*The Annals of Statistics*
2023, Vol. 51, No. 2, 668–690
https://doi.org/10.1214/23-AOS2265


# EFFICIENT FUNCTIONAL ESTIMATION AND THE SUPER-ORACLE PHENOMENON

BY THOMAS B. BERRETT[1,a] AND RICHARD J. SAMWORTH[2,b]

[1]*Department of Statistics, University of Warwick,* [a]*tom.berrett@warwick.ac.uk*
[2]*Statistical Laboratory, University of Cambridge,* [b]*r.samworth@statslab.cam.ac.uk*

We consider the estimation of two-sample integral functionals, of the type that occur naturally, for example, when the object of interest is a divergence between unknown probability densities. Our first main result is that, in wide generality, a weighted nearest neighbour estimator is efficient, in the sense of achieving the local asymptotic minimax lower bound. Moreover, we also prove a corresponding central limit theorem, which facilitates the construction of asymptotically valid confidence intervals for the functional, having asymptotically minimal width. One interesting consequence of our results is the discovery that, for certain functionals, the worst-case performance of our estimator may improve on that of the natural 'oracle' estimator, which itself can be optimal in the related problem where the data consist of the values of the unknown densities at the observations.

**1. Introduction.** This paper concerns the estimation of two-sample density functionals of the form

$$T = T(f,g) := \int_{\mathcal{X}} f(x)\phi\big(f(x), g(x)\big)\,dx, \tag{1}$$

where $\mathcal{X} := \{x \in \mathbb{R}^d : f(x) > 0, g(x) > 0\}$, based on independent $d$-dimensional random vectors $X_1, \ldots, X_m, Y_1, \ldots, Y_n$, where $X_1, \ldots, X_m$ have density $f$ and $Y_1, \ldots, Y_n$ have density $g$. The interest in the estimation of such functionals arises from many applications: for instance, many divergences such as the Kullback–Leibler divergence, total variation and Hellinger distances (or more generally, all $\varphi$-divergences) are of this form. The estimation of such divergences is important for two-sample testing (Wornowizki and Fried (2016)), registration problems in image analysis (Hero et al. (2002)) and generative adversarial networks (Nowozin, Cseke and Tomioka (2016)), to name just a few examples. Of course, we can regard the problem of estimation of one-sample density functionals

$$H(f) := \int_{\{x: f(x)>0\}} f(x)\psi\big(f(x)\big)\,dx, \tag{2}$$

which include Shannon and Rényi entropies, as a special case.

Motivated by these applications, the estimation of the two-sample functional (1) (or closely related quantities) has received considerable attention in the literature recently (e.g., Kandasamy et al. (2015), Krishnamurthy et al. (2014), Moon et al. (2018), Singh and Póczos (2016), Singh, Sriperumbudur and Póczos (2018)). Naturally, the one-sample version of the problem, and special cases of it, have been highly-studied subjects over several decades (e.g., Beirlant et al. (1997), Berrett, Samworth and Yuan (2019), Biau and Devroye (2015), Bickel and Ritov (1988), Birgé and Massart (1995), Han et al. (2020), Kozachenko and Leonenko (1987), Laurent (1996), Leonenko, Pronzato and Savani (2008), Leonenko and Seleznjev

Received April 2020; revised August 2022.
*MSC2020 subject classifications.* 62G05, 62G20.
*Key words and phrases.* Efficiency, functional estimation, weighted nearest neighbours.

(2010)). It turns out that many functionals of interest involve functions $\phi$ in (1) that are nonsmooth as their arguments approach zero, or functions $\psi$ in (2) that are nonsmooth as their argument vanishes. For instance, for the Shannon entropy, $\psi(y) = -\log y$, while the Rényi entropy of order $\kappa$ is essentially equivalent to $\psi(y) = y^{\kappa-1}$, which is nonsmooth as $y \to 0$ when $\kappa \in (0, 1)$. To avoid problems caused by this lack of smoothness, many of the aforementioned authors assume that the density $f$ is bounded away from zero on its (compact) support. In that case, *efficient* estimators can sometimes be obtained; to give just one example, when $f$ is also $s$-Hölder smooth on $\{x : f(x) > 0\}$ with $s > d/4$, Laurent (1996) obtained a Shannon entropy estimator $\widehat{H}_m$ satisfying

$$m\mathbb{E}\big[\big\{\widehat{H}_m - H(f)\big\}^2\big] \to \int_{\{x:f(x)>0\}} f \log^2 f - H(f)^2. \tag{3}$$

The limit in (3) is the asymptotic rescaled mean squared error of the oracle estimator $H_m^* := -m^{-1}\sum_{i=1}^m \log f(X_i)$, and is optimal in a local asymptotic minimax sense (Ibragimov and Khas'minskiĭ (1991), Laurent (1996)).

However, the assumption that the density $f$ is bounded away from zero on its support is made purely for mathematical convenience; it assumes away the essential difficulty of the problem caused by the nonsmoothness and rules out many standard densities of common interest. In the related problem of density estimation, it is known that, depending on the loss function and the smoothness of the densities considered, optimal rates of convergence can be very different when densities with unbounded support are allowed (Donoho et al. (1996), Goldenshluger and Lepski (2014), Juditsky and Lambert-Lacroix (2004)).

It is therefore of great interest to understand the ways in which low density regions interact with the potential nonsmoothness of the functional to determine the behaviour of estimators. Previous works in this direction have tended to focus on specific functionals and on rates of convergence (e.g., Han et al. (2020), Tsybakov and van der Meulen (1996)). By contrast, in this work our aim is to provide a class of estimators that are efficient for a wide spectrum of functionals. Our estimators will be deterministically weighted versions of preliminary estimators based on nearest neighbour distances. To set the scene, for integers $k_X \in \{1, \dots, m-1\}$ and $k_Y \in \{1, \dots, n\}$, write $\rho_{(k_X),i,X}$ for the (Euclidean) distance between $X_i$ and its $k_X$th nearest neighbour in the sample $X_1, \dots, X_{i-1}, X_{i+1}, \dots, X_m$, and write $\rho_{(k_Y),i,Y}$ for the distance between $X_i$ and its $k_Y$th nearest neighbour in the sample $Y_1, \dots, Y_n$. The starting point for the construction of our estimators is the approximation

$$f(X_i) V_d \rho_{(k_X),i,X}^d \approx k_X/m,$$

where $V_d := \pi^{d/2}/\Gamma(1 + d/2)$ denotes the $d$-dimensional Lebesgue measure of the unit Euclidean ball in $\mathbb{R}^d$; this arises by comparing the proportion of points in a ball of radius $\rho_{(k_X),i,X}$ about $X_i$ with a local constant approximation to the probability content of the same ball. This motivates the initial estimator

$$\widetilde{T}_{m,n} = \widetilde{T}_{m,n,k_X,k_Y} := \frac{1}{m}\sum_{i=1}^m \phi\Big(\frac{k_X}{m V_d \rho_{(k_X),i,X}^d}, \frac{k_Y}{n V_d \rho_{(k_Y),i,Y}^d}\Big). \tag{4}$$

Restricting attention for simplicity of exposition to the one-sample analogue $\widetilde{T}_m = \widetilde{T}_{m,k}$ of (4) that simply replaces $\phi(\cdot,\cdot)$ with $\psi(\cdot)$ and $k_X$ with $k$, it has long been known in the special case of the Shannon entropy functional that one should debias $\widetilde{T}_m$ by replacing $k$ with $e^{\Psi(k)}$, where $\Psi(\cdot)$ denotes the digamma function (Kozachenko and Leonenko (1987)). This amounts to adding $\log k - \Psi(k)$ to the original estimator. Ryu et al. (2018) argued that for general two-sample functionals, the estimator (4) can be debiased to leading order via an implicit inverse Laplace transform, and showed that this has an explicit expression in certain examples. It turns out, however, that even the remaining bias is large enough to preclude

efficient estimation when $d \geq 4$, and this motivates us to consider weighted linear combinations of estimators of the form (4) over different choices of $k_X$ and $k_Y$, where the weights are chosen to cancel sufficient terms in the bias expansion. A subtle question concerns the issue of whether to apply our weights to the original estimators (4) or their debiased versions. We address this by using fractional calculus techniques to provide an explicit expression for the leading order remaining bias of the debiased estimators. We conclude that, in general, the gain from the fact that fewer nonzero weights are required to obtain an efficient estimator when applying these weights to the debiased estimator is outweighed by the added complication of the resulting estimator. However, in special cases such as the Kullback–Leibler and Rényi divergences, where the correct explicit debiasing terms are available, the weighting scheme simplifies and we advocate applying the weights to the debiased estimator.

Returning to the general case, our final estimators $\widehat{T}_{m,n}$ are based on weighted averages of estimators of the form $\widetilde{T}_{m,n,k_X,k_Y}$ for different choices of $k_X$ and $k_Y$; such estimators are attractive because they generalise easily to multivariate cases (unlike, for example, estimators based on sample spacings), and because they are straightforward to compute. Our first main result (Theorem 2 in Section 2), reveals that the dominant asymptotic contribution to the squared error risk of $\widehat{T}_{m,n}$ is of the form $v_1/m + v_2/n$ as $m, n \to \infty$, uniformly over appropriate classes of densities $f$, $g$, functions $\phi$ and choices of weights, for certain variance functionals $v_1 = v_1(f, g)$ and $v_2 = v_2(f, g)$ given in (8) below. Theorem 14 in Section 6 complements this by establishing that $v_1$ and $v_2$ are optimal in a local asymptotic minimax sense. We therefore conclude that, under the conditions of these results, the estimators $\widehat{T}_{m,n}$ are efficient.

In addition to studying the efficiency of our estimators $\widehat{T}_{m,n}$, it is also highly desirable to be able to derive their asymptotic distributions; such a result could be used, for instance, to obtain an asymptotically valid confidence interval for $T$. Despite the fact that the summands in our estimator are dependent, for the special case of the one-sample Shannon entropy functional, it is straightforward to derive the asymptotic normality of the weighted nearest neighbour estimator, as it is well approximated by the efficient, 'oracle' estimator $-m^{-1}\sum_{i=1}^{m} \log f(X_i)$. However, for general functionals, the natural oracle estimator may not be efficient, as explained in the next paragraph; this means that deriving the asymptotic distribution of $\widehat{T}_{m,n}$ in such cases remains a significant challenge. In our second main result (Theorem 3 in Section 2), we show how the problem can be reexpressed in a form where we can apply the central limit theorem of Baldi and Rinott (1989) for dependent random variables for which the degrees of the nodes in the pairwise dependency graph are controlled. Thus, the estimators $\widehat{T}_{m,n}$ are indeed asymptotically normal under appropriate conditions.

As a byproduct of our efficiency analysis, we uncover a curious phenomenon that can occur for certain functionals; for ease of exposition here, we focus on the Rényi-type functional

$$H_\kappa := \int_{\mathbb{R}^d} f(x)^\kappa \, dx,$$

with $\kappa \in (1/2, 1)$. Given access to $f(X_1), \ldots, f(X_m)$, the natural oracle estimator in this setting is

$$H_m^* := \frac{1}{m}\sum_{i=1}^{m} f(X_i)^{\kappa-1}.$$

Indeed, Proposition 12 reveals that this oracle estimator can be optimal in a local asymptotic minimax sense for the oracle problem where the practitioner seeks to estimate a one-sample functional such as $H_\kappa$ based on $f(X_1), \ldots, f(X_m)$. Nevertheless, surprisingly, we find that there exists an estimator $\widehat{H}_m$ and general classes $\mathcal{F}$ of densities for which

$$\lim_{m\to\infty} \sup_{f\in\mathcal{F}} \frac{\mathbb{E}_f\{(\widehat{H}_m - H_\kappa)^2\}}{\mathbb{E}_f\{(H_m^* - H_\kappa)^2\}} = \kappa^2 < 1. \tag{5}$$

We refer to this as the *super-oracle phenomenon*. It is important to note that this is very different from the phenomenon of superefficiency, as occurs with, for example, the Hodges estimator (Lehmann and Casella (1998), Example 6.2.5). There, in the case of scalar parameter estimation, asymptotic improvement in mean squared error risk is possible at a set of fixed parameter values, which form a Lebesgue null set (LeCam (1953), van der Vaart (1997)). Moreover, and more importantly from our perspective, the superefficient asymptotic behaviour is necessarily accompanied by worse finite-sample performance in a neighbourhood of points of superefficiency, so that any apparent improvement is really an artefact of the pointwise asymptotic regime considered. By contrast, in (5), the supremum is taken inside the limit, so that the super-oracle improvement for large $m$ can be considered as genuine.

The remainder of the paper is organised as follows: in Section 2, we present our main results on the asymptotic squared error risk and asymptotic normality of our general two-sample functional estimators. Section 3 is devoted to understanding the bias of these estimators and a discussion of the potential benefits of debiasing them before computing our weighted averages, while Section 4 considers their variance properties. In Section 5, we describe the super-oracle phenomenon in greater detail, and in Section 6 we present a local asymptotic minimax lower bound that illustrates the asymptotic optimality of our estimators and justifies referring to them as efficient. Our main theoretical arguments are given in the Supplementary Material (Berrett and Samworth (2023)), as well as various auxiliary results and bounds on remainder terms.

We end this section by introducing some notation used throughout the paper. For $m \in \mathbb{N}_0$, we write $[m] := \{0, 1, \dots, m\}$. If $A$ is a vector, matrix or array, we write $\|A\|$ for its Euclidean vectorised norm. For $x \in \mathbb{R}^d$ and $r \geq 0$, let $B_x(r) := \{y \in \mathbb{R}^d : \|y - x\| \leq r\}$ denote the closed Euclidean ball or radius $r$ about $x$. For vectors $a$ and $b$ of the same dimension, we write $a \circ b$ for their Hadamard product. If $Z$ is a random variable, we write $\mathcal{L}(Z)$ for its law. We write $\mathcal{Z} := (0, \infty)^2$. For a smooth function $\phi : \mathcal{Z} \to \mathbb{R}$, $\mathbf{z} = (u, v) \in \mathcal{Z}$ and $j, l \in \mathbb{N}$, we write $\phi_{jl}(\mathbf{z}) := \frac{\partial^{j+l}\phi}{\partial u^j \partial v^l}$. We also use multi-index notation for derivatives, so that, for a sufficiently smooth density $f^*$ on $\mathbb{R}^d$, $x = (x_1, \dots, x_d)^T \in \mathbb{R}^d$, $t \in \mathbb{N}$ and a multi-index $\boldsymbol{\alpha} = (\alpha_1, \dots, \alpha_d) \in \mathbb{N}_0^d$ with $|\boldsymbol{\alpha}| := \sum_{j=1}^d \alpha_j = t$, we write $\partial^{\boldsymbol{\alpha}} f^* := \frac{\partial^t f^*}{\partial x_1^{\alpha_1} \dots \partial x_d^{\alpha_d}}$. For $\alpha > 0$ and a density $f^*$ on $\mathbb{R}^d$, we write $\mu_\alpha(f^*) := \int_{\mathbb{R}^d} \|x\|^\alpha f^*(x)\, dx$ and $\|f^*\|_\infty := \sup_{x \in \mathbb{R}^d} f^*(x)$. For $r \in [0, \infty)$ and $x \in \mathbb{R}^d$, we also define $h_{x,f^*}(r) := \int_{B_x(r)} f^*(y)\, dy$ and, for $s \in [0, 1)$, let $h^{-1}_{x,f^*}(s) := \inf\{r \geq 0 : h_{x,f^*}(r) \geq s\}$. Recall that, for $a, b > 0$, the beta function is defined by $\mathrm{B}_{a,b} := \int_0^1 t^{a-1}(1-t)^{b-1}\, dt$ and define also the corresponding density $\mathrm{B}_{a,b}(s) := s^{a-1}(1-s)^{b-1}/\mathrm{B}_{a,b}$ for $s \in (0, 1)$.

**2. Main results.** Let $X_1, \dots, X_m$, and $Y_1, \dots, Y_n$ be independent $d$-dimensional random vectors, with $X_1, \dots, X_m$ having density $f$ and with $Y_1, \dots, Y_n$ having density $g$, both with respect to Lebesgue measure on $\mathbb{R}^d$. We consider the estimation of the functional $T(f, g)$ in (1).

Before we can state our main theorems on the asymptotic risk and normality of our functional estimators, we need some preparatory work. This will consist of definitions of the classes of functionals and densities over which our results will hold, the definitions of our weighted nearest neighbour estimators and the corresponding classes of allowable weights, as well as various parameters that will play a role in the statements of our results.

Starting with our classes of functionals, we impose a condition on the function $\phi$ in (1). It will be convenient to introduce the shorthand $x_\wedge := x \wedge 1$ and $x_\vee := x \vee 1$ for $x \geq 0$. Let $\Xi := \mathbb{R}^2 \times (\mathbb{N} \setminus \{1\}) \times (1, \infty)$, and for $\xi = (\kappa_1, \kappa_2, \beta^*, L) \in \Xi$, let $\Phi \equiv \Phi(\xi)$ denote the class of functions $\phi : \mathcal{Z} \to \mathbb{R}$ for which the partial derivatives $\phi_{\ell_1 \ell_2}$ exist for all $\ell_1, \ell_2 \in \mathbb{N}_0$ with

$\ell_1 + \ell_2 \leq \beta^*$ and satisfy

$$\left|u^{\ell_1} v^{\ell_2} \phi_{\ell_1 \ell_2}(u, v)\right| \leq L u_\wedge^{\kappa_1} u_\vee^{L} v_\wedge^{\kappa_2} v_\vee^{L}$$

for all $(u, v) \in \mathcal{Z}$. This is a growth condition on $\phi$ and its partial derivatives of order up to $\beta^*$. The pre-multiplier $u^{\ell_1} v^{\ell_2}$ allows us to control discrepancies of $\phi$ under relative, as opposed to absolute, changes in its arguments. Moreover, the right-hand side of the bound affords additional flexibility regarding the level of regularity required for both small and large values of these first and second arguments, controlled by the parameters $\kappa_1$, $\kappa_2$ and $L$. This latter aspect will allow us to include functionals such as the Kullback–Leibler and Rényi divergences, for which the corresponding $\phi$ is nonsmooth as the densities approach zero; see Examples 1 and 2 below. More generally, for the $\varphi$-divergence functional with $\phi(u, v) = \varphi(v/u)$, it is straightforward to express this condition in terms of a condition on $\varphi$.

For our classes of densities, fix $\beta > 0$, a density $f$ on $\mathbb{R}^d$, and $x \in \mathbb{R}^d$ with $f(x) > 0$ such that $f$ is $\underline{\beta} := \lceil \beta \rceil - 1$-times differentiable at $x$. Write $f^{(t)}(x) \in (\mathbb{R}^d)^{\otimes t}$ for the $t$th derivative array of $f$ at $x$ for $t \in [\underline{\beta}]$, so that $f^{(t)}_{j_1 \ldots j_t}(x) := \frac{\partial^t f}{\partial x_{j_1} \ldots \partial x_{j_t}}(x)$ for $(j_1, \ldots, j_t) \in \{1, \ldots, d\}^t$. Now define

$$M_{f,\beta}(x) := \inf\Bigg\{ M \geq 1 : \max_{t \in [\underline{\beta}]} \left( \frac{\|f^{(t)}(x)\|}{f(x)} \right)^{1/t} \vee \sup_{\substack{y,z \in B_x(1/M), \\ y \neq z}} \left( \frac{\|f^{(\underline{\beta})}(z) - f^{(\underline{\beta})}(y)\|}{f(x)\|z - y\|^{\beta - \underline{\beta}}} \right)^{1/\beta} \leq M \Bigg\};$$

otherwise, we set $M_{f,\beta}(x) := \infty$. The quantity $M_{f,\beta}(x)$ measures the smoothness of derivatives of $f$ in neighbourhoods of $x$, relative to $f(x)$ itself, but does not require $f$ to be smooth everywhere. For instance, if $f$ is the uniform density on the unit ball $B_0(1)$, then $M_{f,\beta}(x) = 1/(1 - \|x\|)$ for $\|x\| < 1$. Now, for $\theta = (\alpha, \beta, \lambda, C) \in (0, \infty)^4$, and writing $\mathcal{F}_d$ for the class of densities on $\mathbb{R}^d$, let

$$\mathcal{G}_{d,\theta} := \left\{ f \in \mathcal{F}_d : \mu_\alpha(f) \leq C, \|f\|_\infty \leq C, \int_{\{x : f(x) > 0\}} f(x) \left\{ \frac{M_{f,\beta}(x)^d}{f(x)} \right\}^\lambda dx \leq C \right\}.$$

Thus, in addition to requiring a moment assumption and a bounded density, the classes $\mathcal{G}_{d,\theta}$ also impose an integrability condition on our local measure of smoothness; to understand this condition, we note that in constructing a nearest-neighbour based estimate of $f(x)$, the crucial quantity that controls the bias is the function

$$s \mapsto \inf\left\{ r \geq 0 : \int_{B_x(r)} f(y)\, dy \geq s \right\} =: h^{-1}_{x,f}(s)$$

on $(0, 1)$. If $f$ is constant in a neighbourhood of $x$ with $f(x) > 0$, then $h^{-1}_{x,f}(s)^d = \frac{s}{V_d f(x)}$ for small $s > 0$. More generally, the error of the approximation of $h^{-1}_{x,f}(s)^d$ by this linear function of $s$ (together with higher-order Taylor expansion terms) is controlled by $\frac{M_{f,\beta}(\cdot)^d}{f(\cdot)}$; see Lemma S4 in the Supplementary Material ((Berrett and Samworth (2023))) for a formal statement. This explains why we ask for a condition on an appropriate norm of $\frac{M_{f,\beta}(\cdot)^d}{f(\cdot)}$ in our classes. It is an attractive feature that the assumption comes in an integral form, as opposed to requiring a boundedness condition on $M_{f,\beta}(x)$, for instance. This integrability condition is our primary tool for avoiding the assumption that the density is bounded away from zero on its support (see the discussion in the Introduction). While Tsybakov and van der Meulen (1996) and Berrett, Samworth and Yuan (2019) made first steps in this direction in the context of Shannon

entropy estimation, the former of these works, which focused on the case $d = 1$, required a strictly positive density on the whole real line; the latter relaxed this condition a little, but made extremely stringent requirements on the behaviour of the density $f$ in neighbourhoods of points $x_0 \in \mathbb{R}^d$ with $f(x_0) = 0$. In particular, no Beta$(a, a)$ density was allowed, for any $a > 0$, and the only densities having points $x_0$ with $f(x_0) = 0$ that were shown to belong to their classes involved all derivatives also vanishing at $x_0$. By contrast, Proposition 1 below shows that a multivariate spherically symmetric generalisation of a Beta$(a, b)$ density belongs to $\mathcal{G}_{d,\theta}$ for suitable $\theta \in (0, \infty)^4$, provided only that $a, b \geq 1$ (though in fact the requirements of our Theorem 2 on efficiency would actually also need $b > d - 1$ for this family).

PROPOSITION 1. *Fix $a, b \in [1, \infty)$, and let $f$ denote the density on $\mathbb{R}^d$ given by*

$$f(x) = C_{d,a,b}\|x\|^{a-1}\big(1 - \|x\|\big)^{b-1}\mathbb{1}_{\{\|x\|\leq 1\}},$$

*where $C_{d,a,b} := \frac{\Gamma(a+b+d-1)}{dV_d\Gamma(a+d-1)\Gamma(b)}$. Then for any $\alpha, \beta > 0$ and any $\lambda \in (0, b/(b+d-1))$, there exists $C_0 > 0$, depending only on $\alpha$, $\beta$ and $\lambda$, such that $f \in \mathcal{G}_{d,(\alpha,\beta,\lambda,C)}$ for any $C \geq C_0$.*

From Proposition 1 we also see that discontinuous densities may also belong to $\mathcal{G}_{d,\theta}$ for suitable $\theta \in (0, \infty)^4$; in particular, the $U[-1, 1]$ density belongs to $\mathcal{G}_{1,(\alpha,\beta,\lambda,C)}$ for any $\alpha, \beta > 0$, $\lambda \in (0, 1)$ and $C \geq 1/(1 - \lambda)$. We also remark that, similar to Berrett, Samworth and Yuan (2019), all Gaussian densities belong to $\mathcal{G}_{d,\theta}$ for any $\alpha, \beta > 0$, $\lambda \in (0, 1)$ and sufficiently large $C > 0$, and multivariate-$t$ densities with $\nu$ degrees of freedom belong to $\mathcal{G}_{d,\theta}$ for any $\alpha \in (0, \nu)$, any $\beta > 0$, $\lambda \in (0, \nu/(\nu + d))$ and $C > 0$ sufficiently large.

To define our main class of densities, then, for $\Theta = (0, \infty)^5$ and $\vartheta = (\alpha, \beta, \lambda_1, \lambda_2, C) \in \Theta$, let $M_\beta(x) \equiv M_{f,g,\beta}(x) := M_{f,\beta}(x) \vee M_{g,\beta}(x)$ and set

$$\mathcal{F}_{d,\vartheta} := \bigg\{(f, g) \in \mathcal{G}_{d,(\alpha,\beta,\lambda_1,C)} \times \mathcal{F}_d : \int_{\mathcal{X}} f(x)\bigg[\bigg\{\frac{M_\beta(x)^d}{f(x)}\bigg\}^{\lambda_1} + \bigg\{\frac{M_\beta(x)^d}{g(x)}\bigg\}^{\lambda_2}\bigg]dx \leq C,$$
$$\mu_{1/C}(g) \leq C, \|g\|_\infty \leq C, \int_{\mathcal{X}} f(x)^{2+2\kappa_1-1/C} g(x)^{2\kappa_2-1-1/C}\,dx \leq C\bigg\}.$$

Note that $\mathcal{F}_{d,\vartheta}$ also depends on $\xi$ through $\kappa_1$ and $\kappa_2$, that is, on the functional we wish to estimate, though we suppress this in our notation. To understand the final integrability condition in $\mathcal{F}_{d,\vartheta}$, we first note that the efficient variance $v_2$, defined in (8) below, can be bounded above as follows:

$$v_2 = \mathrm{Var}\big(f(Y_1)\phi_{01}\big(f(Y_1), g(Y_1)\big)\big) \leq L^2 C^{4L+2(|\kappa_1|+|\kappa_2|)} \int_{\mathcal{X}} f(x)^{2+2\kappa_1} g(x)^{2\kappa_2-1}\,dx,$$

for $C \geq 1$. Thus, for large values of $C$, our condition is only slightly stronger than assuming that $v_2$ is bounded. This slight strengthening of that assumption is made so that the integral over $\mathcal{X}$ in $v_2$ can be approximated by integrals over large subsets of $\mathcal{X}$, uniformly over $(f, g) \in \mathcal{F}_{d,\vartheta}$.

We now introduce the class of weights that we consider for our estimators. To this end, for $k, I \in \mathbb{N}$ and $c \in (0, 1)$, define

$$\mathcal{W}^{(k)}_{I,c} := \bigg\{w = (w_1, \ldots, w_k) \in \mathbb{R}^k : \sum_{j=1}^k w_j = 1 \text{ and } w_j = 0 \text{ for } j < ck, \|w\|_1 \leq 1/c, \sum_{j=1}^k j^{\frac{2\ell}{d}-i} w_j = 0 \text{ for } (\ell, i) \in \big([\lceil d/2\rceil - 1] \times [I]\big) \setminus \{(0, 0)\}\bigg\}. \tag{6}$$

Fixing $\xi = (\kappa_1, \kappa_2, \beta^*, L) \in \Xi$ and $c \in (0,1)$, and for $w_X \in \mathcal{W}^{(k_X)}_{\lceil(\beta^*-1)/2\rceil,c}$ and $w_Y \in \mathcal{W}^{(k_Y)}_{\lceil(\beta^*-1)/2\rceil,c}$, we can now define our weighted functional estimators as

$$\widehat{T}_{m,n} \equiv \widehat{T}^{w_X,w_Y}_{m,n} := \sum_{j_X=1}^{k_X} \sum_{j_Y=1}^{k_Y} w_{X,j_X} w_{Y,j_Y} \widetilde{T}_{m,n,j_X,j_Y}. \tag{7}$$

Note that the constraint on the support of $w_X$ ensures that all component indices with nonzero weight are of the same order as $k_X$, with the corresponding property also holding for $w_Y$. Once this is satisfied, and given appropriate choices of $k_X$, $k_Y$, the remaining constraints in (6) will ensure that the bias of $\widehat{T}_{m,n}$ is asymptotically negligible.

It is convenient to use the shorthand $\phi_x := \phi(f(x), g(x))$, as well as $(f\phi_{10})_x := f(x)\phi_{10}(f(x), g(x))$ and $(f\phi_{01})_x := f(x)\phi_{01}(f(x), g(x))$ for $x \in \mathbb{R}^d$. Our result on the asymptotic risk of $\widehat{T}_{m,n}$ will be expressed in terms of

$$v_1 = v_1(f,g) := \mathrm{Var}\big(\phi_{X_1} + (f\phi_{10})_{X_1}\big) \quad \text{and} \quad v_2 = v_2(f,g) := \mathrm{Var}\big((f\phi_{01})_{Y_1}\big). \tag{8}$$

Fixing $d \in \mathbb{N}$, $\vartheta = (\alpha, \beta, \lambda_1, \lambda_2, C) \in \Theta$ and $\xi = (\kappa_1, \kappa_2, \beta^*, L) \in \Xi$, we will moreover impose requirements on various derived parameters. In particular, writing $\kappa_i^- := \max(-\kappa_i, 0)$ for $i = 1, 2$, it will also be convenient to define

$$\begin{aligned} \zeta &:= \frac{\kappa_1^-}{\lambda_1} + \frac{\kappa_2^-}{\lambda_2} + \frac{d(\kappa_1^- + \kappa_2^-)}{\alpha}, \\ \tau_i &:= 1 - \max\bigg(\frac{d}{2\beta}, \frac{d}{2(2\wedge\beta)+d}, \frac{d}{2(2\wedge\beta)\beta^*}, \frac{1}{2(\lambda_i \wedge 1)(1-\zeta)}\bigg), \quad i = 1, 2. \end{aligned} \tag{9}$$

Finally, then, we are in a position to state our first main result, on the asymptotic squared error risk of $\widehat{T}_{m,n}$.

THEOREM 2. *Fix $d \in \mathbb{N}$, fix $\vartheta = (\alpha, \beta, \lambda_1, \lambda_2, C) \in \Theta$ and fix $\xi = (\kappa_1, \kappa_2, \beta^*, L) \in \Xi$. Assume that $\zeta < 1/2$ and that $\min(\tau_1, \tau_2) > 1/\beta^*$. Let $(k_X^{\mathrm{L}})$, $(k_Y^{\mathrm{L}})$, $(k_X^{\mathrm{U}})$ and $(k_Y^{\mathrm{U}})$ be deterministic sequences of positive integers that satisfying $\min(k_X^{\mathrm{L}} m^{-1/\beta^*}, k_Y^{\mathrm{L}} n^{-1/\beta^*}) \to \infty$ and $\max(k_X^{\mathrm{U}} m^{-(\tau_1-\epsilon)}, k_Y^{\mathrm{U}} n^{-(\tau_2-\epsilon)}) \to 0$ for some $\epsilon > 0$. Then for each $c \in (0,1)$, each $w_X = w_X^{(k_X)} \in \mathcal{W}^{(k_X)}_{\lceil(\beta^*-1)/2\rceil,c}$ and each $w_Y = w_Y^{(k_Y)} \in \mathcal{W}^{(k_Y)}_{\lceil(\beta^*-1)/2\rceil,c}$, we have*

$$\sup_{\phi\in\Phi(\xi)} \sup_{(f,g)\in\mathcal{F}_{d,\vartheta}} \max_{\substack{k_X\in\{k_X^{\mathrm{L}},\dots,k_X^{\mathrm{U}}\} \\ k_Y\in\{k_Y^{\mathrm{L}},\dots,k_Y^{\mathrm{U}}\}}} \bigg| \mathbb{E}_{f,g}\{(\widehat{T}_{m,n} - T)^2\} - \frac{v_1}{m} - \frac{v_2}{n} \bigg| = o\bigg(\frac{1}{m} + \frac{1}{n}\bigg)$$

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

In Proposition 8 in Section 3.2, we will improve Theorem 2 by showing that when $\beta \in (0, 1]$, the same conclusion holds when we replace the term $d/(2\beta)$ in the definitions of $\tau_1, \tau_2$ in (9) with $d/(4\beta)$. This allows us to weaken the smoothness requirement on our densities for the estimators $\widehat{T}_{m,n}$ to be efficient. In particular, we only need $\beta > d/4$ instead of $\beta > d/2$, when $d \in \{1, 2, 3\}$ and when $\beta^*$ may be taken to be arbitrarily large, which is the case in several examples of interest, as illustrated below.

Theorem 2 follows immediately from combining Proposition 6 in Section 3 with Proposition 11 in Section 4, which elucidate the asymptotic bias and variance of $\widehat{T}_{m,n}$ respectively. We therefore defer a description of the main ideas of our proofs until after the statements of these results, and first illustrate Theorem 2 via several examples.

EXAMPLE 1. Consider the Kullback–Leibler divergence, for which we may take $\phi(u,v) = \log(u/v)$. For any $\epsilon \in (0, 1/2)$, any $\beta^* \geq 2$, and any $L > (\beta^* - 1)!$, we have that $\phi \in \Phi(-\epsilon, -\epsilon, \beta^*, L)$. Thus, for any $d \in \mathbb{N}$ and $\vartheta = (\alpha, \beta, \lambda_1, \lambda_2, C) \in \Theta$ such that $\beta > d/2$ and $\min(\lambda_1, \lambda_2) > 1/2$, Theorem 2 tells us that we can find sequences $(k_X)$, $(k_Y)$, $(w_X)$, $(w_Y)$ such that

$$\sup_{(f,g)\in\mathcal{F}_{d,\vartheta}} \left| \mathbb{E}_{f,g}\{(\widehat{T}_{m,n} - T)^2\} - \frac{1}{m}\operatorname{Var}_f \log\left(\frac{f(X_1)}{g(X_1)}\right) - \frac{1}{n}\operatorname{Var}_g\left(\frac{f(Y_1)}{g(Y_1)}\right)\right| = o\left(\frac{1}{m} + \frac{1}{n}\right).$$

If $f$ and $g$ are spherically symmetric beta densities as in Proposition 1 with parameters $(a_f, b_f)$ and $(a_g, b_g)$ respectively, then we see from the proof of Proposition 1 that we have $M_\beta(x) \leq A/\{\|x\|(1 - \|x\|)\}$, where $A > 0$ depends only on $d$, $a_f$, $b_f$, $a_g$ and $b_g$. Thus, $(f, g) \in \mathcal{F}_{d,\vartheta}$ for sufficiently large $C > 0$ whenever

$$\lambda_1 \in \left(0, \frac{b_f}{b_f + d - 1}\right),$$

$$\lambda_2 \in \left(0, \min\left\{\frac{a_f + d - 1}{a_g + d - 1}, \frac{b_f}{b_g + d - 1}\right\}\right),$$

$$2a_f - a_g + d - 1 > 0 \quad \text{and} \quad 2b_f - b_g > 0.$$

It follows from simplifying the condition $\min(\lambda_1, \lambda_2) > 1/2$ that we have efficiency whenever $\beta > d/2$ and

$$\min\left(\frac{b_f}{b_f + d - 1}, \frac{a_f + d - 1}{a_g + d - 1}, \frac{b_f}{b_g + d - 1}\right) > \frac{1}{2}.$$

As mentioned above, in Section 3.2 we will see that here, as in Examples 2 and 3 below, we can weaken the first of these conditions to $\beta > d/4$ whenever $d \in \{1, 2, 3\}$.

EXAMPLE 2. For $\kappa \in (1/2, 3/2)$, consider the $\kappa$-Rényi divergence, for which we may take $\phi(u,v) = (u/v)^{\kappa-1}$. For any $\beta^* \geq 2$ and $L \geq (\beta^*)!$ we have $\phi \in \Phi(\kappa - 1, 1 - \kappa, \beta^*, L)$. Let $d \in \mathbb{N}$ and $\vartheta = (\alpha, \beta, \lambda_1, \lambda_2, C) \in \Theta$ be such that $\beta > d/2$, such that

$$\zeta = \frac{(\kappa - 1)_-}{\lambda_1} + \frac{(1 - \kappa)_-}{\lambda_2} + \frac{d|1 - \kappa|}{\alpha} < \frac{1}{2},$$

and such that $\min(\lambda_1, \lambda_2) > 1/\{2(1 - \zeta)\}$. Then, by Theorem 2, we can find sequences $(k_X)$, $(k_Y)$, $(w_X)$, $(w_Y)$ such that

$$\sup_{(f,g)\in\mathcal{F}_{d,\vartheta}} \left| \mathbb{E}_{f,g}\{(\widehat{T}_{m,n} - T)^2\} - \frac{\kappa^2}{m}\operatorname{Var}_f\left(\frac{f(X_1)^{\kappa-1}}{g(X_1)^{\kappa-1}}\right) - \frac{(\kappa-1)^2}{n}\operatorname{Var}_g\left(\frac{f(Y_1)^\kappa}{g(Y_1)^\kappa}\right)\right|$$
$$= o\left(\frac{1}{m} + \frac{1}{n}\right).$$

As in Example 1, we simplify these conditions for spherically symmetric beta distributions, but here we restrict attention to $d = 1$ and $\beta > 1/4$ for simplicity. When $\kappa \in (1, 3/2)$ we have efficiency when $\min(a_f/a_g, b_f/b_g) > \kappa - 1/2$, and when $\kappa \in (1/2, 1)$ we have efficiency when $\min(a_f/a_g, b_f/b_g) > 1/(2\kappa)$.

EXAMPLE 3. Suppose we would like to estimate $\int_{\mathbb{R}^d}\{f(x) - g(x)\}^2\,dx = \int_{\mathbb{R}^d} f(x)^2\,dx + \int_{\mathbb{R}^d} g(x)^2\,dx - 2\int_{\mathbb{R}^d} f(x)g(x)\,dx$. We may estimate each of these terms separately using one- or two-sample estimators as appropriate. Then, by Theorem 2 and a corresponding one-sample version, we can achieve a mean squared error of $O(1/m + 1/n)$ uniformly over

classes of densities $(f, g)$ such that $\|f\|_\infty, \|g\|_\infty \le C$, such that $\mu_{1/C}(f), \mu_{1/C}(g) \le C$, such that

$$\int_{\mathbb{R}^d} f(x)^{1-\lambda_1} M_\beta(x)^{d\lambda_1}\,dx \le C, \qquad \int_{\mathbb{R}^d} g(x)^{1-\lambda_2} M_\beta(x)^{d\lambda_2}\,dx \le C,$$

and such that

$$\int_{\{x:g(x)>0\}} f(x)\biggl\{\frac{M_\beta(x)^d}{g(x)}\biggr\}^{\lambda_3} dx \le C, \tag{10}$$

for any $C > 0$, for any $\beta > d/2$ and for any $\lambda_1, \lambda_2, \lambda_3 > 1/2$. It may be the case that $f$ has heavier tails than $g$, so that (10) holds with the roles of $f$ and $g$ reversed. In that case, we can obtain the same order of mean squared error by reversing the roles of the two samples in our estimator.

To study the asymptotic normality of $\widehat{T}_{m,n}$, we impose a stronger condition on the pair $(f, g)$: for $\vartheta = (\alpha, \beta, \lambda_1, \lambda_2, C) \in \Theta$, let

$$\begin{aligned}\widetilde{\mathcal{F}}_{d,\vartheta} := \biggl\{&(f,g) \in \mathcal{F}_{d,\vartheta} : \min(v_1, v_2) \ge 1/C,\\ &\max_{p=3,4} \max\biggl(\int_{\mathcal{X}} f(x)^{1+p\kappa_1} g(x)^{p\kappa_2}\,dx,\\ &\int_{\mathcal{X}} g(y)^{1+p(\kappa_2-1)} f(y)^{p+p\kappa_1}\,dy\biggr) \le C\biggr\}.\end{aligned} \tag{11}$$

To explain the lower bounds on $v_1$ and $v_2$ in (11), consider the setting in which $\phi(u, v) = \varphi(v/u)$, as is the case with $\varphi$-divergences. Then, writing $W := g(X_1)/f(X_1)$ and $Z := g(Y_1)/f(Y_1)$, we have that

$$v_1 = \mathrm{Var}\bigl(\varphi(W) - W\varphi'(W)\bigr) \quad \text{and} \quad v_2 = \mathrm{Var}\bigl(\varphi'(Z)\bigr).$$

Now, if $f = g$ then we have $v_1 = v_2 = 0$, and it is possible that estimators will converge to $T$ at a faster rate than $m^{-1/2} + n^{-1/2}$ (with a potentially nonnormal limiting distribution). Thus, in order to state uniform results on the asymptotic normality of $\widehat{T}_{m,n}$, we work over a class of densities for which $v_1$ and $v_2$ are bounded below.

The bounds on the integrals in (11) arise from considering the influence functions given by $\mathrm{IF}_1(x) := \phi_x + (f\phi_{10})_x$ and $\mathrm{IF}_2(y) := (f\phi_{01})_y$. Our conditions on $\phi$ imply that $|\mathrm{IF}_1(x)| \le 2LC^{2L+|\kappa_1|+|\kappa_2|} f(x)^{\kappa_1} g(x)^{\kappa_2}$ and $|\mathrm{IF}_2(y)| \le LC^{2L+|\kappa_1|+|\kappa_2|} f(y)^{\kappa_1+1} g(y)^{\kappa_2-1}$. Under our assumptions, we can therefore obtain bounds on $\mathbb{E}\{|\mathrm{IF}_1(X_1)|^p\}$ and $\mathbb{E}\{|\mathrm{IF}_2(Y_1)|^p\}$ for $p = 3, 4$. This is helpful for the application of the central limit theorem of Baldi and Rinott (1989).

For two random variables $X$ and $Y$ with distribution functions $F$ and $G$ (where for later convenience we allow $X$ and $Y$ to take values in the extended real line), let

$$d_{\mathrm{K}}\bigl(\mathcal{L}(X), \mathcal{L}(Y)\bigr) := \sup_{t\in\mathbb{R}} \bigl|F(t) - G(t)\bigr|$$

denote the Kolmogorov distance between the distributions of $X$ and $Y$.

THEOREM 3. *Suppose that the conditions of Theorem 2 hold. If* $(k_X^{\mathrm{U}})^4 \log^8 m = o(m)$ *and* $(k_Y^{\mathrm{U}})^4 \log^8 n = o(n)$, *then*

$$\sup_{\phi\in\Phi(\xi)} \sup_{(f,g)\in\widetilde{\mathcal{F}}_{d,\vartheta}} \max_{\substack{k_X\in\{k_X^{\mathrm{L}},\dots,k_X^{\mathrm{U}}\}\\ k_Y\in\{k_Y^{\mathrm{L}},\dots,k_Y^{\mathrm{U}}\}}} d_{\mathrm{K}}\biggl(\mathcal{L}\biggl(\frac{\widehat{T}_{m,n} - T}{\{v_1/m + v_2/n\}^{1/2}}\biggr), N(0,1)\biggr) \to 0$$

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

The proof of Theorem 3 relies on a Poissonisation argument. By this, we mean that we initially consider the related problem where instead of observing samples $X_1, \dots, X_m$ and $Y_1, \dots, Y_n$ of fixed size, we first sample $M \sim \mathrm{Poi}(m)$ and $N \sim \mathrm{Poi}(n)$, and, conditional on $M$ and $N$, observe two independent samples $X_1, \dots, X_M \overset{\mathrm{iid}}{\sim} f$ and $Y_1, \dots, Y_N \overset{\mathrm{iid}}{\sim} g$. The main reason for doing this is because in this model, appropriately truncated nearest neighbour distances of $X_i$ and $X_j$ are independent provided that $X_i$ and $X_j$ are sufficiently far apart. One of the key ideas of the proof is the observation that, after Poissonisation and nearest neighbour distance truncation, we can construct a careful partition of $\mathbb{R}^d$ into Voronoi cells, such that the probability content of each cell is roughly the same and decays with the sample size, and yet each cell has only a small number of other cells that are close to it (Proposition S2, Berrett and Samworth (2023)). By decomposing our estimator into contributions from each cell of the partition, we therefore obtain a sum of terms with a sparse dependency graph, which enables us to apply the central limit theorem of Baldi and Rinott (1989).

Another key aspect of the proof of Theorem 3 is an approximation of our unweighted nearest neighbour functional estimators by a sum of two terms, each of which only depends on one of the samples. To describe this decomposition, we write $\rho_{(k),i,\ell}$ for the $k$th nearest neighbour distance of $X_i$ among the sample $X_1, \dots, X_\ell$ whenever $\ell \geq \max(k+1, i)$. We will also write $\rho_{(k),\ell}(x)$ for the $k$th nearest neighbour distance of $x$ among the sample $Y_1, \dots, Y_\ell$ whenever $\ell \geq k$. Now define the random variables

$$T_m^{(1)} := \frac{1}{m} \sum_{i=1}^{m} \phi\Big( \frac{k_X}{m V_d \rho_{(k_X),i,m}^d}, g(X_i) \Big),$$
$$T_n^{(2)} := \int_{\mathcal{X}} f(x) \phi\Big( f(x), \frac{k_Y}{n V_d \rho_{(k_Y),n}(x)^d} \Big) dx. \tag{12}$$

We can think of $T_m^{(1)}$ and $T_n^{(2)}$ as semi-oracle estimators, where in the first case the sample size $n$ from density $g$ is infinite, and in the second case, the sample size $m$ from density $f$ is infinite. In particular, the crucial point is that $T_m^{(1)}$ depends only on $X_1, \dots, X_m$ and $T_n^{(2)}$ depends only on $Y_1, \dots, Y_n$. In fact, our proof reveals the interesting observation that under our conditions,

$$\widetilde{T}_{m,n} - \mathbb{E}(\widetilde{T}_{m,n}) = T_m^{(1)} - \mathbb{E}(T_m^{(1)}) + T_n^{(2)} - \mathbb{E}(T_n^{(2)}) + o_p(m^{-1/2} + n^{-1/2}).$$

The main advantage of this decomposition is that it allows us to establish the asymptotic normality of $\widetilde{T}_{m,n}$ by considering $T_m^{(1)}$ and $T_n^{(2)}$ separately. A further benefit is that it facilitates control of the Poissonisation error more easily than would otherwise be the case, as we now explain. Let $M \sim \mathrm{Poi}(m)$ and $N \sim \mathrm{Poi}(n)$ be independent (and independent of the data), and, when $M \geq (k_X + 1)\log(em)$ and $N \geq k_Y \log(en)$, define

$$T_m^{(1),\mathrm{p}} := \frac{1}{m} \sum_{i=1}^{M} \phi\Big( \frac{k_X}{m V_d \rho_{(k_X),i,M}^d}, g(X_i) \Big) - \Big( \frac{M}{m} - 1 \Big) \int_{\mathcal{X}} f(x) \{ \phi_x + (f\phi_{10})_x \} dx,$$
$$T_n^{(2),\mathrm{p}} := \int_{\mathcal{X}} f(x) \phi\Big( f(x), \frac{k_Y}{n V_d \rho_{(k_Y),N}(x)^d} \Big) dx - \Big( \frac{N}{n} - 1 \Big) \int_{\mathcal{X}} f(x) (g\phi_{01})_x \, dx.$$

If $M < (k_X + 1)\log(em)$, say $T_m^{(1),\mathrm{p}} := 0$, and similarly if $N < k_Y \log(en)$, say $T_n^{(2),\mathrm{p}} := 0$. The following result bounds the mean squared difference of these approximations.

PROPOSITION 4. *Assume that the conditions of Theorem 2 hold and additionally assume that* $k_X^{\mathrm{U}} = o(m^{1/4})$ *and* $k_Y^{\mathrm{U}} = o(n^{1/6})$. *Then*

$$\sup_{\phi \in \Phi(\xi)} \sup_{(f,g) \in \mathcal{F}_{d,\vartheta}} \max_{k_X \in \{k_X^{\mathrm{L}}, \dots, k_X^{\mathrm{U}}\}} \mathbb{E}\big\{ \big( T_m^{(1)} - T_m^{(1),\mathrm{p}} \big)^2 \big\} = o(1/m)$$

*and*

$$\sup_{\phi\in\Phi(\xi)}\sup_{(f,g)\in\mathcal{F}_{d,\vartheta}}\max_{k_Y\in\{k_Y^{\mathrm{L}},\ldots,k_Y^{\mathrm{U}}\}}\mathbb{E}\big\{\big(T_n^{(2)}-T_n^{(2),\mathrm{p}}\big)^2\big\}=o(1/n).$$

Theorem 3 also facilitates the construction of asymptotically valid confidence intervals of asymptotically minimal width, provided we can find consistent estimators of $v_1$ and $v_2$. To describe our methodology here, it is convenient to introduce the shorthand

$$\widehat{f}_{(k_X),i}:=\frac{k_X}{mV_d\rho_{(k_X),i,X}^d}\quad\text{and}\quad\widehat{g}_{(k_Y),i}:=\frac{k_Y}{nV_d\rho_{(k_Y),i,Y}^d} \tag{13}$$

for $i\in\{1,\ldots,m\}$, $k_X\in\{1,\ldots,m-1\}$ and $k_Y\in\{1,\ldots,n\}$. Further, define

$$\widehat{V}_{m,n}^{(1),1}:=\frac{1}{m}\sum_{i=1}^m\min\big[\{\phi(\widehat{f}_{(k_X),i},\widehat{g}_{(k_Y),i})+\widehat{f}_{(k_X),i}\phi_{10}(\widehat{f}_{(k_X),i},\widehat{g}_{(k_Y),i})\}^2,\log m,\log n\big],$$

$$\widehat{V}_{m,n}^{(1),2}:=\widetilde{T}_{m,n}+\frac{1}{m}\sum_{i=1}^m\widehat{f}_{(k_X),i}\phi_{10}(\widehat{f}_{(k_X),i},\widehat{g}_{(k_Y),i}),$$

$$\widehat{V}_{m,n}^{(2),1}:=\frac{1}{m}\sum_{i=1}^m\min\big\{\widehat{f}_{(k_X),i}\widehat{g}_{(k_Y),i}\phi_{01}(\widehat{f}_{(k_X),i},\widehat{g}_{(k_Y),i})^2,\log m,\log n\big\},$$

$$\widehat{V}_{m,n}^{(2),2}:=\frac{1}{m}\sum_{i=1}^m\widehat{g}_{(k_Y),i}\phi_{01}(\widehat{f}_{(k_X),i},\widehat{g}_{(k_Y),i}),$$

as well as $\widehat{V}_{m,n}^{(1)}:=\max\{\widehat{V}_{m,n}^{(1),1}-(\widehat{V}_{m,n}^{(1),2})^2,0\}$ and $\widehat{V}_{m,n}^{(2)}:=\max\{\widehat{V}_{m,n}^{(2),1}-(\widehat{V}_{m,n}^{(2),2})^2,0\}$. It turns out that $\widehat{V}_{m,n}^{(1)}$ and $\widehat{V}_{m,n}^{(2)}$ satisfy the consistency property that we seek, so, writing $z_q$ for the $(1-q)$th quantile of the standard normal distribution, $\widehat{v}_{m,n}:=\widehat{V}_{m,n}^{(1)}/m+\widehat{V}_{m,n}^{(2)}/n$ and

$$I_{m,n,q}:=[\widehat{T}_{m,n}-z_{q/2}\widehat{v}_{m,n}^{1/2},\widehat{T}_{m,n}+z_{q/2}\widehat{v}_{m,n}^{1/2}],$$

we have the following result.

THEOREM 5. *Suppose that the conditions of Theorem 3 hold. Then*

$$\sup_{\phi\in\Phi(\xi)}\sup_{(f,g)\in\widetilde{\mathcal{F}}_{d,\vartheta}}\max_{\substack{k_X\in\{k_X^{\mathrm{L}},\ldots,k_X^{\mathrm{U}}\}\\ k_Y\in\{k_Y^{\mathrm{L}},\ldots,k_Y^{\mathrm{U}}\}}}d_{\mathrm{K}}\bigg(\mathcal{L}\bigg(\frac{\widehat{T}_{m,n}-T}{\widehat{v}_{m,n}^{1/2}}\bigg),N(0,1)\bigg)\to0$$

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

$$\sup_{q\in(0,1)}\sup_{\phi\in\Phi(\xi)}\sup_{(f,g)\in\widetilde{\mathcal{F}}_{d,\vartheta}}\max_{\substack{k_X\in\{k_X^{\mathrm{L}},\ldots,k_X^{\mathrm{U}}\}\\ k_Y\in\{k_Y^{\mathrm{L}},\ldots,k_Y^{\mathrm{U}}\}}}\big|\mathbb{P}\big(I_{m,n,q}\ni T(f,g)\big)-(1-q)\big|\to0$$

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

## 3. Bias.

3.1. *Bias of the naive estimator.* Here we state a result on the bias of the estimator (4). It is in fact an immediate consequence of a more general statement, given as Proposition S1 in the Supplementary Material (Berrett and Samworth (2023)), which considers a wider range of choices of $k_X$ and $k_Y$.

PROPOSITION 6. *Fix* $d \in \mathbb{N}$, $\vartheta = (\alpha, \beta, \lambda_1, \lambda_2, C) \in \Theta$ *and* $\xi = (\kappa_1, \kappa_2, \beta^*, L) \in \Xi$. *Assume that* $\zeta < 1/2$ *and that* $\min(\tau_1, \tau_2) > 1/\beta^*$. *Suppose further that* $\min(k_X^{\mathrm{L}} m^{-1/\beta^*}, k_Y^{\mathrm{L}} n^{-1/\beta^*}) \to \infty$ *and that there exists* $\epsilon > 0$ *with* $\max(k_X^{\mathrm{U}} m^{-\tau_1+\epsilon}, k_Y^{\mathrm{U}} n^{-\tau_2+\epsilon}) \to 0$. *Then for each* $i_1, i_2 \in [\lceil d/2 \rceil - 1]$ *and* $j_1, j_2 \in \mathbb{N}_0$ *such that* $j_1 + j_2 \leq \lceil (\beta^* - 1)/2 \rceil$, *we can find coefficients* $\lambda_{i_1 i_2 j_1 j_2} \equiv \lambda_{i_1 i_2 j_1 j_2}(d, f, g, \phi)$, *with the properties that* $\lambda_{0,0,0,0} = T(f, g)$, *that*

$$\sup_{\phi \in \Phi(\xi)} \sup_{(f,g) \in \mathcal{F}_{d,\vartheta}} |\lambda_{i_1 i_2 j_1 j_2}| < \infty,$$

*and that*

$$\left| \mathbb{E}_{f,g}(\widetilde{T}_{m,n}) - \sum_{i_1, i_2 = 0}^{\lceil d/2 \rceil - 1} \sum_{j_1, j_2 = 0}^{\infty} \mathbb{1}_{\{j_1 + j_2 \leq \lceil (\beta^* - 1)/2 \rceil\}} \frac{\lambda_{i_1 i_2 j_1 j_2}}{k_X^{j_1} k_Y^{j_2}} \left( \frac{k_X}{m} \right)^{\frac{2i_1}{d}} \left( \frac{k_Y}{n} \right)^{\frac{2i_2}{d}} \right| = o\big(m^{-1/2} + n^{-1/2}\big)$$

*as* $m, n \to \infty$, *uniformly for* $\phi \in \Phi(\xi)$, $(f, g) \in \mathcal{F}_{d,\vartheta}$, $k_X \in \{k_X^{\mathrm{L}}, \ldots, k_X^{\mathrm{U}}\}$ *and* $k_Y \in \{k_Y^{\mathrm{L}}, \ldots, k_Y^{\mathrm{U}}\}$.

Proposition 6 provides conditions on the classes of densities and functionals under which we can give a uniform asymptotic expansion of the bias of $\widetilde{T}_{m,n}$, up to terms of negligible order. This expansion also holds uniformly over a range of values of $k_X$ and $k_Y$, which can be chosen adaptively (i.e., without knowledge of the parameters of the underlying densities) to satisfy the conditions of the theorem, for example, by setting $k_X = m^{1/\beta^*} \log m$ and $k_Y = n^{1/\beta^*} \log n$.

As revealed by Corollary 7 below, Proposition 6 allows us to form weighted versions of the estimators $\widetilde{T}_{m,n,k_X,k_Y}$, for different choices of $k_X$ and $k_Y$, so as to cancel the dominant terms in the expression for the bias of the naive estimator. Indeed, it was this result that motivated our choice of the class of weights that we consider in Theorem 2.

COROLLARY 7. *Suppose that the conditions of Proposition 6 hold. Then for each* $c \in (0, 1)$, *each* $w_X = w_X^{(k_X)} \in \mathcal{W}_{\lceil (\beta^* - 1)/2 \rceil, c}^{(k_X)}$ *and each* $w_Y = w_Y^{(k_Y)} \in \mathcal{W}_{\lceil (\beta^* - 1)/2 \rceil, c}^{(k_Y)}$, *we have*

$$\sup_{\phi \in \Phi(\xi)} \sup_{(f,g) \in \mathcal{F}_{d,\vartheta}} \sup_{\substack{k_X \in \{k_X^{\mathrm{L}}, \ldots, k_X^{\mathrm{U}}\} \\ k_Y \in \{k_Y^{\mathrm{L}}, \ldots, k_Y^{\mathrm{U}}\}}} \big| \mathbb{E}_{f,g}\big(\widehat{T}_{m,n}^{w_X, w_Y}\big) - T(f, g) \big| = o\big(m^{-1/2} + n^{-1/2}\big)$$

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

In order to gain intuition about the level of smoothness of the functional required in Corollary 7, it is helpful to consider the following (favourable) case: if our assumptions hold for all $\alpha, \beta, \lambda_2 > 0$ and all $\lambda_1 < 1$, then it suffices that $\kappa_1 > -1/2$ and that $\beta^* > \max\{2, 1 + d/4, \frac{2(1-\kappa_1^-)}{1-2\kappa_1^-}\}$.

The key idea of our bias proofs is a truncation argument that partitions $\mathcal{X}$ as $\mathcal{X}_{m,n} \cup (\mathcal{X} \setminus \mathcal{X}_{m,n})$, where

$$\mathcal{X}_{m,n} := \left\{ x \in \mathcal{X} : \frac{f(x)}{M_\beta(x)^d} \geq \frac{k_X \log m}{m}, \frac{g(x)}{M_\beta(x)^d} \geq \frac{k_Y \log n}{n} \right\}.$$

Further, by Lemma S5 of the Supplementary Material (Berrett and Samworth (2023)), we have that $f$ and $g$ are uniformly well-approximated in a relative sense, over balls of an ap-

propriate radius, by their values at the centres of these balls; more precisely, for every $\vartheta \in \Theta$, and writing $A := (16d)^{1/(\beta - \underline{\beta})}$ and $r_0(x) := 1/\{AM_\beta(x)\}$,

$$\sup_{(f,g)\in\mathcal{F}_{d,\vartheta}} \sup_{y\in B_x(r_0(x))} \left| \frac{f(y)}{f(x)} - 1 \right| \vee \left| \frac{g(y)}{g(x)} - 1 \right| \leq \frac{1}{2}.$$

In particular, this means that

$$\inf_{x\in\mathcal{X}_{m,n}} h_{x,f}\big(r_0(x)\big) \geq \frac{V_d k_X \log m}{2A^d m} \quad \text{and} \quad \inf_{x\in\mathcal{X}_{m,n}} h_{x,g}\big(r_0(x)\big) \geq \frac{V_d k_Y \log n}{2A^d n} \tag{14}$$

whenever $(f, g) \in \mathcal{F}_{d,\vartheta}$. Thus for each $x \in \mathcal{X}_{m,n}$, it is the case that with high probability, the $k_X$ nearest neighbours of $x$ among $X_1, \ldots, X_m$, as well as the $k_Y$ nearest neighbours of $x$ among $Y_1, \ldots, Y_n$, lie in $B_x(r_0(x))$. Moreover, the functions $h_{x,f}(\cdot)$ and $h_{x,g}(\cdot)$ can be approximated by Taylor expansions on $[0, r_0(x)]$, which yield corresponding expansions for their respective inverses. Since $h_{X_i,f}(\rho_{(k),i,X})|X_i \sim \text{Beta}(k, m-k)$ and $h_{X_i,g}(\rho_{(k),i,Y})|X_i \sim \text{Beta}(k, n+1-k)$, these facts, in combination with (14), allow us to deduce a stochastic expansion for $\rho_{(k),i,X}$ and $\rho_{(k),i,Y}$ in terms of powers of the relevant beta random variables. The contribution to the bias from the region $\mathcal{X}_{m,n}$ can then be computed by a Taylor expansion of $\phi$ and using exact formulae for moments of beta random variables. For $x \in \mathcal{X} \setminus \mathcal{X}_{m,n}$, we have no guarantees about the proximity of the $k_X$ nearest neighbours of $x$ among $X_1, \ldots, X_m$, nor the $k_Y$ nearest neighbours of $x$ among $Y_1, \ldots, Y_n$; however,

$$\mathbb{P}(X_1 \in \mathcal{X} \setminus \mathcal{X}_{m,n}) \leq C \left\{ \left( \frac{k_X \log m}{m} \right)^{\lambda_1} \vee \left( \frac{k_Y \log n}{n} \right)^{\lambda_2} \right\},$$

so the integrability conditions in our classes $\mathcal{F}_{d,\vartheta}$ allow us to control the contribution to the bias from this region.

3.2. *Tighter control of the bias when $\beta \leq 1$.* Our general bias result in Proposition S1 of the Supplementary Material (Berrett and Samworth (2023)) has remainder terms of the order $(k_X/m)^{\beta/d}$ and $(k_Y/n)^{\beta/d}$ in the expansion, and leads naturally to the condition $\beta > d/2$ for efficiency. A requirement of this level of smoothness for a parametric rate of convergence (albeit with smoothness measured in different ways) also appears in several other related works on functional estimation, including Leonenko and Seleznjev (2010), Kandasamy et al. (2015) and Singh and Póczos (2016). However, other results show that $d/4$ smoothness (often in the case $d = 1$ or while also requiring this smoothness to be at most 1) may suffice for certain functionals without singularities (Bickel and Ritov (1988), Birgé and Massart (1995), Giné and Nickl (2008), Laurent (1996), Leonenko and Seleznjev (2010)). The purpose of Proposition 8 below, then, is to demonstrate that when $\beta \in (0, 1]$, it is possible to tighten our bias bounds to have terms of the order $(k_X/m)^{2\beta/d}$ and $(k_Y/n)^{2\beta/d}$, so that we only require $\beta > d/4$ for efficiency.

PROPOSITION 8. *Fix $d \in \mathbb{N}$, $\vartheta = (\alpha, \beta, \lambda_1, \lambda_2, C) \in \Theta$ with $\beta \in (0, 1]$ and $\xi = (\kappa_1, \kappa_2, \beta_1^*, \beta_2^*, L) \in \Xi$. Let $k_X^{\mathrm{L}} \leq k_X^{\mathrm{U}}$, $k_Y^{\mathrm{L}} \leq k_Y^{\mathrm{U}}$ be deterministic sequences of positive integers such that $k_X^{\mathrm{L}}/\log m \to \infty$, $k_Y^{\mathrm{L}}/\log n \to \infty$, $k_X^{\mathrm{U}} = O(m^{1-\epsilon})$ and $k_Y^{\mathrm{U}} = O(n^{1-\epsilon})$ for some $\epsilon > 0$. Suppose that $\zeta < 1$. Then for each $j_1 \in [\lceil(\beta^* - 1)/2\rceil]$ and $j_2 \in [\lceil(\beta^* - 1)/2\rceil]$, we can find $\lambda_{j_1 j_2} \equiv \lambda_{j_1 j_2}(d, f, g, \phi)$, with the properties that $\lambda_{0,0} = T(f, g)$,*

$$\sup_{\phi\in\Phi(\xi)} \sup_{(f,g)\in\mathcal{F}_{d,\vartheta}} |\lambda_{j_1 j_2}| < \infty,$$

*and that, for every* $\epsilon > 0$,

$$\sup_{\phi\in\Phi(\xi)}\sup_{(f,g)\in\mathcal{F}_{d,\vartheta}}\left|\mathbb{E}_{f,g}(\widetilde{T}_{m,n}) - \sum_{j_1,j_2=0}^{\infty}\mathbb{1}_{\{j_1+j_2\le\lceil(\beta^*-1)/2\rceil\}}\frac{\lambda_{j_1j_2}}{k_X^{j_1}k_Y^{j_2}}\right|$$

$$= O\left(\max\left\{k_X^{-\beta^*/2}, \left(\frac{k_X}{m}\right)^{2\beta/d}, \left(\frac{k_X}{m}\right)^{\lambda_1(1-\zeta)-\epsilon}, k_Y^{-\beta^*/2},\right.\right. \tag{15}$$

$$\left.\left.\left(\frac{k_Y}{n}\right)^{2\beta/d}, \left(\frac{k_Y}{n}\right)^{\lambda_2(1-\zeta)-\epsilon}, 1/m, 1/n\right\}\right),$$

*as* $m, n \to \infty$, *uniformly for* $k_X \in \{k_X^{\mathrm{L}}, \dots, k_X^{\mathrm{U}}\}$ *and* $k_Y \in \{k_Y^{\mathrm{L}}, \dots, k_Y^{\mathrm{U}}\}$.

The proof of Proposition 8 is given in Section S1.3 of the Supplementary Material (Berrett and Samworth (2023)). The interest in the result arises because it reveals that the bias of nearest-neighbour functional estimators is of smaller order than that of the corresponding density estimators, at least when $\beta \le 1$ and when the function $\phi$ is smooth away from its singularities. This reduced bias is due to the fact that the nearest-neighbour density estimate biases at different values of $x \in \mathcal{X}$ cancel to leading order when we integrate over $\mathcal{X}$. While similar phenomena have been observed for kernel-based density estimates in the context of the estimation of quadratic functionals (Giné and Nickl (2008), Leonenko and Seleznjev (2010)), we are not aware of corresponding results for nearest-neighbour methods or nonsmooth functionals.

An immediate corollary of Proposition 8 is that the conclusions of Theorems 2 and 3 hold with the $d/(2\beta)$ term in the definitions of $\tau_1$ and $\tau_2$ in (9) replaced with $d/(4\beta)$, provided $\beta \le 1$. In particular, in this case it suffices to have $\beta > d/4$ in Examples 1, 2 and 3.

3.3. *Bias of an alternative debiased estimator.* As mentioned in the Introduction, building on the original debiasing idea of Kozachenko and Leonenko (1987), Ryu et al. (2018) proposed a debiasing technique for the naive estimator $\widetilde{T}_{m,n}$ of a general two-sample functional. The initial goal of this subsection is to use fractional calculus techniques to give an informal study of the remaining bias of these resulting estimators, with a view to addressing the question of whether to apply our weighting scheme to the naive estimator (4) or that of Ryu et al. (2018).

For simplicity, we will focus on the one-sample setting in (2), though all of the calculations have analogues in the two-sample setting. Suppose that there exists a sequence of differentiable functions $(\psi_k)$ for which

$$\psi(u) = \int_0^\infty e^{-s}\frac{s^{k-1}}{\Gamma(k)}\psi_k\left(\frac{ku}{s}\right)ds \tag{16}$$

for all $u \in (0, \infty)$; examples in the cases of Shannon and Rényi entropies will be given below. We will consider the debiased estimator of $H(f)$ given by

$$\widetilde{H}_m := \frac{1}{m}\sum_{i=1}^{m}\psi_k(\widehat{f}_{(k),i}).$$

Write $\mathcal{X}_f := \{x : f(x) > 0\}$. Then, under regularity conditions on $f$ and $\psi_k$, since $m\,\mathrm{Beta}(k, m-k)$ can be approximated by a $\Gamma(k, 1)$ random variable, we have that

$$\mathbb{E}\widetilde{H}_m = \int_{\mathcal{X}_f} f(x)\int_0^1 \psi_k\left(\frac{k}{mV_d h_{x,f}^{-1}(s)^d}\right)\mathrm{B}_{k,m-k}(s)\,ds\,dx$$

$$
\begin{aligned}
&\approx \int_{\mathcal{X}_f} f(x) \int_0^1 \biggl\{ \psi_k\biggl(\frac{kf(x)}{ms}\biggr) - \frac{V_d f(x) h_{x,f}^{-1}(s)^d - s}{ms^2/\{kf(x)\}} \psi_k'\biggl(\frac{kf(x)}{ms}\biggr) \biggr\} \\
&\quad \times \mathrm{B}_{k,m-k}(s)\, ds\, dx \\
&\approx \int_{\mathcal{X}_f} f(x) \int_0^\infty \biggl\{ \psi_k\biggl(\frac{kf(x)}{t}\biggr) + \frac{kt^{\frac{2}{d}-1}\Delta f(x)}{2(d+2)\{V_d n f(x)\}^{\frac{2}{d}}} \psi_k'\biggl(\frac{kf(x)}{t}\biggr) \biggr\} \\
&\quad \times \frac{e^{-t}t^{k-1}}{\Gamma(k)}\, dt\, dx \\
&= H(f) + \frac{1}{2(d+2)(V_d n)^{\frac{2}{d}}} \int_{\mathcal{X}_f} \frac{\Delta f(x)}{f(x)^{\frac{2}{d}-1}} \int_0^\infty \frac{e^{-t}t^{k+2/d-2}}{\Gamma(k-1)} \psi_k'\biggl(\frac{kf(x)}{t}\biggr) dt\, dx.
\end{aligned} \tag{17}
$$

In order to understand the behaviour of the dominant bias term on the right-hand side of (17), for $\alpha \in [0,1)$ define the operator $D^\alpha$ by

$$
(D^\alpha g)(u) := -\frac{1}{\Gamma(1-\alpha)} \int_u^\infty \frac{g'(s)}{(s-u)^\alpha}\, ds.
$$

This is closely related to the Caputo fractional derivative (Kilbas, Srivastava and Trujillo (2006), Section 2.4). Then, with $g(s) = e^{-\lambda s}$ for some $\lambda \in (0,\infty)$, we have that

$$
\bigl(D^\alpha g\bigr)(u) = \frac{1}{\Gamma(1-\alpha)} \int_u^\infty \frac{\lambda e^{-\lambda s}}{(s-u)^\alpha}\, ds = \lambda^\alpha e^{-\lambda u} = \lambda^\alpha g(u).
$$

From (16), we can see that

$$
\frac{\Gamma(k-1)}{u^{k-1}} \psi'(u) = u^{-(k-1)} \int_0^\infty e^{-t} t^{k-2} \psi_k'\biggl(\frac{ku}{t}\biggr) dt = \int_0^\infty e^{-su} s^{k-2} \psi_k'\biggl(\frac{k}{s}\biggr) ds. \tag{18}
$$

When $d \geq 3$, we can apply the operator $D^{2/d}$ to both sides of (18) to simplify the inner integral in our expression for the dominant bias term in (17) as follows:

$$
\begin{aligned}
&\frac{1}{\Gamma(k-1)} \int_0^\infty e^{-t} t^{k+\frac{2}{d}-2} \psi_k'\biggl(\frac{ku}{t}\biggr) dt \\
&\quad = \frac{u^{k+2/d-1}}{\Gamma(k-1)} \int_0^\infty e^{-su} s^{k+\frac{2}{d}-2} \psi_k'\biggl(\frac{k}{s}\biggr) ds \\
&\quad = -\frac{u^{k+2/d-1}}{\Gamma(1-2/d)} \int_u^\infty \frac{\frac{d}{ds}(\psi'(s)/s^{k-1})}{(s-u)^{2/d}}\, ds \\
&\quad = \frac{u^{k+2/d-1}}{\Gamma(1-2/d)} \int_u^\infty \frac{(k-1)s^{-k}\psi'(s) - s^{1-k}\psi''(s)}{(s-u)^{2/d}}\, ds \\
&\quad = \frac{\Gamma(k+2/d-1)}{\Gamma(k-1)} \int_0^1 \mathrm{B}_{1-2/d,k+2/d-1}(s) \\
&\qquad \times \biggl\{ \psi'\biggl(\frac{u}{1-s}\biggr) - \frac{u}{(k-1)(1-s)} \psi''\biggl(\frac{u}{1-s}\biggr) \biggr\} ds.
\end{aligned} \tag{19}
$$

For Shannon and Rényi entropies, both $\psi'$ and $\psi''$ are constant multiples of functions $g$ with the property that $g(xy) = g(x)g(y)$ for any $x, y \in (0,\infty)$. In these cases, the leading order bias separates into a coefficient depending only on $d$, $n$ and $f$ and a factor that is a function of $k$. Using weights, this leading order bias may be cancelled out, and it can be seen that, when $f$ is sufficiently regular, the next term is of order $k^{4/d}/n^{4/d}$. However, the only

continuous functions $g$ with this property are $g(x) = x^a$ for some $a \in \mathbb{R}$ (e.g., Dieudonné ((1969), (4.3.7), p. 86)). If the term in braces in (19) is separable for all values of $k$ then both $u \mapsto \psi'(u)$ and $u \mapsto u\psi''(u)$ must be separable individually, and so $\psi'(u) \propto u^a$ for some $a \in \mathbb{R}$. Thus, the Shannon and Rényi entropies are the only functionals with this property. In general, all that can be said is that this term in the bias can be expanded as a series of the form $\frac{k^{2/d}}{n^{2/d}}(c_0 + c_1/k + c_2/k^2 + \cdots)$. For larger values of $d$, to cancel out sufficient bias that the resulting estimator is efficient, the weighting scheme is then only marginally simpler than the weighting scheme for the naive estimator, and the analysis is significantly more complicated.

Despite the general conclusion of our discussion in the previous paragraph, returning to the two-sample functional setting, we now show that in the special case of the Kullback–Leibler and Rényi divergence functionals, the debiasing scheme described above significantly simplifies the weighting scheme, while facilitating the same conclusions regarding efficiency. To this end, for the Kullback–Leibler divergence, we define the following class of weight vectors:

$$\mathcal{W}_c^{(k),\mathrm{KL}} := \Bigg\{ w = (w_1, \ldots, w_k) \in \mathbb{R}^k : \sum_{j=1}^k w_j = 1 \text{ and } w_j = 0 \text{ for } j < ck, \|w\|_1 \le 1/c, \\ \sum_{j=1}^k \frac{\Gamma(j + 2\ell/d)}{\Gamma(j)} w_j = 0 \text{ for } \ell \in \big[\lceil d/2 \rceil - 1\big] \setminus \{0\} \Bigg\}.$$

The analogue of the Kozachenko–Leonenko debiased estimator is

$$\widetilde{D}_{m,n} := \frac{1}{m} \sum_{i=1}^m \log\Bigg( \frac{e^{\Psi(k_X)}}{m\rho_{(k_X),i,X}^d} \frac{n\rho_{(k_Y),i,Y}^d}{e^{\Psi(k_Y)}} \Bigg) = \widetilde{T}_{m,n} + \Psi(k_X) - \log k_X - \Psi(k_Y) + \log k_Y$$

(Ryu et al. (2018)). If the weighted estimator $\widehat{D}_{m,n}^{w_X,w_Y}$ is then formed as in (7) then the following theorem elucidates its asymptotic bias. Since this result uses very similar (in fact, somewhat simpler) arguments to those in Proposition S1 in the Supplementary Material (Berrett and Samworth (2023)), its proof, together with that of Proposition 10 below, is omitted for brevity.

PROPOSITION 9. *Fix $d \in \mathbb{N}$, let $\vartheta = (\alpha, \beta, \lambda_1, \lambda_2, C) \in \Theta$ and let $\phi(u, v) = \log(u/v)$. Assume that*

$$\tau_1 = 1 - \max\Big(\frac{d}{2\beta}, \frac{1}{2\lambda_1}\Big) > 0 \quad \text{and} \quad \tau_2 = 1 - \max\Big(\frac{d}{2\beta}, \frac{1}{2\lambda_2}\Big) > 0,$$

*and that there exists $\epsilon > 0$ such that $\max(k_X^{\mathrm{U}} m^{-\tau_1+\epsilon}, k_Y^{\mathrm{U}} n^{-\tau_2+\epsilon}) \to 0$. Then for each $c \in (0,1)$, each $w_X = w_X^{(k_X)} \in \mathcal{W}_c^{(k_X),\mathrm{KL}}$ and each $w_Y = w_Y^{(k_Y)} \in \mathcal{W}_c^{(k_Y),\mathrm{KL}}$, we have*

$$\sup_{(f,g)\in\mathcal{F}_{d,\vartheta}} \sup_{\substack{k_X \in \{1,\ldots,k_X^{\mathrm{U}}\} \\ k_Y \in \{1,\ldots,k_Y^{\mathrm{U}}\}}} \big|\mathbb{E}_{f,g}\big(\widehat{D}_{m,n}^{w_X,w_Y}\big) - T(f,g)\big| = o\big(m^{-1/2} + n^{-1/2}\big)$$

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

Since $\widetilde{D}_{m,n}$ is simply a deterministic translation of $\widetilde{T}_{m,n}$, our variance results in Section 4 continue to hold, so the corresponding efficiency result for $\widehat{D}_{m,n}^{w_X,w_Y}$ is immediate.

When estimating the Rényi integral $\int_{\mathcal{X}} f^\kappa g^{-(\kappa-1)}$, for $b \in \mathbb{R}$ and $c > 0$, we define

$$\mathcal{W}_{b,c}^{(k),\mathrm{R}} := \Bigg\{ w = (w_1, \ldots, w_k) \in \mathbb{R}^k : \sum_{j=1}^k w_j = 1 \text{ and } w_j = 0 \text{ for } j < ck, \|w\|_1 \le 1/c,$$

$$\sum_{j=1}^{k} \frac{\Gamma(j-b+2\ell/d)}{\Gamma(j-b)} w_j = 0 \text{ for } \ell \in \big[\lceil d/2 \rceil - 1\big] \setminus \{0\}\Bigg\}.$$

The corresponding debiased estimator is

$$\begin{aligned}\check{D}_{m,n} &:= \frac{1}{m}\sum_{i=1}^{m} \frac{\Gamma(k_X)\Gamma(k_Y)}{\Gamma(k_X-\kappa+1)\Gamma(k_Y+\kappa-1)} \bigg(\frac{n\rho^d_{(k_Y),i,Y}}{m\rho^d_{(k_X),i,X}}\bigg)^{\kappa-1} \\ &= \frac{k_X^{1-\kappa}\Gamma(k_X)k_Y^{\kappa-1}\Gamma(k_Y)}{\Gamma(k_X-\kappa+1)\Gamma(k_Y+\kappa-1)} \widetilde{T}_{m,n}\end{aligned}$$

(Ryu et al. (2018)). If the weighted estimator $\widehat{D}^{w_X,w_Y}_{m,n}$ is again formed as in (7) then the following result provides the corresponding bias guarantee.

PROPOSITION 10. *Fix $d \in \mathbb{N}$, let $\vartheta = (\alpha, \beta, \lambda_1, \lambda_2, C) \in \Theta$ and let $\phi(u,v) = (u/v)^{\kappa-1}$ for some $\kappa \in (1/2, \infty)$. With $\zeta$ as defined as in (9), $\kappa_1 = -\kappa_2 = \kappa - 1$,*

$$\tau_1 = 1 - \max\bigg(\frac{d}{2\beta}, \frac{1}{2\lambda_1(1-\zeta)}\bigg) \quad \text{and} \quad \tau_2 = 1 - \max\bigg(\frac{d}{2\beta}, \frac{1}{2\lambda_2(1-\zeta)}\bigg),$$

*assume that $\zeta < 1/2$ and $\min(\tau_1, \tau_2) > 0$. Suppose further that there exists $\epsilon > 0$ such that $\max(k_X^{\mathrm{U}} m^{-\tau_1+\epsilon}, k_Y^{\mathrm{U}} n^{-\tau_2+\epsilon}) \to 0$. Then for each $c \in (0,1)$, each $w_X = w_X^{(k_X)} \in \mathcal{W}^{(k_X),\mathrm{R}}_{\kappa-1,c}$ and each $w_Y = w_Y^{(k_Y)} \in \mathcal{W}^{(k_Y),\mathrm{R}}_{1-\kappa,c}$, we have*

$$\sup_{(f,g)\in\mathcal{F}_{d,\vartheta}} \sup_{\substack{k_X\in\{1,\ldots,k_X^{\mathrm{U}}\} \\ k_Y\in\{1,\ldots,k_Y^{\mathrm{U}}\}}} \big|\mathbb{E}_{f,g}\big(\widehat{D}^{w_X,w_Y}_{m,n}\big) - T(f,g)\big| = o\big(m^{-1/2} + n^{-1/2}\big)$$

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

In this case, with $k_X^{\mathrm{L}}$ and $k_Y^{\mathrm{L}}$ defined as in Theorem 2, we have

$$\frac{\check{D}_{m,n}}{\widetilde{T}_{m,n}} - 1 = \frac{k_X^{1-\kappa}\Gamma(k_X)k_Y^{\kappa-1}\Gamma(k_Y)}{\Gamma(k_X-\kappa+1)\Gamma(k_Y+\kappa-1)} - 1 \to 0$$

uniformly for $k_X \geq k_X^{\mathrm{L}}$ and $k_Y \geq k_Y^{\mathrm{L}}$, so we can again deduce an efficiency result for $\widehat{D}^{w_X,w_Y}_{m,n}$.

**4. Variance.** The following result provides the main asymptotic variance expansion for our weighted estimators. Write $\tau_i' = 1 - \max\{\frac{d}{d+2(2\wedge\beta)}, \frac{1}{2(\lambda_i\wedge 1)(1-\zeta)}\}$ for $i = 1, 2$.

PROPOSITION 11. *Fix $d \in \mathbb{N}$, $\vartheta = (\alpha, \beta, \lambda_1, \lambda_2, C) \in \Theta$ and $\xi = (\kappa_1, \kappa_2, \beta^*, L) \in \Xi$ such that $\zeta < 1/2$, $\tau_1' > 0$, $\tau_2' > 0$. Let $(k_X^{\mathrm{L}})$, $(k_Y^{\mathrm{L}})$, $(k_X^{\mathrm{U}})$ and $(k_Y^{\mathrm{U}})$ be deterministic sequences of positive integers satisfying $\min(k_X^{\mathrm{L}}/\log^5 m, k_Y^{\mathrm{L}}/\log^5 n) \to \infty$ and $\max(k_X^{\mathrm{U}} m^{-(\tau_1'-\epsilon)}, k_Y^{\mathrm{U}} n^{-(\tau_2'-\epsilon)}) \to 0$ for some $\epsilon > 0$. Then for each $c \in (0,1)$, each $w_X = w_X^{(k_X)} \in \mathcal{W}^{(k_X)}_{\lceil(\beta^*-1)/2\rceil,c}$ and each $w_Y = w_Y^{(k_Y)} \in \mathcal{W}^{(k_Y)}_{\lceil(\beta^*-1)/2\rceil,c}$, we have*

$$\sup_{\phi\in\Phi(\xi)} \sup_{(f,g)\in\mathcal{F}_{d,\vartheta}} \max_{\substack{k_X\in\{k_X^{\mathrm{L}},\ldots,k_X^{\mathrm{U}}\} \\ k_Y\in\{k_Y^{\mathrm{L}},\ldots,k_Y^{\mathrm{U}}\}}} \bigg|\mathrm{Var}_{f,g}\big(\widehat{T}^{w_X,w_Y}_{m,n}\big) - \frac{v_1}{m} - \frac{v_2}{n}\bigg| = o\bigg(\frac{1}{m} + \frac{1}{n}\bigg)$$

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

The proof of Proposition 11 is significantly more complicated that those of the bias proofs in Section 3, primarily owing to the need to consider the joint distribution of nearest neighbour distances around two different points, $X_1$ and $X_2$, say. These have an intricate dependence structure because, for instance, $X_1$ may be one of the five nearest neighbours of $X_2$, but not vice-versa. To describe our main strategy for approximating $\mathrm{Var}_{f,g}(\widehat{T}_{m,n}^{w_X,w_Y})$, we write $\widehat{T}_{m,n}^{w_X,w_Y} =: m^{-1}\sum_{i=1}^{m}\widehat{T}_{m,n}^{(i)}$ as shorthand, so that

$$\mathrm{Var}_{f,g}(\widehat{T}_{m,n}^{w_X,w_Y}) = \frac{1}{m}\mathrm{Var}_{f,g}(\widehat{T}_{m,n}^{(1)}) + \frac{m-1}{m}\mathrm{Cov}_{f,g}(\widehat{T}_{m,n}^{(1)}, \widehat{T}_{m,n}^{(2)}). \tag{20}$$

Using similar techniques to those employed in Section 3, it can be shown that

$$\mathrm{Var}_{f,g}(\widehat{T}_{m,n}^{(1)}) \to \mathrm{Var}_f\, \phi_{X_1}.$$

For the covariance term in (20), we first condition on $X_1$ and $X_2$. It turns out that this term can be further decomposed into a sum of two terms, representing the contributions from the events on which $X_1$ and $X_2$ either share or do not share nearest neighbours. Observe that if

$$\|X_1 - X_2\| > \left\{\frac{k_X}{mV_d}\left(1 + \frac{\log^{1/2} m}{k_X^{1/2}}\right)\right\}^{1/d}\{f(X_1)^{-1/d} + f(X_2)^{-1/d}\} =: R(X_1, X_2),$$

say, then, with high probability, $X_1$ and $X_2$ do not share any of their $k_X$ nearest neighbours among $X_3, \ldots, X_m$. This means that the random vector $(h_{X_1}(\rho_{(k_X),1,X}), h_{X_2}(\rho_{(k_X),2,X}), 1 - h_{X_1}(\rho_{(k_X),1,X}) - h_{X_2}(\rho_{(k_X),2,X}))$ has approximately the same distribution as $(Z_1, Z_2, Z_3)$, say, where $(Z_1, Z_2, Z_3) \sim \mathrm{Dirichlet}(k_X, k_X, m - 2k_X - 1)$. Writing $\|\cdot\|_{\mathrm{TV}}$ for the total variation norm on signed measures, we can then exploit the facts that

$$\big\|\mathcal{L}(Z_1, Z_2) - \mathrm{Beta}(k_X, m - k_X) \otimes \mathrm{Beta}(k_X, m - k_X)\big\|_{\mathrm{TV}} = O(k_X/m)$$

and

$$\frac{\widehat{f}_{(k_X),1}}{f(X_1)} = 1 + O_p(k_X^{-1/2}) \tag{21}$$

to show that the contribution to the covariance from this region is $O(1/m)$ (where in fact we also determine the leading constant). On the other hand,

$$\mathbb{P}\big[\{\|X_1 - X_2\| \le R(X_1, X_2)\} \cap \{X_1 \in \mathcal{X}_{m,n}\}\big] = O(k_X/m),$$

and this, together with (21) again, allows us to demonstrate that the contribution to the covariance from this region due to the nearest neighbour distances among $X_3, \ldots, X_m$ is also $O(1/m)$ (with a different leading constant). The terms arising from the nearest neighbour distances of $Y_1, \ldots, Y_n$ from $X_1$ and $X_2$ can be handled similarly, and their contributions can be shown to be $O(1/n)$. Combining these dominant terms results in the expansion

$$\mathrm{Cov}_{f,g}(\widehat{T}_{m,n}^{(1)}, \widehat{T}_{m,n}^{(2)}) = \frac{2}{m}\mathrm{Cov}_f(\phi_{X_1}, (f\phi_{10})_{X_1}) + \frac{1}{m}\mathrm{Var}_f((f\phi_{10})_{X_1}) + \frac{v_2}{n} + o\left(\frac{1}{m} + \frac{1}{n}\right),$$

and the conclusion follows.

**5. The super-oracle phenomenon.** In this section, we consider an alternative estimation problem, where we are still interested in the functional $T(f,g)$ in (1), but where instead of observing data $X_1,\dots,X_m,Y_1,\dots,Y_n$ as before, we instead observe $f(X_1),\dots,f(X_m),g(X_1),\dots,g(X_m)$. Although this latter framework should be considered as an 'oracle' version of the problem, because typically $f(X_1),\dots,f(X_m)$ and $g(X_1),\dots,g(X_m)$ are unknown, it is nevertheless instructive to compare the performance of our efficient estimator $\widehat{T}_{m,n}$ with that of the estimator

$$T_m^* := \frac{1}{m}\sum_{i=1}^{m}\phi\big(f(X_i),g(X_i)\big)$$

in the new problem. The estimator $T_m^*$ is unbiased, and moreover, $m^{1/2}(T_m^*-T)\xrightarrow{d}N(0,\sigma^2)$, where $\sigma^2=\sigma^2(f,g):=\mathrm{Var}_f\,\phi(f(X_1),g(X_1))$. In fact, as we now show, $T_m^*$ can be the optimal estimator, in a local asymptotic minimax sense, of $T$ in our oracle problem. Our aim here is not to seek maximal generality, but instead to give a simple class of examples for which $T_m^*$ has this optimality property.

For simplicity of exposition, we will focus on the one-sample functional (2) with $\psi(u)=u^{-(1-\kappa)}$ for some $\kappa\in(1/2,1)$. Thus, we consider estimation of the Rényi functional

$$H(f)=\int_0^\infty f(x)\psi\big(f(x)\big)\,dx=\int_0^\infty f(x)^\kappa\,dx,$$

based on the observations $f(X_1),\dots,f(X_m)$. Moreover, we take $\mathcal{X}=[0,\infty)$, and assume that $f(x)=e^{-P(x)}$ for some convex, strictly increasing polynomial $P:[0,\infty)\to\mathbb{R}$. Define the function $h:[0,\infty)\to\mathbb{R}$ by

$$h(x):=\frac{f'(x)}{f(x)}\int_0^x\big\{\psi\big(f(y)\big)-H(f)\big\}\,dy=-P'(x)\int_0^x\big\{f(y)^{-(1-\kappa)}-H(f)\big\}\,dy. \tag{22}$$

Now, for $t\in[0,\infty)$, define $f_t:[0,\infty)\to\mathbb{R}$ by

$$f_t(x):=\big\{1-th(x)\big\}f(x);$$

in the proof of Proposition 12 below, we will see that $f_t$ is a bounded probability density for sufficiently small $t\ge 0$. Moreover $f_0=f$, and we will see that $\{f_t:t\in[0,\infty)\}$ constitutes a least favourable sub-model in this problem.

Recall that $(H_m)$ is called an *estimator sequence* if $H_m:\mathbb{R}^{m\times d}\to\mathbb{R}$ is a measurable function for each $m\in\mathbb{N}$. We are now in a position to state a local asymptotic minimax lower bound that reveals the optimality of the one-sample version of $T_m^*$ in this context.

PROPOSITION 12. *Writing $\mathcal{I}$ for the set of all finite subsets of $[0,\infty)$, for any estimator sequence $(\tilde{H}_m)$ we have that*

$$\sup_{I\in\mathcal{I}}\liminf_{m\to\infty}\max_{t\in I}m\mathbb{E}_{f_{t/m^{1/2}}}\big[\big\{H_m-H(f_{t/m^{1/2}})\big\}^2\big]\ge\mathrm{Var}_f\,\psi\big(f(X_1)\big).$$

*Moreover, fixing $\alpha,\beta>0$ and $\lambda\in(0,1)$, there exist $t_0>0$, depending only on $\kappa\in(1/2,1)$ and $f$ as defined above, and $C=C(\alpha,\beta,\lambda,\kappa,f)>0$ such that $f_t\in\mathcal{G}_{1,\theta}$ for $t\in[0,t_0]$, where $\theta=(\alpha,\beta,\lambda,C)$.*

Specialising the estimator $T_m^*$ to this one-sample problem, we see that $T_m^*$ is efficient in the sense of van der Vaart ((1997), Chapter 25), and hence optimal in this local asymptotic minimax sense.

The following result, which is an immediate consequence of Theorem 2, compares the asymptotic worst-case squared error risks of $\widehat{T}_{m,n}$ (in the original problem with data

$X_1, \ldots, X_m, Y_1, \ldots, Y_n$) and $T_m^*$ (in the oracle problem with data $f(X_1), \ldots, f(X_m)$ and $g(X_1), \ldots, g(X_m)$). We first define a slight modification of the class $\mathcal{F}_{d,\vartheta}$, by setting

$$\mathcal{F}_{d,\vartheta}^* := \{(f,g) \in \mathcal{F}_{d,\vartheta} : \min(v_1, v_2) \geq 1/C\}. \tag{23}$$

THEOREM 13. *Assume the conditions of Theorem 2. Then*

$$\sup_{\phi \in \Phi(\xi)} \sup_{(f,g) \in \mathcal{F}_{d,\vartheta}^*} \max_{\substack{k_X \in \{k_X^{\mathrm{L}}, \ldots, k_X^{\mathrm{U}}\} \\ k_Y \in \{k_Y^{\mathrm{L}}, \ldots, k_Y^{\mathrm{U}}\}}} \frac{\mathbb{E}_{f,g}\{(\widehat{T}_{m,n} - T)^2\}}{\mathbb{E}_f\{(T_m^* - T)^2\}} \cdot \frac{\sigma^2/m}{v_1/m + v_2/n} \to 1$$

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

To understand the implications of this theorem, consider the case where $n$ is at least of the same order as $m$, so that $A := \limsup_{n\to\infty} m/n \in [0, \infty)$. If $\sigma^2/(v_1 + Av_2) > 1$, then the worst-case risk of $\widehat{T}_{m,n}$ is asymptotically better than that of $T_m^*$, and we have an illustration of the super-oracle phenomenon. The one-sample functional (2) corresponds to $A = 0$, and the arguments above reveal that for the Rényi-type functional $\int_{\mathbb{R}^d} f(x)^\kappa \, dx$ with $\kappa \in (1/2, 1)$, the efficient variance in the original problem is strictly smaller than that in the oracle problem since $\sigma^2 \equiv \sigma^2(f) = \mathrm{Var}_f(f(X_1)^{\kappa-1})$ and $v_1 = \kappa^2\sigma^2$ (note that $\sup_{f \in \mathcal{F}_{d,\vartheta}^*} \sigma^2(f) < \infty$ whenever $\lambda_1 > 2 - 2\kappa$). In general, the phenomenon occurs if and only if

$$2\,\mathrm{Cov}_f\big(\phi_{X_1}, (f\phi_{10})_{X_1}\big) < -\,\mathrm{Var}_f(f\phi_{10})_{X_1} - Av_2.$$

One of the surprising aspects of the super-oracle phenomenon is the fact that the estimator $\widehat{T}_{m,n}$ is constructed so as to mimic $T_m^*$, by estimating $f(X_1), \ldots, f(X_m)$ and $g(X_1), \ldots, g(X_m)$, but can in some cases outperform $T_m^*$ itself.

**6. A local asymptotic minimax lower bound.** Before we can state our local asymptotic minimax result, we require some further assumptions on the function $\phi$. For $\xi = (\kappa_1, \kappa_2, \beta^*, L) \in \Xi$ let $\tilde{\Phi}(\xi)$ denote the subset of $\Phi(\xi)$ consisting of those $\phi$ for which:

(i) for all $\mathbf{z} = (u, v) \in \mathcal{Z}$ and $\ell_1 \in [\beta^*]$ we have

$$\max_{\ell_2 \in [\beta^* - \ell_1]} \frac{u^{\ell_1} v^{\ell_2} |\phi_{\ell_1\ell_2}(\mathbf{z})|}{|\phi(\mathbf{z}) + u\phi_{10}(\mathbf{z})| \vee 1} \vee \max_{\ell_2 \in [\beta^* - \ell_1] \setminus \{0\}} \frac{u^{\ell_1+1} v^{\ell_2-1} |\phi_{\ell_1\ell_2}(\mathbf{z})|}{(u|\phi_{01}(\mathbf{z})|) \vee 1} \leq L;$$

(ii) for all $\boldsymbol{\epsilon} = (\epsilon_1, \epsilon_2) \in (-1/L, 1/L)^2$, $\mathbf{z} = (u, v) \in \mathcal{Z}$, and $\ell_1, \ell_2 \in \mathbb{N}_0$ with $\ell_1 + \ell_2 \leq \beta^* - 1$, we have

$$\frac{u^{\ell_1} v^{\ell_2} |\phi_{\ell_1\ell_2}(\mathbf{z} + \boldsymbol{\epsilon}) - \phi_{\ell_1\ell_2}(\mathbf{z})|}{|\phi(\mathbf{z}) + u\phi_{10}(\mathbf{z})| \vee 1} \leq L\left(\left|\frac{\epsilon_1}{u_1}\right|^{(\beta^* - \ell_1) \wedge 1} + \left|\frac{\epsilon_2}{u_2}\right|^{(\beta^* - \ell_2) \wedge 1}\right);$$

$$\frac{u^{\ell_1+1} v^{\ell_2-1} |\phi_{\ell_1\ell_2}(\mathbf{z} + \boldsymbol{\epsilon}) - \phi_{\ell_1\ell_2}(\mathbf{z})|}{(u|\phi_{01}(\mathbf{z})|) \vee 1}$$

$$\leq L\left(\left|\frac{\epsilon_1}{u_1}\right|^{(\beta^* - \ell_1) \wedge 1} + \left|\frac{\epsilon_2}{u_2}\right|^{(\beta^* - \ell_2) \wedge 1}\right) \quad \text{when } \ell_2 \geq 1.$$

To understand these conditions it is instructive to consider the case of $\varphi$-divergences, for which $\phi(u, v) = \varphi(v/u)$ for some function $\varphi$. Here, (i) reduces to requiring that

$$\sup_{w > 0} \left\{ \max_{\ell \in [\beta^*]} \frac{w^\ell |\varphi^{(\ell)}(w)|}{|\varphi(w) - w\varphi'(w)| \vee 1}, \max_{\ell \in [\beta^*] \setminus \{0\}} \frac{w^{\ell-1} |\varphi^{(\ell)}(w)|}{|\varphi'(w)| \vee 1} \right\} < \infty,$$

and a similar reduction holds for (ii). This is satisfied for the Kullback–Leibler divergence and all Rényi divergences. Moreover, when $\phi(u,v)=v$, we have $\phi\in\tilde{\Phi}(0,0,\beta^*,1+1/\beta^*)$ for every $\beta^*>0$.

Now fix $(f,g)\in\mathcal{F}_d^2$ and $\phi:\mathcal{Z}\to\mathbb{R}$ and define the functions

$$h_1(x):=\phi_x+(f\phi_{10})_x-\mathbb{E}\{\phi_{X_1}+(f\phi_{10})_{X_1}\}$$
$$h_2(x):=(f\phi_{01})_x-\mathbb{E}\{(f\phi_{01})_{Y_1}\}.$$

This enables us to define, for each $t=(t_1,t_2)\in\mathbb{R}^2$, the densities

$$f_{t_1}(x):=c_1(t_1)K\big(t_1h_1(x)\big)f(x)\quad\text{and}\quad g_{t_2}(x):=c_2(t_2)K\big(t_2h_2(x)\big)g(x),$$

where $K(t):=1/2+1/(1+e^{-4t})$ and $c_1(\cdot)$, $c_2(\cdot)$ are normalising constants. Our choice of $K$ is made so that $K(0)=K'(0)=1$, that $K$ is smooth, and that $K$ is bounded above and below by positive constants. Now, for each $t=(t_1,t_2)\in\mathbb{R}^2$ we define the sequence of probability measures $(P_{n,t})$ on $\mathbb{R}^{(m+n)\times d}$ so that $P_{n,t}$ has density $f_{m^{-1/2}t_1}^{\otimes m}\otimes g_{n^{-1/2}t_2}^{\otimes n}$ (here we think of $m$ as a function of $n$). It turns out that the family $\{P_{n,t}:t\in\mathbb{R}^2\}$ constitutes a least favourable parametric sub-model for this estimation problem. For an arbitrary probability measure $P$ on $\mathbb{R}^{(m+n)\times d}$, we write $\mathbb{E}_P$ to denote expectation over $(X_1,\dots,X_m,Y_1,\dots,Y_n)^T\sim P$.

We can now state our local asymptotic minimax lower bound, and the consequent optimality property of our estimators $\widehat{T}_{m,n}$.

THEOREM 14. *Fix $d\in\mathbb{N}$, $\vartheta=(\alpha,\beta,\lambda_1,\lambda_2,C)\in\Theta$ and $\xi=(\kappa_1,\kappa_2,\beta^*,L)\in\Xi$. Let $m=m_n$ be any sequence of positive integers such that $m\to\infty$ and $m/n\to A$ for some $A\in[0,\infty]$, let $(f,g)\in\mathcal{F}_{d,\vartheta}$, let $\phi\in\tilde{\Phi}(\xi)$ and let $\mathcal{I}$ denote the set of finite subsets of $\mathbb{R}^2$.*

(i) *For any estimator sequence $(T_{m,n})$, we have that*

$$\sup_{I\in\mathcal{I}}\liminf_{n\to\infty}\max_{t=(t_1,t_2)\in I}n\mathbb{E}_{P_{n,t}}\big[\{T_{m,n}-T(f_{m^{-1/2}t_1},g_{n^{-1/2}t_2})\}^2\big]\geq\frac{1}{A}v_1(f,g)+v_2(f,g).$$

(ii) *There exists $t_0=t_0(d,\vartheta,\xi)\in(0,1]$ such that, for any $t_1,t_2\in(-t_0,t_0)$, we have $(f_{t_1},g_{t_2})\in\mathcal{F}_{d,\tilde{\vartheta}}$, where $\tilde{\vartheta}=(\alpha,\tilde{\beta},\lambda_1,\lambda_2,C/t_0)$ and $\tilde{\beta}:=\min\{\beta,(1\wedge\beta)(\beta^*-1)\}$. In particular, when the conditions of Theorem 2 hold and $\tilde{\beta}=\beta$, the estimators $\widehat{T}_{m,n}$ in (7) satisfy*

$$\sup_{I\in\mathcal{I}}\limsup_{n\to\infty}\max_{t=(t_1,t_2)\in I}n\mathbb{E}_{P_{n,t}}\big[\{\widehat{T}_{m,n}-T(f_{m^{-1/2}t_1},g_{n^{-1/2}t_2})\}^2\big]=\frac{1}{A}v_1(f,g)+v_2(f,g).$$

Recall, for example, that for both the Kullback–Leibler divergence and all Rényi-type divergences, we can take $\beta^*$ large enough that $\tilde{\beta}=\beta$. In these and other cases for which the conditions hold, then, the local asymptotic minimax bounds in Theorem 14 justify the claim that suitably chosen versions of our weighted nearest neighbour estimator (7) are efficient over these classes of densities and functionals.

We conclude with a few extensions of Theorem 14. The condition $\tilde{\beta}=\beta$ can be weakened to $\tilde{\beta}>d/2$ (or in fact $\tilde{\beta}>d/4$ when $d\in\{1,2,3\}$), at the expense of slightly stronger conditions on the tuning parameters in the definition of $\widehat{T}_{m,n}$. Theorem 14(i) implies a (non-local) minimax lower bound over the classes $\mathcal{F}_{d,\vartheta}^*\subseteq\mathcal{F}_{d,\vartheta}$ from (23), and this matches the upper bound in Theorem 2 over $\mathcal{F}_{d,\vartheta}$. Theorem 14(i) may also be extended to broader classes of loss functions, namely those that have closed, convex, symmetric sub-level sets; see van der Vaart and Wellner ((1996), Theorem 3.11.5) for details. Finally, Theorem 3 allows us to extend Theorem 2, and consequently Theorem 14(ii), to $L_q$-losses with $q\in(0,2)$. The combination of these results implies that our estimators are asymptotically optimal in a local asymptotic minimax sense for these $L_q$ losses too; we omit formal statements for brevity.

**Acknowledgements.** The authors are very grateful to the anonymous reviewers for their constructive comments, which helped to improve the paper.

**Funding.** The first author was supported by Engineering and Physical Sciences Reseach Council (EPSRC) New Investigator Award EP/W016117/1.

The second author was supported in part by EPSRC Programme grant EP/N031938/1, EPSRC Fellowship EP/P031447/1 and European Research Council Advanced grant 101019498.

## SUPPLEMENTARY MATERIAL

**Supplementary Material to 'Efficient functional estimation and the super-oracle phenomenon'** (DOI: 10.1214/23-AOS2265SUPP; .pdf). Proofs of results from the main text and auxiliary results.

## REFERENCES

BALDI, P. and RINOTT, Y. (1989). On normal approximations of distributions in terms of dependency graphs. *Ann. Probab.* **17** 1646–1650. MR1048950

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. and SAMWORTH, R. J. (2023). Supplement to "Efficient functional estimation and the super-oracle phenomenon." https://doi.org/10.1214/23-AOS2265SUPP

BERRETT, T. B., SAMWORTH, R. J. and YUAN, M. (2019). Efficient multivariate entropy estimation via $k$-nearest neighbour distances. *Ann. Statist.* **47** 288–318. MR3909934 https://doi.org/10.1214/18-AOS1688

BIAU, G. and DEVROYE, L. (2015). *Lectures on the Nearest Neighbor Method. Springer Series in the Data Sciences*. Springer, Cham. MR3445317 https://doi.org/10.1007/978-3-319-25388-6

BICKEL, P. J. and RITOV, Y. (1988). Estimating integrated squared density derivatives: Sharp best order of convergence estimates. *Sankhyā Ser. A* **50** 381–393. MR1065550

BIRGÉ, L. and MASSART, P. (1995). Estimation of integral functionals of a density. *Ann. Statist.* **23** 11–29. MR1331653 https://doi.org/10.1214/aos/1176324452

DIEUDONNÉ, J. (1969). *Foundations of Modern Analysis. Pure and Applied Mathematics*, *Vol.* 10-*I*. Academic Press, New York. MR0349288

DONOHO, D. L., JOHNSTONE, I. M., KERKYACHARIAN, G. and PICARD, D. (1996). Density estimation by wavelet thresholding. *Ann. Statist.* **24** 508–539. MR1394974 https://doi.org/10.1214/aos/1032894451

GINÉ, E. and NICKL, R. (2008). A simple adaptive estimator of the integrated square of a density. *Bernoulli* **14** 47–61. MR2401653 https://doi.org/10.3150/07-BEJ110

GOLDENSHLUGER, A. and LEPSKI, O. (2014). On adaptive minimax density estimation on $R^d$. *Probab. Theory Related Fields* **159** 479–543. MR3230001 https://doi.org/10.1007/s00440-013-0512-1

HAN, Y., JIAO, J., WEISSMAN, T. and WU, Y. (2020). Optimal rates of entropy estimation over Lipschitz balls. *Ann. Statist.* **48** 3228–3250. MR4185807 https://doi.org/10.1214/19-AOS1927

HERO, A. O., MA, B., MICHEL, O. and GORMAN, J. (2002). Applications of entropic spanning graphs. *IEEE Signal Process. Mag.* **19** 85–95.

IBRAGIMOV, I. A. and KHAS'MINSKIĬ, R. Z. (1991). Asymptotically normal families of distributions and efficient estimation. *Ann. Statist.* **19** 1681–1724. MR1135145 https://doi.org/10.1214/aos/1176348367

JUDITSKY, A. and LAMBERT-LACROIX, S. (2004). On minimax density estimation on $\mathbb{R}$. *Bernoulli* **10** 187–220. MR2046772 https://doi.org/10.3150/bj/1082380217

KANDASAMY, K., KRISHNAMURTHY, A., PÓCZOS, B., WASSERMAN, L. and ROBINS, J. M. (2015). Nonparametric von Mises estimators for entropies, divergences and mutual informations. *NeurIPS* **28**.

KILBAS, A. A., SRIVASTAVA, H. M. and TRUJILLO, J. J. (2006). *Theory and Applications of Fractional Differential Equations. North-Holland Mathematics Studies* **204**. Elsevier, Amsterdam. MR2218073

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

KRISHNAMURTHY, A., KANDASAMY, K., POĆZOS, B. and WASSERMAN, L. (2014). Nonparametric estimation of Rényi divergence and friends. In *Proc. 31st Int. Conf. on Mach. Learn* (*ICML*). **32** 919–927.

LAURENT, B. (1996). Efficient estimation of integral functionals of a density. *Ann. Statist.* **24** 659–681. MR1394981 https://doi.org/10.1214/aos/1032894458

LECAM, L. (1953). On some asymptotic properties of maximum likelihood estimates and related Bayes' estimates. *Univ. Calif. Publ. Stat.* **1** 277–329. MR0054913

LEHMANN, E. L. and CASELLA, G. (1998). *Theory of Point Estimation*, 2nd ed. *Springer Texts in Statistics*. Springer, New York. MR1639875

LEONENKO, N., PRONZATO, L. and SAVANI, V. (2008). A class of Rényi information estimators for multidimensional densities. *Ann. Statist.* **36** 2153–2182. MR2458183 https://doi.org/10.1214/07-AOS539

LEONENKO, N. and SELEZNJEV, O. (2010). Statistical inference for the $\epsilon$-entropy and the quadratic Rényi entropy. *J. Multivariate Anal.* **101** 1981–1994. MR2671196 https://doi.org/10.1016/j.jmva.2010.05.009

MOON, K. R., SRICHARAN, K., GREENEWALD, K. and HERO, A. O. III (2018). Ensemble estimation of information divergence. *Entropy* **20** 560. MR3892642 https://doi.org/10.3390/e20080560

NOWOZIN, S., CSEKE, B. and TOMIOKA, R. (2016). F-GAN: Training generative neural samplers using variational divergence minimization. *Adv. Neural Inf. Process. Syst.*

RYU, J., GANGULY, S., KIM, Y., NOH, Y. and LEE, D. D. (2018). Nearest neighbor density functional estimation based on inverse Laplace transform. *IEEE Trans. Inf. Theory* **68** 3511–3551.

SINGH, S. and PÓCZOS, B. (2016). Finite-sample analysis of fixed-$k$ nearest neighbor density functional estimators. In *Annual Conference on Neural Information Processing Systems* (*NIPS*) 1217–1225.

SINGH, S., SRIPERUMBUDUR, B. K. and PÓCZOS, B. (2018). Minimax estimation of quadratic Fourier functionals. Available at https://arxiv.org/abs/1803.11451.

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. (1997). Superefficiency. In *Festscrift for Lucien Le Cam* (D. Pollard, E. Torgersen and G. Yang, eds.) Springer, Berlin.

VAN DER VAART, A. W. and WELLNER, J. A. (1996). *Weak Convergence and Empirical Processes. Springer Series in Statistics*. Springer, New York. MR1385671 https://doi.org/10.1007/978-1-4757-2545-2

WORNOWIZKI, M. and FRIED, R. (2016). Two-sample homogeneity tests based on divergence measures. *Comput. Statist.* **31** 291–313. MR3481806 https://doi.org/10.1007/s00180-015-0633-3