frac{d}{r}\lambda_{\mathrm{Leb}}{}^k(\mathbb{B}(\mathbf{x},r)).\end{aligned}$$

We then bound the density of $U_{km}(\mathbf{x})$ as

$$\begin{aligned}\rho_{U_{km}(\mathbf{x})}(u) &= \rho_{r_{km}(\mathbf{x})}\bigl(\varrho(\frac{u}{m})\bigr)\frac{\mathrm{d}\,\varrho(\frac{u}{m})}{\mathrm{d}u}\\ &\le \frac{(C_p m)^k}{\Gamma(k)}\frac{d}{\varrho(\frac{u}{m})}\bigl(\frac{u}{m}\bigr)^k\frac{\varrho(\frac{u}{m})}{du} = \frac{C_p^k}{\Gamma(k)}u^{k-1},\end{aligned}$$

which concludes the proof. □

D. Bounds on expected values of k-NN statistics

Let $\tilde{f}_{km}(u|\mathbf{x}) := \rho_{\overline{U}_{km}(\mathbf{x})}(v)$ denote the density of the normalized volume $\overline{U}_{km}(\mathbf{x}) = \lambda_{\mathrm{Leb}}(\mathbb{B}(\mathbf{x},r_k(\mathbf{x}|\tilde{\mathbf{X}}_{1:m})))$, where $\tilde{\mathbf{X}}_{1:m}$ is drawn i.i.d. from density $\tilde{p}$. Later, the density $\tilde{p}$ may be identified as the density p for $\mathbf{X}_{1:m}$ or the density q for $\mathbf{Y}_{1:n}$. Pick any numbers $0 \le \tau_m \le 1 \le \nu_m \le \kappa_m < \infty$. Suppose that we are given a nondecreasing function $\xi \in \Xi$. For $(a,b) \in \mathbb{R}^2$ and $k \in \mathbb{N}$, we define, for each $\mathbf{x} \in \mathbb{R}^d$

$$A_{km}(\mathbf{x};\tilde{p};\xi) := \int_0^{\tau_m}\xi(u^a)\tilde{f}_{km}(u|\mathbf{x})\,\mathrm{d}u, \tag{B.8}$$

$$B_{km}^{(1)}(\mathbf{x};\tilde{p};\xi) := \int_1^{\nu_m}\xi(u^b)\tilde{f}_{km}(u|\mathbf{x})\,\mathrm{d}u, \tag{B.9}$$

$$B_{km}^{(2)}(\mathbf{x};\tilde{p};\xi) := \int_{\nu_m}^{\kappa_m}\xi(u^b)\tilde{f}_{km}(u|\mathbf{x})\,\mathrm{d}u, \tag{B.10}$$

and

$$B^{(3)}_{km}(\mathbf{x};\tilde{p};\xi) := \int_{\kappa_m}^{\infty} \xi(u^b)\tilde{f}_{km}(u|\mathbf{x})\,\mathrm{d}u. \tag{B.11}$$

Lemma B.13. *For $r = \varrho(\frac{\tau_m}{m})$, we have*

$$\begin{aligned} &A_{km}(\mathbf{x};\tilde{p};\xi) \\ &\leq \frac{(M_r\tilde{p}(\mathbf{x}))^k}{k!}\Big(\tau_m^k\xi(\tau_m^a) - \mathbb{1}_{(-\infty,0)}(a)\int_0^{\tau_m} u^k\,\mathrm{d}\xi(u^a)\Big). \end{aligned}$$

In particular, if $\tau_m = 1$ and $-\int_0^1 u^k\,\mathrm{d}\xi(u^a) < \infty$, we have for $r = \varrho(\frac{1}{m})$,

$$A_{km}(\mathbf{x};\tilde{p};\xi) \lesssim \frac{(M_r\tilde{p}(\mathbf{x}))^k}{k!}.$$

Proof. Integrating by parts and applying Lemma B.8, we have

$$\begin{aligned} &A_{km}(\mathbf{x};\tilde{p};\xi) \\ &= \int_0^{\tau_m} \xi(u^a)\,\mathrm{d}\overline{F}_{km}(u|\mathbf{x}) \\ &\leq \xi(\tau_m^a)\overline{F}_{km}(\tau_m|\mathbf{x}) - \int_0^{\tau_m} \overline{F}_{km}(u|\mathbf{x})\,\mathrm{d}\xi(u^a) \\ &\leq \frac{(M_{\varrho(\frac{\tau_m}{m})}\tilde{p}(\mathbf{x}))^k}{k!}\tau_m^k\xi(\tau_m^a) - \int_0^{\tau_m} \overline{F}_{km}(u|\mathbf{x})\,\mathrm{d}\xi(u^a). \end{aligned}$$

If $a < 0$, we again apply Lemma B.8 again to the remaining integral and obtain

$$\begin{aligned} &A_{km}(\mathbf{x};\tilde{p};\xi) \\ &\leq \frac{(M_{\varrho(\frac{\tau_m}{m})}\tilde{p}(\mathbf{x}))^k}{k!}\Big(\tau_m^k\xi(\tau_m^a) - \int_0^{\tau_m} u^k\,\mathrm{d}\xi(u^a)\Big). \qquad \square \end{aligned}$$

Lemma B.14. *If $b \leq 0$, we have*

$$B^{(1)}_{km}(\mathbf{x};\tilde{p};\xi) \lesssim 1.$$

If $b > 0$ and $\int_0^\infty e^{-t}\xi(t^b)\,\mathrm{d}t < \infty$, then for any $0 < D < 1$ and $r = \varrho(\frac{\nu_m}{m})$, we have

$$B^{(1)}_{km}(\mathbf{x};\tilde{p};\xi) \lesssim_{k,D} \xi(\nu_m^b)e^{-D\nu_m(m_r p(\mathbf{x}))} + \xi((Dm_r\tilde{p}(\mathbf{x}))^{-b}).$$

Proof. By definition, if $b \leq 0$, we have

$$\begin{aligned} B^{(1)}_{km}(\mathbf{x};\tilde{p};\xi) &= \int_1^{\nu_m} \xi(u^b)\tilde{f}_{km}(u|\mathbf{x})\,\mathrm{d}u \\ &\leq \xi(1)\int_1^{\nu_m} \tilde{f}_{km}(u|\mathbf{x})\,\mathrm{d}u \\ &\leq \xi(1). \end{aligned}$$

We now assume $b > 0$. Integrating by parts, we have

$$\begin{aligned} &B^{(1)}_{km}(\mathbf{x};\tilde{p};\xi) \\ &= -\int_1^{\nu_m} \xi(u^b)\,\mathrm{d}(1 - \overline{F}_{km}(u|\mathbf{x})) \\ &\leq \xi(1)(1 - \overline{F}_{km}(u|\mathbf{x})) + \int_1^{\nu_m} (1 - \overline{F}_{km}(u|\mathbf{x}))\,\mathrm{d}\xi(u^b). \end{aligned}$$

Applying Lemma B.9 yields, for any $0 < D < 1$, that

$$\begin{aligned} &B^{(1)}_{km}(\mathbf{x};\tilde{p};\xi) \\ &\leq \xi(1) + (1-D)^{-k+1}\int_1^{\nu_m} e^{-Dm\tilde{\mathcal{P}}_m(u|\mathbf{x})}\,\mathrm{d}\xi(u^b) \\ &\leq \xi(1) + (1-D)^{-k+1}\int_1^{\nu_m} e^{-Du(m_r\tilde{p}(\mathbf{x}))}\,\mathrm{d}\xi(u^b) \qquad \text{(B.12)} \end{aligned}$$

for $r = \varrho(\frac{\nu_m}{m})$. Integrating by parts again, we thus obtain

$$\begin{aligned} &\int_1^{\nu_m} e^{-Du(m_r\tilde{p}(\mathbf{x}))}\,\mathrm{d}\xi(u^b) \\ &\leq \xi(\nu_m^b)e^{-D\nu_m(m_r\tilde{p}(\mathbf{x}))} \\ &\quad + D(m_r\tilde{p}(\mathbf{x}))\int_1^{\nu_m} e^{-Du(m_r\tilde{p}(\mathbf{x}))}\xi(u^b)\,\mathrm{d}u \\ &\leq \xi(\nu_m^b)e^{-D\nu_m(m_r\tilde{p}(\mathbf{x}))} \\ &\quad + \int_{D(m_r\tilde{p}(\mathbf{x}))}^{D\nu_m(m_r\tilde{p}(\mathbf{x}))} e^{-t}\xi(t^b(Dm_r\tilde{p}(\mathbf{x}))^{-b})\,\mathrm{d}t. \qquad \text{(B.13)} \end{aligned}$$

Here, using the property that $\xi(xy) \leq \xi(x)\xi(y)$ for any $x, y > t_0$ for some $t_0 \geq 0$, it is easy to show that

$$\begin{aligned} &\int_{D(m_r\tilde{p}(\mathbf{x}))}^{D\nu_m(m_r\tilde{p}(\mathbf{x}))} e^{-t}\xi(t^b(Dm_r\tilde{p}(\mathbf{x}))^{-b})\,\mathrm{d}t \\ &\leq \Big(t_0\xi(t_0^b) + \int_0^\infty e^{-t}\xi(t^b)\,\mathrm{d}t\Big)\xi((Dm_r\tilde{p}(\mathbf{x}))^{-b}) \\ &\qquad + \xi(t_0)\int_0^\infty e^{-t}\xi(t^b)\,\mathrm{d}\mathbf{x} \\ &\lesssim 1 + \xi((Dm_r\tilde{p}(\mathbf{x}))^{-b}). \qquad \text{(B.14)} \end{aligned}$$

Putting (B.12), (B.13), and (B.14) together, we obtain the desired bound. □

Lemma B.15. *For any $0 < D < 1$ and $r = \varrho(\frac{\nu_m}{m})$, we have*

$$B^{(2)}_{km}(\mathbf{x};\tilde{p};\xi) \lesssim_{k,D} \xi(\nu_m^b \vee \kappa_m^b)e^{-D\nu_m(m_r\tilde{p}(\mathbf{x}))}.$$

Proof. Integrating by parts, we have

$$\begin{aligned} &B^{(2)}_{km}(\mathbf{x};\tilde{p};\xi) \\ &= -\int_{\nu_m}^{\kappa_m} \xi(u^b)\,\mathrm{d}(1 - \overline{F}_{km}(u|\mathbf{x})) \\ &\leq \xi(\nu_m^b)(1 - \overline{F}_{km}(\nu_m|\mathbf{x})) + \int_{\nu_m}^{\kappa_m} (1 - F_{km}(u|\mathbf{x}))\,\mathrm{d}\xi(u^b) \\ &\leq 2\xi(\nu_m^b \vee \kappa_m^b)(1 - \overline{F}_{km}(\nu_m|\mathbf{x})). \qquad \text{(B.15)} \end{aligned}$$

Applying Lemma B.9, we have that for any $0 < D < 1$ and $r = \varrho(\frac{\nu_m}{m})$

$$\begin{aligned} &B^{(2)}_{km}(\mathbf{x};\tilde{p};\xi) \\ &\leq 2(1-D)^{-k+1}\xi(\nu_m^b \vee \kappa_m^b)e^{-Dm\tilde{\mathcal{P}}_m(\nu_m|\mathbf{x})} \\ &\leq 2(1-D)^{-k+1}\xi((\nu_m^b \vee \kappa_m^b)e^{-D\nu_m(m_r p(\mathbf{x}))}. \qquad \square \end{aligned}$$

Lemma B.16. *For any $\delta > 0$ and m sufficiently large, we have*

$$\begin{aligned} &B^{(3)}_{km}(\mathbf{x};\tilde{p};\xi) \\ &\lesssim_\delta \xi(m^b)\int p(\mathbf{y})\xi(\upsilon^b(\rho(\mathbf{x},\mathbf{y})))\mathbb{1}_{\{\rho(\mathbf{x},\mathbf{y})>\varrho(\frac{\kappa_m}{m})\}}\,\mathrm{d}\mathbf{y}. \end{aligned}$$

Proof. We recall the following bound (B.7) on the complementary cdf $1 - \overline{F}_{km}(u|\mathbf{x})$ from Lemma B.8: for any $\delta > 0$ and $m \geq (1 + 1/\delta)(k-1)$, we have

$$\begin{aligned} 1 - \overline{F}_{km}(u|\mathbf{x}) &\leq (1+\delta)(1 - \tilde{\mathcal{P}}_m(u|\mathbf{x})) \\ &= (1+\delta)\int \tilde{p}(\mathbf{y})\mathbb{1}_{\{\rho(\mathbf{x},\mathbf{y})>\varrho(\frac{u}{m})\}}\,\mathrm{d}\mathbf{y}. \end{aligned}$$

Integrating by parts, we first obtain

$$\begin{aligned}
&B^{(3)}_{km}(\mathbf{x};\tilde{p};\xi)\\
&= -\int_{\kappa_m}^{\infty} \xi(u^b)\,\mathrm{d}(1-\overline{F}_{km}(u|\mathbf{x}))\\
&\le \xi(\kappa_m^b)(1-\overline{F}_{km}(\kappa_m|\mathbf{x}))\\
&\qquad + \int_{\kappa_m}^{\infty} (1-\overline{F}_m(u|\mathbf{x}))\,\mathrm{d}\xi(u^b)\\
&\le \xi(\kappa_m^b)(1-\overline{F}_{km}(\kappa_m|\mathbf{x}))\\
&\qquad + (1+\delta)\int_{\kappa_m}^{\infty} (1-\tilde{\mathcal{P}}_m(u|\mathbf{x}))\,\mathrm{d}\xi(u^b). \qquad \text{(B.16)}
\end{aligned}$$

Integrating the second term by parts leads to

$$\begin{aligned}
&\int_{\kappa_m}^{\infty} (1-\tilde{\mathcal{P}}_m(u|\mathbf{x}))\,\mathrm{d}\xi(u^b) \qquad \text{(B.17)}\\
&\le \lim_{u\to\infty} \xi(u^b)(1-\tilde{\mathcal{P}}_m(u|\mathbf{x})) + \int_{\kappa_m}^{\infty} \xi(u^b)\,\mathrm{d}\tilde{\mathcal{P}}_m(u|\mathbf{x}).
\end{aligned}$$

For the first term in (B.17), since for m sufficiently large with $m^b > t_0$ and $(\kappa_m/m)^b > t_0$, we have $\xi(u^b) \le \xi(m^b)\xi((u/m)^b)$ for $u \ge \kappa_m$, it follows that

$$\begin{aligned}
&\xi(u^b)(1-\tilde{\mathcal{P}}_m(u|\mathbf{x}))\\
&= \xi(u^b)\int \tilde{p}(\mathbf{y})\mathbb{1}_{\{\rho(\mathbf{x},\mathbf{y})>\varrho(\frac{u}{m})\}}\,\mathrm{d}\mathbf{y}\\
&\le \xi(m^b)\int \tilde{p}(\mathbf{y})\xi\Big(\Big(\frac{u}{m}\Big)^b\Big)\mathbb{1}_{\{\rho(\mathbf{x},\mathbf{y})>\varrho(\frac{u}{m})\}}\,\mathrm{d}\mathbf{y}\\
&\le \xi(m^b)\int \tilde{p}(\mathbf{y})\xi(\upsilon^b(\rho(\mathbf{x},\mathbf{y})))\mathbb{1}_{\{\rho(\mathbf{x},\mathbf{y})>\varrho(\frac{u}{m})\}}\,\mathrm{d}\mathbf{y}\\
&\le \xi(m^b)\int \tilde{p}(\mathbf{y})\xi(\upsilon^b(\rho(\mathbf{x},\mathbf{y})))\mathbb{1}_{\{\rho(\mathbf{x},\mathbf{y})>\varrho(\frac{\kappa_m}{m})\}}\,\mathrm{d}\mathbf{y}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\lim_{u\to\infty} \xi(u^b)(1-\tilde{\mathcal{P}}_m(u|\mathbf{x})) \qquad \text{(B.18)}\\
&\le \xi(m^b)\int \tilde{p}(\mathbf{y})\xi(\upsilon^b(\rho(\mathbf{x},\mathbf{y})))\mathbb{1}_{\{\rho(\mathbf{x},\mathbf{y})>\varrho(\frac{\kappa_m}{m})\}}\,\mathrm{d}\mathbf{y}.
\end{aligned}$$

The second term in (B.17) can be bounded similarly as

$$\begin{aligned}
&\int_{\kappa_m}^{\infty} \xi(u^b)\,\mathrm{d}\tilde{\mathcal{P}}_m(u|\mathbf{x}) \qquad \text{(B.19)}\\
&= \int \tilde{p}(\mathbf{y})\xi((m\upsilon(\rho(\mathbf{x},\mathbf{y})))^b)\mathbb{1}_{\{\rho(\mathbf{x},\mathbf{y})>\varrho(\frac{\kappa_m}{m})\}}\,\mathrm{d}\mathbf{y}\\
&\le \xi(m^b)\int \tilde{p}(\mathbf{y})\xi(\upsilon^b(\rho(\mathbf{x},\mathbf{y})))\mathbb{1}_{\{\rho(\mathbf{x},\mathbf{y})>\varrho(\frac{\kappa_m}{m})\}}\,\mathrm{d}\mathbf{y}.
\end{aligned}$$

Plugging (B.17), (B.18), and (B.19) into (B.16) establishes the desired bound. □

The following is the key lemma in establishing vanishing bias and vanishing variance for single- and double-density cases.

Lemma B.17. *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 densities* p *and* $\tilde{p}$ *satisfy* $\tilde{\mathcal{P}} \ll \mathcal{P}$, *(*$\mathbf{U}_{p\tilde{p}}$*;* k, a*), and (*$\mathbf{L}_{p\tilde{p}}$*;* ξ, b*), we have*

$$\limsup_{m\to\infty} \int p(\mathbf{x})\int_0^\infty \xi(\psi_{a,b}(u))\,\mathrm{d}\overline{F}_{km}(u|\mathbf{x})\,\mathrm{d}\mathbf{x} < \infty.$$

Proof. Let $\tau_m = 1$ and $\kappa_m = e^{o(m)}$. Then, there exists ν_m such that $\nu_m \to \infty$, $\nu_m/m \to 0$, and for any $c > 0$, $e^{-c\nu_m}\xi(\kappa_m^b) \to 0$, as $m \to \infty$. Consider

$$\begin{aligned}
&\int_0^\infty \xi(\psi_{a,b}(u))\,\mathrm{d}F_{km}(u|\mathbf{x})\\
&= \int_0^1 \xi(u^a)\,\mathrm{d}\overline{F}_{km}(u|\mathbf{x}) + \int_1^\infty \xi(u^b)\,\mathrm{d}\overline{F}_{km}(u|\mathbf{x})\\
&= A_{km}(\mathbf{x};\tilde{p};\xi) + B_{km}(\mathbf{x};\tilde{p};\xi),
\end{aligned}$$

where

$$B_{km}(\mathbf{x};\tilde{p};\xi) := B^{(1)}_{km}(\mathbf{x};\tilde{p};\xi) + B^{(2)}_{km}(\mathbf{x};\tilde{p};\xi) + B^{(3)}_{km}(\mathbf{x};\tilde{p};\xi).$$

Recall the definitions of $A_{km}(\mathbf{x};\tilde{p};\xi)$, $B^{(1)}_{km}(\mathbf{x};\tilde{p};\xi)$, $B^{(2)}_{km}(\mathbf{x};\tilde{p};\xi)$, and $B^{(3)}_{km}(\mathbf{x};\tilde{p};\xi)$ in (B.8), (B.9), (B.10), and (B.11), respectively. Letting

$$A_{km}(p,\tilde{p};\xi) := \int p(\mathbf{x})A_{km}(\mathbf{x};\tilde{p};\xi)\,\mathrm{d}\mathbf{x}$$

and

$$B_{km}(p,\tilde{p};\xi) := \int p(\mathbf{x})B_{km}(\mathbf{x};\tilde{p};\xi)\,\mathrm{d}\mathbf{x},$$

we show separately that $\limsup_{m\to\infty} A_{km}(p,\tilde{p};\xi) < \infty$ and $\limsup_{m\to\infty} B_{km}(p,\tilde{p};\xi) < \infty$.

Step 1. Bounding $A_{km}(p,\tilde{p};\xi)$**.** If $a \ge 0$, we trivially have $A_{km}(p,\tilde{p};\xi) \le \xi(1)$. If $a < 0$, by Lemma B.13, we have

$$\begin{aligned}
A_{km}(p,\tilde{p};\xi) &\le \frac{W(p,\tilde{p};k,\varrho(\frac{1}{m}))}{k!}\Big(\xi(1) - \int_0^1 u^k\,\mathrm{d}\xi(u^a)\Big)\\
&\lesssim_k W\Big(p,\tilde{p};k,\varrho\Big(\frac{1}{m}\Big)\Big).
\end{aligned}$$

Hence, since there exists $r' > 0$ such that $W(p,\tilde{p};k,r') < \infty$ by the the condition ($\mathbf{U}_{p\tilde{p}}$; k, a) and $W(p,\tilde{p};k,r)$ is nonincreasing as $r \to 0$, we conclude that $A_{km}(p,\tilde{p};\xi) < \infty$ for m sufficiently large such that $\varrho(1/m) < r'$.

Step 2. Bounding $B_{km}(p,\tilde{p};\xi)$**.** If $b \le 0$, then we trivially have $B_{km}(p,\tilde{p};\xi) \le \xi(1)$. If $b > 0$, by applying Lemmas B.14, B.15, and B.16, we have that for any $0 < D < 1$ and m sufficiently large

$$\begin{aligned}
B_{km}(p,\tilde{p};\xi) \lesssim \xi(\kappa_m^b)&\int e^{-D\nu_m(m_{r_1}\tilde{p}(\mathbf{x}))}p(\mathbf{x})\,\mathrm{d}\mathbf{x}\\
&+ w(p,\tilde{p};\xi,b,r_1)\\
&+ \xi(m^b)R(p,\tilde{p};\xi,b,r_2),
\end{aligned}$$

where $r_1 = \varrho(\nu_m/m)$ and $r_2 = \varrho(\kappa_m/m)$.

- For the first term, since $\mathcal{P} \ll \tilde{\mathcal{P}}$ implies that $\mathcal{P}(\{\mathbf{x}\colon m_{r_1}\tilde{p}(\mathbf{x}) > 0\}) = 1$ (Lemma B.1), we have $\xi(\kappa_m^b)e^{-\nu_m(m_{r_1}\tilde{p}(\mathbf{x}))} \to 0$ as $m \to \infty$ for $\mathcal{P}$-a.e. $\mathbf{x}$ by definition of ν_m and κ_m. Therefore, by the dominated convergence theorem,

$$\lim_{m\to\infty} \int \xi(\kappa_m^b)e^{-\nu_m(m_{r_1}\tilde{p}(\mathbf{x}))}p(\mathbf{x})\,\mathrm{d}\mathbf{x} = 0.$$

- Since there exists $r'' > 0$ such that $w(p,\tilde{p};\xi,b,r'') < \infty$ by the condition ($\mathbf{L}_{p\tilde{p}}$; ξ, b) and $w(p,\tilde{p};\xi,b,r)$ is nonincreasing as $r \to 0$, the second term is bounded for m sufficiently large such that $\varrho(\frac{\kappa_m}{m}) < r''$.

This article has been accepted for publication in a future issue of this journal, but has not been fully edited. Content may change prior to final publication. Citation information: DOI 10.1109/TIT.2022.3151231, IEEE Transactions on Information Theory

- The limit superior of the last term $\xi(m^b)R(p,\tilde{p};\xi,b,r_2)$ as $m\to\infty$ is bounded by the condition **($\mathbf{L}_{p\tilde{p}}$; ξ, b)**.

Overall, we conclude that

$$\limsup_{m\to\infty} B_{km}(p,\tilde{p};\xi) < \infty. \qquad \square$$

Following the proof of Lemma B.17 with the stronger assumptions establishes the following bound.

Lemma B.18. *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* $\tilde{p}$ *satisfies the conditions* ***(U***$_p$***)***, ***(L1***$_p$***)***, ***(L2***$_p$***)***, *and* ***(L3***$_p$***)***, *we have*

$$\int_0^\infty \xi(\psi_{a,b}(u))\,\mathrm{d}\overline{F}_{km}(u|\mathbf{x}) \lesssim 1$$

for $\mathcal{P}$*-a.e.* $\mathbf{x}$.

Continuing from (B.15) and applying (B.6) in Lemma B.9 yield the following bound, which is required for establishing performance guarantees with adaptive choices of k and l.

Lemma B.19. *For* $r=\varrho(\nu_m/m)$*, we have*

$$B_{km}^{(2)}(\mathbf{x};\tilde{p};\xi) \le 2\xi(\nu_m^b \vee \kappa_m^b)\Bigl(\frac{e\nu_m M_r\tilde{p}(\mathbf{x})}{k}\Bigr)^k e^{-\nu_m m_r \tilde{p}(\mathbf{x})}.$$

Lemma B.20. *If* $k+a>0$*, for* $\mathbf{x}\in\mathrm{supp}(p)$*, we have*

$$\begin{aligned}\int_0^\infty &\psi_{a,b}(u)\rho_{U_{k\infty}(\mathbf{x})}(u)\,\mathrm{d}u\\ &\le \frac{p^k(\mathbf{x})}{(k+a)\Gamma(k)} + \frac{\Gamma((k+b)\vee 1)}{\Gamma(k)}(p(\mathbf{x}))^{(k-1)\wedge(-b)}.\end{aligned}$$

In particular, if $c_p \le p(\mathbf{x}) \le C_p$*, then*

$$\int_0^\infty \psi_{a,b}(u)\rho_{U_{k\infty}(\mathbf{x})}(u)\,\mathrm{d}u \lesssim 1.$$

Proof. First, consider

$$\begin{aligned}\int_0^1 u^a\rho_{U_{k\infty}(\mathbf{x})}(u)\,\mathrm{d}u &= \frac{p^k(\mathbf{x})}{\Gamma(k)}\int_0^1 u^{k+a-1}e^{-up(\mathbf{x})}\,\mathrm{d}u\\ &= \frac{(p(\mathbf{x}))^{-a}}{\Gamma(k)}\int_0^{p(\mathbf{x})} t^{k+a-1}e^{-t}\,\mathrm{d}t\\ &\le \frac{p^k(\mathbf{x})}{(k+a)\Gamma(k)},\end{aligned}$$

where the last inequality follows from the bound on the lower incomplete gamma in Lemma B.2. Similarly, we consider

$$\int_0^1 u^b\rho_{U_{k\infty}(\mathbf{x})}(u)\,\mathrm{d}u = \frac{(p(\mathbf{x}))^{-b}}{\Gamma(k)}\int_0^{p(\mathbf{x})} t^{k+b-1}e^{-t}\,\mathrm{d}t.$$

On the one hand, if $k+b>1$, by bounding the integral by $\Gamma(k+b)$, we have

$$\int_0^1 u^b\rho_{U_{k\infty}(\mathbf{x})}(u)\,\mathrm{d}u \le \frac{\Gamma(k+b)}{\Gamma(k)}(p(\mathbf{x}))^{-b}.$$

On the other hand, if $k+b\le 1$, we have

$$\begin{aligned}\int_0^1 u^b\rho_{U_{k\infty}(\mathbf{x})}(u)\,\mathrm{d}u &\le \frac{(p(\mathbf{x}))^{k-1}}{\Gamma(k)}\int_{p(\mathbf{x})}^\infty e^{-t}\,\mathrm{d}t\\ &\le \frac{(p(\mathbf{x}))^{k-1}}{\Gamma(k)}.\end{aligned}$$

Therefore, we obtain

$$\int_0^1 u^b\rho_{U_{k\infty}(\mathbf{x})}(u)\,\mathrm{d}u \le \frac{\Gamma((k+b)\vee 1)}{\Gamma(k)}(p(\mathbf{x}))^{(k-1)\wedge(-b)},$$

which completes the proof. $\square$

E. Generic bias bounds

Lemma B.21 (Generic inner bias bound)**.** *Suppose that the density* p *satisfies the conditions* ***(U***$_p$***)***, ***(S***$_p$***)***, *and* ***(B***$_p$***)***, *and let* $k=o(\sqrt{m})$ *as* $m\to\infty$.

1) We have

$$\begin{aligned}I_{in,1} = O\Bigl(&\frac{\tau_m^{(a+\frac{\sigma_p}{d}+1)\wedge 0}}{m^{\frac{\sigma_p}{d}}} + \frac{k^{-k}}{m}\\ &+ \Bigl(\frac{1}{m}\Bigr)^{\frac{1}{d}}\tau_m^{(a+1)\wedge 0}\Bigr).\end{aligned} \tag{B.20}$$

2) If $\nu_m = o(\sqrt{m})$ *as* $m\to\infty$*, we have*

$$\begin{aligned}I_{in,2} = O\Bigl(&\frac{\nu_m^{(b+\frac{\sigma_p}{d}+2)\vee 0}}{m^{\frac{\sigma_p}{d}}} + \frac{k^{-k}\nu_m^{(b+k+2)\vee 0}}{m}\\ &+ \Bigl(\frac{\nu_m}{m}\Bigr)^{\frac{1}{d}}\nu_m^{(b+2)\vee 0}\Bigr).\end{aligned} \tag{B.21}$$

Proof. We establish each bound separately.

Bounding the lower inner bias $I_{\mathrm{in},1}$. For each $r>0$, define a set

$$\begin{aligned}S_p(r) := \{\mathbf{x}\in\mathrm{supp}(p)\colon &\, p \text{ is } \sigma_p\text{-Hölder continuous}\\ &\text{over } \mathbb{B}(\mathbf{x},r)\}.\end{aligned}$$

By the smoothness assumption **(S$_p$)**, we can bound the inner bias incurred at the "smooth region", i.e.,

$$I_{\mathrm{in},1,\mathrm{smooth}} = \int_{S_p(\varrho(\frac{1}{m}))} I_{\mathrm{in},1}(\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x},$$

by applying Lemma B.4. Since $p(\mathbf{x})\le C_p<\infty$ for $\mathcal{P}$-a.e. $\mathbf{x}$, this lemma holds for m sufficiently large uniformly over $\mathcal{P}$-a.e. $\mathbf{x}$. Applying Lemma B.4 for $\mathbf{x}\in S_p(\varrho(\frac{1}{m}))$, we have

$$\begin{aligned}&I_{\mathrm{in},1}(\mathbf{x})\\ &\lesssim_{\sigma_p,L,C_p,C_0,d} \int_{\tau_m}^1 u^a\Bigl\{(1+u)\Bigl(\frac{u}{m}\Bigr)^{\frac{\sigma_p}{d}}\\ &\qquad\qquad + k^{-k}\frac{(k^2+u^2)u^{k-1}e^{-up(\mathbf{x})}}{m}\Bigr\}\,\mathrm{d}u.\end{aligned} \tag{B.22}$$

It is easy to see that the first term is bounded by $O\bigl(\tau_m^{(a+\frac{\sigma_p}{d}+1)\wedge 0}m^{-\frac{\sigma_p}{d}}\bigr)$.[4] To bound the second term, we use the upper bound on the lower incomplete gamma function (Lemma B.2). Since we always assume that $k+a>0$, we have

$$\begin{aligned}\int_{\tau_m}^1 &\frac{k^{-k}}{m}(k^2+u^2)u^{k+a-1}e^{-up(\mathbf{x})}\,\mathrm{d}u\\ &\le \frac{k^{-k}}{m}\bigl\{k^2p(\mathbf{x})^{-(k+a)}\gamma(k+a,p(\mathbf{x}))\\ &\quad + p(\mathbf{x})^{-(k+a+2)}\gamma(k+a+2,p(\mathbf{x}))\bigr\} = O\Bigl(\frac{k^{-k}}{m}\Bigr).\end{aligned}$$

[4] Here $a+\frac{\sigma_p}{d}+1\ne 0$ is implicitly assumed. If $a+\frac{\sigma_p}{d}+1=0$, then the first term behaves as $O((\ln\tau_m)m^{-\frac{\sigma_p}{d}})$

Hence, we conclude that

$$\begin{aligned} &I_{\text{in},1,\text{smooth}} \qquad\qquad \text{(B.23)}\\ &\quad = O(\tau_m^{(a+\frac{\sigma_p}{d}+1)\wedge 0} m^{-\frac{\sigma_p}{d}} + k^{-k}m^{-1}). \end{aligned}$$

To control the inner bias incurred at $\mathbf{x} \in \text{supp}(p)\backslash S_p(\varrho(m^{-1}))$, i.e.,

$$I_{\text{in},1,\text{nonsmooth}} = \int_{\text{supp}(p)\backslash S_p(\varrho(\frac{1}{m}))} I_{\text{in},1}(\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x},$$

we first note that the bound (B.22) on $I_{\text{in},1}(\mathbf{x})$ holds with $\sigma_p = 0$ from the upper boundedness assumption **($\mathbf{U}_p$)**, which implies that

$$\begin{aligned} &I_{\text{in},1,\text{nonsmooth}} \qquad\qquad \text{(B.24)}\\ &= O(\lambda_{\text{Leb}}(\text{supp}(p)\backslash S_p(\varrho(m^{-1})))(\tau_m^{(a+1)\wedge 0} + k^{-k}m^{-1})). \end{aligned}$$

We now only need to bound the Lebesgue measure of the set where $\text{supp}(p)\backslash S_p(\varrho(m^{-1}))$. Observe that for any $r > 0$

$$\text{supp}(p)\backslash S_p(r) \subseteq \{\mathbf{x} \in \mathbb{R}^d : \mathbb{B}(\mathbf{x}, r) \cap \partial(\text{supp}(p)) \neq \emptyset\},$$

where ∂A denotes the boundary of a set A. Using the following lemma with the condition **($\mathbf{B}_p$)** on the finiteness of the Hausdorff measure of the boundary of the support, we can bound the Lebesgue measure of $\mathbb{R}^d\backslash S_p(\varrho(m^{-1}))$ by $O(\varrho(1/m)) = O(m^{-\frac{1}{d}})$.

Lemma B.22 ([30, Section A])**.** *For $S \subset \mathbb{R}^d$, suppose that $0 < H^{d-1}(S) < \infty$. Let $T(r) := \{\mathbf{x} \in \mathbb{R}^d : \mathbb{B}(\mathbf{x}, r) \cap S \neq \varnothing\}$ for $r > 0$. Then $\lambda_{\text{Leb}}(T(r)) = 2rH^{d-1}(S) + o(r)$ for r sufficiently small.*

Combining (B.23) and (B.24) establishes the desired bound (B.20).

Bounding the upper inner bias $I_{\text{in},2}$. The proof follows a similar line of argument as that of (B.20). We first apply Lemma B.4 for $\mathbf{x} \in S_p(\varrho(\frac{\nu_m}{m}))$ and obtain

$$\begin{aligned} &I_{\text{in},2}(\mathbf{x})\\ &\lesssim_{\sigma_p,L,C_p,C_0,d} \int_1^{\nu_m} u^b\Big\{(1+u)\Big(\frac{u}{m}\Big)^{\frac{\sigma_p}{d}}\\ &\qquad\qquad + k^{-k}\frac{(k^2+u^2)u^{k-1}e^{-up(\mathbf{x})}}{m}\Big\}\,\mathrm{d}u. \end{aligned}$$

The first term is bounded by $O(m^{-\frac{\sigma_p}{d}}\nu_m^{(b+\frac{\sigma_p}{d}+2)\vee 0})$. The second term is again bounded by the upper bound on the lower incomplete gamma function. If $b + k > 0$, we have

$$\begin{aligned} &\int_1^{\nu_m} \frac{k^{-k}}{m}(k^2+u^2)u^{b+k-1}e^{-up(\mathbf{x})}\,\mathrm{d}u\\ &\le \frac{k^{-k}}{m}(k^2p^{-(b+k)}(\mathbf{x})\gamma(b+k, \nu_m p(\mathbf{x}))\\ &\qquad + p^{-(b+k+2)}(\mathbf{x})\gamma(b+k+2, \nu_m p(\mathbf{x})))\\ &= O\Big(k^{-k}\frac{(k^2\nu_m^{(b+k)\vee 0} + \nu_m^{(b+k+2)\vee 0})}{m}\Big)\\ &= O\Big(k^{-k}\frac{\nu_m^{(b+k+2)\vee 0}}{m}\Big). \end{aligned}$$

One can easily show that the bound also holds when $b+k \le 0$. Hence, we conclude that

$$\begin{aligned} &I_{\text{in},2,\text{smooth}} \qquad\qquad \text{(B.25)}\\ &= \int_{S_p(\varrho(\frac{\nu_m}{m}))} I_{\text{in},2}(\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x}\\ &= O(m^{-\frac{\sigma_p}{d}}\nu_m^{(b+\frac{\sigma_p}{d}+2)\vee 0} + m^{-1}\nu_m^{(b+k+2)\vee 0}). \end{aligned}$$

Similar to (B.24), we have

$$\begin{aligned} &I_{\text{in},2,\text{nonsmooth}} \qquad\qquad \text{(B.26)}\\ &= \int_{\text{supp}(p)\backslash S_p(\varrho(\frac{\nu_m}{m}))} I_{\text{in},2}(\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x},\\ &= O((\nu_m/m)^{\frac{1}{d}}(\nu_m^{(b+2)\vee 0} + m^{-1}\nu_m^{(b+k+2)\vee 0})), \end{aligned}$$

since $\lambda_{\text{Leb}}(\text{supp}(p)\backslash S_p(\varrho(\nu_m/m))) = O(\varrho(\nu_m/m)) = O((\nu_m/m)^{\frac{1}{d}})$ by Lemma B.22. Putting (B.25) and (B.26) together establishes the desired bound (B.21). □

Lemma B.23 (Generic outer bias bound)**.** *Suppose that the density p satisfies **($\mathbf{U}_p$)**.*

1) *If $k > -a$, we have*

$$I_{\textit{out},1} = O\big(k^{-k}\tau_m^{k+a}\big). \qquad \text{(B.27)}$$

2) *If p satisfies **($\mathbf{L1}_p$)**, **($\mathbf{L2}_p$)**, and **($\mathbf{L3}_p$)**, then, for m sufficiently large, we have*

$$\begin{aligned} I_{\textit{out},2} = O\big(&k^b\nu_m^{b+k-1}e^{-c_p\nu_m}\\ &+ (\nu_m^b \vee \kappa_m^b)\Big(\frac{\nu_m}{k}\Big)^k e^{-\eta_p c_p\nu_m}\big). \qquad \text{(B.28)} \end{aligned}$$

Proof. Recall that

$$\rho_{U_{k\infty}(\mathbf{x})}(u) = \frac{p^k(\mathbf{x})}{\Gamma(k)}u^{k-1}e^{-up(\mathbf{x})}.$$

Define

$$A_{k\infty}(\mathbf{x}; p) := \int_0^{\tau_m} u^a\rho_{U_{k\infty}(\mathbf{x})}(u)\,\mathrm{d}u$$

and

$$B_{k\infty}(\mathbf{x}; p) := \int_{\nu_m}^{\infty} u^b\rho_{U_{k\infty}(\mathbf{x})}(u)\,\mathrm{d}u.$$

For some $\kappa_m = \omega(m)$ such that $\kappa_m \ge \nu_m$, we also let $A_{km}(\mathbf{x}; p) := A_{km}(\mathbf{x}; p; \xi)$, $B_{km}^{(2)}(\mathbf{x}; p) := B_{km}^{(2)}(\mathbf{x}; p; \xi)$, and $B_{km}^{(3)}(\mathbf{x}; p) := B_{km}^{(3)}(\mathbf{x}; p; \xi)$ for $\xi(t) = t$; recall the definitions in Appendix B-D. Now we can write the lower outer bias as

$$I_{\text{out},1} = \int p(\mathbf{x})(A_{km}(\mathbf{x}; p) + A_{k\infty}(\mathbf{x}; p))\,\mathrm{d}\mathbf{x}$$

and the upper outer bias as

$$I_{\text{out},2} = \int p(\mathbf{x})(B_{km}^{(2)}(\mathbf{x}; p) + B_{km}^{(3)}(\mathbf{x}; p) + B_{k\infty}(\mathbf{x}; p))\,\mathrm{d}\mathbf{x}$$

Bounding the lower outer bias $I_{\mathrm{out},1}$. On the one hand, by invoking the lower incomplete gamma function in Lemma B.2, we obtain

$$\begin{aligned}A_{k\infty}(\mathbf{x};p) &= \frac{p^k(\mathbf{x})}{\Gamma(k)}\int_0^{\tau_m} u^{k+a-1}e^{-up(\mathbf{x})}\,\mathrm{d}u\\ &= \frac{p^{-a}(\mathbf{x})}{\Gamma(k)}\gamma(k+a,\tau_m p(\mathbf{x}))\\ &\le \frac{p^k(\mathbf{x})\tau_m^{k+a}}{\Gamma(k)(k+a)}\\ &\le \frac{C_p^k\tau_m^{k+a}}{\Gamma(k)(k+a)} = O(k^{-k}\tau_m^{k+a}).\end{aligned}$$

On the other hand, by applying Lemma B.13 with the upper boundedness condition $(\mathbf{U}_p)$, we obtain

$$\begin{aligned}\int p(\mathbf{x})A_{km}(\mathbf{x};p)\,\mathrm{d}\mathbf{x} &\le \frac{C_p^k\tau_m^{k+a}}{k!}\Big(1\vee\frac{k}{k+a}\Big)\\ &= O(k^{-k}\tau_m^{k+a}).\end{aligned}$$

Combining the two bounds, we conclude that $I_{\mathrm{out},1} = O(k^{-k}\tau_m^{k+a})$.

Bounding the upper outer bias $I_{\mathrm{out},2}$. For the $B_{k\infty}(\mathbf{x};p)$ term in the upper outer bias $I_{\mathrm{out},2}$, we apply the bound (B.2) on the upper incomplete gamma function in Lemma B.2. Consider

$$\begin{aligned}B_{k\infty}(\mathbf{x};p) &= \frac{p^k(\mathbf{x})}{\Gamma(k)}\int_{\nu_m}^{\infty} u^{k+b-1}e^{-up(\mathbf{x})}\,\mathrm{d}u\\ &= \frac{p^{-b}(\mathbf{x})}{\Gamma(k)}\int_{\nu_m p(\mathbf{x})}^{\infty} t^{k+b-1}e^{-t}\,\mathrm{d}t.\end{aligned}$$

If $\nu_m p(\mathbf{x})<1$, we have

$$\begin{aligned}B_{k\infty}(\mathbf{x};p) &\le \frac{p^{-b}(\mathbf{x})}{\Gamma(k)}\int_0^{\infty} t^{k+b-1}e^{-t}\,\mathrm{d}t\\ &\le \frac{\Gamma((k+b)\vee 1)}{\Gamma(k)}p^{-b}(\mathbf{x}).\end{aligned}$$

We now assume that $\nu_m p(\mathbf{x})\ge 1$. If $k+b\ge 1$, we have

$$\begin{aligned}B_{k\infty}(\mathbf{x};p) &= \frac{p^{-b}(\mathbf{x})}{\Gamma(k)}\Gamma(k+b,\nu_m p(\mathbf{x}))\\ &\le \frac{p^{-b}(\mathbf{x})}{\Gamma(k)}\Gamma(k+b)(\nu_m p(\mathbf{x}))^{k+b-1}e^{-\nu_m p(\mathbf{x})+1}\\ &= \frac{\Gamma(k+b)}{\Gamma(k)}\nu_m^{k+b-1}p^{k-1}(\mathbf{x})e^{-\nu_m p(\mathbf{x})+1},\end{aligned}$$

where the inequality follows from Lemma B.2. For $k+b<1$, a similar bound can be derived:

$$\begin{aligned}B_{k\infty}(\mathbf{x};p) &= \frac{p^{-b}(\mathbf{x})}{\Gamma(k)}(\nu_m p(\mathbf{x}))^{k+b-1}\int_{\nu_m p(\mathbf{x})}^{\infty} e^{-t}\,\mathrm{d}t\\ &= \frac{1}{\Gamma(k)}\nu_m^{k+b-1}p^{k-1}(\mathbf{x})e^{-\nu_m p(\mathbf{x})}.\end{aligned}$$

To sum up, we can bound $B_{k\infty}(\mathbf{x};p)$ as

$$\begin{aligned}&B_{k\infty}(\mathbf{x};p)\\ &\le \frac{\Gamma((k+b)\vee 1)}{\Gamma(k)}\big(p^{-b}(\mathbf{x})\mathbb{1}_{\{\nu_m p(\mathbf{x})<1\}}\\ &\qquad\qquad + \nu_m^{k+b-1}p^{k-1}(\mathbf{x})e^{-\nu_m p(\mathbf{x})+1}\big)\\ &\overset{(a)}{\le} \frac{\Gamma((k+b)\vee 1)}{\Gamma(k)}(p^{-b}(\mathbf{x})+\nu_m^{k+b-1}p^{k-1}(\mathbf{x}))e^{-\nu_m p(\mathbf{x})+1}\\ &\overset{(b)}{\le} \frac{\Gamma((k+b)\vee 1)}{\Gamma(k)}((C_p^{-b}\vee c_p^{-b})+\nu_m^{k+b-1}C_p^{k-1})e^{-\nu_m c_p+1}\\ &= O(k^b\nu_m^{k+b-1}e^{-c_p\nu_m}).\end{aligned}$$

Here, (a) follows from the inequality $\mathbb{1}_{\{t\le 1\}}\le e^{-t+1}$, and (b) follows from the boundedness conditions $(\mathbf{U}_p)$ and $(\mathbf{L1}_p)$. Therefore, we conclude that

$$\int p(\mathbf{x})B_{k\infty}(\mathbf{x};p)\,\mathrm{d}\mathbf{x} = O(k^b\nu_m^{k+b-1}e^{-c_p\nu_m}). \tag{B.29}$$

Next, we bound $\int p(\mathbf{x})(B_{km}^{(2)}(\mathbf{x};p)+B_{km}^{(3)}(\mathbf{x};p))\,\mathrm{d}\mathbf{x}$. On the one hand, applying Lemma B.19 with the upper boundedness condition $(\mathbf{U}_p)$, we first have

$$\begin{aligned}&\int p(\mathbf{x})B_{km}^{(2)}(\mathbf{x};p)\,\mathrm{d}\mathbf{x}\\ &\le 2(\nu_m^b\vee\kappa_m^b)\Big(\frac{eC_p\nu_m}{k}\Big)^k\int p(\mathbf{x})e^{-\nu_m m_r p(\mathbf{x})}\,\mathrm{d}\mathbf{x}\end{aligned}$$

for $r=\varrho(\frac{\nu_m}{m})$. Further, since we have

$$\eta_p = \inf_{\mathbf{x}\in\mathrm{supp}(p)}\inf_{r'\in(0,r]}\frac{\lambda_{\mathrm{Leb}}(\mathbb{B}(\mathbf{x},r)\cap\mathrm{supp}(p))}{\lambda_{\mathrm{Leb}}(\mathbb{B}(\mathbf{x},r))}>0$$

from condition $(\mathbf{L3}_p)$, it follows that $m_r p(\mathbf{x})\ge c_p\eta_p$ for $\mathbf{x}\in\mathrm{supp}(p)$, leading to

$$\int p(\mathbf{x})B_{km}^{(2)}(\mathbf{x};p)\,\mathrm{d}\mathbf{x} \le 2(\nu_m^b\vee\kappa_m^b)\Big(\frac{e\nu_m C_p}{k}\Big)^k e^{-\eta_p c_p\nu_m}.$$

On the other hand, since the support of the density p is bounded by the condition $(\mathbf{L2}_p)$, $R(p,p;\xi,b,\varrho(\kappa_m/m))$ becomes 0 for m sufficiently large, since $\kappa_m/m\to\infty$ as $m\to\infty$. Hence, by applying Lemma B.16 for a fixed $\delta>0$, we have

$$\begin{aligned}&\int p(\mathbf{x})B_{km}^{(3)}(\mathbf{x};p)\,\mathrm{d}\mathbf{x}\\ &\le 3(1+\delta)m^bR\big(p,p;\xi,b,\varrho\big(\frac{\kappa_m}{m}\big)\big)=0\end{aligned}$$

for m sufficiently large. Therefore, we conclude that

$$\begin{aligned}&\int p(\mathbf{x})(B_{km}^{(2)}(\mathbf{x};p)+B_{km}^{(3)}(\mathbf{x};p))\,\mathrm{d}\mathbf{x}\\ &= O\Big((\nu_m^b\vee\kappa_m^b)\Big(\frac{\nu_m}{k}\Big)^k e^{-\eta_p c_p\nu_m}\Big).\end{aligned} \tag{B.30}$$

Combining the bounds (B.29) and (B.30) establishes the desired bound (B.28). □

Remark B.3. A more general condition, namely, that

$(\mathbf{B1}'_p)$ there exists $E_0,E_1>0$ such that $\int p(\mathbf{x})e^{-\beta p(\mathbf{x})}\,\mathrm{d}\mathbf{x}\le E_0e^{-E_1\beta}$ for all $\beta>1$,

was originally assumed in [30]. Known examples of densities that satisfy the condition $(\mathbf{B1}'_p)$ satisfy the more intuitive condition **(L1$_p$)**. We remark, however, that it is nontrivial to adapt the proofs in this paper to work with **(B1$'_p$)** in place of **(L1$_p$)**, as the lower boundedness condition **(L1$_p$)** is explicitly utilized to remove the upper truncation of the estimator in the analysis of [30].

F. Generic variance bounds

Lemma B.24. *For a given function $\phi\colon \mathbb{R}_+ \to \mathbb{R}$, let $\zeta_k(\mathbf{x}|\mathbf{x}_{1:m}) := \phi(r_k(\mathbf{x}|\mathbf{x}_{1:m}))$ for any points $\mathbf{x}, \mathbf{x}_{1:m}$ in the d-dimensional Euclidean space $(\mathbb{R}^d, \|\cdot\|)$. Let*

$$\Phi(\mathbf{x}_{1:m}) = \frac{1}{m}\sum_{i=1}^m \zeta_k(\mathbf{x}_i|\mathbf{x}_{1:m}^{\sim i}). \tag{B.31}$$

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

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

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

Before we prove Lemma B.24, we introduce two technical lemmas.

Lemma B.25 (Efron–Stein inequality [79, 80])**.** *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*

$$\begin{aligned}&\mathrm{Var}(g(X_{1:n}))\\ &\le \frac{1}{2}\sum_{i=1}^n \mathbb{E}\bigl[|g(X_{1:n}) - g(X_{1:i-1}X'_iX_{i+1:n})|^2\bigr].\end{aligned}$$

The proof of this lemma can be found in [80].

We need another fact on k-nearest neighbors in the Euclidean space, stated below in Lemma B.24. Informally speaking, given a finite collection S of points in $\mathbb{R}^d$, each fixed point in $\mathbb{R}^d$ can be one of the k nearest neighbors of at most γ_d points in S, where γ_d depends only on d. Henceforth, for a set of points A such that $\mathbf{x} \notin A$, we use $N_k(\mathbf{x}|A)$ to denote the k-nearest neighbors of $\mathbf{x}$ in A.

Lemma B.26 ([37, Lemma 20.6], [81, Ch. 5.3])**.** *In the d-dimensional Euclidean space $(\mathbb{R}^d, \|\cdot\|)$ 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 $\mathbf{x}, \mathbf{x}_1, \ldots, \mathbf{x}_m \in \mathbb{R}^d$,*

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

Proof. We follow the proof of Stone's lemma in Devroye et al. [81, Ch. 5.3]. For $\mathbf{z} \in \mathbb{R}^d\backslash\{\mathbf{0}\}$ and $\theta \in (0, \pi/2]$, we define a cone $\mathcal{C}(\mathbf{z}, \theta) := \{\mathbf{y} \in \mathbb{R}^d\colon \mathbf{y} = \mathbf{0} \text{ or } \angle(\mathbf{z}, \mathbf{y}) \le \theta\}$. It is well known [37, Theorem 20.16] that there exists a constant $\gamma_d > 0$, which depends only on the dimension d, such that there exist γ_d cones $\mathcal{C}(\mathbf{z}_1, \pi/6), \ldots, \mathcal{C}(\mathbf{z}_{\gamma_d}, \pi/6)$ which cover the entire space $\mathbb{R}^d$. Furthermore, it is easy to see that ($\star$) if $\mathbf{y}_1, \mathbf{y}_2 \in \mathcal{C}(\mathbf{x}, \pi/6)$ and $\|\mathbf{y}_1\| < \|\mathbf{y}_2\|$, then $\|\mathbf{y}_1 - \mathbf{y}_2\| < \|\mathbf{y}_2\|$; see, e.g., [37, Lemma 20.5].

Now, for each $j \in [\gamma_d]$, *mark* all $\mathbf{x}_i$'s (if any) among the k-nearest neighbors of $\mathbf{x}$ in $\mathbf{x} + \mathcal{C}(\mathbf{z}_j, \pi/6)$. If $\mathbf{x}_i \in \mathbf{x} + \mathcal{C}(\mathbf{z}_j, \pi/6)$ for some $j \in [\gamma_d]$ and $\mathbf{x}_i$ is not marked, then $\mathbf{x}$ is not among the k-nearest neighbors of $\mathbf{x}_i$ in $\mathbf{x}_{1:i-1}, \mathbf{x}_{i+1:m}, \mathbf{x}$, i.e., $\mathbf{x} \notin N_k(\mathbf{x}_i|\mathbf{x}_{1:m}^{\sim i}, \mathbf{x})$, by the property ($\star$). Therefore, we have

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

since there exist at most $k\gamma_d$ marked points. □

We are now ready to prove Lemma B.24.

Proof of Lemma B.24. Let $\mathbf{X}'_1$ be an independent copy of $\mathbf{X}_1$. Then, by applying the Efron–Stein inequality (Lemma B.25), we have

$$\begin{aligned}&\mathrm{Var}\bigl(\Phi(\mathbf{X}_{1:m})\bigr)\\ &\le \frac{m}{2}\mathbb{E}\bigl[\bigl(\Phi(\mathbf{X}_{1:m}) - \Phi(\mathbf{X}'_1\mathbf{X}_{2:m})\bigr)^2\bigr]\\ &\overset{(a)}{\le} m\mathbb{E}\Bigl[\Bigl(\Phi(\mathbf{X}_{1:m}) - \frac{m-1}{m}\Phi(\mathbf{X}_{2:m})\Bigr)^2 \qquad \text{(B.32)}\\ &\qquad + \Bigl(\Phi(\mathbf{X}'_1\mathbf{X}_{2:m}) - \frac{m-1}{m}\Phi(\mathbf{X}_{2:m})\Bigr)^2\Bigr]\\ &= 2m\mathbb{E}\Bigl[\Bigl(\Phi(\mathbf{X}_{1:m}) - \frac{m-1}{m}\Phi(\mathbf{X}_{2:m})\Bigr)^2\Bigr], \qquad \text{(B.33)}\end{aligned}$$

where (a) follows from the elementary inequality $(a-b)^2 \le 2((a-x)^2 + (b-x)^2)$.

Define

$$E_i := \{\mathbf{X}_1 \text{ is one of the } k\text{-NNs of } \mathbf{X}_i \text{ in } \mathbf{X}_{1:m}^{\sim i}\}$$

for $2 \le i \le m$. Applying Lemma B.26, we obtain

$$\sum_{i=2}^m \mathbb{1}_{E_i} \le k\gamma_d.$$

Further, note that if E_i^c occurs, i.e., $\mathbf{X}_1$ is not among the k nearest neighbors of $\mathbf{X}_i$ in $\mathbf{X}_{1:m}^{\sim i}$, then $\zeta_k(\mathbf{X}_i|\mathbf{X}_{1:m}^{\sim i}) = \zeta_k(\mathbf{X}_i|\mathbf{X}_{2:m}^{\sim i})$. We thus obtain (B.34), where (b) follows from Cauchy–Schwarz inequality. By taking expectations with respect to $\mathbf{X}_{1:m}$ on both sides and multiplying by $2/m$, we can continue from (B.33) to obtain

$$\begin{aligned}&\mathrm{Var}\bigl(\Phi(\mathbf{X}_{1:m})\bigr) \qquad \text{(B.35)}\\ &\le \frac{2(1+k\gamma_d)}{m}\\ &\quad \times \Bigl\{\mathbb{E}\bigl[\zeta_k^2(\mathbf{X}_1|\mathbf{X}_{2:m})\bigr]\\ &\qquad + 2\mathbb{E}\Bigl[\sum_{i=2}^m \mathbb{1}_{E_i}(\zeta_k^2(\mathbf{X}_i|\mathbf{X}_{1:m}^{\sim i}) + \zeta_k^2(\mathbf{X}_i|\mathbf{X}_{2:m}^{\sim i}))\Bigr]\Bigr\}.\end{aligned}$$

$$\begin{aligned} m^2\big(\Phi(\mathbf{X}_{1:m}) - \frac{m-1}{m}\Phi(\mathbf{X}_{2:m})\big)^2 &= \Big(\zeta_k(\mathbf{X}_1|\mathbf{X}_{2:m}) + \sum_{i=2}^m \mathbb{1}_{E_i}\big(\zeta_k(\mathbf{X}_i|\mathbf{X}_{1:m}^{\sim i}) - \zeta_k(\mathbf{X}_i|\mathbf{X}_{2:m}^{\sim i})\big)\Big)^2 \\ &\overset{(b)}{\le} \Big(1+\sum_{i=2}^m \mathbb{1}_{E_i}\Big)\Big(\zeta_k^2(\mathbf{X}_1|\mathbf{X}_{2:m}) + \sum_{i=2}^m \mathbb{1}_{E_i}\big(\zeta_k(\mathbf{X}_i|\mathbf{X}_{1:m}^{\sim i}) - \zeta_k(\mathbf{X}_i|\mathbf{X}_{2:m}^{\sim i})\big)^2\Big) \\ &\le (1+k\gamma_d)\Big(\zeta_k^2(\mathbf{X}_1|\mathbf{X}_{2:m}) + 2\sum_{i=2}^m \mathbb{1}_{E_i}\big(\zeta_k^2(\mathbf{X}_i|\mathbf{X}_{1:m}^{\sim i}) + \zeta_k^2(\mathbf{X}_i|\mathbf{X}_{2:m}^{\sim i})\big)\Big). \end{aligned} \tag{B.34}$$

Note that if E_i occurs, i.e., $\mathbf{X}_1$ is among the k nearest neighbors of $\mathbf{X}_i$ in $\mathbf{X}_{1:m}^{\sim i}$, we have $\zeta_k(\mathbf{X}_i|\mathbf{X}_{2:m}^{\sim i}) = \zeta_{k+1}(\mathbf{X}_i|\mathbf{X}_{1:m}^{\sim i})$. Therefore, it follows that

$$\begin{aligned} &\mathbb{E}\Big[\sum_{i=2}^m \mathbb{1}_{E_i}(\zeta_k^2(\mathbf{X}_i|\mathbf{X}_{1:m}^{\sim i}) + \zeta_k^2(\mathbf{X}_i|\mathbf{X}_{2:m}^{\sim i}))\Big] \\ &= \mathbb{E}\Big[\sum_{i=2}^m \mathbb{1}_{E_i}(\zeta_k^2(\mathbf{X}_i|\mathbf{X}_{1:m}^{\sim i}) + \zeta_{k+1}^2(\mathbf{X}_i|\mathbf{X}_{1:m}^{\sim i}))\Big] \\ &\overset{(c)}{=} \mathbb{E}\Big[\sum_{i=2}^m \mathbb{1}_{\{\mathbf{X}_i \text{ is among the } k\text{-NNs of } \mathbf{X}_1 \text{ in } \mathbf{X}_{2:m}\}} \\ &\qquad\qquad \times (\zeta_k^2(\mathbf{X}_1|\mathbf{X}_{2:m}) + \zeta_{k+1}^2(\mathbf{X}_1|\mathbf{X}_{2:m}))\Big] \\ &= k\mathbb{E}[\zeta_k^2(\mathbf{X}_1|\mathbf{X}_{2:m}) + \zeta_{k+1}^2(\mathbf{X}_1|\mathbf{X}_{2:m})], \end{aligned} \tag{B.36}$$

where (c) follows by exchanging $\mathbf{X}_1$ and $\mathbf{X}_i$ in each summand $2 \le i \le m$. Therefore, plugging the equation in (B.36) into (B.35) proves the desired bound. □

For the double-density case, we can establish a similar variance bound.

Lemma B.27. *For a given function $\phi\colon \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$, let $\zeta_{kl}(\mathbf{x}|\mathbf{x}_{1:m}, \mathbf{y}_{1:n}) := \phi(r_k(\mathbf{x}|\mathbf{x}_{1:m}), r_l(\mathbf{x}|\mathbf{y}_{1:n}))$ for any points $\mathbf{x}, \mathbf{x}_{1:m}, \mathbf{y}_{1:n}$ in the d-dimensional Euclidean space $(\mathbb{R}^d, \|\cdot\|)$. Let*

$$\Phi(\mathbf{x}_{1:m}, \mathbf{y}_{1:n}) := \frac{1}{m}\sum_{i=1}^m \zeta_{kl}(\mathbf{x}_i|\mathbf{x}_{1:m}^{\sim i}, \mathbf{y}_{1:n}). \tag{B.37}$$

If $\mathbf{X}_{1:m}$ and $\mathbf{Y}_{1:n}$ are independent i.i.d. samples, we have

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

Proof. Given $\mathbf{Y}_{1:n} = \mathbf{y}_{1:n}$, we can show that

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

by following the same line of reasoning as in the proof of Lemma B.24. Since $\mathbf{Y}_{1:n}$ is independent of $\mathbf{X}_{1:m}$, taking expectation on both sides with respect to $\mathbf{Y}_{1:n}$ establishes the desired bound. □

Appendix C
Deferred Proofs of Main Results

A. Detailed proof of Theorem III.5

We continue the proof from (III.12).

$$\big|\mathbb{E}[\hat{T}_f^{(k)}] - T_f(p)\big| \lesssim I_{\text{out},1} + I_{\text{in},1} + I_{\text{in},2} + I_{\text{out},2}. \tag{III.12}$$

Applying the bounds in Lemmas B.21 and B.23, we obtain the following bias bound for an underlying density p satisfying the conditions **($\mathbf{U}_p$)**, **($\mathbf{L1}_p$)**, **($\mathbf{S}_p$)**, and **($\mathbf{B}_p$)**, provided that $\nu_m = o(\sqrt{m})$ as $m \to \infty$ and $k \in \mathbb{N}$ is fixed:

$$\begin{aligned} &|\mathbb{E}[\hat{T}_f^{(k)}] - T_f(p)| \lesssim_{\sigma_p, L, C_p, C_0, d, k} \\ &\quad m^{-\frac{\sigma_p}{d}}\tau_m^{(a+\frac{\sigma_p}{d}+1)\wedge 0} + m^{-1} + m^{-\frac{1}{d}}\tau_m^{(a+1)\wedge 0} \\ &\quad + m^{-\frac{\sigma_p}{d}}\nu_m^{(b+\frac{\sigma_p}{d}+2)\vee 0} + m^{-1}\nu_m^{(b+k+2)\vee 0} + m^{-\frac{1}{d}}\nu_m^{(b+2)\vee 0+\frac{1}{d}} \\ &\quad + \tau_m^{k+a} + \nu_m^{b+k-1}e^{-c_p\nu_m}. \end{aligned}$$

First, by choosing $\nu_m = \Theta((\ln m)^{1+\delta})$ for some $\delta > 0$, we make the last term $\nu_m^{b+k-1}e^{-c_p\nu_m}$ decay faster than any polynomial rate. With this choice, the bound can be simplified as

$$\begin{aligned} &|\mathbb{E}[\hat{T}_f^{(k)}] - T_f(p)| \\ &= \tilde{O}_{\sigma_p, L, C_p, C_0, d, k}(\tau_m^{(a+\frac{\sigma_p}{d}+1)\wedge 0}m^{-\frac{\sigma_p}{d}} + \tau_m^{(a+1)\wedge 0}m^{-\frac{1}{d}} \\ &\qquad + m^{-\frac{\sigma_p\wedge 1}{d}} + \tau_m^{k+a}). \end{aligned}$$

We consider three different ranges of the lower tail exponent a.

1) If $a \le -\sigma_p/d - 1$, we have

$$|\mathbb{E}[\hat{T}_f^{(k)}] - T_f(p)| = \tilde{O}(\tau_m^{a+1}m^{-\frac{\sigma_p\wedge 1}{d}} + \tau_m^{k+a})$$

as a suboptimal bound. By equating the two terms, we establish a rate $\tilde{O}(m^{-\frac{(\sigma_p\wedge 1)}{d}\frac{k+a}{k-1}})$ with $\tau_m = \Theta(m^{-\frac{(\sigma_p\wedge 1)}{d}\frac{1}{k-1}})$.

2) If $-\sigma_p/d - 1 < a \le -1$, the rate becomes

$$|\mathbb{E}[\hat{T}_f^{(k)}] - T_f(p)| = \tilde{O}(\tau_m^{a+1}m^{-\frac{1}{d}} + m^{-\frac{\sigma_p\wedge 1}{d}} + \tau_m^{k+a}).$$

Equating $\tau_m^{a+1}m^{-\frac{1}{d}}$ and τ_m^{k+a} as a suboptimal choice, we obtain $\tau_m = \Theta(m^{-\frac{1}{d}\frac{1}{k-1}})$, which results in the final rate

$$\begin{aligned} |\mathbb{E}[\hat{T}_f^{(k)}] - T_f(p)| &= \tilde{O}(m^{-\frac{1}{d}\frac{k+a}{k-1}} + m^{-\frac{\sigma_p\wedge 1}{d}}) \\ &= \tilde{O}(m^{-\frac{1}{d}(\sigma_p\wedge\frac{k+a}{k-1})}) \end{aligned}$$

3) If $a > -1$, we can attain the bias rate $\tilde{O}(m^{-\frac{\sigma_p\wedge 1}{d}})$ by using $\tau_m = O(m^{-\frac{1}{d(a+1)}})$.

To sum up, by choosing

$$\begin{aligned}\tau_m &= \tau(m,d,\sigma_p,a,k) \qquad \text{(C.1)}\\ &= \begin{cases} \Theta\big(m^{-\frac{\sigma_p\wedge 1}{d(k-1)}}\big) & \text{if } a \le -\frac{\sigma_p}{d}-1,\\ \Theta\big(m^{-\frac{1}{d(k-1)}}\big) & \text{if } -\frac{\sigma_p}{d}-1 < a \le -1,\\ O\big(m^{-\frac{1}{d(a+1)}}\big) & \text{if } a > -1,\end{cases}\end{aligned}$$

we establish the bias bound in Theorem III.5. □

B. Proof of Theorem IV.1

Following a similar line of reasoning as in the proof of Proposition I.1 and using the continuous mapping theorem, it is easy to show that $\phi_k(U_{k,m-1}(\mathbf{X}_m), V_{ln}(\mathbf{X}_m))$ converges to $\phi_{kl}(U_{k\infty}(\mathbf{X}), V_{l\infty}(\mathbf{X}))$ in distribution as $m,n\to\infty$, where $U_{k\infty}(\mathbf{x})$ and $V_{l\infty}(\mathbf{x})$ are a $\mathsf{G}(k,p(\mathbf{x}))$ random variable and a $\mathsf{G}(l,q(\mathbf{x}))$ random variable, respectively, which are independent of each other and of $\mathbf{X}\sim p$, for $\mathcal{P}$-a.e. $\mathbf{x}$. Hence, if we can only show that the collection of random variables $(\phi_{kl}(U_{k,m-1}(\mathbf{X}_m), V_{ln}(\mathbf{X}_m)))_{m,n\ge 1}$ is uniformly integrable, we can readily establish the asymptotic unbiasedness as follows:

$$\begin{aligned}&\lim_{m,n\to\infty}\mathbb{E}[\hat{T}_f^{(kl)}(\mathbf{X}_{1:m},\mathbf{Y}_{1:n})]\\ &= \lim_{m,n\to\infty}\mathbb{E}[\phi_{kl}(U_{k,m-1}(\mathbf{X}_m), V_{ln}(\mathbf{X}_m))]\\ &= \mathbb{E}[\phi_{kl}(U_{k\infty}(\mathbf{X}), V_{l\infty}(\mathbf{X}))]\\ &= T_f(p,q).\end{aligned}$$

Consider

$$\begin{aligned}&\mathbb{E}[\xi(|\phi_{kl}(U_{k,m-1}(\mathbf{X}_m), V_{ln}(\mathbf{X}_m))|)]\\ &= \int p(\mathbf{x})\mathbb{E}[\xi(|\phi_{kl}(U_{k,m-1}(\mathbf{x}), V_{ln}(\mathbf{x}))|)]\,\mathrm{d}\mathbf{x}.\end{aligned}$$

By invoking the polynomial bound $|\phi_{kl}(u,v)| \lesssim \psi_{a,b}(u)\psi_{\tilde{a},\tilde{b}}(v)$ and using the independence of $U_{k,m-1}(\mathbf{x})$ and $V_{ln}(\mathbf{x})$, we have

$$\begin{aligned}&\mathbb{E}[\xi(|\phi_{kl}(U_{k,m-1}(\mathbf{X}_m), V_{ln}(\mathbf{X}_m))|)] \qquad \text{(C.2)}\\ &\lesssim_{\xi(t_0)} 1 + \mathbb{E}[\xi(\psi_{a,b}(U_{k,m-1}(\mathbf{X}_m)))]\\ &\qquad + \mathbb{E}[\xi(\psi_{\tilde{a},\tilde{b}}(V_{ln}(\mathbf{X}_m)))]\\ &\qquad + \{\mathbb{E}[(\mathbb{E}[\xi(\psi_{a,b}(U_{km}(\mathbf{X}_m)))|\mathbf{X}_m]\\ &\qquad\qquad \times \mathbb{E}[\xi(\psi_{\tilde{a},\tilde{b}}(V_{ln}(\mathbf{X}_m)))|\mathbf{X}_m])]\}^2,\end{aligned}$$

since $\xi(xy)\le\xi(x)\xi(y)$ for any $x,y>t_0$. We can bound the last term as

$$\begin{aligned}&\{\mathbb{E}[(\mathbb{E}[\xi(\psi_{a,b}(U_{km}(\mathbf{X}_m)))|\mathbf{X}_m]\\ &\qquad \times \mathbb{E}[\xi(\psi_{\tilde{a},\tilde{b}}(V_{ln}(\mathbf{X}_m)))|\mathbf{X}_m])]\}^2\\ &\overset{(a)}{\le} \mathbb{E}[(\mathbb{E}[\xi^2(\psi_{a,b}(U_{km}(\mathbf{X}_m)))|\mathbf{X}_m])^2]\\ &\qquad \times \mathbb{E}[(\mathbb{E}[\xi^2(\psi_{\tilde{a},\tilde{b}}(V_{ln}(\mathbf{X}_m)))|\mathbf{X}_m])^2]\\ &\overset{(b)}{\le} \mathbb{E}[\xi^2(\psi_{a,b}(U_{k,m-1}(\mathbf{X}_m)))]\mathbb{E}[\xi^2(\psi_{\tilde{a},\tilde{b}}(V_{ln}(\mathbf{X}_m)))],\end{aligned}$$

where (a) and (b) follow from Cauchy–Schwarz inequality and Jensen's inequality. We thus only need to show that

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

and

$$\limsup_{n\to\infty}\mathbb{E}[\xi^2(\psi_{\tilde{a},\tilde{b}}(V_{ln}(\mathbf{X}_m)))] < \infty,$$

since they would imply that all the terms in (C.2) are bounded. By applying Lemma B.17 to both integrals for $k > -2a\omega(\xi)$ and $l > -2\tilde{a}\omega(\xi)$, we conclude the proof by the de la Vallée Poussin theorem (Lemma III.4). □

C. Proof of Theorem IV.2

Recall from the generic variance bound (Lemma B.27) that we have

$$\begin{aligned}&\mathrm{Var}(T_f^{(kl)})\\ &\le \frac{2(1+k\gamma_d)}{m}\{(2k+1)\mathbb{E}[\phi_{kl}^2(U_{k,m-1}(\mathbf{X}_m), V_{ln}(\mathbf{X}_m))]\\ &\qquad + 2k\mathbb{E}[\phi_{kl}^2(U_{k+1,m-1}(\mathbf{X}_m), V_{ln}(\mathbf{X}_m))]\}.\end{aligned}$$

Hence, following the same logic as in Section C-B, in order to ensure that $\mathrm{Var}(\hat{T}_f^{(kl)}) = O(m^{-1})$ for m and n sufficiently large, it is enough to show that

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

and

$$\limsup_{n\to\infty}\mathbb{E}[\xi^2(\psi_{\tilde{a},\tilde{b}}(V_{ln}(\mathbf{X}_m)))] < \infty$$

for $\xi(t)=t^2$ and for $k'\in\{k,k+1\}$. By applying Lemma B.17 to both integrals for $k>-4a$ and $l>-4\tilde{a}$ with $\xi(t)=t^2$, we conclude the proof. □

D. Proof of Theorem IV.4

Let $k>-a$ and $l>-\tilde{a}$ be fixed. First, following similar steps as in (III.10), we can write the expected value of $\hat{T}_f^{(kl)}(\mathbf{X}_{1:m},\mathbf{Y}_{1:n})$ as

$$\begin{aligned}&\mathbb{E}\big[\hat{T}_f^{(kl)}(\mathbf{X}_{1:m},\mathbf{Y}_{1:n})\big]\\ &= \int p(\mathbf{x})\mathbb{E}\big[\phi_{kl}(U_{k,m-1}(\mathbf{x}), V_{ln}(\mathbf{x}))\big]\,\mathrm{d}\mathbf{x},\end{aligned}$$

since $U_{k,m-1}(\mathbf{x})$ and $V_{ln}(\mathbf{x})$ are independent of $\mathbf{X}_m=\mathbf{x}$ for $\mathcal{P}$-a.e. $\mathbf{x}$. Moreover, similar to (III.11), we can write the target density functional as

$$T_f(p,q) = \int p(\mathbf{x})\mathbb{E}[\phi_{kl}(U_{k\infty}(\mathbf{x}), V_{l\infty}(\mathbf{x}))]\,\mathrm{d}\mathbf{x},$$

where $U_{k\infty}(\mathbf{x})\sim\mathsf{G}(k,p(\mathbf{x}))$ and $V_{l\infty}(\mathbf{x})\sim\mathsf{G}(l,q(\mathbf{x}))$ are independent each other, and of $\mathbf{X}\sim p$ for $\mathcal{P}$-a.e. $\mathbf{x}$. Consider real numbers $\tau_m,\nu_m,\tilde{\tau}_n$, and $\tilde{\nu}_n$, to be determined later, such that $0\le\tau_m\le 1\le\nu_m<\infty$ and $0\le\tilde{\tau}_n\le 1\le\tilde{\nu}_n<\infty$. Using the polynomial bound $|\phi_{kl}(u,v)|\lesssim\psi_{a,b}(u)\psi_{\tilde{a},\tilde{b}}(v)$ and the triangle inequality, we then have

$$\begin{aligned}|\mathbb{E}[\hat{T}_f^{(kl)}] - T_f(p,q)| &\lesssim \int (I_{\mathrm{in}}(\mathbf{x}) + I_{\mathrm{out}}(\mathbf{x}))p(\mathbf{x})\,\mathrm{d}\mathbf{x}\\ &= I_{\mathrm{in}} + I_{\mathrm{out}}, \qquad \text{(C.3)}\end{aligned}$$

where $I_{\mathrm{in}}(\mathbf{x})$ and $I_{\mathrm{out}}(\mathbf{x})$ are defined in (C.4) and (C.5), where $\square_{m,n} := (\tau_m,\nu_m)\times(\tilde{\tau}_n,\tilde{\nu}_n)$. We bound the inner bias $I_{\mathrm{in}} = \int I_{\mathrm{in}}(\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x}$ and the outer bias $I_{\mathrm{out}} = \int I_{\mathrm{out}}(\mathbf{x})p(\mathbf{x})\,\mathrm{d}\mathbf{x}$

$$I_{\text{in}}(\mathbf{x}) := \int_{\square_{m,n}} \psi_{a,b}(u)\psi_{\tilde{a},\tilde{b}}(v)\big|\rho_{U_{k\infty}(\mathbf{x})}(u)\rho_{V_{l\infty}(\mathbf{x})}(v) - \rho_{U_{k,m-1}(\mathbf{x})}(u)\rho_{V_{ln}(\mathbf{x})}(v)\big|\,\mathrm{d}u\,\mathrm{d}v, \tag{C.4}$$

$$I_{\text{out}}(\mathbf{x}) := \int_{\mathbb{R}^2_+\setminus\square_{m,n}} \psi_{a,b}(u)\psi_{\tilde{a},\tilde{b}}(v)(\rho_{U_{k\infty}(\mathbf{x})}(u)\rho_{V_{l\infty}(\mathbf{x})}(v) + \rho_{U_{k,m-1}(\mathbf{x})}(u)\rho_{V_{ln}(\mathbf{x})}(v))\,\mathrm{d}u\,\mathrm{d}v. \tag{C.5}$$

separately. Henceforth, we use the following shorthand notation:

$$\overline{\psi}_{a,b}(u;\tau,\nu) = \psi_{a,b}(u)\mathbb{1}_{(\tau,\nu)}(u)$$

and

$$\overline{\overline{\psi}}_{a,b}(u;\tau,\nu) = \psi_{a,b}(u)(1 - \mathbb{1}_{(\tau,\nu)}(u)).$$

Step 1: Bounding the inner bias. For $\mathbf{x} \in \mathbb{R}^d$, let $\delta^{(p)}_{km}(u|\mathbf{x}) := |\rho_{U_{k,m-1}(\mathbf{x})}(u) - \rho_{U_{k\infty}(\mathbf{x})}(u)|$ and $\delta^{(q)}_{ln}(v|\mathbf{x}) := |\rho_{V_{ln}(\mathbf{x})}(v) - \rho_{V_{l\infty}(\mathbf{x})}(v)|$. By the triangle inequality, we have

$$\begin{aligned}
&|\rho_{U_{k,m-1}(\mathbf{x})}(u)\rho_{V_{ln}(\mathbf{x})}(v) - \rho_{U_{k\infty}(\mathbf{x})}(u)\rho_{V_{l\infty}(\mathbf{x})}(v)|\\
&\le \delta^{(p)}_{km}(u|\mathbf{x})\rho_{V_{ln}(\mathbf{x})}(v) + \delta^{(q)}_{ln}(v|\mathbf{x})\rho_{U_{k\infty}(\mathbf{x})}(v)\\
&\le \delta^{(p)}_{km}(u|\mathbf{x})\delta^{(q)}_{ln}(v|\mathbf{x}) + \delta^{(p)}_{km}(u|\mathbf{x})\rho_{V_{l\infty}}(\mathbf{x})\\
&\qquad + \delta^{(q)}_{ln}(v|\mathbf{x})\rho_{U_{k\infty}}(\mathbf{x}).
\end{aligned}$$

Therefore, for each $\mathbf{x} \in \operatorname{supp}(p)$, we can bound $I_{\text{in}}(\mathbf{x})$ as

$$\begin{aligned}
&I_{\text{in}}(\mathbf{x})\\
&\le \int_{\tau_m}^{\nu_m} \psi_{a,b}(u)\delta^{(p)}_{km}(u|\mathbf{x})\,\mathrm{d}u \int_{\tilde{\tau}_n}^{\tilde{\nu}_n} \psi_{\tilde{a},\tilde{b}}(v)\delta^{(q)}_{ln}(v|\mathbf{x})\,\mathrm{d}v\\
&\quad + \mathbb{E}[\overline{\psi}_{\tilde{a},\tilde{b}}(V_{l\infty}(\mathbf{x});\tilde{\tau}_n,\tilde{\nu}_n)] \int_{\tau_m}^{\nu_m} \psi_{a,b}(u)\delta^{(p)}_{km}(u|\mathbf{x})\,\mathrm{d}u\\
&\quad + \mathbb{E}[\overline{\psi}_{a,b}(U_{k\infty}(\mathbf{x});\tau_m,\nu_m)] \int_{\tilde{\tau}_n}^{\tilde{\nu}_n} \psi_{\tilde{a},\tilde{b}}(v)\delta^{(q)}_{ln}(v|\mathbf{x})\,\mathrm{d}v\\
&\overset{(a)}{\lesssim} \int_{\tau_m}^{\nu_m} \psi_{a,b}(u)\delta^{(p)}_{km}(u|\mathbf{x})\,\mathrm{d}u + \int_{\tilde{\tau}_n}^{\tilde{\nu}_n} \psi_{\tilde{a},\tilde{b}}(v)\delta^{(q)}_{ln}(v|\mathbf{x})\,\mathrm{d}v,
\end{aligned}$$

where (a) follows by applying Lemma B.20 with the assumptions **(U$_p$)** and **(L1$_p$)**. Therefore, we have

$$\begin{aligned}
I_{\text{in}} \lesssim \int p(\mathbf{x})\Big(&\int_{\tau_m}^{\nu_m} \psi_{a,b}(u)\delta^{(p)}_{km}(u|\mathbf{x})\,\mathrm{d}u\\
&+ \int_{\tilde{\tau}_n}^{\tilde{\nu}_n} \psi_{\tilde{a},\tilde{b}}(v)\delta^{(q)}_{ln}(v|\mathbf{x})\,\mathrm{d}v\Big)\,\mathrm{d}\mathbf{x},
\end{aligned}$$

and we can now apply the generic inner bias bounds in Lemma B.21 to bound the inner bias.

Step 2: Bounding the outer bias. We first consider the upper bound of $I_{\text{out}}(\mathbf{x})$ in (C.6). For the first integral, we have

$$\begin{aligned}
&\int_{\mathbb{R}\setminus(\tau_m,\nu_m)} \int_{\tilde{\tau}_n}^{\tilde{\nu}_n} \{\rho_{U_{k\infty}(\mathbf{x})}(u)\rho_{V_{l\infty}(\mathbf{x})}(v)\\
&\qquad\qquad + \rho_{U_{k,m-1}(\mathbf{x})}(u)\rho_{V_{ln}(\mathbf{x})}(v)\}\\
&\qquad\qquad \times \psi_{a,b}(u)\psi_{\tilde{a},\tilde{b}}(v)\,\mathrm{d}u\,\mathrm{d}v\\
&= \mathbb{E}[\overline{\overline{\psi}}(U_{k\infty}(\mathbf{x});\tau_m,\nu_m) + \overline{\overline{\psi}}(U_{k,m-1}(\mathbf{x});\tau_m,\nu_m)]\\
&\quad \times \mathbb{E}[\overline{\psi}(V_{l\infty}(\mathbf{x});\tilde{\tau}_n,\tilde{\nu}_n) + \overline{\psi}(V_{ln}(\mathbf{x});\tilde{\tau}_n,\tilde{\nu}_n)]\\
&\overset{(b)}{\lesssim} \mathbb{E}[\overline{\overline{\psi}}(U_{k\infty}(\mathbf{x});\tau_m,\nu_m) + \overline{\overline{\psi}}(U_{k,m-1}(\mathbf{x});\tau_m,\nu_m)],
\end{aligned}$$

where (b) follows from Lemmas B.20 and B.18. The second integral can be bounded similarly. Overall, we have

$$\begin{aligned}
I_{\text{out}} \lesssim &\int p(\mathbf{x})\mathbb{E}[\overline{\overline{\psi}}(U_{k\infty}(\mathbf{x});\tau_m,\nu_m)\\
&\qquad\qquad + \overline{\overline{\psi}}(U_{k,m-1}(\mathbf{x});\tau_m,\nu_m)]\,\mathrm{d}\mathbf{x}\\
&+ \int p(\mathbf{x})\mathbb{E}[\overline{\overline{\psi}}(V_{l\infty}(\mathbf{x});\tilde{\tau}_n,\tilde{\nu}_n)\\
&\qquad\qquad + \overline{\overline{\psi}}(V_{ln}(\mathbf{x});\tilde{\tau}_n,\tilde{\nu}_n)]\,\mathrm{d}\mathbf{x},
\end{aligned}$$

and we can now apply the generic outer bias bounds in Lemma B.23.

Step 3: Choosing break points. Putting the bounds on the inner and outer bias together and choosing the break points $(\tau_m,\nu_m,\tilde{\tau}_n,\tilde{\nu}_n)$ as in the proof of Theorem IV.4, we obtain the desired bias rates. □

E. Proof of Theorem IV.5

By Lemma B.27, we have

$$\begin{aligned}
&\operatorname{Var}(T^{(kl)}_f(\mathbf{X}_{1:m},\mathbf{Y}_{1:n}))\\
&\le \frac{2(1+k\gamma_d)}{m}\{(2k-2)\mathbb{E}[\phi^2_{kl}(U_{k-1,m-1}(\mathbf{X}_m),V_{ln}(\mathbf{X}_m))]\\
&\qquad + (2k+1)\mathbb{E}[\phi^2_{kl}(U_{k,m-1}(\mathbf{X}_m),V_{ln}(\mathbf{X}_m))]\\
&\qquad + \mathbb{E}[\phi^2_{kl}(U_{k+1,m-1}(\mathbf{X}_m),V_{ln}(\mathbf{X}_m))]\}.
\end{aligned}$$

Using Lemma B.18, we have

$$\begin{aligned}
&\mathbb{E}[\phi^2_{kl}(U_{k',m-1}(\mathbf{X}_m),V_{ln}(\mathbf{X}_m))]\\
&= \int p(\mathbf{x})\mathbb{E}[\phi^2_{kl}(U_{k',m-1}(\mathbf{x}),V_{ln}(\mathbf{x}))]\,\mathrm{d}\mathbf{x}\\
&\lesssim \int p(\mathbf{x})\mathbb{E}[\psi^2_{a,b}(U_{k',m-1}(\mathbf{x}))]\mathbb{E}[\psi^2_{\tilde{a},\tilde{b}}(V_{ln}(\mathbf{x}))]\,\mathrm{d}\mathbf{x}\\
&\lesssim 1
\end{aligned}$$

for

$$k \in \begin{cases} \{1,2\} & \text{if } k = 1,\\ \{k-1,k,k+1\} & \text{if } k \ge 2,\end{cases}$$

and for m and n sufficiently large, which concludes the proof. □

Appendix D
Deferred Proofs of Auxiliary Results

A. Proof of Proposition III.8

Similar to Lemmas B.21 and B.23, we establish the following bounds.

Lemma D.1 (Generic inner bias bound under **(S$'_p$)**). *Suppose that the density p satisfies the conditions* **(U$_p$)** *and* **(S$'_p$)** *, and let $k = o(\sqrt{m})$ as $m \to \infty$.*

$$I_{\text{out}}(\mathbf{x}) \le \int_{\mathbb{R}\setminus(\tau_m,\nu_m)} \int_{\tilde{\tau}_n}^{\tilde{\nu}_n} (\rho_{U_{k\infty}(\mathbf{x})}(u)\rho_{V_{l\infty}(\mathbf{x})}(v) + \rho_{U_{k,m-1}(\mathbf{x})}(u)\rho_{V_{ln}(\mathbf{x})}(v))\psi_{a,b}(u)\psi_{\tilde{a},\tilde{b}}(v)\,\mathrm{d}u\,\mathrm{d}v \\ + \int_{\tau_m}^{\nu_m} \int_{\mathbb{R}\setminus(\tilde{\tau}_n,\tilde{\nu}_n)} (\rho_{U_{k\infty}(\mathbf{x})}(u)\rho_{V_{l\infty}(\mathbf{x})}(v) + \rho_{U_{k,m-1}(\mathbf{x})}(u)\rho_{V_{ln}(\mathbf{x})}(v))\psi_{a,b}(u)\psi_{\tilde{a},\tilde{b}}(v)\,\mathrm{d}u\,\mathrm{d}v. \tag{C.6}$$

1) *We have*

$$I_{in,1} = O\Big(\frac{\tau_m^{(a+\frac{\sigma_p}{d}+1)\wedge 0}}{m^{\frac{\sigma_p}{d}}} + \frac{k^{-k}}{m}\Big).$$

2) *Suppose that $\phi_k(u)$ is differentiable at every $u > 0$ and $|\phi_k'(u)| \lesssim \psi_{a-1,b-1}(u)$. If $\nu_m = o(\sqrt{m})$ as $m \to \infty$, then we have*

$$I_{in,2} = O\Big(k\frac{\nu_m^{(b+\frac{\sigma_p}{d}+1)\vee 0}}{m^{\frac{\sigma_p}{d}}} + \frac{\nu_m^{(b+k+1)\vee 0}}{m}\Big).$$

Proof. To establish the second bound, we invoke Lemma B.7 instead of Lemma B.4; this helps us obtain a tighter bias bound by reducing the exponent of ν_m by at most 1, which comes at the cost of additional factors in k. Let

$$\Delta_{km}(u) := |\mathrm{P}_{U_{km}(\mathbf{x})}(u) - \mathrm{P}_{U_{k\infty}(\mathbf{x})}(u)|.$$

Since we assume that $\phi_k(u)$ is differentiable at any $u > 0$ and $|\phi_k'(u)| \lesssim \psi_{a-1,b-1}(u)$, integration by parts leads to

$$\begin{aligned} I_{\text{in},2}(\mathbf{x}) &= \Big|[\phi_k(u)\Delta_{km}(u)]_1^{\nu_m} + \int_1^{\nu_m} \phi_k'(u)\Delta_{km}(u)\,\mathrm{d}u\Big| \\ &\le |\phi_k(\nu_m)|\cdot\Delta_{km}(\nu_m) + |\phi_k(1)|\cdot\Delta_{km}(1) \\ &\quad + \int_1^{\nu_m} |\phi_k'(u)|\cdot\Delta_{km}(u)\,\mathrm{d}u \\ &= \tilde{O}_{\sigma_p,L,d}\Big(k\frac{\nu_m^{(b+\frac{\sigma_p}{d}+1)\vee 0}}{m^{\frac{\sigma_p}{d}}} + \frac{\nu_m^{(k+b+1)\vee 0}}{m}\Big) \end{aligned}$$

for $\mathbf{x} \in \operatorname{supp}(p)$, establishing the second bound. □

Assuming **(L1$_p'$)** in place of **(L1$_p$)**, we obtain a different generic bound on the upper outer bias $I_{\text{out},2}$ than that of Lemma B.23; see also Remark B.3.

Lemma D.2 (Generic outer bias bound under **(L1$_p'$)** and **(L4$_p$)**)**.** *Suppose that the density p satisfies the conditions **(U$_p$)**, **(L1$_p'$)**, and **(L4$_p$)**, we have*

$$I_{out,2} = O(\nu_m^{b+k-1-\theta}).$$

For any density p satisfying the conditions **(U$_p$)**, **(L1$_p'$)**, **(L4$_p$)**, and **(S$_p'$)**, if $\nu_m = o(\sqrt{m})$ and k is fixed, we have the bias bound from Lemmas D.1 and D.2:

$$\begin{aligned} &\big|\mathbb{E}[\hat{T}_f^{(k)}] - T_f(p)\big| \\ &\lesssim_{\sigma_p,L,C_p,C_0,d,k} \frac{\tau_m^{(a+\frac{\sigma_p}{d}+1)\wedge 0}}{m^{\frac{\sigma_p}{d}}} + \frac{\nu_m^{(b+\frac{\sigma_p}{d}+1)\vee 0}}{m^{\frac{\sigma_p}{d}}} \\ &\qquad + \frac{\nu_m^{(b+k+1)\vee 0}}{m} + \tau_m^{k+a} + \nu_m^{b+k-1-\theta}. \end{aligned}$$

Since $\nu_m \to \infty$ as $m \to \infty$, we require $b + k - 1 - \theta < 0$ to guarantee that the bias vanishes in our analysis, which forces us to choose a fixed k.

We first choose τ_m. If $a + \frac{\sigma_p}{d} + 1 > 0$, we can take $\tau_m = O(m^{-\frac{\sigma_p}{d}\frac{1}{k+a}})$. Otherwise, we take $\tau_m = \Theta(m^{-\frac{\sigma_p}{d}\frac{1}{k-1-\frac{\sigma_p}{d}}})$ to make the first and the fourth terms decay with the same speed. To summarize, we choose

$$\tau_m = \begin{cases} \Theta(m^{-\frac{\sigma_p}{d}\frac{1}{k-\frac{\sigma_p}{d}-1}}) & \text{if } a \le -\frac{\sigma_p}{d} - 1, \\ O(m^{-\frac{\sigma_p}{d}\frac{1}{k+a}}) & \text{o.w.} \end{cases} \tag{D.1}$$

to bound the first and the fourth terms as

$$\begin{aligned} &\frac{\tau_m^{(a+\frac{\sigma_p}{d}+1)\wedge 0}}{m^{\frac{\sigma_p}{d}}} + \tau_m^{k+a} \\ &= \begin{cases} O(m^{-\frac{\sigma_p}{d}\frac{k+a}{k-\frac{\sigma_p}{d}-1}}) & \text{if } a \le -\frac{\sigma_p}{d} - 1, \\ O(m^{-\frac{\sigma_p}{d}}) & \text{o.w.} \end{cases} \end{aligned}$$

Similarly, by choosing ν_m as defined in (D.2) with $\nu_m = o(\sqrt{m})$ as $m \to \infty$, we bound the second, third, and last terms as

$$\frac{1}{m^{\frac{\sigma_p}{d}}} + \frac{\nu_m^{(b+k+2)\vee 0}}{m} + \nu_m^{b+k-\theta-1} = O(m^{-\lambda_\nu}),$$

where λ_ν is as defined in (III.17). □

B. *Proof of Proposition V.1*

For any density p satisfying the conditions **(U$_p$)**, **(L1$_p$)**, **(S$_p$)**, and **(B$_p$)**, if $\nu_m = o(\sqrt{m})$ and $k \to \infty$ with $k = o(\sqrt{m})$ as $m \to \infty$, we have the bias bound from Lemma B.21:

$$\begin{aligned} &\big|\mathbb{E}[\hat{T}_f^{(k)}] - T_f(p)\big| \\ &\lesssim_{\sigma_p,L,C_p,C_0,d} \frac{\tau_m^{(a+\frac{\sigma_p}{d}+1)\wedge 0}}{m^{\frac{\sigma_p}{d}}} + \frac{k^{-k}}{m} + \frac{\tau_m^{(a+1)\wedge 0}}{m^{\frac{1}{d}}} \\ &\qquad + \frac{\nu_m^{(b+\frac{\sigma_p}{d}+2)\vee 0}}{m^{\frac{\sigma_p}{d}}} + k^{-k}\frac{\nu_m^{(b+k+2)\vee 0}}{m} + \frac{\nu_m^{(b+2)\vee 0+\frac{1}{d}}}{m^{\frac{1}{d}}} \\ &\qquad + k^{-k}\tau_m^{k+a} + k^{(b\vee 0)}\nu_m^{b+k-1}e^{-c_p\nu_m}. \end{aligned}$$

Setting $\nu_m = \Theta((\ln m)^{1+\delta})$ and $k = \Theta((\ln m)^{1+\delta'})$ for some $0 < \delta' < \delta$, the last term $k^{(b\vee 0)}\nu_m^{b+k-1}e^{-c_p\nu_m}$ decays faster than any polynomial rate, that is, for any $C > 0$,

$$(b \vee 0)\ln k + (b + k - 1)\ln\nu_m - c_p\nu_m < -C\ln m$$

for m sufficiently large. With these choices of ν_m and k, the bias bound then can be simplified as

$$\begin{aligned} &\big|\mathbb{E}[\hat{T}_f^{(k)}] - T_f(p)\big| \\ &= \tilde{O}_{\sigma_p,L,C_p,C_0,d}\Big(\frac{\tau_m^{(a+\frac{\sigma_p}{d}+1)\wedge 0}}{m^{\frac{\sigma_p}{d}}} + \frac{\tau_m^{(a+1)\wedge 0}}{m^{\frac{1}{d}}} + \frac{1}{m^{\frac{\sigma_p\wedge 1}{d}}}\Big). \end{aligned}$$

By choosing

$$\begin{aligned} \tau_m &= \tau'(m, a_k) \\ &= \begin{cases} O((\operatorname{poly}\ln m)^{-1}) & \text{if } a_k \le -1 \\ 0 & \text{if } a_k > -1, \end{cases} \end{aligned} \tag{D.3}$$

$$\nu_m = \begin{cases} \Theta(m^{(\frac{\sigma_p}{d}\wedge 1)\frac{1}{\theta-k-b+1}}) & \text{if } k \le -b-1, b \le -\frac{\sigma_p}{d}-1, \\ \Theta(m^{\frac{\sigma_p}{d}\frac{1}{\theta-k+\frac{\sigma_p}{d}+2}}) & \text{if } k \le -b-1, b > -\frac{\sigma_p}{d}-1, \\ \Theta(m^{\frac{1}{\theta+2}}) & \text{if } k > -b-1, b \le -\frac{\sigma_p}{d}-1, \\ \Theta(m^{(\frac{\sigma_p}{d}\wedge 1)\frac{1}{\theta+2}}) & \text{if } k > -b-1, b > -\frac{\sigma_p}{d}-1 \end{cases} \tag{D.2}$$

we obtain

$$\left|\mathbb{E}[\hat{T}_f^{(k)}] - T_f(p)\right| = \tilde{O}_{\sigma_p,L,C_p,C_0,d}\left(m^{-\frac{\sigma_p\wedge 1}{d}}\right).$$

Now, we show that $\mathrm{Var}(\hat{T}_f^{(k)}) = \tilde{O}(m^{-1})$ if $k = \Theta((\ln m)^{1+\delta})$ as $m \to \infty$ for some $\delta > 0$. Using Lemmas B.13, B.11, B.19, and B.16, if we choose ν_m and κ_m such that $\nu_m/m \to 0$ and $\kappa_m/m \to \infty$ as $m \to \infty$, we have

$$\begin{aligned} &\mathrm{Var}(\hat{T}_f^{(k)}) \\ &= O\Bigl(\frac{k^2}{m}\Bigl\{\frac{C_p^k}{k!} + \nu_m^{2b\vee 0} \\ &\qquad + (\nu_m^{2b} \vee \kappa_m^{2b})e^{-\nu_m\eta_p c_p}\Bigl(\frac{eC_p\nu_m}{k}\Bigr)^k\Bigr\}\Bigr) \end{aligned}$$

for m sufficiently large. Letting $\nu_m = (2b/(\eta_p c_p))(\ln m)^{1+\delta/2}$ and $\kappa_m = e^{(\ln m)^{1+\delta/4}}$ ensures that the bound is $\tilde{O}(m^{-1})$. □

Appendix E
Derivation of Estimator Functions

In this section, we present derivations of some selected examples of estimator functions $\phi_{kl}(u,v)$ for some functions $f(p,q)$ in Table II. Estimator functions $\phi_k(u)$ for the single-density case can be computed in a similar manner. In particular, we present the examples of KL divergence (Example E.1), logarithmic α-divergences (Example E.3), entropy difference (Example E.5), reverse KL divergence (Example E.6), polynomial functionals (Example E.2), Le Cam distance (Example E.4), and Jensen–Shannon divergence (Example E.7).

We remark that as alluded to in the main text, the estimator function $\phi_{kl}(u,v)$ is a function of u/v if $f(p,q)$ is a function of q/p.

Proposition E.1. *If $f(p,q)$ is a function of q/p, then there exists a function $\varphi_{kl}\colon \mathbb{R}_+ \to \mathbb{R}$ such that $\phi_{kl}(u,v) = \varphi_{kl}(u/v)$.*

Proof. Suppose that we can write $f(p,q) = g(q/p)$ for some function $g\colon \mathbb{R}_+ \to \mathbb{R}$. Recall that we have

$$\begin{aligned} &\mathcal{L}\{u^{k-1}v^{l-1}\phi_{kl}(u,v)\}(p,q) \\ &= \iint_{\mathbb{R}_+^2} u^{k-1}v^{l-1}e^{-pu}e^{-qv}\phi_{kl}(u,v)\,\mathrm{d}u\,\mathrm{d}v \\ &= \frac{\Gamma(k)\Gamma(l)}{p^k q^l}g\Bigl(\frac{q}{p}\Bigr). \end{aligned}$$

Now, for any $c > 0$, we consider

$$\begin{aligned} &\mathcal{L}\{u^{k-1}v^{l-1}\phi_{kl}(cu,cv)\}(p,q) \\ &= \iint_{\mathbb{R}_+^2} u^{k-1}v^{l-1}e^{-pu}e^{-qv}\phi_{kl}(cu,cv)\,\mathrm{d}u\,\mathrm{d}v \\ &= \frac{1}{c^{k+l}}\iint_{\mathbb{R}_+^2} \tilde{u}^{k-1}\tilde{v}^{l-1}e^{-p\tilde{u}/c}e^{-q\tilde{v}/c}\phi_{kl}(\tilde{u},\tilde{v})\,\mathrm{d}\tilde{u}\,\mathrm{d}\tilde{v} \\ &= \frac{1}{c^{k+l}}\cdot\frac{\Gamma(k)\Gamma(l)}{(p/c)^k(q/c)^l}g\Bigl(\frac{q/c}{p/c}\Bigr) \\ &= \frac{\Gamma(k)\Gamma(l)}{p^k q^l}g\Bigl(\frac{q}{p}\Bigr). \end{aligned}$$

Thus, by the (a.e.) uniqueness of Laplace transform, we have $\phi_{kl}(cu,cv) = \phi_{kl}(u,v)$, whence $\phi_{kl}(u,v)$ can be written as $\phi_{kl}(u,v) = \varphi_{kl}(u/v)$ for some function $\varphi\colon \mathbb{R}_+ \to \mathbb{R}$. □

In what follows, for the one-dimensional inverse Laplace transform of two-variable functions, we will specify the transformed variable by a subscript of the inverse Laplace operator. For example, $\mathcal{L}_p^{-1}\{G(p,q)\}(u)$ denotes the inverse Laplace transform of $G(p,q)$ along the p-axis with a corresponding time-domain variable u.

Example E.1 (KL divergence; Example IV.1). For $f(p,q) = \ln(p/q)$, the corresponding functional $T_f(p,q) = D(p \,\|\, q)$ is the KL divergence. This is one of the simplest cases, as we only need to deal with one-dimensional inverse Laplace transforms by linearity:

$$\begin{aligned} &\mathcal{L}^{-1}\Bigl\{\frac{1}{p^k q^l}\ln\frac{p}{q}\Bigr\} \\ &= \mathcal{L}^{-1}\Bigl\{\frac{\ln p}{p^k}\Bigr\}\mathcal{L}^{-1}\Bigl\{\frac{1}{q^l}\Bigr\} - \mathcal{L}^{-1}\Bigl\{\frac{1}{p^k}\Bigr\}\mathcal{L}^{-1}\Bigl\{\frac{\ln q}{q^l}\Bigr\}. \end{aligned}$$

Note that for any $\kappa > 0$,

$$\mathcal{L}^{-1}\Bigl\{\frac{\ln p}{p^\kappa}\Bigr\} = \frac{u^{\kappa-1}}{\Gamma(\kappa)}\bigl(\Psi(\kappa) - \ln u\bigr). \tag{E.1}$$

This can be verified by taking Laplace transform of the right-hand expression. From the definition of the estimator function $\phi_{kl}(u,v)$ in (I.9), we obtain

$$\phi_{kl}(u,v) = \ln\frac{v}{u} + \Psi(k) - \Psi(l). \tag{E.2}$$

Example E.2 (Polynomial functionals; Example IV.2). Consider $f(p,q) = p^{\alpha-1}q^\beta$ for some $\alpha,\beta \in \mathbb{R}$, which corresponds to the functional

$$T_f(p,q) = \mathbb{E}\bigl[p^{\alpha-1}(\mathbf{X})q^\beta(\mathbf{X})\bigr] = \int p^\alpha(\mathbf{x})q^\beta(\mathbf{x})\,\mathrm{d}\mathbf{x}.$$

This includes many special cases such as Rényi entropies, Rényi divergences, Hellinger distance, and χ^2-divergence. The estimator function is

$$\phi_{kl}(u,v) = \frac{\Gamma(k)\Gamma(l)}{\Gamma(k-\alpha+1)\Gamma(l-\beta)}u^{1-\alpha}v^{-\beta}$$

for $k > \alpha - 1$ and $l > \beta$. We remark that our estimator recovers the bias-corrected estimator presented in [51].

Example E.3 (Logarithmic α-divergence; Example IV.3). For $\alpha \in \mathbb{R}$, consider a function $f(p,q) = (p/q)^{\alpha-1} \ln \frac{p}{q}$, which corresponds to the functional

$$\begin{aligned} T_f(p,q) &= \mathbb{E}\Big[\Big(\frac{p(\mathbf{X})}{q(\mathbf{X})}\Big)^{\alpha-1} \ln \frac{p(\mathbf{X})}{q(\mathbf{X})}\Big] \\ &= \int p^{\alpha}(\mathbf{x}) q^{1-\alpha}(\mathbf{x}) \ln \frac{p(\mathbf{x})}{q(\mathbf{x})} \, \mathrm{d}\mathbf{x}. \end{aligned}$$

Similar to KL divergence, the estimator function can be found immediately from (E.1), i.e.,

$$\begin{aligned} \phi_{kl}(u,v) = {}& \frac{\Gamma(k)\Gamma(l)}{\Gamma(k-\alpha+1)\Gamma(l+\alpha-1)} \Big(\frac{v}{u}\Big)^{\alpha-1} \\ & \times \Big(\ln \frac{v}{u} + \Psi(k-\alpha+1) - \Psi(l+\alpha-1)\Big), \end{aligned}$$

for $k > \alpha - 1$ and $l > -\alpha + 1$. Note that $\alpha = 1$ recovers the estimator function for the KL divergence (E.2).

Example E.4 (Le Cam distance; Example IV.4). For $f(p,q) = 1 - 2q/(p+q)$, we wish to compute the estimator function $\phi_{kl}(u,v)$, that is,

$$\phi_{kl}(u,v) = 2\frac{\Gamma(k)\Gamma(l)}{u^{k-1}v^{l-1}} \mathcal{L}^{-1}\Big\{\frac{1}{p^k q^l} \frac{1}{1+\frac{q}{p}}\Big\} - 1.$$

The two-dimensional inverse Laplace transform can be peeled off dimension by dimension as follows:

$$\begin{aligned} & \mathcal{L}_{p,q}^{-1}\Big\{\frac{1}{p^k q^l} \frac{1}{1+\frac{q}{p}}\Big\}(u,v) \\ &= \mathcal{L}_p^{-1}\Big\{\frac{1}{p^{k+l}} \mathcal{L}_q^{-1}\Big\{\frac{1}{(\frac{q}{p})^l(1+\frac{q}{p})}\Big\}(v)\Big\}(u). \qquad \text{(E.3)} \end{aligned}$$

Letting $\tilde{q} = q/p$, we first find the inverse Laplace transform of

$$\frac{1}{\tilde{q}^l(1+\tilde{q})} = (-1)^l \Big(\sum_{i=1}^{l} \frac{(-1)^i}{\tilde{q}^i} + \frac{1}{1+\tilde{q}}\Big), \qquad \text{(E.4)}$$

which is

$$\mathcal{L}_{\tilde{q}}^{-1}\Big\{\frac{1}{\tilde{q}^l(1+\tilde{q})}\Big\}(v) = (-1)^l \Big(e^{-v} - \sum_{i=0}^{l-1} \frac{(-v)^i}{i!}\Big),$$

since we have

$$\mathcal{L}_p^{-1}\Big\{\frac{1}{p^{n+1}}\Big\}(u) = \frac{u^n}{n!} \mathbb{1}_{[0,\infty)}(u)$$

for $n \in \mathbb{N} \cup \{0\}$ and

$$\mathcal{L}_p^{-1}\Big\{\frac{1}{s+a}\Big\}(u) = e^{-au} \mathbb{1}_{[0,\infty)}(u).$$

Moreover, by the time-scaling property, we have

$$\begin{aligned} & \mathcal{L}_q^{-1}\Big\{\frac{1}{(\frac{q}{p})^l(1+\frac{q}{p})}\Big\}(v) \\ &= (-1)^l \Big(p e^{-pv} - \sum_{i=0}^{l-1} \frac{(-v)^i}{i!} p^{i+1}\Big). \end{aligned}$$

Now, continuing from (E.3), we have (E.5), which leads to the estimator function (E.6). As a bound on the estimator function $\phi_{kl}(u,v)$, we observe that

$$\begin{aligned} |\phi_{kl}(u,v)| &\lesssim \Big(\frac{u}{v}\Big)^{l-1} \Big(\sum_{i=0}^{l-1} \Big(\frac{v}{u}\Big)^i + \sum_{j=0}^{k+l-2} \Big(\frac{v}{u}\Big)^j\Big) \\ &\lesssim \psi_{-k+1,l-1}(u) \psi_{-l+1,k-1}(v). \end{aligned}$$

For the remaining examples, we assume that $\mathcal{Q} \ll \mathcal{P}$.

Example E.5 (Entropy difference). For $f(p,q) = \ln(1/p) - (q/p)\ln(1/q)$, the corresponding functional $T_f(p,q) = h(p) - h(q)$ becomes the difference of the differential entropies $h(p)$ and $h(q)$. It is easy to show that

$$\phi_{kl}(u,v) = \frac{(l-1)}{k} \frac{u}{v} (\Psi(l-1) - \ln v) - (\Psi(k) - \ln u).$$

As a bound on the estimator function $\phi_{kl}(u,v)$, we have

$$\begin{aligned} |\phi_{kl}(u,v)| &\lesssim \frac{u}{v}(1 + |\ln v|) + (1 + |\ln u|) \\ &\lesssim \psi_{1,1}(u) \psi_{-1-\epsilon,-1+\epsilon}(v) + \psi_{-\epsilon,\epsilon}(u) \\ &\lesssim \psi_{-\epsilon,1}(u) \psi_{-1-\epsilon,-1+\epsilon}(v). \end{aligned}$$

Example E.6 (Reverse KL divergence). When $\mathcal{Q} \ll \mathcal{P}$, we can write the reverse KL divergence as

$$\begin{aligned} D(q \,\|\, p) &= \int q(\mathbf{x}) \ln \frac{q(\mathbf{x})}{p(\mathbf{x})} \, \mathrm{d}\mathbf{x} \\ &= \int p(\mathbf{x}) \frac{q(\mathbf{x})}{p(\mathbf{x})} \ln \frac{q(\mathbf{x})}{p(\mathbf{x})} \, \mathrm{d}\mathbf{x} = T_f(p,q) \end{aligned}$$

for $f(p,q) = (q/p)\ln(q/p)$. Then, for $k \ge 1$ and $l \ge 2$, we have

$$\begin{aligned} \mathcal{L}^{-1}\Big\{\frac{f(p,q)}{p^k q^l}\Big\} &= \mathcal{L}^{-1}\Big\{\frac{1}{p^{k+1}}\Big\} \mathcal{L}_q^{-1}\Big\{\frac{\ln q}{q^{l-1}}\Big\} \\ &\quad - \mathcal{L}^{-1}\Big\{\frac{\ln p}{p^{k+1}}\Big\} \mathcal{L}_q^{-1}\Big\{\frac{1}{q^{l-1}}\Big\} \\ &= \frac{u^k}{\Gamma(k+1)} \frac{v^{l-2}}{\Gamma(l-1)} \big(\Psi(l-1) - \ln v\big) \\ &\quad - \frac{u^k}{\Gamma(k+1)} \big(\Psi(k+1) - \ln u\big) \frac{v^{l-2}}{\Gamma(l-1)}. \end{aligned}$$

Here, the case $l = 1$ is excluded, since $\mathcal{L}^{-1}\{\ln s\}$ is ill-defined. Finally, we have

$$\begin{aligned} \phi_{kl}(u,v) &= \frac{\Gamma(k)\Gamma(l)}{u^{k-1}v^{l-1}} \frac{u^k}{\Gamma(k+1)} \frac{v^{l-2}}{\Gamma(l-1)} \\ &\quad \times \big\{\big(\Psi(l-1) - \ln v\big) - \big(\Psi(k+1) - \ln u\big)\big\} \\ &= \frac{l-1}{k} \frac{u}{v} \Big(\ln \frac{u}{v} + \Psi(l-1) - \Psi(k+1)\Big). \end{aligned}$$

As a bound on the estimator function $\phi_{kl}(u,v)$, we have

$$\begin{aligned} |\phi_{kl}(u,v)| &\lesssim \frac{u}{v}(1 + |\ln u| + |\ln v|) \\ &\lesssim \frac{u}{v}(1 + |\ln u|)(1 + |\ln v|) \\ &\lesssim \psi_{1-\epsilon,1+\epsilon}(u) \psi_{-1-\epsilon,-1+\epsilon}(v). \end{aligned}$$

$$\begin{aligned}\mathcal{L}_{p,q}^{-1}\Big\{\frac{1}{p^kq^l}\frac{1}{1+\frac{q}{p}}\Big\}(u,v) &= \mathcal{L}_{p}^{-1}\Big\{(-1)^l\Big(\frac{e^{-pv}}{p^{k+l-1}}-\sum_{i=0}^{l-1}\frac{(-v)^i}{i!}\frac{1}{p^{k+l-i-1}}\Big)\Big\}(u)\\ &= (-1)^l\Big(\frac{(u-v)^{k+l-2}}{(k+l-2)!}\mathbb{1}_{[v,\infty)}(u)-\sum_{i=0}^{l-1}\frac{(-v)^i}{i!}\frac{u^{k+l-i-2}}{(k+l-i-2)!}\Big)\\ &= (-1)^l\frac{u^{k+l-2}}{(k+l-2)!}\Big(\Big(1-\frac{v}{u}\Big)^{k+l-2}\mathbb{1}_{[v,\infty)}(u)-\sum_{i=0}^{l-1}\binom{k+l-2}{i}\Big(\frac{-v}{u}\Big)^i\Big). \qquad \text{(E.5)}\end{aligned}$$

$$\phi_{kl}(u,v)=2\binom{k+l-2}{k-1}^{-1}\Big(-\frac{u}{v}\Big)^{l-1}\Big(\sum_{i=0}^{l-1}\binom{k+l-2}{i}\Big(-\frac{v}{u}\Big)^i-\Big(1-\frac{v}{u}\Big)^{k+l-2}\mathbb{1}_{[v,\infty)}(u)\Big)-1. \qquad \text{(E.6)}$$

TABLE III
INVERSE LAPLACE TRANSFORMS OF FEW ELEMENTARY FUNCTIONS AND BASIC OPERATIONS.

Frequency domain $F(p)=\mathcal{L}\{f(u)\}$	Time domain $f(u)=\mathcal{L}^{-1}\{F(p)\}$
p^{-k} $(k>0)$	$u^{k-1}/\Gamma(k)$
$\ln p/p$	$-(\ln u+\gamma)$
$1/(p+\alpha)$	$e^{-\alpha u}$
$F(ap)$	$f(u/a)/a$
$e^{-ap}F(p)$	$f(u-a)\mathbb{1}_{[a,\infty)}(u)$
$F^{(n)}(p)$	$(-1)^n u^n f(u)$
$F(p)/p$	$\int_0^u f(t)\,\mathrm{d}t$
$F(p)G(p)$	$(f*g)(u)=\int_0^u f(\tilde{u})g(u-\tilde{u})\,\mathrm{d}\tilde{u}$
$pF(p)$	$f'(u)-f(0)$

Example E.7 (Jensen–Shannon divergence; Example IV.5). We wish to compute the estimator function $\phi_{kl}(u,v)$ for

$$f(p,q)=\frac{1}{2}\Big(\frac{q}{p}+1\Big)\ln\frac{2}{(q/p)+1}+\frac{q}{2p}\ln\frac{q}{p}.$$

For $l\geq 2$, we have

$$\begin{aligned}\frac{2f(p,q)}{p^kq^l} &= \Big(\frac{1}{p^{k+1}q^{l-1}}+\frac{1}{p^kq^l}\Big)\ln 2+\frac{1}{p^{k+1}q^{l-1}}\ln\frac{q}{p}\\ &\quad+\frac{G_{l-1}(\frac{q}{p})+G_l(\frac{q}{p})}{p^{k+l}},\end{aligned}$$

where we define $G_l(q):=-\ln(q+1)/q^l$. Using the identity (E.4), we can show that for $l\in\mathbb{N}$

$$\begin{aligned}g_l(v) &= \mathcal{L}_q^{-1}\{G_l(q)\}(v)\\ &= (-1)^l\Big(\int_1^\infty\frac{e^{-vx}}{x^l}\,\mathrm{d}x-\sum_{j=0}^{l-2}\frac{(-v)^j}{(l-1-j)j!}\Big).\end{aligned}$$

Now the desired estimator function can be written as

$$\begin{aligned}2\phi_{kl}(u,v) &= \frac{\Gamma(k)\Gamma(l)}{u^{k-1}v^{l-1}}\mathcal{L}^{-1}\Big\{\frac{f(p,q)}{p^kq^l}\Big\}(u,v)\\ &= \frac{l-1}{k}\frac{u}{v}\Big(\Psi(l-1)-\Psi(k+1)+\ln\frac{u}{v}\Big)\\ &\quad+\Big(\frac{l-1}{k}\frac{u}{v}+1\Big)\ln 2+A_{kl}(u,v), \qquad \text{(E.7)}\end{aligned}$$

where we define

$$\begin{aligned}&A_{kl}(u,v)\\ &= \frac{\Gamma(k)\Gamma(l)}{u^{k-1}v^{l-1}}\mathcal{L}_p^{-1}\Big\{\frac{\mathcal{L}_q^{-1}\{G_{l-1}(\frac{q}{p})+G_l(\frac{q}{p})\}(v)}{p^{k+l}}\Big\}(u)\\ &\overset{(a)}{=} \frac{\Gamma(k)\Gamma(l)}{u^{k-1}v^{l-1}}\mathcal{L}_p^{-1}\Big\{\frac{g_{l-1}(pv)+g_l(pv)}{p^{k+l-1}}\Big\}(u)\\ &= B_{kl}(u,v)+\frac{l-1}{k}\frac{u}{v}B_{k+1,l-1}(u,v), \qquad \text{(E.8)}\end{aligned}$$

where

$$B_{kl}(u,v)=\frac{\Gamma(k)\Gamma(l)}{u^{k-1}v^{l-1}}\mathcal{L}_p^{-1}\Big\{\frac{g_l(pv)}{p^{k+l-1}}\Big\}(u).$$

Here, (a) follows by the time scaling property, that is, $\mathcal{L}_q^{-1}\{G_l(q/p)\}(v)=pg_l(pv)$. Now, since we have (E.9), it follows that

$$\begin{aligned}&\binom{k+l-2}{k-1}B_{kl}(u,v)\\ &= -\mathbb{1}_{[1,\infty)}(w)(-w)^{-k+1}\int_1^w\frac{(x-w)^{k+l-2}}{x^l}\,\mathrm{d}x\\ &\quad+\sum_{j=0}^{l-2}\binom{k+l-2}{j}\frac{(-w)^{l-1-j}}{l-1-j},\end{aligned}$$

where $w:=u/v$.

Rearranging the integral in the parenthesis as

$$\begin{aligned}&(-w)^{k+1}\int_1^w\frac{(x-w)^{k+l-2}}{x^l}\,\mathrm{d}x\\ &= \sum_{\substack{i=0\\ i\neq k-1}}^{k+l-2}\binom{k+l-2}{i}\frac{(-1)^{k-1-i}-(-w^{-1})^{k-1-i}}{k-1-i}\\ &\quad+\binom{k+l-2}{k-1}\ln w,\end{aligned}$$

we finally obtain

$$\begin{aligned}&B_{kl}(u,v) \qquad \text{(E.10)}\\ &= \binom{k+l-2}{k-1}^{-1}\sum_{j=0}^{l-2}\binom{k+l-2}{j}\frac{(-u/v)^{l-1-j}}{l-1-j}\end{aligned}$$

$$\begin{aligned}\mathcal{L}_p^{-1}\Big\{\frac{g_l(pv)}{p^{k+l-1}}\Big\} &= \int_1^\infty \frac{1}{x^l}\mathcal{L}_p^{-1}\Big\{\frac{e^{-pvx}}{p^{k+l-1}}\Big\}\,\mathrm{d}x - \sum_{j=0}^{l-2}\frac{(-v)^j}{(l-1-j)j!}\mathcal{L}_p^{-1}\Big\{\frac{1}{p^{k+l-1-j}}\Big\}\\ &= \int_1^\infty \frac{1}{x^l}\mathbb{1}_{[vx,\infty)}(u)\frac{(u-vx)^{k+l-2}}{(k+l-2)!}\,\mathrm{d}x - \sum_{j=0}^{l-2}\frac{(-v)^j}{(l-1-j)j!}\frac{u^{k+l-2-j}}{(k+l-2-j)!},\end{aligned} \tag{E.9}$$

if $\frac{u}{v} < 1$, and

$$\begin{aligned}B_{kl}(u,v) &= -\ln\frac{u}{v} + \binom{k+l-2}{k-1}^{-1}\\ &\times\Bigg\{\sum_{i=0}^{k-2}\binom{k+l-2}{i}\frac{(-v/u)^{k-1-i}}{k-1-i}\\ &\quad-\sum_{\substack{i=0\\ i\neq k-1}}^{k+l-2}\binom{k+l-2}{i}\frac{(-1)^{k-1-i}}{k-1-i}\Bigg\}\end{aligned} \tag{E.11}$$

if $\frac{u}{v} \geq 1$. Substituting the expressions for $B_{kl}(u,v)$ from (E.10) and (E.11) into (E.8) and then into (E.7) yields the final expression for the estimator function as

$$\begin{aligned}&\phi_{kl}(u,v)\\ &= \frac{1}{2}\Big\{\ln 2 + \frac{l-1}{k}\frac{u}{v}\Big(\ln 2 + \Psi(l-1) - \Psi(k+1) + \ln\frac{u}{v}\Big)\\ &\quad + B_{kl}(u,v) + \frac{l-1}{k}\frac{u}{v}B_{k+1,l-1}(u,v)\Big\}.\end{aligned}$$

As a bound on the estimator function $\phi_{kl}(u,v)$, we have

$$|\phi_{kl}(u,v)| \lesssim \psi_{-k+1,l-1}(u)\psi_{-l+1,k-1}(v).$$

Appendix F
Examples of Smooth Densities

In this section, we show that the d-dimensional truncated Gaussian, Cauchy, and exponential distributions, as well as the uniform distribution and the d-dimensional product of identical beta distributions with parameters $\alpha \geq 3$ and $\beta \geq 3$ satisfy the conditions **(U$_p$)**, **(L1$_p$)**, **(S$_p$)**, and **(B$_p$)** with $\sigma_p = 2$, and the d-dimensional truncated Laplace distribution satisfies the conditions with $\sigma_p = 1$. We remark that the boundedness of the Hessian of the density p over a compact set implies 2-Hölder continuity, if the Hessian is integrable. Since we have considered that the Hessian is integrable, we only need to prove the boundedness of the Hessian in order to demonstrate the 2-Hölder continuity.

Example F.1 (Truncated Gaussian). Consider the *truncated d-dimensional Gaussian distribution* defined by the density

$$p(\mathbf{x}) := \frac{\Gamma(d/2+1)}{\pi^{d/2}K_d(R)}e^{-\|\mathbf{x}\|_2^2/2}\mathbb{1}_{(-\infty,R]}(\|\mathbf{x}\|_2),$$

where $K_d(R) := \int_0^R dr^{d-1}e^{-r^2/2}\,\mathrm{d}r$. Then, $\mathrm{supp}(p) = \{\mathbf{x} \in \mathbb{R}^d : \|\mathbf{x}\| \leq R\}$ and

$$\frac{\Gamma(d/2+1)}{\pi^{d/2}K_d(R)}e^{-R^2/2} \leq p(\mathbf{x}) \leq \frac{\Gamma(d/2+1)}{\pi^{d/2}K_d(R)}$$

for $\mathbf{x} \in \mathrm{supp}(p)$. Moreover, on $\mathrm{supp}(p)^{\circ}$,

$$\nabla^2 p(\mathbf{x})_{ij} = \frac{\Gamma(d/2+1)}{\pi^{d/2}K_d(R)}(x_i x_j - \delta_{ij})e^{-\|\mathbf{x}\|_2^2/2},$$

whence,

$$\|\nabla^2 p(\mathbf{x})\| \leq \|\nabla^2 p(\mathbf{x})\|_F \leq \frac{\Gamma(d/2+1)}{\pi^{d/2}K_d(R)}\sqrt{R^4+d}.$$

Finally, $\partial\mathrm{supp}(p) = \mathbb{S}(\mathbf{0},R)$ satisfies

$$H^{d-1}(\mathbb{S}(\mathbf{0},R)) = dv_d R^{d-1}.$$

Therefore, this density satisfies the conditions **(U$_p$)**, **(L1$_p$)**, **(S$_p$)**, and **(B$_p$)** with $\sigma_p = 2$ and

$$\begin{aligned}\sup_{\mathbf{x}} p(\mathbf{x}) &= \frac{\Gamma(d/2+1)}{\pi^{d/2}K_d(R)},\\ L(p;\mathrm{supp}(p)^{\circ}) &= \frac{\Gamma(d/2+1)}{\pi^{d/2}K_d(R)}\sqrt{R^4+d},\\ H^{d-1}(\partial\mathrm{supp}(p)) &= dv_d R^{d-1}.\end{aligned}$$

Example F.2 (Truncated exponential). Let $S_R := \{\mathbf{x} \in \mathbb{R}^d \colon x_1,\ldots,x_d \geq 0, x_1+\ldots+x_d \leq R\}$. The *truncated d-dimensional exponential distribution* defined by the density

$$p(\mathbf{x}) := \frac{e^{-(x_1+\cdots+x_d)}}{1-\big(\sum_{i=0}^{d-1}\frac{R^i}{i!}\big)e^{-R}}\mathbb{1}_{S_R}(\mathbf{x})$$

is 2-Hölder continuous over $\mathrm{supp}(p)$ and satisfies

$$\begin{aligned}\sup_{\mathbf{x}} p(\mathbf{x}) &= \Big(1-\Big(\sum_{i=0}^{d-1}\frac{R^i}{i!}\Big)e^{-R}\Big)^{-1},\\ L(p;\mathrm{supp}(p)^{\circ}) &= d\sup_{\mathbf{x}} p(\mathbf{x}),\end{aligned}$$

and

$$H^{d-1}(\partial\mathrm{supp}(p)) = \Big(\frac{\sqrt{d}}{(d-1)!}+d\Big)R^{d-1},$$

as can be seen by an analysis similar to that in the previous example.

Example F.3 (Truncated Laplace). Consider the *truncated d-dimensional Laplace distribution* defined by the density

$$p(\mathbf{x}) := \frac{e^{-(|x_1|+\cdots+|x_d|)}}{2^d\big(1-\big(\sum_{i=0}^{d-1}\frac{R^i}{i!}\big)e^{-R}\big)}\mathbb{1}_{(-\infty,R]}(\|\mathbf{x}\|_1).$$

Then, **(U$_p$)**, **(L1$_p$)**, and **(B$_p$)** can be demonstrated similarly to the previous examples. For **(S$_p$)**, note that for $x, y \in \mathbb{R}$,

$$\big|e^{-|x|}-e^{-|y|}\big| \leq |x-y|.$$

Generalizing this to d dimensions, we have

$$\begin{aligned}\big|e^{-(|x_1|+\cdots+|x_d|)}-e^{-(|y_1|+\cdots+|y_d|)}\big| &\leq \|\mathbf{x}-\mathbf{y}\|_1\\ &\leq \sqrt{d}\|\mathbf{x}-\mathbf{y}\|_2.\end{aligned}$$

Therefore, the truncated d-dimensional Laplace distribution is 1-Hölder continuous over $\operatorname{supp}(p)$ and satisfies

$$\begin{aligned}\sup_{\mathbf{x}} p(\mathbf{x}) &= \Big(2^d\Big(1-\Big(\sum_{i=0}^{d-1}\frac{R^i}{i!}\Big)e^{-R}\Big)\Big)^{-1},\\ L(p;\operatorname{supp}(p)^{\circ}) &= \sqrt{d}\sup_{\mathbf{x}} p(\mathbf{x}),\end{aligned}$$

and

$$H^{d-1}(\partial\operatorname{supp}(p)) = \frac{2^d\sqrt{d}}{(d-1)!}R^{d-1}.$$

Example F.4 (Truncated Cauchy). Consider the *truncated d-dimensional Cauchy distribution* defined by the density

$$p(\mathbf{x}) := \frac{\Gamma\big((d+1)/2\big)}{\pi^{(d+1)/2}L_d(R)\big(1+\|\mathbf{x}\|_2^2\big)^{(d+1)/2}}\mathbb{1}_{(-\infty,R]}(\|\mathbf{x}\|_2),$$

where

$$L_d(R) := \frac{\int_0^{\arctan R}\sin^{d-1}\theta\,\mathrm{d}\theta}{\int_0^{\pi/2}\sin^{d-1}\theta\,\mathrm{d}\theta} \in [0,1].$$

Then, we have

$$\begin{aligned}\nabla^2 p(\mathbf{x})_{ij} = {}& \frac{(d+1)\Gamma((d+1)/2)}{\pi^{(d+1)/2}L_d(R)\big(1+\|\mathbf{x}\|_2^2\big)^{(d+5)/2}}\\ &\times\big((d+3)x_ix_j - \big(1+\|\mathbf{x}\|_2^2\big)\delta_{ij}\big),\end{aligned}$$

which leads to the bound

$$\|\nabla^2 p(\mathbf{x})\| \le \frac{(d+1)\Gamma((d+1)/2)}{\pi^{(d+1)/2}L_d(R)}\sqrt{R^4(d+1)(d+3)+d}$$

on $\operatorname{supp}(p)^{\circ}$. Therefore, the truncated d-dimensional Cauchy distribution is 2-Hölder continuous over $\operatorname{supp}(p)$ and satisfies

$$\begin{aligned}\sup_{\mathbf{x}} p(\mathbf{x}) &= \frac{\Gamma\big((d+1)/2\big)}{\pi^{(d+1)/2}L_d(R)},\\ L(p;\operatorname{supp}(p)^{\circ}) &= \frac{(d+1)\Gamma((d+1)/2)}{\pi^{(d+1)/2}L_d(R)}\\ &\quad\times\sqrt{R^4(d+1)(d+3)+d},\end{aligned}$$

and

$$H^{d-1}(\partial\operatorname{supp}(p)) = d\upsilon_d R^{d-1}.$$

References

[1] H. L. Weidemann and E. B. Stear, "Entropy analysis of parameter estimation," *Inf. Control*, vol. 14, no. 6, pp. 493–506, 1969.

[2] E. Wolsztynski, E. Thierry, and L. Pronzato, "Minimum-entropy estimation in semi-parametric models," *Signal Process.*, vol. 85, pp. 937–949, 2005.

[3] V. Girardin and J. Lequesne, "Entropy-based goodness-of-fit tests – a unifying framework: Application to DNA replication," *Commun. Statist. Theory Methods*, pp. 1–13, 2017.

[4] P. Crzcgorzewski and R. Wirczorkowski, "Entropy-based goodness-of-fit test for exponentiality," *Commun. Statist. Theory Methods*, vol. 28, no. 5, pp. 1183–1202, 1999.

[5] M. N. Goria, N. N. Leonenko, V. V. Mergel, and P. L. Inverardi, "A new class of random vector entropy estimators and its applications in testing statistical hypotheses," *J. Nonparametr. Statist.*, vol. 17, no. 3, pp. 277–297, 2005.

[6] S. Marano, V. Matta, and P. Willett, "Asymptotic design of quantizers for decentralized MMSE estimation," *IEEE Trans. Signal Process.*, vol. 55, no. 11, pp. 5485–5496, 2007.

[7] A. Kraskov, H. Stögbauer, and P. Grassberger, "Estimating mutual information," *Phys. Rev. E. Statist. Phys. Plasmas Fluids Relat. Interdiscip. Topics*, vol. 69, no. 6, p. 066138, 2004.

[8] E. G. Learned-Miller and J. W. Fisher III, "ICA using spacings estimates of entropy," *J. Mach. Learn. Res.*, vol. 4, no. December, pp. 1271–1295, 2003.

[9] Z. Boukouvalas, R. Mowakeaa, G.-S. Fu, and T. Adali, "Independent Component Analysis by Entropy Maximization with Kernels," *arXiv preprint arXiv:1610.07104*, 2016.

[10] A. O. Hero, B. Ma, O. J. J. Michel, and J. Gorman, "Applications of entropic spanning graphs," *IEEE Signal Process. Mag.*, vol. 19, no. 5, pp. 85–95, 2002.

[11] S. Susan and M. Hanmandlu, "A non-extensive entropy feature and its application to texture classification," *Neurocomputing*, vol. 120, pp. 214–225, 2013.

[12] J. Liepe, S. Filippi, K. Michał, and M. P. H. Stumpf, "Maximizing the information content of experiments in systems biology," *PLoS Comput. Biol.*, vol. 9, no. 1, p. e1002888, 2013.

[13] J. Lewi, R. Butera, and L. Paninski, "Real-time adaptive information-theoretic optimization of neurophysiology experiments," in *Adv. Neural Inf. Proc. Syst.*, vol. 20, 2007, pp. 857–864.

[14] A. O. Hero and O. J. J. Michel, "Asymptotic theory of greedy approximations to minimal k-point random graphs," *IEEE Trans. Inf. Theory*, vol. 45, no. 6, pp. 1921–1938, 1999.

[15] H. Neemuchwala, A. Hero, and P. Carson, "Image matching using alpha-entropy measures and entropic graphs," *Signal Process.*, vol. 85, no. 2, pp. 277–296, 2005.

[16] S. M. Lajevardi and Z. M. Hussain, "Feature extraction for facial expression recognition based on hybrid face regions," *Adv. Electr. Comput. Eng.*, vol. 9, no. 3, pp. 63–67, 2009.

[17] C. Shan, S. Gong, and P. W. McOwan, "Conditional Mutual Information Based Boosting for Facial Expression Recognition," in *Proc. British Mach. Vis. Conf.*, 2005.

[18] M. Aghagolzadeh, H. Soltanian-Zadeh, B. Araabi, and A. Aghagolzadeh, "A hierarchical clustering based on mutual information maximization," in *Proc. IEEE Int. Conf. Image Proc.*, vol. 1, 2007.

[19] H. Peng, F. Long, and C. Ding, "Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 27, no. 8, pp. 1226–1238, 2005.

[20] J. M. Sotoca and F. Pla, "Supervised feature selection by clustering using conditional mutual information-based

distances," *Pattern Recogni.*, vol. 43, no. 6, pp. 2068–2081, 2010.

[21] L. Giet and M. Lubrano, "A minimum Hellinger distance estimator for stochastic differential equations: An application to statistical inference for continuous time interest rate models," *Comput. Statist. Data Anal.*, vol. 52, no. 6, pp. 2945–2965, 2008.

[22] J. Oliva, B. Póczos, and J. Schneider, "Distribution to distribution regression," in *Proc. Int. Conf. Mach. Learn.*, 2013, pp. 1049–1057.

[23] K. Henderson, B. Gallagher, and T. Eliassi-Rad, "EP-MEANS: An efficient nonparametric clustering of empirical probability distributions," in *Proc. Symp. Appl. Comput.* ACM, 2015, pp. 893–900.

[24] G. A. Korn and T. M. Korn, *Mathematical Handbook for Scientists and Engineers: Definitions, Theorems, and Formulas for Reference and Review*. Courier Corporation, 2000.

[25] L. F. Kozachenko and N. N. Leonenko, "Sample estimate of the entropy of a random vector," *Probl. Inf. Transm.*, vol. 23, no. 2, pp. 9–16, 1987, (Russian).

[26] H. Singh, N. Misra, V. Hnizdo, A. Fedorowicz, and E. Demchuk, "Nearest neighbor estimates of entropy," *Am. J. Math. Manag. Sci.*, vol. 23, no. 3-4, pp. 301–321, 2003.

[27] N. Leonenko, L. Pronzato, and V. Savani, "A class of Rényi information estimators for multidimensional densities," *Ann. Statist.*, vol. 36, no. 5, pp. 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.

[28] A. Bulinski and D. Dimitrov, "Statistical estimation of the shannon entropy," *Acta Mathematica Sinica, English Series*, vol. 35, no. 1, pp. 17–46, 2019.

[29] ——, "Statistical estimation of the Kullback–Leibler divergence," *arXiv preprint arXiv:1907.00196*, 2019.

[30] W. Gao, S. Oh, and P. Viswanath, "Demystifying fixed k-nearest neighbor information estimators," *IEEE Trans. Inf. Theory*, vol. 64, no. 8, pp. 5629–5661, August 2018.

[31] Q. Wang, S. R. Kulkarni, and S. Verdú, "Divergence estimation for multidimensional densities via k-nearest-neighbor distances," *IEEE Trans. Inf. Theory*, vol. 55, no. 5, pp. 2392–2405, 2009.

[32] B. Póczos and J. G. Schneider, "On the Estimation of alpha-Divergences," *Int. Conf. Artif. Int. Statist.*, pp. 609–617, 2011.

[33] A. M. Cohen, *Numerical Methods for Laplace Transform Inversion*. Springer Science & Business Media, 2007, vol. 5.

[34] Y.-K. Noh, "Generative metric learning and dimensionality reduction with f-divergences," Ph.D. dissertation, Seoul National University, August 2011. [Online]. Available: http://s-space.snu.ac.kr/handle/10371/159245

[35] T. H. Cormen, C. E. Leiserson, R. L. Rivest, and C. Stein, *Introduction to Algorithms*. MIT Press, 2009.

[36] D. O. Loftsgaarden and C. P. Quesenberry, "A nonparametric estimate of a multivariate density function," *Ann. Math. Statist.*, vol. 36, no. 3, pp. 1049–1051, 1965.

[37] G. Biau and L. Devroye, *Lectures on the Nearest Neighbor Method*. Springer International Publishing, 2015.

[38] K. Sricharan, R. Raich, and A. O. Hero, "Estimation of nonlinear functionals of densities with confidence," *IEEE Trans. Inf. Theory*, vol. 58, no. 7, pp. 4135–4159, 2012.

[39] K. Sricharan, D. Wei, and A. O. Hero, "Ensemble estimators for multivariate entropy estimation," *IEEE Trans. Inf. Theory*, vol. 59, no. 7, pp. 4374–4388, 2013.

[40] K. R. Moon and A. O. Hero, "Ensemble estimation of multivariate f-divergence," in *Proc. IEEE Int. Symp. Inf. Theory*. IEEE, 2014, pp. 356–360.

[41] ——, "Multivariate f-divergence estimation with confidence," in *Adv. Neural Inf. Proc. Syst.*, vol. 27, 2014, pp. 2420–2428.

[42] T. B. Berrett and R. J. Samworth, "Efficient two-sample functional estimation and the super-oracle phenomenon," *arXiv preprint arXiv:1904.09347*, 2019.

[43] R. L. Dobrushin, "A simplified method of experimentally evaluating the entropy of a stationary sequence," *Theory of Probability & Its Applications*, vol. 3, no. 4, pp. 428–430, 1958.

[44] 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.

[45] S. Delattre and N. Fournier, "On the Kozachenko–Leonenko entropy estimator," *J. Statist. Plan. Inference*, vol. 185, pp. 69–93, 2017.

[46] T. B. Berrett, R. J. Samworth, and M. Yuan, "Efficient multivariate entropy estimation via k-nearest neighbour distances," *Ann. Statist.*, vol. 47, no. 1, pp. 288–318, 2019.

[47] Y. Han, J. Jiao, T. Weissman, and Y. Wu, "Optimal rates of entropy estimation over Lipschitz balls," *Ann. Statist.*, vol. 48, no. 6, pp. 3228–3250, 2020.

[48] J. Jiao, W. Gao, and Y. Han, "The Nearest Neighbor Information Estimator is Adaptively Near Minimax Rate-Optimal," in *Adv. Neural Inf. Proc. Syst.*, vol. 31, December 2018.

[49] D. Pál, B. Póczos, and C. Szepesvári, "Estimation of rényi entropy and mutual information based on generalized nearest-neighbor graphs," in *Adv. Neural Inf. Proc. Syst.*, vol. 23, 2010, pp. 1849–1857.

[50] N. N. Leonenko and L. Pronzato, "Correction: A class of Rényi information estimators for multidimensional densities," *Ann. Statist.*, vol. 38, no. 6, pp. 3837–3838, 2010.

[51] B. Póczos, L. Xiong, D. J. Sutherland, and J. Schneider, "Nonparametric kernel estimators for image classification," in *Proc. IEEE Comput. Soc. Conf. Comput. Vis. Pattern Recognit.* IEEE, June 2012, pp. 2989–2996.

[52] S. Singh and B. Póczos, "Finite-sample analysis of fixed-k nearest neighbor density functional estimators," in *Adv. Neural Inf. Proc. Syst.* Curran Associates, Inc., 2016, vol. 29, pp. 1217–1225.

[53] L. Birge and P. Massart, "Estimation of integrals functionals of a density," *Ann. Statist.*, vol. 23, no. 1, pp. 11–29, 1995.

[54] A. Krishnamurthy, K. Kandasamy, B. Póczos, and L. Wasserman, "Nonparametric estimation of Rényi divergence and friends," in *Proc. Int. Conf. Mach. Learn.*, 2014, pp. 919–927.

[55] K. Kandasamy, A. Krishnamurthy, B. Poczos, L. A. Wasserman, and J. M. Robins, "Nonparametric von Mises estimators for entropies, divergences and mutual informations." in *Adv. Neural Inf. Proc. Syst.*, vol. 28, 2015, pp. 397–405.

[56] H. Liu, J. Lafferty, and L. Wasserman, "Exponential concentration inequality for mutual information estimation," in *Adv. Neural Inf. Proc. Syst.*, vol. 25, 2012.

[57] S. Singh and B. Póczos, "Generalized exponential concentration inequality for rényi divergence estimation," in *Proc. Int. Conf. Mach. Learn.* PMLR, 2014, pp. 333–341.

[58] ——, "Exponential concentration of a density functional estimator," in *Adv. Neural Inf. Proc. Syst.*, vol. 27, 2014, pp. 3032–3040.

[59] K. R. Moon, K. Sricharan, and A. O. Hero, "Ensemble estimation of mutual information," in *Proc. IEEE Int. Symp. Inf. Theory*. IEEE, June 2017, pp. 3030–3034.

[60] K. R. Moon, K. Sricharan, K. Greenewald, and A. O. Hero, "Ensemble estimation of information divergence," *Entropy*, vol. 20, no. 8, p. 560, 2018.

[61] M. Noshad, K. R. Moon, S. Y. Sekeh, and A. O. Hero, "Direct estimation of information divergence using nearest neighbor ratios," in *Proc. IEEE Int. Symp. Inf. Theory*. IEEE, 2017, pp. 903–907.

[62] A. Wisler, K. Moon, and V. Berisha, "Direct ensemble estimation of density functionals," in *Int. Conf. Acoust. Speech Signal Process.* IEEE, 2018, pp. 2866–2870.

[63] X. Nguyen, M. J. Wainwright, and M. I. Jordan, "Estimating divergence functionals and the likelihood ratio by convex risk minimization," *IEEE Trans. Inf. Theory*, vol. 56, no. 11, pp. 5847–5861, 2010.

[64] A. Rényi, "On measures of entropy and information," in *Proc. 4th Berkeley Sympos. Math. Statist. Probab.*, vol. 1. Univ. California Press, Berkeley, 1961, pp. 547–761.

[65] J. Harvda and F. Charvat, "Quantification method of classification processes. concept of structural α-entropy," *Kybernetika (Prague)*, vol. 3, pp. 30–35, 1967.

[66] C. Tsallis, "Possible generalization of boltzmann-gibbs statistics," *J. of Statist. Phys.*, vol. 52, no. 1-2, pp. 479–487, 1988.

[67] V. S. Borkar, *Probability theory: an advanced course*. Springer Science & Business Media, 1995.

[68] J.-Y. Audibert, A. B. Tsybakov *et al.*, "Fast learning rates for plug-in classifiers," *Ann. Statist.*, vol. 35, no. 2, pp. 608–633, 2007.

[69] G. B. Folland, *Real Analysis: Modern Techniques and Their Applications*. John Wiley & Sons, 2013.

[70] A. B. Tsybakov, *Introduction to Nonparametric Estimation*, ser. Springer Series in Statistics. New York, NY: Springer New York, 2009.

[71] A. Cichocki, H. Lee, Y.-D. Kim, and S. Choi, "Non-negative matrix factorization with α-divergence," *Pattern Recogni. Letters*, vol. 29, no. 9, pp. 1433–1440, 2008.

[72] L. Le Cam, *Asymptotic methods in statistical decision theory*. Springer Science & Business Media, 2012.

[73] Y. Polyanskiy and Y. Wu, "Lecture notes on information theory," 2019. [Online]. Available: http://www.stat.yale.edu/~yw562/teaching/itlectures.pdf

[74] T. Cover and P. Hart, "Nearest neighbor pattern classification," *IEEE Trans. Inf. Theory*, vol. 13, no. 1, pp. 21–27, 1967.

[75] F. Pérez-Cruz, "Estimation of information theoretic measures for continuous random variables," in *Adv. Neural Inf. Proc. Syst.*, vol. 22, 2009, pp. 1257–1264.

[76] K.-T. Sturm, "On the geometry of metric measure spaces," *Acta Math.*, vol. 196, no. 1, pp. 65–131, 2006.

[77] W. Rudin, *Real and Complex Analysis*. McGraw-Hill Education, 1987.

[78] L. C. Evans and R. F. Gariepy, *Measure Theory and Fine Properties of Functions*. CRC Press, 2015.

[79] B. Efron and C. Stein, "The Jackknife Estimate of Variance," *Ann. Statist.*, vol. 9, no. 3, pp. 586–596, 1981.

[80] J. M. Steele, "An Efron–Stein Inequality for Nonsymmetric Statistics," *Ann. Statist.*, vol. 14, no. 2, pp. 753–758, June 1986.

[81] L. Devroye, L. Györfi, and G. Lugosi, *A probabilistic theory of pattern recognition*. Springer Science & Business Media, 2013, vol. 31.

J. Jon Ryu J. Jon Ryu (S'18) received the B.S. (Hons.) degrees in electrical and computer engineering and mathematical science (double major) from Seoul National University, Seoul, South Korea, in 2015. He is pursuing the Ph.D. degree in the Department of Electrical and Computer Engineering from the University of California San Diego (UCSD), La Jolla, CA, USA. He was a recipient of Kwanjeong Scholarship for graduate study from 2015 to 2020. His research interests include information theory, data science, and statistical machine learning.

Shouvik Ganguly Shouvik Ganguly (S'17–M'21) received the B.Tech. degree in electrical engineering from Indian Institute of Technology, Kanpur in 2013, and the Ph.D. degree in electrical engineering from the University of California San Diego (UCSD) in 2020. In 2020, he joined XCOM Labs, San Diego, CA, USA, where he is currently a Member, Technical Staff. His research interests include network information theory and communication theory.

Young-Han Kim Young-Han Kim (S'99–M'06–SM'12–F'15) received the B.S. degree (Hons.) in electrical engineering from Seoul National University, Seoul, South Korea, in 1996, and the M.S. degrees in electrical engineering and in statistics and the Ph.D. degree in electrical engineering from Stanford University, Stanford, CA, USA, in 2001, 2006, and 2006, respectively. In 2006, he joined the University of California San Diego, La Jolla, CA USA, where he is currently a Professor in the Department of Electrical and Computer Engineering. Since 2020, he has also been a founding CEO of Gauss Labs Inc., an industrial AI startup company in Silicon Valley and Seoul, South Korea. He has co-authored the book Network Information Theory (Cambridge University Press, 2011) and the monograph Fundamentals of Index Coding (Now Publishers, 2018). His current research interests include data science, machine learning, information theory, and their applications in manufacturing, microelectronics, communications, networking, cryptography, and bioinformatics. Prof. Kim was a recipient of the 2008 NSF Faculty Early Career Development Award, the 2009 US–Israel Binational Science Foundation Bergmann Memorial Award, the 2012 IEEE Information Theory Paper Award, and the 2015 IEEE Information Theory Society James L. Massey Research and Teaching Award for Young Scholars. He served as an Associate Editor of the IEEE Transactions on Information Theory and a Distinguished Lecturer for the IEEE Information Theory Society. He is a foreign member of the National Academy of Engineering of Korea.

Yung-Kyun Noh Yung-Kyun Noh (M'19) is an Associate Professor in the Department of Computer Science at Hanyang University and an Affiliate Professor in the School of Computational Sciences at the Korea Institute for Advanced Study. He received the BS degree in physics from POSTECH, and the PhD degree in computer science from Seoul National University. His research interests include metric learning and dimensionality reduction in machine learning, and he is especially interested in applying statistical theory of nearest neighbors to real, large datasets. He worked in the GRASP Robotics Laboratory, University of Pennsylvania in Philadelphia as a visiting researcher. He is currently a visiting scientist at the RIKEN Center for Advanced Intelligence Project in Tokyo and a visiting scholar at the Mayo Clinic Gastroenterology and Hepatology in Rochester.

Daniel D. Lee Dr. Daniel Dongyuel Lee (F'14) is the Tisch University Professor in Electrical and Computer Engineering at Cornell Tech and Executive Vice President and Head of the Global AI Center for Samsung Research. He received his B.A. summa cum laude in Physics from Harvard University and his Ph.D. in Condensed Matter Physics from the Massachusetts Institute of Technology. He was also a researcher at Bell Labs in the Theoretical Physics and Biological Computation departments. He is a Fellow of the IEEE and AAAI and has received the NSF CAREER award and the Lindback award for distinguished teaching. He was also a fellow of the Hebrew University Institute of Advanced Studies in Jerusalem, an affiliate of the Korea Advanced Institute of Science and Technology, and organized the US-Japan National Academy of Engineering Frontiers of Engineering symposium and Neural Information Processing Systems (NeurIPS) conference. His group focuses on understanding general computational principles in biological systems and on applying that knowledge to build autonomous systems.