The Annals of Statistics
2019, Vol. 47, No. 1, 288–318
https://doi.org/10.1214/18-AOS1688

EFFICIENT MULTIVARIATE ENTROPY ESTIMATION VIA k-NEAREST NEIGHBOUR DISTANCES

BY THOMAS B. BERRETT[*,1], RICHARD J. SAMWORTH[*,2] AND MING YUAN[†,3]

University of Cambridge[*] and University of Wisconsin–Madison[†]

Many statistical procedures, including goodness-of-fit tests and methods for independent component analysis, rely critically on the estimation of the entropy of a distribution. In this paper, we seek entropy estimators that are efficient and achieve the local asymptotic minimax lower bound with respect to squared error loss. To this end, we study weighted averages of the estimators originally proposed by Kozachenko and Leonenko [*Probl. Inform. Transm.* **23** (1987), 95–101], based on the k-nearest neighbour distances of a sample of n independent and identically distributed random vectors in $\mathbb{R}^d$. A careful choice of weights enables us to obtain an efficient estimator in arbitrary dimensions, given sufficient smoothness, while the original unweighted estimator is typically only efficient when $d \leq 3$. In addition to the new estimator proposed and theoretical understanding provided, our results facilitate the construction of asymptotically valid confidence intervals for the entropy of asymptotically minimal width.

1. Introduction. The concept of entropy plays a central role in information theory, and has found a wide array of uses in other disciplines, including statistics, probability and combinatorics. The *(differential) entropy* of a random vector X with density function f is defined as

$$H = H(X) = H(f) := -\mathbb{E}\{\log f(X)\} = -\int_{\mathcal{X}} f(x) \log f(x)\, dx,$$

where $\mathcal{X} := \{x : f(x) > 0\}$. It represents the average information content of an observation, and is usually thought of as a measure of unpredictability.

In statistical contexts, it is often the estimation of entropy that is of primary interest, for instance in goodness-of-fit tests of normality [Vasicek (1976)] or uniformity [Cressie (1976)], tests of independence [Goria et al. (2005)], independent component analysis [Learned-Miller and Fisher (2004)] and feature selection in classification [Kwak and Choi (2002)]. See, for example, Beirlant et al. (1997)

Received June 2017; revised November 2017.
[1]Supported by a Ph.D. scholarship from the SIMS fund.
[2]Supported by an EPSRC Early Career Fellowship and a grant from the Leverhulme Trust.
[3]Supported by NSF FRG Grant DMS-1265202 and NIH Grant 1-U54AI117924-01.
MSC2010 subject classifications. 62G05, 62G20.
Key words and phrases. Efficiency, entropy estimation, Kozachenko–Leonenko estimator, weighted nearest neighbours.

and Paninski (2003) for other applications and an overview of nonparametric techniques, which include methods based on sample spacings in the univariate case [e.g., El Haje Hussein and Golubev (2009)], histograms [Hall and Morton (1993)] and kernel density estimates [Paninski and Yajima (2008), Sricharan, Wei and Hero (2013)], among others. The estimator of Kozachenko and Leonenko (1987) is particularly attractive as a starting point, both because it generalises easily to multivariate cases, and because, since it only relies on the evaluation of kth-nearest neighbour distances, it is straightforward to compute.

To introduce this estimator, for $n \geq 2$, let $X_1, \dots, X_n$ be independent random vectors with density f on $\mathbb{R}^d$. Write $\|\cdot\|$ for the Euclidean norm on $\mathbb{R}^d$, and for $i = 1, \dots, n$, let $X_{(1),i}, \dots, X_{(n-1),i}$ denote a permutation of $\{X_1, \dots, X_n\} \setminus \{X_i\}$ such that $\|X_{(1),i} - X_i\| \leq \cdots \leq \|X_{(n-1),i} - X_i\|$. For conciseness, we let

$$\rho_{(k),i} := \|X_{(k),i} - X_i\|$$

denote the distance between X_i and the kth nearest neighbour of X_i. The Kozachenko–Leonenko estimator of the entropy H is given by

$$\hat{H}_n = \hat{H}_n(X_1, \dots, X_n) := \frac{1}{n} \sum_{i=1}^{n} \log\Bigg(\frac{\rho_{(k),i}^d V_d (n-1)}{e^{\Psi(k)}} \Bigg), \tag{1}$$

where $V_d := \pi^{d/2} / \Gamma(1 + d/2)$ denotes the volume of the unit d-dimensional Euclidean ball and where Ψ denotes the digamma function. In fact, this is a generalisation of the estimator originally proposed by Kozachenko and Leonenko (1987), which was defined for $k = 1$. For integers k, we have $\Psi(k) = -\gamma + \sum_{j=1}^{k-1} 1/j$ where $\gamma := 0.577216\dots$ is the Euler–Mascheroni constant, so that $e^{\Psi(k)}/k \to 1$ as $k \to \infty$. This estimator can be regarded as an attempt to mimic the "oracle" estimator $H_n^* := -n^{-1} \sum_{i=1}^{n} \log f(X_i)$, based on a k-nearest neighbour density estimate that relies on the approximation

$$\frac{k}{n-1} \approx V_d \rho_{(k),1}^d f(X_1).$$

It turns out that, when $d \leq 3$ and other regularity conditions hold, the estimator $\hat{H}_n$ in (1) has the same asymptotic behaviour as H_n^*, in that

$$n^{1/2}(\hat{H}_n - H) \xrightarrow{d} N\big(0, \operatorname{Var} \log f(X_1)\big).$$

We will see that in such settings this estimator is asymptotically efficient, in the sense of, for example, van der Vaart [(1998), p. 367]. However, when $d \geq 4$, a non-trivial bias typically precludes its efficiency. Our main object of interest, therefore, will be a generalisation of the estimator (1), formed as a weighted average of Kozachenko–Leonenko estimators for different values of k, where the weights

are chosen to try to cancel the dominant bias terms. More precisely, for a weight vector $w = (w_1, \ldots, w_k)^T \in \mathbb{R}^k$ with $\sum_{j=1}^k w_j = 1$, we consider the estimator

$$\hat{H}_n^w := \frac{1}{n} \sum_{i=1}^n \sum_{j=1}^k w_j \log \xi_{(j),i},$$

where $\xi_{(j),i} := e^{-\Psi(j)} V_d (n-1) \rho_{(j),i}^d$. Weighted estimators of this general type have been considered recently [e.g., Moon et al. (2016), Sricharan, Wei and Hero (2013)], though our construction of the weights and our analysis is new. In particular, we show that under stronger smoothness assumptions, and with a suitable choice of weights, the weighted Kozachenko–Leonenko estimator is efficient in arbitrary dimensions.

There have been several previous studies of the (unweighted) Kozachenko–Leonenko estimator, but results on the rate of convergence have until now confined either to the case $k = 1$ or (very recently) the case where k is fixed as n diverges. The original Kozachenko and Leonenko (1987) paper proved consistency of the estimator under mild conditions in the case $k = 1$. Tsybakov and van der Meulen (1996) proved that the mean squared error of a truncated version of the estimator is $O(n^{-1})$ when $k = 1$ and $d = 1$ under a condition that is almost equivalent to an exponential tail; Biau and Devroye (2015) showed that the bias vanishes asymptotically while the variance is $O(n^{-1})$ when $k = 1$ and f is compactly supported and bounded away from zero on its support. Very recently, in independent work and under regularity conditions, Delattre and Fournier (2017) derived the asymptotic normality of the estimator when $k = 1$, confirming the suboptimal asymptotic variance in this case. Previous works on the general k case include Singh et al. (2003), where heuristic arguments were presented to suggest the estimator is consistent for general d and general fixed k and has variance $O(n^{-1})$ for $d = 1$ and general fixed k. Gao, Oh and Viswanath (2016) obtain a mean squared error bound of $O(n^{-1})$ up to polylogarithmic factors for fixed k and $d \leq 2$, though the only densities which the authors can show satisfy their tail condition have bounded support. Singh and Póczos (2016) obtain a similar bound (without the polylogarithmic factors, but explicitly assuming bounded support) for fixed k and $d \leq 4$. Mnatsakanov et al. (2008) allow k to diverge with n, and show that the estimator is consistent for general d.

Plug-in kernel methods are also popular for entropy estimation. Paninski and Yajima (2008), for example, show that a smaller bandwidth than would be required for a consistent density estimator can still yield a consistent entropy estimator. A k-nearest neighbour density estimate can be regarded as a kernel estimator with a bandwidth that depends both on the data and on the point at which the estimate is required. Sricharan, Wei and Hero (2013) obtain the parametric rate of convergence for a plug-in kernel method, assuming bounded support and at least d derivatives in the interior of the support.

Importantly, the class of densities considered in our results allows the support of the density to be unbounded; for instance, it may be the whole of $\mathbb{R}^d$. Such settings present significant new challenges and lead to different behaviour compared with more commonly-studied situations where the underlying density is compactly supported and bounded away from zero on its support. To gain intuition, consider the following second-order Taylor expansion of $H(f)$ around a density estimator $\hat{f}$:

$$H(f) \approx -\int_{\mathbb{R}^d} f(x) \log \hat{f}(x)\, dx - \frac{1}{2}\biggl(\int_{\mathbb{R}^d} \frac{f^2(x)}{\hat{f}(x)}\, dx - 1\biggr).$$

When f is bounded away from zero on its support, one can estimate the (smaller order) second term on the right-hand side, thereby obtaining efficient estimators of entropy in higher dimensions [Laurent (1996)]; however, when f is not bounded away from zero on its support such procedures are no longer effective. To the best of our knowledge, therefore, this is the first time that a nonparametric entropy estimator has been shown to be efficient in multivariate settings for densities having unbounded support. [We remark that when $d = 1$, the histogram estimator of Hall and Morton (1993) is known to be efficient under fairly strong tail conditions.]

The outline of the rest of the paper is as follows. In Section 2, we give our main results on the mean squared error and asymptotic normality of weighted Kozachenko–Leonenko estimators, and discuss confidence interval construction. These main results arise from asymptotic expansions for the bias and variance, which are stated in Section 3. Here, we also give examples to illustrate densities satisfying our conditions, discuss how they may be weakened, and address the fixed k case. Corresponding lower bounds are presented in Section 4. Proofs of main results are presented in Section 5 with auxiliary material and detailed bounds for various error terms deferred to the Appendix, which appears as the Supplementary Material [Berrett, Samworth and Yuan (2019)].

We conclude the Introduction with some notation used throughout the paper. For $x \in \mathbb{R}^d$ and $r > 0$, let $B_x(r)$ be the closed Euclidean ball of radius r about x, and let $B_x^\circ(r) := B_x(r) \setminus \{x\}$ denote the corresponding punctured ball. We write $\|A\|_{\mathrm{op}}$ and $|A|$ for the operator norm and determinant, respectively, of $A \in \mathbb{R}^{d\times d}$, and let $\|A\|$ denote the vectorised Euclidean norm of a vector, matrix or array. For a smooth function $f : \mathbb{R}^d \to [0,\infty)$, we write $\dot{f}(x)$, $\ddot{f}(x)$ and $f^{(m)}(x)$, respectively, for the gradient vector of f at x, Hessian matrix of f at x and the array with $(j_1,\dots,j_m)$th entry $\frac{\partial^m f(x)}{\partial x_{j_1}\dots\partial x_{j_m}}$. We also write $\Delta f(x) := \sum_{j=1}^d \frac{\partial^2 f}{\partial x_j^2}(x)$ for its Laplacian, and $\|f\|_\infty := \sup_{x\in\mathbb{R}^d} f(x)$ for its uniform norm.

2. Main results. We begin by introducing the class of densities over which our results will hold. Let $\mathcal{F}_d$ denote the class of all density functions with respect to Lebesgue measure on $\mathbb{R}^d$. For $f \in \mathcal{F}_d$ and $\alpha > 0$, let

$$\mu_\alpha(f) := \int_{\mathbb{R}^d} \|x\|^\alpha f(x)\, dx.$$

Now let $\mathcal{A}$ denote the class of decreasing functions $a : (0, \infty) \to [1, \infty)$ satisfying $a(\delta) = o(\delta^{-\varepsilon})$ as $\delta \searrow 0$, for every $\varepsilon > 0$. If $a \in \mathcal{A}$, $\beta > 0$ and $f \in \mathcal{F}_d$ is $m := \lceil \beta \rceil - 1$-times differentiable and $x \in \mathcal{X}$, we define $r_a(x) := \{8d^{1/2}a(f(x))\}^{-1/(\beta \wedge 1)}$ and

$$M_{f,a,\beta}(x) := \max\left\{ \max_{t=1,\ldots,m} \frac{\|f^{(t)}(x)\|}{f(x)}, \sup_{y \in B_x^\circ(r_a(x))} \frac{\|f^{(m)}(y) - f^{(m)}(x)\|}{f(x)\|y - x\|^{\beta - m}} \right\}.$$

The quantity $M_{f,a,\beta}(x)$ measures the smoothness of derivatives of f in neighbourhoods of x, relative to $f(x)$ itself. Note that these neighbourhoods of x are allowed to become smaller when $f(x)$ is small. Finally, for $\Theta := (0, \infty)^4 \times \mathcal{A}$, and $\theta = (\alpha, \beta, \nu, \gamma, a) \in \Theta$, let

$$\mathcal{F}_{d,\theta} := \left\{ f \in \mathcal{F}_d : \mu_\alpha(f) \le \nu, \|f\|_\infty \le \gamma, \sup_{x : f(x) \ge \delta} M_{f,a,\beta}(x) \le a(\delta) \ \forall \delta > 0 \right\}.$$

We note here that Lemma 2 in the online Supplementary Material [Berrett, Samworth and Yuan (2019)] can be used to derive a nestedness property of the classes with respect to the smoothness parameter, namely that if $\theta = (\alpha, \beta, \gamma, \nu, a) \in \Theta$, $\beta' \in (0, \beta)$ and $a'(\delta) = 15d^{\lceil \beta \rceil/2}a(\delta)$, then $\mathcal{F}_{d,\theta} \subseteq \mathcal{F}_{d,\theta'}$, where $\theta' = (\alpha, \beta', \gamma, \nu, a') \in \Theta$. In Section 3.2 below, we discuss the requirements of the class $\mathcal{F}_{d,\theta}$ in greater detail, and give several examples, including Gaussian and multivariate-t densities, which belong to $\mathcal{F}_{d,\theta}$ for suitable θ.

We now introduce the class of weights $w = (w_1, \ldots, w_k)^T$ that we consider. For $k \in \mathbb{N}$, let

$$\mathcal{W}^{(k)} := \left\{ w \in \mathbb{R}^k : \sum_{j=1}^k w_j \frac{\Gamma(j + 2\ell/d)}{\Gamma(j)} = 0 \text{ for } \ell = 1, \ldots, \lfloor d/4 \rfloor \right. \tag{2}$$
$$\left. \sum_{j=1}^k w_j = 1 \text{ and } w_j = 0 \text{ if } j \notin \{\lfloor k/d \rfloor, \lfloor 2k/d \rfloor, \ldots, k\} \right\}.$$

Our main result below shows that for appropriately chosen weight vectors in $\mathcal{W}^{(k)}$, the normalised risk of the weighted Kozachenko–Leonenko estimator $\hat{H}_n^w$ converges in a uniform sense to that of the oracle estimator $H_n^* := -n^{-1}\sum_{i=1}^n \log f(X_i)$. Theorem 8 in Section 4 shows that this limiting risk is optimal.

THEOREM 1. *Fix $d \in \mathbb{N}$ and $\theta = (\alpha, \beta, \nu, \gamma, a) \in \Theta$ with $\alpha > d$ and with $\beta > d/2$. Let $k_0^* = k_{0,n}^*$ and $k_1^* = k_{1,n}^*$ denote any two deterministic sequences of positive integers with $k_0^* \le k_1^*$, with $k_0^*/\log^5 n \to \infty$ and with $k_1^* = O(n^{\tau_1})$ and $k_1^* = o(n^{\tau_2})$, where*

$$\tau_1 < \min\left(\frac{2\alpha}{5\alpha + 3d}, \frac{\alpha - d}{2\alpha}, \frac{4\beta^*}{4\beta^* + 3d} \right),$$
$$\tau_2 := \min\left(1 - \frac{d/4}{1 + \lfloor d/4 \rfloor}, 1 - \frac{d}{2\beta} \right)$$

and $\beta^* := \beta \wedge 1$. *There exists* $k_d \in \mathbb{N}$, *depending only on* d, *such that for each* $k \geq k_d$, *we can find* $w = w^{(k)} \in \mathcal{W}^{(k)}$ *with* $\sup_{k \geq k_d} \|w^{(k)}\| < \infty$. *For such* w,

$$\sup_{k \in \{k_0^*, \ldots, k_1^*\}} \sup_{f \in \mathcal{F}_{d,\theta}} n\mathbb{E}_f\{(\hat{H}_n^w - H_n^*)^2\} \to 0 \tag{3}$$

as $n \to \infty$. *In particular,*

$$\sup_{k \in \{k_0^*, \ldots, k_1^*\}} \sup_{f \in \mathcal{F}_{d,\theta}} |n\mathbb{E}_f\{(\hat{H}_n^w - H(f))^2\} - V(f)| \to 0,$$

where $V(f) := \mathrm{Var}_f \log f(X_1) = \int_{\mathcal{X}} f \log^2 f - H(f)^2$.

We remark that the level of smoothness we require for efficiency in Theorem 1, namely $\beta > d/2$, is more than is needed for the two-stage estimator of Laurent (1996) in the case where f is compactly supported and bounded away from zero on its support, where $\beta > d/4$ suffices. As alluded to in the Introduction, the fact that the function $x \mapsto -x \log x$ is nondifferentiable at $x = 0$ means that the entropy functional is no longer smooth when f has full support, so the arguments of Laurent (1996) can no longer be applied and very different behaviour may occur [Cai and Low (2011), Lepski, Nemirovski and Spokoiny (1999)].

It is also useful, for example, for the purposes of constructing confidence intervals for the entropy, to understand the asymptotic normality of the estimator. First, note that the asymptotic variance $V(f)$ can be estimated analogously to $H(f)$ by $\hat{V}_n^w := \max(\tilde{V}_n^w, 0)$, where

$$\tilde{V}_n^w := \frac{1}{n} \sum_{i=1}^{n} \sum_{j=1}^{k} w_j \log^2 \xi_{(j),i} - (\hat{H}_n^w)^2.$$

Fixing $q \in (0, 1)$, this suggests that a natural asymptotic $(1 - q)$-level confidence interval for $H(f)$ is given by

$$I_{n,q} := [\hat{H}_n^w - n^{-1/2} z_{q/2} (\hat{V}_n^w)^{1/2}, \hat{H}_n^w + n^{-1/2} z_{q/2} (\hat{V}_n^w)^{1/2}],$$

where z_q is the $(1 - q)$th quantile of the standard normal distribution; see also Delattre and Fournier (2017). For $p \geq 1$, write $\mathcal{P}_p$ for the class of probability measures P on $\mathbb{R}$ with $\int_{-\infty}^{\infty} |x|^p \, dP(x) < \infty$. For $P, Q \in \mathcal{P}_p$, we define the pth Wasserstein distance between P and Q by

$$d_p(P, Q) := \inf_{(X,Y) \sim (P,Q)} \{\mathbb{E}(|X - Y|^p)\}^{1/p},$$

where the infimum is taken over all pairs (X, Y) defined on a common probability space with $X \sim P$ and $Y \sim Q$. Recall that if $P \in \mathcal{P}_p$ and (P_n) is a sequence in $\mathcal{P}_p$, then $d_p(P_n, P) \to 0$ if and only if both $P_n \xrightarrow{d} P$ and $\int_{-\infty}^{\infty} |x|^p \, dP_n(x) \to \int_{-\infty}^{\infty} |x|^p \, dP(x)$. Write $\mathcal{L}(Z)$ for the distribution of a random variable Z.

THEOREM 2. *Under the conditions of Theorem 1, we have*

$$\sup_{k\in\{k_0^*,\ldots,k_1^*\}} \sup_{f\in\mathcal{F}_{d,\theta}} d_2\big(\mathcal{L}\big(n^{1/2}\{\hat{H}_n^w - H(f)\}\big), N\big(0, V(f)\big)\big) \to 0$$

as $n \to \infty$. *Consequently,*

$$\sup_{q\in(0,1)} \sup_{k\in\{k_0^*,\ldots,k_1^*\}} \sup_{f\in\mathcal{F}_{d,\theta}} \big|\mathbb{P}_f\big(I_{n,q} \ni H(f)\big) - (1-q)\big| \to 0.$$

We remark that the choice $k = k_n = \lceil \log^6 n \rceil$ with $w = w^{(k)} \in \mathcal{W}^{(k)}$ satisfying $\sup_{k\geq k_d} \|w^{(k)}\| < \infty$ for the weighted Kozachenko–Leonenko estimator satisfies the conditions for efficiency in Theorem 1 whenever $f \in \mathcal{F}_{d,\theta}$ with $\theta = (\alpha, \beta, \gamma, \nu, a) \in \Theta$ satisfying $\alpha > d$ and $\beta > d/2$; knowledge of the precise values of α and β is not required. Moreover, the uniformity of the asymptotics in k means that if $\hat{k}_n = \hat{k}_n(X_1, \ldots, X_n)$ is a data-driven choice of k, the conclusions Theorem 2 remain valid provided that $\mathbb{P}(\hat{k}_n < k_0^*) + \mathbb{P}(\hat{k}_n > k_1^*) \to 0$.

3. Bias and variance expansions for Kozachenko–Leonenko estimators.

3.1. *Bias.* The proof of (3) is derived from separate expansions for the bias and variance of the weighted Kozachenko–Leonenko estimator, and we treat the bias in this subsection. To gain intuition, we initially focus for simplicity of exposition on the unweighted estimator

$$\hat{H}_n = \frac{1}{n}\sum_{i=1}^{n} \log \xi_i,$$

where we have written ξ_i as shorthand for $\xi_{(k),i}$. For $x \in \mathbb{R}^d$ and $u \in [0, \infty)$, we introduce the sequence of distribution functions

$$F_{n,x}(u) := \mathbb{P}(\xi_i \leq u | X_i = x) = \sum_{j=k}^{n-1} \binom{n-1}{j} p_{n,x,u}^j (1 - p_{n,x,u})^{n-1-j},$$

where

$$p_{n,x,u} := \int_{B_x(r_{n,u})} f(y)\,dy \quad \text{and} \quad r_{n,u} := \left\{\frac{e^{\Psi(k)}u}{V_d(n-1)}\right\}^{1/d}.$$

Further, for $u \in [0, \infty)$, define the limiting (Gamma) distribution function

$$F_x(u) := \exp\{-uf(x)e^{\Psi(k)}\} \sum_{j=k}^{\infty} \frac{1}{j!}\{uf(x)e^{\Psi(k)}\}^j = e^{-\lambda_{x,u}} \sum_{j=k}^{\infty} \frac{\lambda_{x,u}^j}{j!},$$

where $\lambda_{x,u} := uf(x)e^{\Psi(k)}$. That this is the limit distribution for each fixed k follows from a Poisson approximation to the Binomial distribution and the Lebesgue

differentiation theorem. We therefore expect that

$$\mathbb{E}(\hat{H}_n) = \int_{\mathcal{X}} f(x) \int_0^\infty \log u \, dF_{n,x}(u) \, dx \approx \int_{\mathcal{X}} f(x) \int_0^\infty \log u \, dF_x(u) \, dx$$
$$= \int_{\mathcal{X}} f(x) \int_0^\infty \log\Big(\frac{te^{-\Psi(k)}}{f(x)}\Big) e^{-t} \frac{t^{k-1}}{(k-1)!} \, dt \, dx = H.$$

Although we do not explicitly use this approximation in our asymptotic analysis of the bias, it motivates much of our development. It also explains the reason for using $e^{\Psi(k)}$ in the definition of $\xi_{(k),i}$, rather than simply k. Lemma 3 below gives an expression for the asymptotic bias of the unweighted Kozachenko–Leonenko estimator.

LEMMA 3. *Fix $d \in \mathbb{N}$ and $\theta = (\alpha, \beta, \nu, \gamma, a) \in \Theta$. Let $k^* = k_n^*$ denote any deterministic sequence of positive integers with $k^* = O(n^{1-\varepsilon})$ as $n \to \infty$ for some $\varepsilon > 0$. Then there exist $\lambda_1, \ldots, \lambda_{\lceil \beta/2 \rceil - 1} \in \mathbb{R}$, depending only on f and d, such that $\sup_{f \in \mathcal{F}_{d,\theta}} \max_{l=1,\ldots,\lceil \beta/2 \rceil - 1} |\lambda_l| < \infty$ and for each $\varepsilon > 0$,*

$$\sup_{f \in \mathcal{F}_{d,\theta}} \Big| \mathbb{E}_f(\hat{H}_n) - H - \sum_{l=1}^{\lceil \beta/2 \rceil - 1} \frac{\Gamma(k + 2l/d)\Gamma(n)}{\Gamma(k)\Gamma(n + 2l/d)} \lambda_l \Big| = O\bigg(\max\Big\{ \frac{k^{\frac{\alpha}{\alpha+d} - \varepsilon}}{n^{\frac{\alpha}{\alpha+d} - \varepsilon}}, \frac{k^{\frac{\beta}{d}}}{n^{\frac{\beta}{d}}} \Big\} \bigg)$$

as $n \to \infty$, uniformly for $k \in \{1, \ldots, k^*\}$, where $\lambda_l = 0$ if $2l \geq d\alpha/(\alpha + d)$.

When $d \geq 3$, $\alpha > 2d/(d-2)$ and $\beta > 2$, we have

$$\lambda_1 = -\frac{1}{2(d+2)V_d^{2/d}} \int_{\mathcal{X}} \frac{\Delta f(x)}{f(x)^{2/d}} \, dx,$$

which is finite under these assumptions; cf. the second part of Proposition 9 in Section 5.1. Moreover, since, for each $l > 0$, we have $\frac{\Gamma(n)}{\Gamma(n+2l/d)} = n^{-2l/d}\{1 + O(n^{-1})\}$, we deduce from Lemma 3 that in this setting,

$$\sup_{f \in \mathcal{F}_{d,\theta}} \Big| \mathbb{E}_f(\hat{H}_n) - H + \frac{\Gamma(k + 2/d)}{2(d+2)V_d^{2/d}\Gamma(k)n^{2/d}} \int_{\mathcal{X}} \frac{\Delta f(x)}{f(x)^{2/d}} \, dx \Big| = o\Big(\frac{k^{2/d}}{n^{2/d}} \Big).$$

In particular, when $d \geq 4$ and $\int_{\mathcal{X}} \frac{\Delta f(x)}{f(x)^{2/d}} \, dx \neq 0$, the bias of the unweighted Kozachenko–Leonenko estimator precludes its efficiency.

On the other hand, Lemma 3 motivates the definition of the class of weight vectors $\mathcal{W}^{(k)}$ in (2), and facilitates the expansion for the bias of the weighted Kozachenko–Leonenko estimator in Corollary 4 below. In particular, since $2(\lfloor d/4 \rfloor + 1)/d > 1/2$, we see that this result provides conditions under which the bias is $o(n^{-1/2})$ for suitably chosen k. This explains why we let ℓ take values in the range $\{1, \ldots, \lfloor d/4 \rfloor\}$ in (2).

COROLLARY 4. *Assume the conditions of Lemma 3. If $w = w^{(k)} \in \mathcal{W}^{(k)}$ for $k \geq k_d$ and $\sup_{k \geq k_d} \|w^{(k)}\| < \infty$, then for every $\varepsilon > 0$,*

$$\sup_{f \in \mathcal{F}_{d,\theta}} |\mathbb{E}_f(\hat{H}_n^w) - H(f)| = O\biggl(\max\biggl\{\frac{k^{\frac{\alpha}{\alpha+d}-\varepsilon}}{n^{\frac{\alpha}{\alpha+d}-\varepsilon}}, \frac{k^{\frac{2(\lfloor d/4 \rfloor+1)}{d}}}{n^{\frac{2(\lfloor d/4 \rfloor+1)}{d}}}, \frac{k^{\frac{\beta}{d}}}{n^{\frac{\beta}{d}}}\biggr\}\biggr),$$

uniformly for $k \in \{1, \ldots, k^*\}$.

The proof of Lemma 3 is given in Section 5.1, but we present here some of the main ideas that are particularly relevant for the case $d \geq 3$, $\alpha > 2d/(d-2)$ and $\beta \in (2, 4]$. First, note that

$$\frac{dF_{n,x}(u)}{du} = \mathrm{B}_{k,n-k}(p_{n,x,u}) \frac{\partial p_{n,x,u}}{\partial u}, \tag{4}$$

where $\mathrm{B}_{a,b}(s) := \mathrm{B}_{a,b}^{-1} s^{a-1}(1-s)^{b-1}$ denotes the density of a Beta(a, b) random variable at $s \in (0, 1)$, with $\mathrm{B}_{a,b} := \Gamma(a)\Gamma(b)/\Gamma(a+b)$. For $x \in \mathcal{X}$ and $r > 0$, define $h_x(r) := \int_{B_x(r)} f(y)\,dy$. Since $h_x(r)$ is a continuous, non-decreasing function of r, we can define a left-continuous inverse for $s \in (0, 1)$ by

$$h_x^{-1}(s) := \inf\{r > 0 : h_x(r) \geq s\} = \inf\{r > 0 : h_x(r) = s\}, \tag{5}$$

so that $h_x(r) \geq s$ if and only if $r \geq h_x^{-1}(s)$. We use the approximation

$$V_d f(x) h_x^{-1}(s)^d \approx s - \frac{s^{1+2/d} \Delta f(x)}{2(d+2) V_d^{2/d} f(x)^{1+2/d}}$$

for small $s > 0$, which is formalised in Lemma 10(ii) in Section 5.1. In the case $d \geq 3$, $\alpha > 2d/(d-2)$ and $\beta \in (2, 4]$, the proof of Lemma 3 can be seen as justifying the use of the above approximation in the following:

$$\begin{aligned}
\mathbb{E}(\hat{H}_n) &= \int_{\mathcal{X}} f(x) \int_0^\infty \log u \, dF_{n,x}(u)\, dx \\
&= \int_{\mathcal{X}} f(x) \int_0^1 \log\biggl(\frac{V_d(n-1)h_x^{-1}(s)^d}{e^{\Psi(k)}}\biggr) \mathrm{B}_{k,n-k}(s)\, ds\, dx \\
&\approx \int_{\mathcal{X}} f(x) \int_0^1 \biggl\{\log\biggl(\frac{(n-1)s}{e^{\Psi(k)} f(x)}\biggr) - \frac{V_d^{-2/d} s^{2/d} \Delta f(x)}{2(d+2) f(x)^{1+2/d}}\biggr\} \mathrm{B}_{k,n-k}(s)\, ds\, dx \\
&= \log(n-1) - \Psi(n) + H - \frac{V_d^{-2/d} \Gamma(k+2/d)\Gamma(n)}{2(d+2)\Gamma(k)\Gamma(n+2/d)} \int_{\mathcal{X}} \frac{\Delta f(x)}{f(x)^{2/d}}\, dx.
\end{aligned}$$

Note that $\log(n-1) - \Psi(n) = -1/(2n) + o(1/n)$, which leads to the given bias expression. The proof in other cases proceeds along similar lines. These heuristics make clear that the function $h_x^{-1}(\cdot)$ plays a key role in understanding the bias. This function is in general complicated, though some understanding can be gained from

the following uniform density example, where it can be evaluated explicitly. This leads to an exact expression for the bias, even though the discontinuities mean that the density does not belong to $\mathcal{F}_{1,\theta}$ for any $\theta \in \Theta$.

EXAMPLE 1. Consider the uniform distribution, $U[0,1]$. For $x \le 1/2$, we have

$$h_x^{-1}(s) = \begin{cases} s/2, & \text{if } s \le 2x \\ s - x, & \text{if } 2x < s \le 1. \end{cases}$$

It therefore follows that

$$\begin{aligned} \mathbb{E}(\hat{H}_n) - H &= 2\int_0^{1/2}\int_0^{\infty} \log u \, dF_{n,x}(u)\, dx \\ &= 2\int_0^{1/2}\int_0^{1} \log\Big(\frac{2(n-1)h_x^{-1}(s)}{e^{\Psi(k)}}\Big) \mathrm{B}_{k,n-k}(s)\, ds\, dx \\ &= 2\int_0^{1} \mathrm{B}_{k,n-k}(s)\Big\{\int_0^{s/2} \log\big(2(s-x)\big)\, dx + \int_{s/2}^{1/2} \log s \, dx\Big\} ds \\ &\quad + \log\Big(\frac{n-1}{e^{\Psi(k)}}\Big) \\ &= \frac{k}{n}(\log 4 - 1) + \log(n-1) - \Psi(n). \end{aligned}$$

3.2. *Discussion of conditions and weakening of conditions.* Recall the definitions of the quantity $M_{f,a,\beta}(x)$ and $\mathcal{A}$ from Section 2. In addition to standard moment and boundedness assumptions, the condition $f \in \mathcal{F}_{d,\theta}$ requires that

$$\sup_{x: f(x) \ge \delta} M_{f,a,\beta}(x) \le a(\delta) \qquad \text{for all } \delta > 0 \text{ and some } a \in \mathcal{A}. \tag{6}$$

In this subsection, we explore the condition (6) further, with the aid of several examples.

The condition (6) is reminiscent of more standard Hölder smoothness assumptions, though we also require that the partial derivatives of the density vary less where f is small. On the other hand, we also allow the neighbourhoods of x in the definition of $M_{f,a,\beta}(x)$ to shrink where $f(x)$ is small. Roughly speaking, the condition requires that the partial derivatives of the density decay nearly as fast as the density itself in the tails of the distribution. As a simple stability property, if (6) holds for a density f_0, then it also holds for any density from the location-scale family:

$$\big\{f_\Sigma(\cdot) = |\Sigma|^{-1/2} f_0\big(\Sigma^{-1/2}(\cdot - \mu)\big) : \mu \in \mathbb{R}^d, \Sigma = \Sigma^T \in \mathbb{R}^{d\times d} \text{ positive definite}\big\}.$$

This observation allows us to consider canonical representatives of location-scale families in the examples below.

PROPOSITION 5. *For each of the following densities f, and for each $d \in \mathbb{N}$, there exists $\theta \in \Theta$ such that $f \in \mathcal{F}_{d,\theta}$:*

(i) $f(x) = f(x_1, \ldots, x_d) = (2\pi)^{-d/2} e^{-\|x\|^2/2}$, *the standard normal density*;

(ii) $f(x) = f(x_1, \ldots, x_d) \propto (1 + \|x\|^2/\rho)^{-\frac{d+\rho}{2}}$, *the multivariate-t distribution with $\rho > 0$ degrees of freedom.*

Moreover, the following univariate density f also belongs to $\mathcal{F}_{1,\theta}$ for suitable $\theta \in \Theta$:

$$f(x) \propto \exp\left(-\frac{1}{1-x^2}\right)\mathbb{1}_{\{x\in(-1,1)\}}.$$

The final part of Proposition 5 is included because it provides an example of a density f that belongs to $\mathcal{F}_{1,\theta}$ for suitable $\theta \in \Theta$, even though there exist points $x_0 \in \mathbb{R}$ with $f(x_0) = 0$.

On the other hand, there are also examples, such as Example 2 below, where the behaviour of f near a point x_0 with $f(x_0) = 0$ precludes f belonging to $\mathcal{F}_{d,\theta}$ for any $\theta \in \Theta$. To provide some guarantees in such settings, we now give a very general condition under which our approach to studying the bias can be applied.

PROPOSITION 6. *Assume that f is bounded, that $\mu_\alpha(f) < \infty$ for some $\alpha > 0$ and let k^* be as in Lemma 3. Let $a_n := 3(k+1)\log(n-1)$, let $r_x := \{\frac{2a_n}{V_d(n-1)f(x)}\}^{1/d}$ and assume further that there exists $\beta > 0$ such that the function on $\mathcal{X}$ given by*

$$C_{n,\beta}(x) := \begin{cases} \sup_{y\in B_x^\circ(r_x)} |f(y) - f(x)|/\|y-x\|^\beta, & \text{if } \beta \le 1, \\ \sup_{y\in B_x^\circ(r_x)} \|\dot f(y) - \dot f(x)\|/\|y-x\|^{\beta-1}, & \text{if } \beta > 1, \end{cases}$$

is real-valued. Suppose that $\mathcal{X}_n \subseteq \mathcal{X}$ is such that

$$\sup_{x\in\mathcal{X}_n} \left(\frac{a_n}{n-1}\right)^{\tilde\beta/d} \frac{C_{n,\tilde\beta}(x)}{f(x)^{1+\tilde\beta/d}} \to 0 \tag{7}$$

as $n \to \infty$, where $\tilde\beta := \beta \wedge 2$. Then writing $q_n := \int_{\mathcal{X}_n^c} f$, we have for every $\varepsilon > 0$ that

$$\mathbb{E}_f(\hat H_n) - H = O\left(\max\left\{\frac{k^{\tilde\beta/d}}{n^{\tilde\beta/d}} \int_{\mathcal{X}_n} \frac{C_{n,\tilde\beta}(x)}{f(x)^{\tilde\beta/d}}\,dx, q_n^{1-\varepsilon}, q_n \log n, \frac{1}{n}\right\}\right), \tag{8}$$

uniformly for $k \in \{1, \ldots, k^*\}$.

To aid interpretation of Proposition 6, we first remark that if $f \in \mathcal{F}_{d,\theta}$ for some $\theta = (\alpha, \beta, \gamma, \nu, a) \in \Theta$, then (7) holds, with $\mathcal{X}_n := \{x \in \mathcal{X} : f(x) \ge \delta_n\}$, where

δ_n is defined in (12) below. On the other hand, if $f \notin \mathcal{F}_{d,\theta}$, we may still be able to obtain explicit bounds on the terms in (8) on a case-by-case basis, as in the following example.

EXAMPLE 2. For $a > 1$, consider $f(x) = \Gamma(a)^{-1} x^{a-1} e^{-x} \mathbb{1}_{\{x>0\}}$, the density of the $\Gamma(a, 1)$ distribution. Then for any $\tau \in (0, 1)$ small enough, we may take

$$\mathcal{X}_n = \left[\left(\frac{k}{n} \right)^{\frac{1}{a} - \tau}, (1 - \tau) \log \frac{n}{k} \right]$$

to deduce from Proposition 6 that for every $\varepsilon > 0$,

$$\mathbb{E}_f(\hat{H}_n) - H = o\left(\frac{k^{1-\varepsilon}}{n^{1-\varepsilon}} \right),$$

uniformly for $k \in \{1, \ldots, k^*\}$.

Similar calculations show that the bias is of the same order for Beta(a, b) distributions with $a, b > 1$.

3.3. *Asymptotic variance and normality.* We now study the asymptotic variance of Kozachenko–Leonenko estimators under the assumption that the tuning parameter k is diverging with n; the fixed k case is deferred to the next subsection.

LEMMA 7. *Let $\theta = (\alpha, \beta, \gamma, \nu, a) \in \Theta$ with $\alpha > d$ and $\beta > 0$. Let $k_0^* = k_{0,n}^*$ and $k_1^* = k_{1,n}^*$ denote any two deterministic sequences of positive integers with $k_0^* \le k_1^*$, with $k_0^* / \log^5 n \to \infty$ and with $k_1^* = O(n^{\tau_1})$, where τ_1 satisfies the condition in Theorem 1. Then for any $w = w^{(k)} \in \mathcal{W}^{(k)}$ with $\sup_{k \ge k_d} \|w^{(k)}\| < \infty$, we have*

$$\sup_{k \in \{k_0^*, \ldots, k_1^*\}} \sup_{f \in \mathcal{F}_{d,\theta}} \left| n \operatorname{Var}_f \hat{H}_n^w - V(f) \right| \to 0$$

as $n \to \infty$.

The proof of this lemma is lengthy, and involves many delicate error bounds, so we outline the main ideas in the unweighted case here. First, we argue that

$$\begin{aligned} \operatorname{Var} \hat{H}_n &= n^{-1} \operatorname{Var} \log \xi_1 + \left(1 - n^{-1}\right) \operatorname{Cov}(\log \xi_1, \log \xi_2) \\ &= n^{-1} V(f) + \operatorname{Cov}\left(\log\left(\xi_1 f(X_1)\right), \log\left(\xi_2 f(X_2)\right)\right) + o\left(n^{-1}\right), \end{aligned}$$

where we hope to exploit the fact that $\xi_1 f(X_1) \xrightarrow{p} 1$. The main difficulties in the argument are caused by the fact that handling the covariance above requires us to study the joint distribution of (ξ_1, ξ_2), and this is complicated by the fact that X_2

may be one of the k nearest neighbours of X_1 or vice versa, and more generally, X_1 and X_2 may have some of their k nearest neighbours in common. Dealing carefully with the different possible events requires us to consider separately the cases where $f(X_1)$ is small and large, as well as the proximity of X_2 to X_1. Finally, however, we can apply a normal approximation to the relevant multinomial distribution (which requires that $k \to \infty$) to deduce the result. We remark that under stronger conditions on k, it should also be possible to derive the same conclusion about the asymptotic variance of $\hat{H}_n$ while only assuming similar conditions on the density to those required in Proposition 6, but we do not pursue this here.

3.4. *Fixed k.* A crucial step in the proof of Lemma 7 is the normal approximation to a certain multinomial distribution (cf. the bound on the term W_4). This normal approximation is only valid when $k \to \infty$ as $n \to \infty$. In this subsection, we present evidence to suggest that, when k is fixed (i.e., not depending on n), then Kozachenko–Leonenko estimators are inefficient. For simplicity, we focus on the unweighted version of estimator.

Define the functions

$$\alpha_r(s,t) := \frac{1}{V_d}\mu_d\bigl(B_0(s^{1/d}) \cap B_{r^{1/d}e_1}(t^{1/d})\bigr),$$

where $e_1 = (1, 0, \ldots, 0)^T$ is the first element of the standard basis for $\mathbb{R}^d$ and μ_d denotes Lebesgue measure on $\mathbb{R}^d$. Also define the functions T_k on $[0,\infty)^3$ by

$$T_k(r,s,t) := e^{\alpha_r(s,t)} \sum_{\ell=0}^{L(r,s,t)} \sum_{i=0}^{I(r,s)-\ell} \sum_{j=0}^{J(r,t)-\ell} \frac{\{s-\alpha_r(s,t)\}^i \{t-\alpha_r(s,t)\}^j \alpha_r^\ell(s,t)}{i!j!\ell!} - \sum_{i=0}^{I(r,s)} \sum_{j=0}^{J(r,t)} \frac{s^i t^j}{i!j!},$$

where $L(r,s,t) := k-1-\mathbb{1}_{\{r<\max(s,t)\}}$, $I(r,s) := k-1-\mathbb{1}_{\{r<s\}}$, $J(r,t) := k-1-\mathbb{1}_{\{r<t\}}$.

In the case $k=1$, this function appears in Delattre and Fournier (2017), where the authors show that, under certain regularity conditions,

$$\lim_{n\to\infty} n \operatorname{Var} \hat{H}_n - V(f) = \Psi'(1) + \int_{[0,\infty)^3} e^{-s-t} \frac{T_1(r,s,t)}{st}\, dr\, ds\, dt - 1 + 2\log 2.$$

Since $T_1(r,s,t) = \{e^{\alpha_r(s,t)} - 1\}\mathbb{1}_{\{r \geq s \vee t\}} \geq 0$, this limit is strictly positive. More generally, Poisson approximation to the same multinomial distribution mentioned above, together with analysis similar to the proof of Lemma 7, suggests that for

TABLE 1
Asymptotic variance inflation (9) of the Kozachenko–Leonenko estimator for fixed k

$d\backslash k$	1	2	3	4	5
1	2.14	0.97	0.64	0.48	0.39
2	2.29	1.01	0.64	0.47	0.38
3	2.42	1.03	0.64	0.47	0.37
5	2.61	1.05	0.65	0.47	0.37
10	2.85	1.10	0.68	0.50	0.40

(fixed) $k \geq 2$,

$$\begin{aligned}
&\lim_{n\to\infty} n \operatorname{Var} \hat{H}_n - V(f) \\
&\quad = \Psi'(k) + \int_{[0,\infty)^3} e^{-s-t} \frac{T_k(r,s,t)}{st}\, dr\, ds\, dt - 1 \\
&\quad\quad + 2^{-(2k-2)} \binom{2k-2}{k-1} \{\Psi(2k-1) - \Psi(k) - \log 2\} \\
&\quad\quad + \frac{1}{k-1} \sum_{j=0}^{k-2} 2^{-k-j} \binom{k+j-1}{j} \\
&\quad\quad \times [1 - (k-j)\{\Psi(k+j) - \log 2 - \Psi(k)\}].
\end{aligned} \tag{9}$$

Here, the $\Psi'(k)$ term arises as in (18), the integral term arises from the Poisson approximation, the -1 arises as in (27) and the remaining terms come from the fact that X_1 can be one of the k nearest neighbours of X_2, or vice-versa, which induces a singular component into the joint distribution function $F_{n,x,y}$ of (ξ_1, ξ_2) given $(X_1, X_2) = (x, y)$. It is interesting to observe that this asymptotic inflation of the variance is distribution-free; by Theorem 8 in Section 4 below, any distribution-free upper bound on $\limsup_{n\to\infty} n \operatorname{Var} \hat{H}_n - V(f)$ is necessarily nonnegative. We conjecture that it is in fact strictly positive for each fixed k. Evidence for this is provided in Table 1, where we tabulate numerical values for (9) for a few values of d and k. These agree with those obtained by Delattre and Fournier (2017) for the case $k = 1$.

4. Lower bounds. In this section, we address the optimality in a local asymptotic minimax sense of the limiting normalised risk $V(f)$ given in Theorem 1 using ideas of semiparametric efficiency [e.g., van der Vaart (1998), Chapter 25]. For $f \in \mathcal{F}_{d,\theta}$, $t \geq 0$ and a Borel measurable function $g : \mathbb{R}^d \to \mathbb{R}$, define $f_{t,g} : \mathbb{R}^d \to [0, \infty)$ by

$$f_{t,g}(x) := \frac{2c(t)}{1 + e^{-2tg(x)}} f(x), \tag{10}$$

where $c(t) := (\int_{\mathbb{R}^d} \frac{2}{1+e^{-2tg(x)}} f(x)\,dx)^{-1}$. This definition ensures that $\{f_{t,g} : t \geq 0\}$ is differentiable in quadratic mean at $t = 0$ with score function g [e.g., van der Vaart (1998), Example 25.16]. We say $(\tilde{H}_n)$ is an estimator sequence if $\tilde{H}_n : (\mathbb{R}^d)^{\times n} \to \mathbb{R}$ is a measurable function for each $n \in \mathbb{N}$.

THEOREM 8. *Fix $d \in \mathbb{N}$, $\theta = (\alpha, \beta, \gamma, \mu, a) \in \Theta$ and $f \in \mathcal{F}_{d,\theta}$. For $\lambda \in \mathbb{R}$, let $g_\lambda := \lambda\{\log f + H(f)\}$. Then, writing $\mathcal{I}$ for the set of finite subsets of $\mathbb{R}$, we have for any estimator sequence $(\tilde{H}_n)$ that*

$$\sup_{I\in\mathcal{I}} \liminf_{n\to\infty} \max_{\lambda\in I} n\mathbb{E}_{f_{n^{-1/2},g_\lambda}}\big[\{\tilde{H}_n - H(f_{n^{-1/2},g_\lambda})\}^2\big] \geq V(f). \tag{11}$$

Moreover, whenever $t|\lambda| \leq \min(1, \{144V(f)\}^{-1/2})$, we have $f_{t,g_\lambda} \in \mathcal{F}_{d,\tilde{\theta}}$, where $\tilde{\theta} := (\alpha, \beta, 4\gamma, 4\mu, \tilde{a}) \in \Theta$, and $\tilde{a} \in \mathcal{A}$ is defined in (61) *in the Supplementary Material.*

The proof of Theorem 8 reveals that, at every $f \in \mathcal{F}_{d,\theta}$, the entropy functional H is differentiable relative to the tangent set $\{g_\lambda : \lambda \in \mathbb{R}\}$ with efficient influence function

$$\tilde{\psi}_f := -\log f - H(f).$$

This observation, together with (3) in Theorem 1, confirms that under the assumptions on θ, w and k in that result, the weighted Kozachenko–Leonenko estimator $\hat{H}_n^w$ is (asymptotically) efficient at $f \in \mathcal{F}_{d,\theta}$ in the sense that

$$n^{1/2}\{\hat{H}_n^w - H(f)\} = \frac{1}{n^{1/2}} \sum_{i=1}^n \tilde{\psi}_f(X_i) + o_p(1)$$

[cf. van der Vaart (1998), Lemma 25.23]. Moreover, the second part of Theorem 8 and Theorem 1 imply in particular that, under these same conditions on θ, w and k, the estimator $\hat{H}_n^w$ attains the local asymptotic minimax lower bound, in the sense that

$$\sup_{I\in\mathcal{I}} \lim_{n\to\infty} \max_{\lambda\in I} n\mathbb{E}_{f_{n^{-1/2},g_\lambda}}\big[\{\hat{H}_n^w - H(f_{n^{-1/2},g_\lambda})\}^2\big] = V(f).$$

5. Proofs of main results.

5.1. *Auxiliary results and proofs of Lemma 3 and Corollary 4.* Throughout the proofs, we write $a \lesssim b$ to mean that there exists $C > 0$, depending only on $d \in \mathbb{N}$ and $\theta \in \Theta$, such that $a \leq Cb$. The proof of Lemma 3 relies on the following two auxiliary results, whose proofs are given in Appendix A.1.

PROPOSITION 9. *Let $\theta = (\alpha, \beta, \gamma, \nu, a) \in \Theta$, $d \in \mathbb{N}$ and $\tau \in (\frac{d}{\alpha+d}, 1]$. Then*

$$\sup_{f \in \mathcal{F}_{d,\theta}} \int_{\{x: f(x) < \delta\}} a\big(f(x)\big) f(x)^{\tau} \, dx \to 0$$

as $\delta \searrow 0$. Moreover, for every $\rho > 0$,

$$\sup_{f \in \mathcal{F}_{d,\theta}} \int_{\mathcal{X}} a\big(f(x)\big)^{\rho} f(x)^{\tau} < \infty.$$

Recall the definition of $h_x^{-1}(\cdot)$ in (5). The first part of Lemma 10 below provides crude but general bounds; the second gives much sharper bounds in a more restricted region.

LEMMA 10. (i) *Let $f \in \mathcal{F}_d$ and let $\alpha > 0$. Then for every $s \in (0,1)$ and $x \in \mathbb{R}^d$,*

$$\left(\frac{s}{V_d \|f\|_{\infty}}\right)^{1/d} \le h_x^{-1}(s) \le \|x\| + \left(\frac{\mu_{\alpha}(f)}{1-s}\right)^{1/\alpha}.$$

(ii) *Fix $\theta = (\alpha, \beta, \gamma, \nu, a) \in \Theta$, and let $\mathcal{S}_n \subseteq (0,1)$, $\mathcal{X}_n \subseteq \mathbb{R}^d$ be such that*

$$C_n := \sup_{f \in \mathcal{F}_{d,\theta}} \sup_{s \in \mathcal{S}_n} \sup_{x \in \mathcal{X}_n} \frac{a(f(x))^{d/(1 \wedge \beta)} s}{f(x)} \to 0.$$

Then there exists $n_* = n_*(d, \theta) \in \mathbb{N}$ such that for all $n \ge n_*$, $s \in \mathcal{S}_n$, $x \in \mathcal{X}_n$ and $f \in \mathcal{F}_{d,\theta}$, we have

$$\left| V_d f(x) h_x^{-1}(s)^d - \sum_{l=0}^{\lceil \beta/2 \rceil - 1} b_l(x) s^{1+2l/d} \right| \lesssim s \left\{ \frac{a(f(x))^{d/(2 \wedge \beta)} s}{f(x)} \right\}^{\beta/d},$$

where $b_0(x) = 1$ and $|b_l(x)| \lesssim a(f(x))^l f(x)^{-2l/d}$ for $l \ge 1$. Moreover, if $\beta > 2$, then

$$b_1(x) = -\frac{\Delta f(x)}{2(d+2) V_d^{2/d} f(x)^{1+2/d}}.$$

We are now in a position to prove Lemma 3.

PROOF OF LEMMA 3. (i) We initially prove the result in the case $d \ge 3$, $\alpha > 2d/(d-2)$ and $\beta \in (2, 4]$, where it suffices to show that

$$\sup_{f \in \mathcal{F}_{d,\theta}} \left| \mathbb{E}_f(\hat{H}_n) - H + \frac{\Gamma(k + 2/d)\Gamma(n)}{2(d+2) V_d^{2/d} \Gamma(k) \Gamma(n + 2/d)} \int_{\mathcal{X}} \frac{\Delta f(x)}{f(x)^{2/d}} \, dx \right|$$

$$= O\left(\max\left\{ \frac{k^{\frac{\alpha}{\alpha+d} - \varepsilon}}{n^{\frac{\alpha}{\alpha+d} - \varepsilon}}, \frac{k^{\frac{\beta}{d}}}{n^{\frac{\beta}{d}}} \right\} \right)$$

as $n \to \infty$, uniformly for $k \in \{1, \dots, k^*\}$. Fix $f \in \mathcal{F}_{d,\theta}$. Define $c_n := a(k/(n-1))^{1/(1\wedge\beta)}$, let

$$\delta_n := k c_n^d \log^2(n-1)/(n-1) \tag{12}$$

and let $\mathcal{X}_n := \{x : f(x) \geq \delta_n\}$. Recall that $a_n := 3(k+1)\log(n-1)$ and let

$$u_{x,s} := \frac{V_d(n-1)h_x^{-1}(s)^d}{e^{\Psi(k)}}.$$

The proof is based on (4) and Lemma 10(ii), which allows us to make the transformation $s = p_{n,x,u} = h_x(r_{n,u})$. Writing $R_i, i = 1, \dots, 5$ for remainder terms to be bounded at the end of the proof, we can write

$$\begin{aligned}
\mathbb{E}(\hat{H}_n) &= \int_{\mathcal{X}} f(x) \int_0^\infty \log u \, dF_{n,x}(u)\, dx \\
&= \int_{\mathcal{X}_n} f(x) \int_0^1 \mathrm{B}_{k,n-k}(s) \log u_{x,s}\, ds\, dx + R_1 \\
&= \int_{\mathcal{X}_n} f(x) \int_0^{\frac{a_n}{n-1}} \mathrm{B}_{k,n-k}(s) \log u_{x,s}\, ds\, dx + R_1 + R_2 \\
&= \int_{\mathcal{X}_n} f(x) \int_0^{\frac{a_n}{n-1}} \biggl\{ \log\Bigl(\frac{(n-1)s}{e^{\Psi(k)} f(x)}\Bigr) \\
&\qquad - \frac{V_d^{-2/d} s^{2/d} \Delta f(x)}{2(d+2) f(x)^{1+2/d}} \biggr\} \mathrm{B}_{k,n-k}(s)\, ds\, dx + \sum_{i=1}^3 R_i \\
&= \int_{\mathcal{X}_n} f(x) \biggl\{ \log\Bigl(\frac{n-1}{f(x)}\Bigr) - \Psi(n) - \frac{V_d^{-2/d} \mathrm{B}_{k+2/d,n-k} \Delta f(x)}{2(d+2)\mathrm{B}_{k,n-k} f(x)^{1+2/d}} \biggr\} dx \\
&\qquad + \sum_{i=1}^4 R_i \\
&= H + \log(n-1) - \Psi(n) - \frac{V_d^{-2/d} \Gamma(k+2/d)\Gamma(n)}{2(d+2)\Gamma(k)\Gamma(n+2/d)} \int_{\mathcal{X}_n} \frac{\Delta f(x)}{f(x)^{2/d}}\, dx \\
&\qquad + \sum_{i=1}^5 R_i .
\end{aligned}$$

After multiplying the integrand by an appropriate positive power of $\delta_n / f(x)$, the first part of Proposition 9 tells us that for every $\varepsilon > 0$,

$$\sup_{k \in \{1,\dots,k^*\}} \frac{k^{2/d}}{n^{2/d}} \sup_{f \in \mathcal{F}_{d,\theta}} \int_{\mathcal{X}_n^c} \frac{\Delta f(x)}{f(x)^{2/d}}\, dx = O\Bigl(\frac{k^{\frac{\alpha}{\alpha+d} - \varepsilon}}{n^{\frac{\alpha}{\alpha+d} - \varepsilon}} \Bigr)$$

as $n \to \infty$. Since $\log(n-1) - \Psi(n) = O(1/n)$, it now remains to bound $R_1, \dots, R_5$. Henceforth, to save repetition, we adopt without further mention the convention that whenever an error term inside $O(\cdot)$ or $o(\cdot)$ depends on k, this error is uniform for $k \in \{1, \dots, k^*\}$; thus $g(n,k) = h(n,k) + o(1)$ as $n \to \infty$ means $\sup_{k \in \{1,\dots,k^*\}} |g(n,k) - h(n,k)| \to 0$ as $n \to \infty$.

To bound R_1. By Lemma 10(i), we have $V_d^\alpha \mu_\alpha(f)^d \|f\|_\infty^\alpha \geq \alpha^\alpha d^d/(\alpha+d)^{\alpha+d}$. Hence

$$
\begin{aligned}
|\log u_{x,s}| &\leq \log(n-1) + \bigl|\Psi(k)\bigr| - \log s + \bigl|\log \|f\|_\infty\bigr| + |\log V_d| \\
&\quad + \frac{d}{\alpha}\bigl|\log \mu_\alpha(f)\bigr| - \frac{d}{\alpha}\log(1-s) + d\log\biggl(1 + \frac{\|x\|}{\mu_\alpha^{1/\alpha}(f)}\biggr) \\
&\leq \log(n-1) + \bigl|\Psi(k)\bigr| - \log s \\
&\quad + \max\biggl\{\log\gamma, \frac{1}{\alpha}\log\biggl(\frac{V_d^\alpha \nu^d (\alpha+d)^{\alpha+d}}{\alpha^\alpha d^d}\biggr)\biggr\} \\
&\quad + |\log V_d| + \frac{d}{\alpha}\max\biggl\{\log\nu, \frac{1}{d}\log\biggl(\frac{V_d^\alpha \gamma^\alpha (\alpha+d)^{\alpha+d}}{\alpha^\alpha d^d}\biggr)\biggr\} \\
&\quad - \frac{d}{\alpha}\log(1-s) + d\log\biggl(1 + \frac{\|x\|(\alpha+d)^{\frac{1}{\alpha}+\frac{1}{d}} V_d^{1/d}\gamma^{1/d}}{\alpha^{1/d} d^{1/\alpha}}\biggr).
\end{aligned} \tag{13}
$$

Moreover, for any $C_0, C_1 \geq 0$, $\varepsilon \in (0,\alpha)$ and $\varepsilon' \in (0,\varepsilon)$, we have by Hölder's inequality that

$$
\begin{aligned}
&\sup_{f\in\mathcal{F}_{d,\theta}} \int_{\mathcal{X}_n^c} f(x)\bigl\{C_0 + \log\bigl(1 + C_1\|x\|\bigr)\bigr\}\,dx \\
&\quad \leq \delta_n^{\frac{\alpha-\varepsilon'}{\alpha+d}} \sup_{f\in\mathcal{F}_{d,\theta}} \int_{\mathcal{X}} f(x)^{\frac{d+\varepsilon'}{\alpha+d}} \bigl\{C_0 + \log\bigl(1 + C_1\|x\|\bigr)\bigr\}\,dx \\
&\quad \leq \delta_n^{\frac{\alpha-\varepsilon'}{\alpha+d}} (1+\nu)^{\frac{d+\varepsilon'}{\alpha+d}} \biggl[\int_{\mathbb{R}^d} \frac{\{C_0 + \log(1 + C_1\|x\|)\}^{\frac{\alpha+d}{\alpha-\varepsilon'}}}{(1+\|x\|^\alpha)^{\frac{d+\varepsilon'}{\alpha-\varepsilon'}}}\,dx\biggr]^{\frac{\alpha-\varepsilon'}{\alpha+d}} \\
&\quad = o\biggl(\frac{k^{\frac{\alpha}{\alpha+d}-\varepsilon}}{n^{\frac{\alpha}{\alpha+d}-\varepsilon}}\biggr).
\end{aligned}
$$

Since $|\mathbb{E}(\log \mathrm{B})| = \Psi(a+b) - \Psi(a)$ when $\mathrm{B} \sim \mathrm{Beta}(a,b)$, we deduce that for each $\varepsilon > 0$,

$$
R_1 = \int_{\mathcal{X}_n^c} f(x) \int_0^1 \mathrm{B}_{k,n-k}(s) \log u_{x,s}\,ds\,dx = o\biggl(\frac{k^{\frac{\alpha}{\alpha+d}-\varepsilon}}{n^{\frac{\alpha}{\alpha+d}-\varepsilon}}\biggr)
$$

as $n \to \infty$, uniformly for $f \in \mathcal{F}_{d,\theta}$.

To bound R_2. For random variables $\mathrm{B}_1 \sim \mathrm{Beta}(k, n-k)$ and $B_2 \sim \mathrm{Bin}(n-1, a_n/(n-1))$ we have that for every $\varepsilon > 0$,

$$\begin{aligned}\mathbb{P}\big(\mathrm{B}_1 \geq a_n/(n-1)\big) &= \mathbb{P}(B_2 \leq k-1) \leq \exp\biggl(-\frac{(a_n-k+1)^2}{2a_n}\biggr) \\ &= o\big(n^{-(3-\varepsilon)}\big),\end{aligned} \tag{14}$$

where the inequality follows from standard bounds on the left-hand tail of the binomial distribution [see, e.g., Shorack and Wellner (2009), equation (6), p. 440]. Now, for any $C_1 > 0$, we have $\alpha \log(1 + C_1\|x\|) \leq (1 + C_1\|x\|)^\alpha - 1$, so that $\sup_{f\in\mathcal{F}_{d,\theta}} \int_{\mathcal{X}} f(x)\log(1 + C_1\|x\|)\,dx < \infty$. Moreover,

$$-\int_{\frac{a_n}{n-1}}^1 \log(1-s)\mathrm{B}_{k,n-k}(s)\,ds \leq \frac{n-1}{n-k-1}\int_{\frac{a_n}{n-1}}^1 \mathrm{B}_{k,n-k-1}(s)\,ds = o\big(n^{-(3-\varepsilon)}\big),$$

for every $\varepsilon > 0$, by a virtually identical argument to (14). We therefore deduce from these facts and (13) that for each $\varepsilon > 0$,

$$R_2 = \int_{\mathcal{X}_n} f(x)\int_{\frac{a_n}{n-1}}^1 \mathrm{B}_{k,n-k}(s)\log u_{x,s}\,ds\,dx = o\big(n^{-(3-\varepsilon)}\big), \tag{15}$$

which again holds uniformly in $f \in \mathcal{F}_{d,\theta}$.

To bound R_3. We can write

$$\begin{aligned}R_3 &= \int_{\mathcal{X}_n} f(x)\int_0^{\frac{a_n}{n-1}}\biggl[\biggl\{\log\biggl(\frac{V_d f(x)h_x^{-1}(s)^d}{s}\biggr) - \frac{V_d f(x)h_x^{-1}(s)^d - s}{s}\biggr\} \\ &\quad + \biggl\{\frac{V_d f(x)h_x^{-1}(s)^d - s}{s} + \frac{V_d^{-2/d}s^{2/d}\Delta f(x)}{2(d+2)f(x)^{1+2/d}}\biggr\}\biggr]\mathrm{B}_{k,n-k}(s)\,ds\,dx \\ &=: R_{31} + R_{32},\end{aligned}$$

say. Now, note that

$$\sup_{k\in\{1,\ldots,k^*\}}\sup_{f\in\mathcal{F}_{d,\theta}}\sup_{s\in(0,a_n/(n-1)]}\sup_{x\in\mathcal{X}_n}\frac{a(f(x))^d s}{f(x)} \leq \frac{6}{\log(n-1)} \to 0.$$

It follows by Lemma 10(ii) that there exist a constant $C = C(d,\theta) > 0$ and $n_1 = n_1(d,\theta) \in \mathbb{N}$ such that for $n \geq n_1$, $k \in \{1,\ldots,k^*\}$, $s \leq a_n/(n-1)$ and $x \in \mathcal{X}_n$,

$$\biggl|\frac{V_d f(x)h_x^{-1}(s)^d - s}{s} + \frac{s^{2/d}\Delta f(x)}{2(d+2)V_d^{2/d}f(x)^{1+2/d}}\biggr| \leq C\biggl\{\frac{sa(f(x))^{d/2}}{f(x)}\biggr\}^{\beta/d},$$

and

$$\biggl|\frac{V_d f(x)h_x^{-1}(s)^d - s}{s}\biggr| \leq \frac{d^{1/2}V_d^{-2/d}s^{2/d}a(f(x))}{2(d+2)f(x)^{2/d}} + C\biggl\{\frac{sa(f(x))^{d/2}}{f(x)}\biggr\}^{\beta/d} \leq \frac{1}{2}.$$

Thus, for $n \geq n_1$ and $k \in \{1, \dots, k^*\}$, using the fact that $|\log(1+z) - z| \leq z^2$ for $|z| \leq 1/2$,

$$\begin{aligned}|R_{31}| &\leq 2\int_{\mathcal{X}_n} f(x) \int_0^1 \biggl[\frac{dV_d^{-4/d} s^{4/d} a(f(x))^2}{4(d+2)^2 f(x)^{4/d}} \\ &\quad + C^2 \biggl\{\frac{sa(f(x))^{d/2}}{f(x)}\biggr\}^{2\beta/d}\biggr] \mathrm{B}_{k,n-k}(s)\,ds\,dx \\ &\leq \frac{dV_d^{-4/d}\Gamma(k+4/d)\Gamma(n)}{2(d+2)^2\Gamma(k)\Gamma(n+4/d)} \int_{\mathcal{X}_n} a\bigl(f(x)\bigr)^2 f(x)^{1-4/d}\,dx \\ &\quad + \frac{2C^2\Gamma(k+2\beta/d)\Gamma(n)}{\Gamma(k)\Gamma(n+2\beta/d)} \int_{\mathcal{X}_n} a\bigl(f(x)\bigr)^{\beta} f(x)^{1-2\beta/d}\,dx.\end{aligned}$$

On the other hand, we also have for $n \geq n_1$ and $k \in \{1, \dots, k^*\}$ that

$$\begin{aligned}|R_{32}| &\leq C\int_{\mathcal{X}_n} f(x) \int_0^1 \biggl\{\frac{sa(f(x))^{d/2}}{f(x)}\biggr\}^{\beta/d} \mathrm{B}_{k,n-k}(s)\,ds\,dx \\ &\leq \frac{C\Gamma(k+\beta/d)\Gamma(n)}{\Gamma(k)\Gamma(n+\beta/d)} \int_{\mathcal{X}_n} a\bigl(f(x)\bigr)^{\beta/2} f(x)^{1-\beta/d}\,dx.\end{aligned}$$

Multiplying each of the integrals by $f(x)/\delta_n$ to an appropriate positive power if necessary and by the second part of Proposition 9, for every $\varepsilon > 0$,

$$\max\bigl(|R_{31}|, |R_{32}|\bigr) = O\biggl(\max\biggl\{\frac{k^{\frac{\alpha}{\alpha+d}-\varepsilon}}{n^{\frac{\alpha}{\alpha+d}-\varepsilon}}, \frac{k^{\frac{\beta}{d}}}{n^{\frac{\beta}{d}}}\biggr\}\biggr),$$

uniformly for $f \in \mathcal{F}_{d,\theta}$.

To bound R_4. We have

$$R_4 = \int_{\mathcal{X}_n} f(x) \int_{\frac{a_n}{n-1}}^1 \biggl\{\log\biggl(\frac{(n-1)s}{e^{\Psi(k)} f(x)}\biggr) - \frac{V_d^{-2/d} s^{2/d} \Delta f(x)}{2(d+2) f(x)^{1+2/d}}\biggr\} \mathrm{B}_{k,n-k}(s)\,ds\,dx.$$

Consider the random variable $\mathrm{B}_1 \sim \mathrm{Beta}(k, n-k)$. Then, using (14) and the fact that $(n-1)s/e^{\Psi(k)} \geq 1$ for $s \geq a_n/(n-1)$ and $n \geq 3$, we conclude that for every $\varepsilon > 0$ and $n \geq 3$,

$$\begin{aligned}|R_4| &\leq \biggl\{\log\biggl(\frac{n-1}{e^{\Psi(k)}}\biggr) + \int_{\mathcal{X}_n} f(x)\biggl(\bigl|\log f(x)\bigr| + \frac{a(f(x))}{f(x)^{\frac{2}{d}} V_d^{\frac{2}{d}}}\biggr)\,dx\biggr\} \mathbb{P}\biggl(\mathrm{B}_1 \geq \frac{a_n}{n-1}\biggr) \\ &= o\bigl(n^{-(3-\varepsilon)}\bigr),\end{aligned}$$

uniformly for $f \in \mathcal{F}_{d,\theta}$, where, by Lemma 1(i) in the Supplementary Material, we have $\sup_{f \in \mathcal{F}_{d,\theta}} \int_{\mathcal{X}_n} f(x)|\log f(x)|\,dx < \infty$.

To bound R_5. We use the fact that for $f \in \mathcal{F}_{d,\theta}$, $x \in \mathcal{X}$ and $\varepsilon' > 0$,

$$\begin{aligned}|\log f(x)| &\leq |\log \|f\|_\infty| + \log\biggl(\frac{\|f\|_\infty}{f(x)}\biggr) \\ &\leq \max\biggl\{\log\gamma, \log V_d + \frac{1}{\alpha}\log\biggl(\frac{\nu^d(\alpha+d)^{\alpha+d}}{\alpha^\alpha d^d}\biggr)\biggr\} + \frac{1}{\varepsilon'}\biggl(\frac{\gamma}{f(x)}\biggr)^{\varepsilon'}.\end{aligned}$$

It follows from the first part of Proposition 9 (having replaced $a(\delta)$ with $\max\{a(\delta), |\log\delta|\}$ if necessary) that for each $\varepsilon > 0$,

$$R_5 = \int_{\mathcal{X}_n^c} f(x)\bigl\{\log(n-1) - \Psi(n) - \log f(x)\bigr\}\,dx = o\biggl(\frac{k^{\frac{\alpha}{\alpha+d}-\varepsilon}}{n^{\frac{\alpha}{\alpha+d}-\varepsilon}}\biggr)$$

uniformly in $f \in \mathcal{F}_{d,\theta}$. The claim follows when $d \geq 3$, $\alpha > 2d/(d-2)$ and $\beta \in (2,4]$.

We now consider the case where either $d \leq 2$ or $\alpha \leq 2d/(d-2)$ or $\beta \in (0,2]$, for which we need only show that

$$\sup_{f\in\mathcal{F}_{d,\theta}} |\mathbb{E}_f(\hat{H}_n) - H| = O\biggl(\max\biggl\{\frac{k^{\frac{\alpha}{\alpha+d}-\varepsilon}}{n^{\frac{\alpha}{\alpha+d}-\varepsilon}}, \frac{k^{\frac{\beta}{d}}}{n^{\frac{\beta}{d}}}\biggr\}\biggr).$$

The calculation here is very similar, but we approximate $\log u_{x,s}$ simply by $\log(\frac{(n-1)s}{e^{\Psi(k)}f(x)})$. Writing $R_1', \ldots, R_5'$ for the modified error terms, we obtain

$$\mathbb{E}_f(\hat{H}_n) = H + \log(n-1) - \Psi(n) + \sum_{i=1}^5 R_i'.$$

Here, $R_1' = R_1 = o\{(\frac{k}{n})^{\alpha/(\alpha+d)-\varepsilon}\}$, and $R_2' = R_2 = o(n^{-(3-\varepsilon)})$, for every $\varepsilon > 0$ in both cases. On the other hand,

$$\begin{aligned}R_3' &= \int_{\mathcal{X}_n} f(x) \int_0^{\frac{a_n}{n-1}} \log\biggl(\frac{V_d f(x) h_x^{-1}(s)^d}{s}\biggr)\mathrm{B}_{k,n-k}(s)\,ds\,dx \\ &= O\biggl(\max\biggl\{\frac{k^{\frac{\alpha}{\alpha+d}-\varepsilon}}{n^{\frac{\alpha}{\alpha+d}-\varepsilon}}, \frac{k^{\beta/d}}{n^{\beta/d}}\biggr\}\biggr)\end{aligned}$$

for every $\varepsilon > 0$, by Lemma 10(ii). Similarly, for every $\varepsilon > 0$,

$$R_4' = \int_{\mathcal{X}_n} f(x) \int_{\frac{a_n}{n-1}}^1 \log\biggl(\frac{(n-1)s}{e^{\Psi(k)}f(x)}\biggr)\mathrm{B}_{k,n-k}(s)\,ds\,dx = o\bigl(n^{-(3-\varepsilon)}\bigr),$$

and $R_5' = R_5 = o\{(\frac{k}{n})^{\alpha/(\alpha+d)-\varepsilon}\}$. All of these bounds hold uniformly in $f \in \mathcal{F}_{d,\theta}$, so the claim is established for this setting.

Finally, consider now the case $d \geq 3$, $\alpha > 2d/(d-2)$ and $\beta > 4$. Again the calculation is very similar to the earlier cases, with the main difference being that

in bounding the error corresponding to R_3, we require a higher-order Taylor expansion of

$$\log\biggl(1+\frac{V_d f(x) h_x^{-1}(s)^d - s}{s}\biggr).$$

This can be done using Lemma 10(ii); we omit the details for brevity. □

PROOF OF COROLLARY 4. It is convenient to write $d' := \lfloor d/4 \rfloor + 1$ and $\beta' := \lceil \beta/2 \rceil - 1$. We have

$$\begin{aligned}
|\mathbb{E}_f(\hat{H}_n^w) - H| &= \biggl|\sum_{j=1}^k w_j \biggl\{\mathbb{E}_f(\log \xi_{(j),1}) - H - \sum_{l=1}^{\lfloor d/4 \rfloor} \frac{\Gamma(j+2l/d)\Gamma(n)}{\Gamma(j)\Gamma(n+2l/d)}\lambda_l\biggr\}\biggr| \\
&\le \biggl|\sum_{j=1}^k w_j \biggl\{\mathbb{E}_f(\log \xi_{(j),1}) - H - \sum_{l=1}^{\beta'} \frac{\Gamma(j+2l/d)\Gamma(n)}{\Gamma(j)\Gamma(n+2l/d)}\lambda_l\biggr\}\biggr| \\
&\quad + \biggl|\sum_{j=1}^k w_j \sum_{l=d'}^{\beta'} \frac{\Gamma(j+2l/d)\Gamma(n)}{\Gamma(j)\Gamma(n+2l/d)}\lambda_l\biggr|.
\end{aligned}$$

The first term can be bounded, uniformly for $f \in \mathcal{F}_{d,\theta}$ and $k \in \{1,\dots,k^*\}$, using Lemma 3. For the second term, we can use monotonicity properties of ratios of gamma functions to write

$$\begin{aligned}
&\biggl|\sum_{j=1}^k w_j \sum_{l=d'}^{\beta'} \frac{\Gamma(j+2l/d)\Gamma(n)}{\Gamma(j)\Gamma(n+2l/d)}\lambda_l\biggr| \\
&\quad\le \max_{d' \le \ell \le \beta'} |\lambda_\ell| \sum_{j=1}^k |w_j| \sum_{l=d'}^{\beta'} \frac{\Gamma(k+2l/d)\Gamma(n)}{\Gamma(k)\Gamma(n+2l/d)} \\
&\quad\le d^{1/2}\|w\|(\beta'-d'+1)\frac{\Gamma(k+2d'/d)\Gamma(n)}{\Gamma(k)\Gamma(n+2d'/d)} \max_{d' \le l \le \beta'} |\lambda_l| = O\biggl(\frac{k^{2d'/d}}{n^{2d'/d}}\biggr),
\end{aligned}$$

uniformly for $f \in \mathcal{F}_{d,\theta}$. The result follows. □

5.2. *Proof of Lemma 7.* Since this proof is long, we focus here on the main argument, and defer proofs of bounds on the many error terms to Appendix A.5.

PROOF OF LEMMA 7. We employ the same notation as in the proof of Lemma 3, except that we redefine δ_n so that $\delta_n := k c_n^d \log^3(n-1)/(n-1)$. We write $\mathcal{X}_n := \{x : f(x) \ge \delta_n\}$ for this newly-defined δ_n. Similar to the proof of Lemma 3, all error terms inside $O(\cdot)$ and $o(\cdot)$ that depend on k are uniform for $k \in \{k_0^*,\dots,k_1^*\}$, and we now adopt the additional convention that, where relevant,

these error terms are also uniform for $f \in \mathcal{F}_{d,\theta}$. By the nested properties of the classes $\mathcal{F}_{d,\theta}$ with respect to the smoothness parameter β, we may assume without loss of generality that $\beta \in (0, 1]$. We first deal with the variance of the unweighted estimator $\hat{H}_n$, and note that

$$\begin{aligned}\operatorname{Var}\hat{H}_n &= n^{-1}\operatorname{Var}\log\xi_1 + (1-n^{-1})\operatorname{Cov}(\log\xi_1, \log\xi_2)\\ &= n^{-1}\operatorname{Var}\log\xi_1 + (1-n^{-1})\big\{\operatorname{Cov}\big(\log(\xi_1 f(X_1)), \log(\xi_2 f(X_2))\big)\\ &\quad - 2\operatorname{Cov}\big(\log(\xi_1 f(X_1)), \log f(X_2)\big)\big\}. \end{aligned} \tag{16}$$

We claim that for every $\varepsilon > 0$,

$$\operatorname{Var}\log\xi_1 = V(f) + \frac{1}{k}\{1+o(1)\} + O\bigg\{\max\bigg(\frac{k^{\beta/d}}{n^{\beta/d}}\log n, \frac{k^{\frac{\alpha}{\alpha+d}-\varepsilon}}{n^{\frac{\alpha}{\alpha+d}-\varepsilon}}\bigg)\bigg\} \tag{17}$$

as $n \to \infty$. The proof of this claim uses similar methods to those in the proof of Lemma 3. In particular, writing $S_1, \ldots, S_5$ for remainder terms to be bounded later, we have

$$\begin{aligned}\mathbb{E}(\log^2\xi_1) &= \int_{\mathcal{X}} f(x)\int_0^\infty \log^2 u\, dF_{n,x}(u)\, dx\\ &= \int_{\mathcal{X}_n} f(x)\int_0^1 \mathrm{B}_{k,n-k}(s)\log^2 u_{x,s}\, ds\, dx + S_1\\ &= \int_{\mathcal{X}_n} f(x)\int_0^{\frac{a_n}{n-1}} \mathrm{B}_{k,n-k}(s)\log^2 u_{x,s}\, ds\, dx + S_1 + S_2\\ &= \int_{\mathcal{X}_n} f(x)\int_0^{\frac{a_n}{n-1}} \log^2\bigg(\frac{(n-1)s}{e^{\Psi(k)}f(x)}\bigg)\mathrm{B}_{k,n-k}(s)\, ds\, dx + S_1 + S_2 + S_3\\ &= \int_{\mathcal{X}_n} f(x)\big[\log^2 f(x) - 2\{\log(n-1) - \Psi(n)\}\log f(x)\\ &\qquad + \Psi'(k) - \Psi'(n) + \{\log(n-1) - \Psi(n)\}^2\big]dx + \sum_{i=1}^4 S_i\\ &= \int_{\mathcal{X}} f(x)\log^2 f(x)\, dx + \sum_{i=1}^5 S_i + \frac{1}{k}\{1+o(1)\}, \end{aligned} \tag{18}$$

as $n \to \infty$. In Appendix A.5.1, we show that for every $\varepsilon > 0$,

$$\sum_{i=1}^5 |S_i| = O\bigg\{\max\bigg(\frac{k^{\beta/d}}{n^{\beta/d}}\log n, \frac{k^{\frac{\alpha}{\alpha+d}-\varepsilon}}{n^{\frac{\alpha}{\alpha+d}-\varepsilon}}\bigg)\bigg\} \tag{19}$$

as $n \to \infty$. Combining (18) with (19) and Lemma 3, we deduce that (17) holds.

The next step of our proof consists of showing that for every $\varepsilon > 0$,

$$\begin{aligned}&\operatorname{Cov}\bigl(\log(\xi_1 f(X_1)), \log f(X_2)\bigr)\\ &\quad = O\biggl(\max\biggl\{\frac{k^{-\frac{1}{2}+\frac{2\alpha-\varepsilon}{\alpha+d}}}{n^{\frac{2\alpha-\varepsilon}{\alpha+d}}}, \frac{k^{\frac{1}{2}+\frac{\beta}{d}}}{n^{1+\frac{\beta}{d}}}\log^{2+\beta/d} n\biggr\}\biggr)\end{aligned} \tag{20}$$

as $n \to \infty$. Define

$$F^-_{n,x}(u) := \sum_{j=k}^{n-2} \binom{n-2}{j} p^j_{n,x,u}(1 - p_{n,x,u})^{n-2-j},$$

$$\tilde{F}_{n,x}(u) := \sum_{j=k-1}^{n-2} \binom{n-2}{j} p^j_{n,x,u}(1 - p_{n,x,u})^{n-2-j},$$

so that

$$\mathbb{P}(\xi_1 \le u | X_1 = x, X_2 = y) = \begin{cases} F^-_{n,x}(u) & \text{if } \|x-y\| > r_{n,u},\\ \tilde{F}_{n,x}(u) & \text{if } \|x-y\| \le r_{n,u}.\end{cases}$$

Writing $\tilde{u}_{n,x,y} := V_d(n-1)\|x-y\|^d e^{-\Psi(k)}$, we therefore have that

$$\begin{aligned}&\operatorname{Cov}\bigl(\log(\xi_1 f(X_1)), \log f(X_2)\bigr)\\ &\quad = \int_{\mathcal{X}\times\mathcal{X}} f(x)f(y)\log f(y) \int_{\tilde{u}_{n,x,y}}^{\infty} \log\bigl(uf(x)\bigr) d\bigl(\tilde{F}_{n,x} - F^-_{n,x}\bigr)(u)\,dx\,dy\\ &\qquad - H(f)\int_{\mathcal{X}} f(x)\int_0^{\infty} \log\bigl(uf(x)\bigr) d\bigl(F^-_{n,x} - F_{n,x}\bigr)(u)\,dx.\end{aligned} \tag{21}$$

To deal with the first term in (21), we make the substitution

$$y = y_{x,z} := x + \frac{r_{n,1}}{f(x)^{1/d}} z, \tag{22}$$

and let $d_n := (24\log n)^{1/d}$. Writing T_1, T_2, T_3 for remainder terms to be bounded later, for every $\varepsilon > 0$ and for $k \ge 2$,

$$\begin{aligned}&\int_{\mathcal{X}\times\mathcal{X}} f(x)f(y)\log f(y)\int_{\tilde{u}_{n,x,y}}^{\infty} \log\bigl(uf(x)\bigr) d\bigl(\tilde{F}_{n,x} - F^-_{n,x}\bigr)(u)\,dy\,dx\\ &\quad = r^d_{n,1}\int_{\mathcal{X}_n}\int_{B_0(d_n)} f(y_{x,z})\log f(y_{x,z})\\ &\qquad \times \int_{\frac{\|z\|^d}{f(x)}}^{\infty} \log\bigl(uf(x)\bigr) d\bigl(\tilde{F}_{n,x} - F^-_{n,x}\bigr)(u)\,dz\,dx + T_1\\ &\quad = r^d_{n,1}\int_{\mathcal{X}_n} f(x)\log f(x)\int_{B_0(d_n)}\int_{\frac{\|z\|^d}{f(x)}}^{\infty} \log\bigl(uf(x)\bigr) d\bigl(\tilde{F}_{n,x} - F^-_{n,x}\bigr)(u)\,dz\,dx\end{aligned} \tag{23}$$

$$
\begin{aligned}
&\quad + T_1 + T_2 \\
&= \frac{k-1}{n-k-1} \int_{\mathcal{X}_n} f(x) \log f(x)\, dx \\
&\quad \times \int_0^{\frac{a_n}{n-1}} \log\Big(\frac{(n-1)s}{e^{\Psi(k)}}\Big) \mathrm{B}_{k,n-k-1}(s) \Big(1 - \frac{(n-2)s}{k-1}\Big)\, ds + \sum_{i=1}^{3} T_i \\
&= \frac{H(f)}{n} + O\big(n^{-2}\big) + o\Big(\frac{k^{\frac{\alpha}{\alpha+d}-\varepsilon}}{n^{1+\frac{\alpha}{\alpha+d}-\varepsilon}}\Big) + \sum_{i=1}^{3} T_i.
\end{aligned}
$$

In Appendix A.5.2, we show that for every $\varepsilon > 0$,

$$
\sum_{i=1}^{3} |T_i| = O\Big(\max\Big\{\frac{k^{-\frac{1}{2}+\frac{2\alpha}{\alpha+d}-\varepsilon}}{n^{\frac{2\alpha}{\alpha+d}-\varepsilon}}, \frac{k^{\frac{1}{2}+\frac{\beta}{d}}}{n^{1+\frac{\beta}{d}}} \log^{2+\beta/d} n\Big\}\Big) \tag{24}
$$

as $n \to \infty$. We now deal with the second term in (21). Writing U_1, U_2 for remainder terms to be bounded later, for every $\varepsilon > 0$,

$$
\begin{aligned}
&\int_{\mathcal{X}} f(x) \int_0^{\infty} \log\big(uf(x)\big) d\big(F^-_{n,x} - F_{n,x}\big)(u)\, dx \\
&\quad = \int_{\mathcal{X}_n} f(x) \int_0^{\frac{a_n}{n-1}} \log\big(u_{x,s} f(x)\big) \mathrm{B}_{k,n-k-1}(s) \Big\{\frac{(n-1)s-k}{n-k-1}\Big\}\, ds\, dx + U_1 \\
&\quad = \int_{\mathcal{X}_n} f(x) \int_0^{1} \log\Big(\frac{(n-1)s}{e^{\Psi(k)}}\Big) \mathrm{B}_{k,n-k-1}(s) \Big\{\frac{(n-1)s-k}{n-k-1}\Big\}\, ds\, dx \qquad (25) \\
&\qquad + U_1 + U_2 \\
&\quad = \frac{1}{n-1} + U_1 + U_2 + o\Big(\frac{k^{\frac{\alpha}{\alpha+d}-\varepsilon}}{n^{1+\frac{\alpha}{\alpha+d}-\varepsilon}}\Big).
\end{aligned}
$$

In Appendix A.5.3, we show that for every $\varepsilon > 0$,

$$
|U_1| + |U_2| = O\Big(\frac{k^{1/2}}{n} \max\Big\{\frac{k^{\beta/d}}{n^{\beta/d}}, \frac{k^{\frac{\alpha}{\alpha+d}-\varepsilon}}{n^{\frac{\alpha}{\alpha+d}-\varepsilon}}\Big\}\Big). \tag{26}
$$

From (21), (23), (24), (25) and (26), we conclude that (20) holds.

By (16), it remains to consider $\mathrm{Cov}(\log(\xi_1 f(X_1)), \log(\xi_2 f(X_2)))$. We require some further notation. Let $F_{n,x,y}$ denote the conditional distribution function of (ξ_1, ξ_2) given $X_1 = x$, $X_2 = y$. Let $a_n^- := (k - 3k^{1/2}\log^{1/2} n) \vee 0$, $a_n^+ := (k + 3k^{1/2}\log^{1/2} n) \wedge (n-1)$, and let

$$
\begin{aligned}
v_x &:= \inf\{u \geq 0 : (n-1)p_{n,x,u} = a_n^+\}, \\
l_x &:= \inf\{u \geq 0 : (n-1)p_{n,x,u} = a_n^-\},
\end{aligned}
$$

so that $\mathbb{P}\{\xi_1 \le l_{X_1}\} = o(n^{-(9/2-\varepsilon)})$ and $\mathbb{P}\{\xi_1 \ge v_{X_1}\} = o(n^{-(9/2-\varepsilon)})$ for every $\varepsilon > 0$. For pairs (u, v) with $u \le v_x$ and $v \le v_y$, let $(M_1, M_2, M_3) \sim \text{Multi}(n-2; p_{n,x,u}, p_{n,y,v}, 1 - p_{n,x,u} - p_{n,y,v})$, and write

$$G_{n,x,y}(u, v) := \mathbb{P}(M_1 \ge k, M_2 \ge k),$$

so that $F_{n,x,y}(u, v) = G_{n,x,y}(u, v)$ for $\|x - y\| > r_{n,u} + r_{n,v}$. Write

$$\Sigma := \begin{pmatrix} 1 & \alpha_z \\ \alpha_z & 1 \end{pmatrix}$$

with $\alpha_z := V_d^{-1}\mu_d(B_0(1) \cap B_z(1))$ for $z \in \mathbb{R}^d$, let $\Phi_\Sigma(s, t)$ denote the distribution function of a $N_2(0, \Sigma)$ random vector at (s, t), and let Φ denote the standard univariate normal distribution function. Writing W_i for remainder terms to be bounded later, and writing $h(u, v) := \log(uf(x))\log(vf(y))$ as shorthand, we have

$$\begin{aligned}
&\operatorname{Cov}\big(\log(\xi_1 f(X_1)), \log(\xi_2 f(X_2))\big) \\
&\quad = \int_{\mathcal{X}\times\mathcal{X}} f(x)f(y) \int_0^\infty \int_0^\infty h(u, v)d(F_{n,x,y} - F_{n,x}F_{n,y})(u, v)\,dx\,dy \\
&\quad = \int_{\mathcal{X}\times\mathcal{X}} f(x)f(y) \int_{[l_x,v_x]\times[l_y,v_y]} h(u, v)d(F_{n,x,y} - F_{n,x}F_{n,y})(u, v)\,dx\,dy \\
&\qquad + W_1 \\
&\quad = \int_{\mathcal{X}\times\mathcal{X}} f(x)f(y) \int_{[l_x,v_x]\times[l_y,v_y]} h(u, v)d(F_{n,x,y} - G_{n,x,y})(u, v)\,dx\,dy \\
&\qquad - \frac{1}{n} + \sum_{i=1}^{2} W_i \qquad (27) \\
&\quad = \int_{\mathcal{X}_n\times\mathcal{X}} f(x)f(y) \int_{l_x}^{v_x} \int_{l_y}^{v_y} \frac{(F_{n,x,y} - G_{n,x,y})(u, v)}{uv}\,du\,dv\,dx\,dy - \frac{1}{n} \\
&\qquad + \sum_{i=1}^{3} W_i \\
&\quad = \frac{r_{n,1}^d}{k} \int_{B_0(2)} \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \big\{\Phi_\Sigma(s, t) - \Phi(s)\Phi(t)\big\}\,ds\,dt\,dz - \frac{1}{n} + \sum_{i=1}^{4} W_i \\
&\quad = \frac{e^{\Psi(k)}}{k(n-1)} - \frac{1}{n} + \sum_{i=1}^{4} W_i = O\Big(\frac{1}{nk}\Big) + \sum_{i=1}^{4} W_i.
\end{aligned}$$

The proof in the unweighted case is completed by showing in Appendix A.5.4 that for every $\varepsilon > 0$,

$$\sum_{i=1}^{4} |W_i| = O\bigg(\max\bigg\{\frac{\log^{\frac{5}{2}} n}{nk^{\frac{1}{2}}}, \frac{k^{\frac{3}{2}+\frac{\alpha-\varepsilon}{\alpha+d}}}{n^{1+\frac{\alpha-\varepsilon}{\alpha+d}}}, \frac{k^{\frac{3}{2}+\frac{2\beta}{d}}}{n^{1+\frac{2\beta}{d}}}, \frac{k^{(1+\frac{d}{2\beta})\frac{\alpha-\varepsilon}{\alpha+d}}}{n^{1+\frac{\alpha-\varepsilon}{\alpha+d}}}, \frac{k^{\frac{1}{2}+\frac{\beta}{d}}\log n}{n^{1+\frac{\beta}{d}}}, \frac{k^{\frac{2\alpha-\varepsilon}{\alpha+d}}}{n^{\frac{2\alpha-\varepsilon}{\alpha+d}}}\bigg\}\bigg)$$

as $n \to \infty$.

The proof in the weighted case uses similar arguments; details are deferred to Appendix A.5.4. □

5.3. *Proofs of Theorems 1 and 2.*

PROOF OF THEOREM 1. Writing $j_t := \lfloor tk/d \rfloor$ for $t = 1, \dots, d$ and $d' := \lfloor d/4 \rfloor + 1$ for convenience, a sufficient condition for $\mathcal{W}^{(k)} \neq \varnothing$ is that the matrix $A^{(k)} \in \mathbb{R}^{d' \times d'}$ with (l,t)th entry

$$A^{(k)}_{lt} = \Gamma(j_t)^{-1}\Gamma\bigl(j_t + 2(l-1)/d\bigr)k^{-2(l-1)/d},$$

is invertible. This follows because, writing $e_1 := (1, 0, \dots, 0)^T \in \mathbb{R}^{d'}$ we can then define $w = w^{(k)} \in \mathcal{W}^{(k)}$ by setting

$$(w_{j_t})_{t=1}^{\lfloor d/4 \rfloor + 1} := \bigl(A^{(k)}\bigr)^{-1} e_1$$

and setting all other entries of w to be zero. Now define $A \in \mathbb{R}^{d' \times d'}$ to have (l,t)th entry $A_{lt} := (t/d)^{2(l-1)/d}$. Since $x^{-a}\Gamma(x)^{-1}\Gamma(x+a) \to 1$ as $x \to \infty$ for $a \in \mathbb{R}$, we have $\|A^{(k)} - A\| \to 0$ as $k \to \infty$. Now, A is a Vandermonde matrix (depending only on d) and as such has determinant

$$|A| = \prod_{1 \le t_1 < t_2 \le d'} d^{-2/d}\bigl(t_2^{2/d} - t_1^{2/d}\bigr) > 0.$$

Hence, by the continuity of the determinant and eigenvalues of a matrix, we have that there exists $k_d > 0$ such that, for $k \ge k_d$, the matrix $A^{(k)}$ is invertible and

$$\bigl\|\bigl(A^{(k)}\bigr)^{-1} e_1\bigr\| \le \bigl|\lambda_{\min}\bigl(A^{(k)}\bigr)\bigr|^{-1} \le 2\bigl|\lambda_{\min}(A)\bigr|^{-1},$$

where $\lambda_{\min}(\cdot)$ denotes the eigenvalue of a matrix with smallest absolute value. It follows that, for each $k \ge k_d$, there exists $w^{(k)} \in \mathcal{W}^{(k)}$ satisfying $\sup_{k \ge k_d} \|w^{(k)}\| < \infty$, as required.

Now, by Corollary 4 and the fact that $w \in \mathcal{W}^{(k)}$, we have for $\varepsilon > 0$ sufficiently small,

$$\mathbb{E}_f\bigl(\hat{H}_n^w\bigr) - H(f) = O\biggl(\max\biggl\{\frac{k^{\frac{\alpha}{\alpha+d}-\varepsilon}}{n^{\frac{\alpha}{\alpha+d}-\varepsilon}}, \frac{k^{\frac{2d'}{d}}}{n^{\frac{2d'}{d}}}, \frac{k^{\frac{\beta}{d}}}{n^{\frac{\beta}{d}}}\biggr\}\biggr) = o\bigl(n^{-1/2}\bigr),$$

uniformly for $f \in \mathcal{F}_{d,\theta}$, under our conditions on k_1^*, α and β. By Lemma 7, we have $\operatorname{Var} \hat{H}_n^w = n^{-1}V(f) + o(n^{-1})$ uniformly for $f \in \mathcal{F}_{d,\theta}$. Note that by Cauchy–Schwarz, very similar arguments to those used at (18) and Lemma 1 in the Supplementary Material we have that, for $j \in \operatorname{supp}(w)$,

$$\bigl|\operatorname{Cov}_f\bigl(\log\bigl(\xi_{(j),1} f(X_1)\bigr), \log f(X_1)\bigr)\bigr| \le \bigl\{V(f)\mathbb{E}_f\bigl[\log^2\bigl(\xi_{(j),1} f(X_1)\bigr)\bigr]\bigr\}^{1/2} \to 0$$

uniformly for $f \in \mathcal{F}_{d,\theta}$. Therefore, also using (20), we have that

$$\begin{aligned}\operatorname{Var}_f\bigl(\hat{H}_n^w - H_n^*\bigr) &= \operatorname{Var}_f \hat{H}_n^w + 2\operatorname{Cov}_f\bigl(\hat{H}_n^w, \log f(X_1)\bigr) + n^{-1}V(f)\\ &= \operatorname{Var}_f \hat{H}_n^w - n^{-1}V(f)\\ &\quad + \frac{2}{n}\sum_{j=1}^{k} w_j \operatorname{Cov}_f\bigl(\log\bigl(\xi_{(j),1} f(X_1)\bigr), \log f(X_1)\bigr)\\ &\quad + 2\bigl(1-n^{-1}\bigr)\sum_{j=1}^{k} w_j \operatorname{Cov}\bigl(\log\bigl(\xi_{(j),2} f(X_2)\bigr), \log f(X_1)\bigr)\\ &= o\bigl(n^{-1}\bigr)\end{aligned}$$

as $n \to \infty$, uniformly for $f \in \mathcal{F}_{d,\theta}$. The conclusion (3) follows on writing

$$\mathbb{E}_f\bigl\{\bigl(\hat{H}_n^w - H_n^*\bigr)^2\bigr\} = \operatorname{Var}_f\bigl(\hat{H}_n^w - H_n^*\bigr) + \bigl(\mathbb{E}_f \hat{H}_n^w - H(f)\bigr)^2,$$

and the final conclusion is then immediate. □

PROOF OF THEOREM 2. First, note that by the final conclusion of Theorem 1 and by Lemma 1(i) of the Supplementary Material, there exists $n_0 \in \mathbb{N}$, depending only on d and θ, such that

$$\sup_{k\in\{k_0^*,\dots,k_1^*\}} \sup_{f\in\mathcal{F}_{d,\theta}} n\mathbb{E}\bigl[\bigl\{\hat{H}_n^w - H(f)\bigr\}^2\bigr] < \infty$$

for $n \ge n_0$. Now let $\mathcal{H}$ denote the class of Lipschitz functions $h : \mathbb{R} \to \mathbb{R}$ with $|h(x) - h(y)| \le |x - y|$ for all $x, y \in \mathbb{R}$. By the Kantorovič–Rubinšteĭn theorem [Kantorovič and Rubinšteĭn (1958), Kellerer (1985)], we have for $n \ge n_0$ that

$$\begin{aligned}&d_1\bigl(\mathcal{L}\bigl(n^{1/2}\bigl\{\hat{H}_n^w - H(f)\bigr\}\bigr), \mathcal{L}\bigl(n^{1/2}\bigl\{H_n^* - H(f)\bigr\}\bigr)\bigr)\\ &\quad= \sup_{h\in\mathcal{H}}\bigl|\mathbb{E}_f h\bigl(n^{1/2}\bigl\{\hat{H}_n^w - H(f)\bigr\}\bigr) - \mathbb{E}h\bigl(n^{1/2}\bigl\{H_n^* - H(f)\bigr\}\bigr)\bigr|\\ &\quad\le \sup_{h\in\mathcal{H}}\mathbb{E}_f\bigl|h\bigl(n^{1/2}\bigl\{\hat{H}_n^w - H(f)\bigr\}\bigr) - h\bigl(n^{1/2}\bigl\{H_n^* - H(f)\bigr\}\bigr)\bigr|\\ &\quad\le n^{1/2}\mathbb{E}_f\bigl|\hat{H}_n^w - H_n^*\bigr| \le n^{1/2}\bigl[\mathbb{E}_f\bigl\{\bigl(\hat{H}_n^w - H_n^*\bigr)^2\bigr\}\bigr]^{1/2}.\end{aligned} \tag{28}$$

Now let $Z \sim N(0, V(f))$. Then by the Wasserstein central limit theorem [e.g., Barbour and Chen (2005), Theorem 3.2],

$$d_1\bigl(\mathcal{L}\bigl(n^{1/2}\bigl\{H_n^* - H(f)\bigr\}\bigr), N\bigl(0, V(f)\bigr)\bigr) \le \frac{3\beta_3(f)}{n^{1/2}V(f)}, \tag{29}$$

where

$$\beta_3(f) := \mathbb{E}_f\bigl\{\bigl|\log f(X_1) + H(f)\bigr|^3\bigr\} = \int_{\mathcal{X}} f(x)\bigl|\log f(x) + H(f)\bigr|^3\,dx.$$

We deduce from (28) and (29), together with Theorem 1 and Lemma 1 in the the Supplementary Material, that

$$\sup_{k\in\{k_0^*,\ldots,k_1^*\}}\sup_{f\in\mathcal{F}_{d,\theta}} d_1\big(\mathcal{L}\big(n^{1/2}\{\hat{H}_n^w - H(f)\}\big), N\big(0, V(f)\big)\big) \to 0$$

as $n\to\infty$. But the final conclusion of Theorem 1 then allows us to replace d_1 with d_2 in this convergence statement, and this completes the proof of the first part of the theorem.

For the second part, set

$$\varepsilon_n = \varepsilon_n^w(d,\theta) := \frac{\sup_{k\in\{k_0^*,\ldots,k_1^*\}}\sup_{f\in\mathcal{F}_{d,\theta}}(2\mathbb{E}_f[\{\tilde{V}_n^w - V(f)\}^2])^{1/3}}{\inf_{f\in\mathcal{F}_{d,\theta}} V(f)^{2/3}},$$

so that $\varepsilon_n\to 0$, by Lemmas 1(ii) and 3 in the Supplementary Material. Then, by two applications of Markov's inequality, for n large enough that $\varepsilon_n\le 1$,

$$\begin{aligned}\mathbb{P}_f\bigg(\bigg|\frac{(\hat{V}_n^w)^{1/2}}{V^{1/2}(f)} - 1\bigg| \ge \varepsilon_n\bigg) &\le \mathbb{P}_f\bigg(\bigg|\frac{\tilde{V}_n^w}{V(f)} - 1\bigg| \ge \varepsilon_n\bigg) + \mathbb{P}_f\big(\tilde{V}_n^w \le 0\big)\\ &\le \frac{\mathbb{E}_f[\{\tilde{V}_n^w - V(f)\}^2]}{V(f)^2}\bigg(\frac{1}{\varepsilon_n^2} + 1\bigg) \le \varepsilon_n.\end{aligned}$$

For $n\in\mathbb{N}$ and $L\ge 1$, define $h_{n,L}:\mathbb{R}\to[0,1]$ by

$$h_{n,L}(x) := \begin{cases} 0 & \text{if } |x| > z_{q/2}(1+\varepsilon_n) + 1/L,\\ L\{z_{q/2}(1+\varepsilon_n) + 1/L - |x|\} & \text{if } 0 < |x| - z_{q/2}(1+\varepsilon_n) \le 1/L,\\ 1 & \text{if } |x| \le z_{q/2}(1+\varepsilon_n).\end{cases}$$

Thus $h_{n,L}$ has Lipschitz constant L and $h_{n,L}(x)\ge \mathbb{1}_{\{|x|\le z_{q/2}(1+\varepsilon_n)\}}$. Then, with $Z\sim N(0,1)$ and for large n,

$$\begin{aligned}&\mathbb{P}_f\big(I_{n,q}\ni H(f)\big)\\ &\quad\le \mathbb{P}_f\bigg(\frac{n^{1/2}|\hat{H}_n^w - H(f)|}{V^{1/2}(f)} \le z_{q/2}(1+\varepsilon_n)\bigg) + \mathbb{P}_f\bigg(\frac{V^{1/2}(f)}{(\hat{V}_n^w)^{1/2}} \le \frac{1}{1+\varepsilon_n}\bigg)\\ &\quad\le \mathbb{E}_f h_{n,L}\bigg(\frac{n^{1/2}\{\hat{H}_n^w - H(f)\}}{V^{1/2}(f)}\bigg) + \varepsilon_n\\ &\quad\le \mathbb{E}_f h_{n,L}(Z) + \varepsilon_n + L d_1\bigg(\mathcal{L}\bigg(\frac{n^{1/2}\{\hat{H}_n^w - H(f)\}}{V^{1/2}(f)}\bigg), \mathcal{L}(Z)\bigg)\\ &\quad\le \mathbb{P}\big(|Z| \le z_{q/2}(1+\varepsilon_n) + L^{-1}\big) + \varepsilon_n\\ &\qquad + \frac{L}{V^{1/2}(f)} d_1\big(\mathcal{L}\big(n^{1/2}\big(\hat{H}_n^w - H(f)\big)\big), N\big(0, V(f)\big)\big).\end{aligned}$$

Since $L \geq 1$ was arbitrary, we deduce from the first part of the theorem and Lemma 1(ii) in the Supplementary Material that

$$\limsup_{n\to\infty} \sup_{q\in(0,1)} \sup_{k\in\{k_0^*,\ldots,k_1^*\}} \sup_{f\in\mathcal{F}_{d,\theta}} \mathbb{P}_f\big(I_{n,q} \ni H(f)\big) - (1-q) \leq \inf_{L\geq 1} \frac{2}{L(2\pi)^{1/2}} = 0.$$

The lower bound is obtained by a similar argument, omitted for brevity. □

Acknowledgements. We thank the reviewers for constructive feedback on an earlier draft. The second author is grateful to Sebastian Nowozin for introducing him to this problem, and to Gérard Biau for helpful discussions.

SUPPLEMENTARY MATERIAL

Supplement to "Efficient multivariate entropy estimation via k-nearest neighbour distances". (DOI: 10.1214/18-AOS1688SUPP; .pdf). Auxiliary results and remaining proofs.

REFERENCES

BARBOUR, A. D. and CHEN, L. H. Y., eds. (2005). *An Introduction to Stein's Method. Lecture Notes Series. Institute for Mathematical Sciences. National University of Singapore* **4**. Singapore Univ. Press, Singapore. MR2235447

BEIRLANT, J., DUDEWICZ, E. J., GYÖRFI, L. and VAN DER MEULEN, E. C. (1997). Nonparametric entropy estimation: An overview. *Int. J. Math. Stat. Sci.* **6** 17–39. MR1471870

BERRETT, T. B., SAMWORTH, R. J. and YUAN, M. (2019). Supplement to "Efficient multivariate entropy estimation via k-nearest neighbour distances." DOI:10.1214/18-AOS1688SUPP.

BIAU, G. and DEVROYE, L. (2015). *Lectures on the Nearest Neighbor Method*. Springer, Cham. MR3445317

CAI, T. T. and LOW, M. G. (2011). Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *Ann. Statist.* **39** 1012–1041. MR2816346

CRESSIE, N. (1976). On the logarithms of high-order spacings. *Biometrika* **63** 343–355. MR0428583

DELATTRE, S. and FOURNIER, N. (2017). On the Kozachenko–Leonenko entropy estimator. *J. Statist. Plann. Inference* **185** 69–93. MR3612672

EL HAJE HUSSEIN, F. and GOLUBEV, YU. (2009). On entropy estimation by m-spacing method. *Zap. Nauchn. Sem. S.-Peterburg. Otdel. Mat. Inst. Steklov.* (*POMI*) **363** 151–181, 189. MR2749122

GAO, W., OH, S. and VISWANATH, P. (2016). Demystifying fixed k-nearest neighbor information estimators. Available at arXiv:1604.03006.

GORIA, M. N., LEONENKO, N. N., MERGEL, V. V. and NOVI INVERARDI, P. L. (2005). A new class of random vector entropy estimators and its applications in testing statistical hypotheses. *J. Nonparametr. Stat.* **17** 277–297. MR2129834

HALL, P. and MORTON, S. C. (1993). On the estimation of entropy. *Ann. Inst. Statist. Math.* **45** 69–88. MR1220291

KANTOROVIČ, L. V. and RUBINŠTEĬN, G. Š. (1958). On a space of completely additive functions. *Vestnik Leningrad Univ. Math.* **13** 52–59. MR0102006

KELLERER, H. G. (1985). Duality theorems and probability metrics. In *Proceedings of the Seventh Conference on Probability Theory* (*Braşov*, 1982) 211–220. VNU Sci. Press, Utrecht. MR0867434

KOZACHENKO, L. F. and LEONENKO, N. N. (1987). Sample estimate of the entropy of a random vector. *Probl. Inf. Transm.* **23** 95–101.

KWAK, N. and CHOI, C. (2002). Input feature selection by mutual information based on Parzen window. *IEEE Trans. Pattern Anal. Mach. Intell.* **24** 1667–1671.

LAURENT, B. (1996). Efficient estimation of integral functionals of a density. *Ann. Statist.* **24** 659–681. MR1394981

LEARNED-MILLER, E. G. and FISHER, J. W. III (2004). ICA using spacings estimates of entropy. *J. Mach. Learn. Res.* **4** 1271–1295. MR2103630

LEPSKI, O., NEMIROVSKI, A. and SPOKOINY, V. (1999). On estimation of the L_r norm of a regression function. *Probab. Theory Related Fields* **113** 221–253. MR1670867

MNATSAKANOV, R. M., MISRA, N., LI, SH. and HARNER, E. J. (2008). k_n-nearest neighbor estimators of entropy. *Math. Methods Statist.* **17** 261–277. MR2448950

MOON, K. R., SRICHARAN, K., GREENEWALD, K. and HERO, A. O. (2016). Nonparametric ensemble estimation of distributional functionals. https://arxiv.org/abs/1601.06884v2.

PANINSKI, L. (2003). Estimation of entropy and mutual information. *Neural Comput.* **15** 1191–1253.

PANINSKI, L. and YAJIMA, M. (2008). Undersmoothed kernel entropy estimators. *IEEE Trans. Inform. Theory* **54** 4384–4388. MR2451978

SHORACK, G. R. and WELLNER, J. A. (2009). *Empirical Processes with Applications to Statistics. Classics in Applied Mathematics* **59**. Society for Industrial and Applied Mathematics (SIAM), Philadelphia, PA. MR3396731

SINGH, S. and PÓCZOS, B. (2016). Analysis of k nearest neighbor distances with application to entropy estimation. *NIPS* **29** 1217–1225.

SINGH, H., MISRA, N., HNIZDO, V., FEDOROWICZ, A. and DEMCHUK, E. (2003). Nearest neighbor estimates of entropy. *Amer. J. Math. Management Sci.* **23** 301–321. MR2045530

SRICHARAN, K., WEI, D. and HERO, A. O. III (2013). Ensemble estimators for multivariate entropy estimation. *IEEE Trans. Inform. Theory* **59** 4374–4388. MR3071335

TSYBAKOV, A. B. and VAN DER MEULEN, E. C. (1996). Root-n consistent estimators of entropy for densities with unbounded support. *Scand. J. Stat.* **23** 75–83. MR1380483

VAN DER VAART, A. W. (1998). *Asymptotic Statistics. Cambridge Series in Statistical and Probabilistic Mathematics* **3**. Cambridge Univ. Press, Cambridge. MR1652247

VASICEK, O. (1976). A test for normality based on sample entropy. *J. Roy. Statist. Soc. Ser. B* **38** 54–59. MR0420958

T. B. BERRETT
R. J. SAMWORTH
STATISTICAL LABORATORY
UNIVERSITY OF CAMBRIDGE
WILBERFORCE ROAD
CAMBRIDGE
CB3 0WB
UNITED KINGDOM
E-MAIL: t.berrett@statslab.cam.ac.uk
r.samworth@statslab.cam.ac.uk
URL: http://www.statslab.cam.ac.uk/~tbb26
http://www.statslab.cam.ac.uk/~rjs57

M. YUAN
DEPARTMENT OF STATISTICS
UNIVERSITY OF WISCONSIN–MADISON
MEDICAL SCIENCES CENTER
1300 UNIVERSITY AVENUE
MADISON, WISCONSI 53706
USA
E-MAIL: myuan@stat.wisc.edu
URL: http://pages.stat.wisc.edu/~myuan/