

Efficient two-sample functional estimation and the super-oracle phenomenon

Richard J. Samworth

University of Cambridge

Conference in honour of Aad van der Vaart's 60th birthday
Leiden

21 June 2019



Tom Berrett

Two-sample functionals

We consider estimation of functionals of the form

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

where $\mathcal{X} := \{x \in \mathbb{R}^d : f(x) > 0, g(x) > 0\}$, using independent d -dimensional observations $X_1, \dots, X_m, Y_1, \dots, Y_n$, where $X_1, \dots, X_m \sim f, Y_1, \dots, Y_n \sim g$.

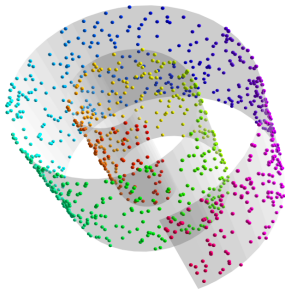
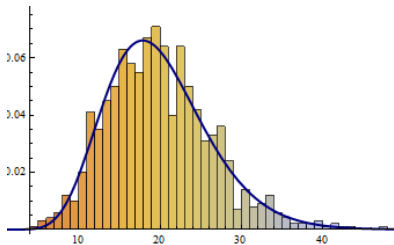
Examples:

- (i) φ -divergences, including TV, Hellinger distances, KL and Rényi divergences
- (ii) $\int_{\mathbb{R}^d} fg$
- (iii) One-sample functionals of the form

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

such as Shannon and Rényi entropies.

Tests of normality (Vasicek, 1976), two-sample testing (Kanamori et al., 2012), dimension reduction (Huber, 1985), image alignment (Viola and Wells 1997), independent component analysis (Comon, 1994), estimation of intrinsic dimension (Carter et al., 2010) and estimation of information flows in deep neural networks (Goldfeld et al., 2018).





Many functionals of interest are non-smooth because of behaviour in low-density regions. E.g.

- For Shannon entropy, $\psi(u, x) = -\log u$
- For Rényi entropy, we take $\psi(u, x) = u^{\kappa-1}$.

When f is bounded away from zero on its support, *efficient* estimators (van der Vaart, 1998) are sometimes obtainable. E.g. when f is s -Hölder smooth on $\{x : f(x) > 0\}$ with $s > d/4$, Laurent (1996) constructed a Shannon entropy estimator $\hat{H}_m$ with

$$m\mathbb{E}[\{\hat{H}_m - H(f)\}^2] \rightarrow \int_{\{x: f(x) > 0\}} f \log^2 f - H(f)^2.$$

The limit is also obtained by the oracle estimator $H_m^* := -\frac{1}{m} \sum_{i=1}^m \log f(X_i)$, and is optimal in a local asymptotic minimax sense (Ibragimov and Khasminskii, 1991; Laurent 1996).



- ▶ **Higher-order influence functions** (Robins, Li, Tchetgen and van der Vaart, 2008; Robins, Li, Mukherjee, Tchetgen and van der Vaart, 2017)
- ▶ **Polynomial approximation** (Cai and Low, 2011; Han, Jiao, Weissman and Wu, 2018)
- ▶ **Sample spacings** (El Haje Hussein and Golubev, 2009)
- ▶ **Histograms** (Hall and Morton, 1993)
- ▶ **Kernel density estimates** (Paninski and Yajima, 2008; Sricharan, Wei and Hero, 2013)



- ▶ **Higher-order influence functions** (Robins, Li, Tchetgen and van der Vaart, 2008; Robins, Li, Mukherjee, Tchetgen and van der Vaart, 2017)
- ▶ **Polynomial approximation** (Cai and Low, 2011; Han, Jiao, Weissman and Wu, 2018)
- ▶ **Sample spacings** (El Haje Hussein and Golubev, 2009)
- ▶ **Histograms** (Hall and Morton, 1993)
- ▶ **Kernel density estimates** (Paninski and Yajima, 2008; Sricharan, Wei and Hero, 2013)

In density estimation, optimal rates of convergence can be very different when densities with unbounded support are allowed (Donoho et al, 1996; Juditsky and Lambert-Lacroix, 2004;

Goldenshluger and Lepski, 2014).

Nearest neighbour estimators

For $k_X \in [m - 1]$, $k_Y \in [n]$, write $\rho_{(k_X),i,X}$ for the distance between X_i and its k_X th nearest neighbour among $\{X_1, \dots, X_m\} \setminus \{X_i\}$, and $\rho_{(k_Y),i,Y}$ for the distance between X_i and its k_Y th nearest neighbour among $\{Y_1, \dots, Y_n\}$.

Our starting point 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 motivates the initial estimator

$$\tilde{T}_{m,n} = \tilde{T}_{m,n,k_X,k_Y} := \frac{1}{m} \sum_{i=1}^m \phi\left(\frac{k_X}{mV_d\rho_{(k_X),i,X}^d}, \frac{k_Y}{nV_d\rho_{(k_Y),i,Y}^d}, X_i\right).$$

Initial classes of densities

For $\beta > 0$, let $\underline{\beta} := \lceil \beta \rceil - 1$, and when $f(x) > 0$ let

$$M_{f,\beta}(x) := \inf \left\{ M \geq 1 : \max_{t \in [\underline{\beta}]} \left(\frac{\|f^{(t)}(x)\|}{f(x)} \right)^{1/t} \right. \\ \left. \bigvee_{\substack{y, z \in B_x(\{2d^{1/2}M\}^{-1}), \\ z \neq y}} \sup \left(\frac{\|f^{(\underline{\beta})}(z) - f^{(\underline{\beta})}(y)\|}{f(x)\|z - y\|^{\beta - \underline{\beta}}} \right)^{1/\beta} \leq M \right\};$$

E.g. if f is uniform on $B_0(1)$, then $M_{f,\beta}(x) = \{2d^{1/2}(1 - \|x\|)\}^{-1}$ for $\|x\| < 1$.

For $\theta = (\alpha, \beta, \lambda, C) \in (0, \infty)^4$, let

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



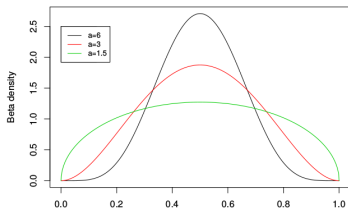
- ▶ Gaussian densities belong to $\mathcal{G}_{d,\theta}$ for any $\alpha, \beta > 0$, $\lambda \in (0, 1)$ and sufficiently large $C > 0$.



- ▶ Gaussian densities belong to $\mathcal{G}_{d,\theta}$ for any $\alpha, \beta > 0$, $\lambda \in (0, 1)$ and sufficiently large $C > 0$.
- ▶ Multivariate- t densities with ν degrees of freedom belong to $\mathcal{G}_{d,\theta}$ for any $\alpha \in (0, \nu)$, any $\beta > 0$, $\lambda \in (0, \frac{\nu}{\nu+d})$ and $C > 0$ sufficiently large.

Examples

- ▶ Gaussian densities belong to $\mathcal{G}_{d,\theta}$ for any $\alpha, \beta > 0, \lambda \in (0, 1)$ and sufficiently large $C > 0$.
- ▶ Multivariate- t densities with ν degrees of freedom belong to $\mathcal{G}_{d,\theta}$ for any $\alpha \in (0, \nu)$, any $\beta > 0, \lambda \in (0, \frac{\nu}{\nu+d})$ and $C > 0$ sufficiently large.



- ▶ The Beta-type density $f(x) \propto_d {}_{a,b} \|x\|^{a-1} (1 - \|x\|)^{b-1} \mathbb{1}_{\{\|x\| \leq 1\}}$ belongs to $\mathcal{G}_{d,\theta}$ for any $\alpha, \beta > 0, \lambda \in (0, \frac{b}{b+d-1})$, and $C \geq C_0 = C_0(\alpha, \beta, \lambda) > 0$.

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 $C \geq 1/(1 - \lambda)$.

Let $\mathcal{F}_d$ denote the class of all densities on $\mathbb{R}^d$.

For $\Theta = (0, \infty)^6$, $\vartheta = (\alpha, \beta, \lambda_1, \lambda_2, \gamma, C) \in \Theta$, let $M_\beta(x) := M_{f,\beta}(x) \vee M_{g,\beta}(x)$ and set

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

The third requirement can be thought of as a quantitative absolute continuity condition.

Classes of functionals

Let $\Xi := [0, \infty)^2 \times (0, \infty)^2 \times (1, \infty)$ and $\mathcal{Z} := (0, \infty)^2 \times \mathbb{R}^d$. For

$\xi = (\kappa_1, \kappa_2, \beta_1^*, \beta_2^*, L) \in \Xi$, let $\Phi \equiv \Phi(\xi)$ denote all $\phi : \mathcal{Z} \rightarrow \mathbb{R}$ for which

- (i) writing $\underline{\beta}_1^* := \lceil \beta_1^* \rceil - 1$ and $\underline{\beta}_2^* := \lceil \beta_2^* \rceil - 1$, the partial derivative $\phi_{\underline{\beta}_1^* \underline{\beta}_2^*}$ exists; also, for $\epsilon = (\epsilon_1, \epsilon_2, 0) \in \left(-\frac{1}{L}, \frac{1}{L}\right)^2 \times \{0\}$ and $\mathbf{z} = (u, v, x) \in \mathcal{Z}$,

$$\left| \phi(\mathbf{z} + \epsilon \circ \mathbf{z}) - \mathcal{T}_{\underline{\beta}_1^*, \underline{\beta}_2^*}^{\phi}(\mathbf{z}; u\epsilon_1, v\epsilon_2) \right| \leq L(u^{-\kappa_1} \vee u^L)(v^{-\kappa_2} \vee v^L)(|\epsilon_1|^{\beta_1^*} + |\epsilon_2|^{\beta_2^*});$$

Classes of functionals

Let $\Xi := [0, \infty)^2 \times (0, \infty)^2 \times (1, \infty)$ and $\mathcal{Z} := (0, \infty)^2 \times \mathbb{R}^d$. For

$\xi = (\kappa_1, \kappa_2, \beta_1^*, \beta_2^*, L) \in \Xi$, let $\Phi \equiv \Phi(\xi)$ denote all $\phi : \mathcal{Z} \rightarrow \mathbb{R}$ for which

- (i) writing $\underline{\beta}_1^* := \lceil \beta_1^* \rceil - 1$ and $\underline{\beta}_2^* := \lceil \beta_2^* \rceil - 1$, the partial derivative $\phi_{\underline{\beta}_1^* \underline{\beta}_2^*}$ exists; also, for $\epsilon = (\epsilon_1, \epsilon_2, 0) \in \left(-\frac{1}{L}, \frac{1}{L}\right)^2 \times \{0\}$ and $\mathbf{z} = (u, v, x) \in \mathcal{Z}$,

$$\left| \phi(\mathbf{z} + \epsilon \circ \mathbf{z}) - \mathcal{T}_{\underline{\beta}_1^*, \underline{\beta}_2^*}^\phi(\mathbf{z}; u\epsilon_1, v\epsilon_2) \right| \leq L(u^{-\kappa_1} \vee u^L)(v^{-\kappa_2} \vee v^L)(|\epsilon_1|^{\beta_1^*} + |\epsilon_2|^{\beta_2^*});$$

- (ii) for $\mathbf{z} = (u, v, x) \in \mathcal{Z}$, $l_1 \in [\underline{\beta}_1^*]$ and $l_2 \in [\underline{\beta}_2^*]$,

$$\left| u^{l_1} v^{l_2} \phi_{l_1 l_2}(\mathbf{z}) \right| \leq L(u^{-\kappa_1} \vee u^L)(v^{-\kappa_2} \vee v^L);$$

Classes of functionals

Let $\Xi := [0, \infty)^2 \times (0, \infty)^2 \times (1, \infty)$ and $\mathcal{Z} := (0, \infty)^2 \times \mathbb{R}^d$. For

$\xi = (\kappa_1, \kappa_2, \beta_1^*, \beta_2^*, L) \in \Xi$, let $\Phi \equiv \Phi(\xi)$ denote all $\phi : \mathcal{Z} \rightarrow \mathbb{R}$ for which

- (i) writing $\underline{\beta}_1^* := \lceil \beta_1^* \rceil - 1$ and $\underline{\beta}_2^* := \lceil \beta_2^* \rceil - 1$, the partial derivative $\phi_{\underline{\beta}_1^* \underline{\beta}_2^*}$ exists; also, for $\epsilon = (\epsilon_1, \epsilon_2, 0) \in \left(-\frac{1}{L}, \frac{1}{L}\right)^2 \times \{0\}$ and $\mathbf{z} = (u, v, x) \in \mathcal{Z}$,

$$\left| \phi(\mathbf{z} + \epsilon \circ \mathbf{z}) - \mathcal{T}_{\underline{\beta}_1^*, \underline{\beta}_2^*}^{\phi}(\mathbf{z}; u\epsilon_1, v\epsilon_2) \right| \leq L(u^{-\kappa_1} \vee u^L)(v^{-\kappa_2} \vee v^L)(|\epsilon_1|^{\beta_1^*} + |\epsilon_2|^{\beta_2^*});$$

- (ii) for $\mathbf{z} = (u, v, x) \in \mathcal{Z}$, $l_1 \in [\underline{\beta}_1^*]$ and $l_2 \in [\underline{\beta}_2^*]$,

$$\left| u^{l_1} v^{l_2} \phi_{l_1 l_2}(\mathbf{z}) \right| \leq L(u^{-\kappa_1} \vee u^L)(v^{-\kappa_2} \vee v^L);$$

- (iii) writing $\mathcal{E}_r := \{0\}^2 \times B_0(r)$, as $r \searrow 0$ we have

$$\sup_{\epsilon \in \mathcal{E}_r} \sup_{\mathbf{z} \in \mathcal{Z}} \max \left\{ \left| \frac{\phi(\mathbf{z} + \epsilon)}{\phi(\mathbf{z})} - 1 \right|, \left| \frac{\phi_{10}(\mathbf{z} + \epsilon)}{\phi_{10}(\mathbf{z})} - 1 \right|, \left| \frac{\phi_{01}(\mathbf{z} + \epsilon)}{\phi_{01}(\mathbf{z})} - 1 \right| \right\} \rightarrow 0.$$

Classes of functionals

Let $\Xi := [0, \infty)^2 \times (0, \infty)^2 \times (1, \infty)$ and $\mathcal{Z} := (0, \infty)^2 \times \mathbb{R}^d$. For $\xi = (\kappa_1, \kappa_2, \beta_1^*, \beta_2^*, L) \in \Xi$, let $\Phi \equiv \Phi(\xi)$ denote all $\phi : \mathcal{Z} \rightarrow \mathbb{R}$ for which

- (i) writing $\underline{\beta}_1^* := \lceil \beta_1^* \rceil - 1$ and $\underline{\beta}_2^* := \lceil \beta_2^* \rceil - 1$, the partial derivative $\phi_{\underline{\beta}_1^* \underline{\beta}_2^*}$ exists; also, for $\epsilon = (\epsilon_1, \epsilon_2, 0) \in \left(-\frac{1}{L}, \frac{1}{L}\right)^2 \times \{0\}$ and $\mathbf{z} = (u, v, x) \in \mathcal{Z}$,

$$\left| \phi(\mathbf{z} + \epsilon \circ \mathbf{z}) - \mathcal{T}_{\underline{\beta}_1^*, \underline{\beta}_2^*}^\phi(\mathbf{z}; u\epsilon_1, v\epsilon_2) \right| \leq L(u^{-\kappa_1} \vee u^L)(v^{-\kappa_2} \vee v^L)(|\epsilon_1|^{\beta_1^*} + |\epsilon_2|^{\beta_2^*});$$

- (ii) for $\mathbf{z} = (u, v, x) \in \mathcal{Z}$, $l_1 \in [\underline{\beta}_1^*]$ and $l_2 \in [\underline{\beta}_2^*]$,

$$\left| u^{l_1} v^{l_2} \phi_{l_1 l_2}(\mathbf{z}) \right| \leq L(u^{-\kappa_1} \vee u^L)(v^{-\kappa_2} \vee v^L);$$

- (iii) when $\phi(u, v, x) = w(x)\tilde{\phi}(u, v)$,

$$\sup_{\epsilon \in (-r, r)} \sup_{x \in \mathbb{R}^d} \left| \frac{w(x + \epsilon)}{w(x)} - 1 \right| \xrightarrow{r \searrow 0} 0.$$

For $k, I \in \mathbb{N}$ and $c \in (0, 1)$, define $[I]^+ := \{0, 1, \dots, I\}$ and

$$\mathcal{W}_{I,c}^{(k)} := \left\{ w = (w_1, \dots, w_k) \in \mathbb{R}^k : \sum_{j=1}^k w_j = 1 \text{ and } w_j = 0 \text{ for } j < ck, \right. \\ \left. \sum_{j=1}^k j^{\frac{2l}{d}-i} w_j = 0 \text{ for } (l, i) \in ([\lfloor d/2 \rfloor - 1]^+ \times [I]^+) \setminus \{(0, 0)\} \right\}.$$

Fixing $\xi = (\kappa_1, \kappa_2, \beta_1^*, \beta_2^*, L) \in \Xi$ and $c \in (0, 1)$, and for $w_X \in \mathcal{W}_{[\frac{\beta_1^*}{2}],c}^{(k_X)}$ and $w_Y \in \mathcal{W}_{[\frac{\beta_2^*}{2}],c}^{(k_Y)}$, we can now define our weighted functional estimators as

$$\hat{T}_{m,n} = \hat{T}_{m,n}^{w_X, w_Y} := \sum_{j_X=1}^{k_X} \sum_{j_Y=1}^{k_Y} w_{X,j_X} w_{Y,j_Y} \tilde{T}_{m,n,j_X,j_Y}.$$

For $x \in \mathbb{R}^d$, let $\phi_x := \phi(f(x), g(x), x)$, $(f\phi_{10})_x := f(x)\phi_{10}(f(x), g(x), x)$ and $(f\phi_{01})_x := f(x)\phi_{01}(f(x), g(x), x)$. Set

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

Fixing $d \in \mathbb{N}$, $\vartheta = (\alpha, \beta, \lambda_1, \lambda_2, \gamma, C) \in \Theta$ and $\xi = (\kappa_1, \kappa_2, \beta_1^*, \beta_2^*, L) \in \Xi$, set

$$\begin{aligned} \zeta &:= \frac{\kappa_1}{\lambda_1} + \frac{\kappa_2}{\lambda_2} + \frac{d(\kappa_1 + \kappa_2)}{\alpha} \\ \tau_1 &:= 1 - \max\left(\frac{d}{2\beta}, \frac{d}{2(2 \wedge \beta) + d}, \frac{d}{4\beta_1^*}, \frac{1}{2\lambda_1(1 - \zeta)}\right) \\ \tau_2 &:= 1 - \max\left(\frac{d}{2\beta}, \frac{d}{2(2 \wedge \beta) + d}, \frac{d}{4\beta_2^*}, \frac{1}{2\lambda_2(1 - \zeta)}\right) \\ \gamma_* &:= \frac{(2\alpha + d)(1 + 2\kappa_2)}{2\alpha + d - 2(\alpha + d)\kappa_1}. \end{aligned}$$

Asymptotic risk

Theorem. Fix $d \in \mathbb{N}$, $\vartheta = (\alpha, \beta, \lambda_1, \lambda_2, \gamma, C) \in \Theta$, $\xi = (\kappa_1, \kappa_2, \beta_1^*, \beta_2^*, L) \in \Xi$.

Assume that $\zeta < 1/2$, $\tau_1 > 1/\beta_1^*$, $\tau_2 > 1/\beta_2^*$ and that $\gamma > \gamma_* > 0$.

Let (k_X^L) , (k_Y^L) , (k_X^U) and (k_Y^U) satisfy $k_X^L m^{-1/\beta_1^*} \wedge k_Y^L n^{-1/\beta_2^*} \rightarrow \infty$ and $k_X^U m^{-(\tau_1-\epsilon)} \vee k_Y^U n^{-(\tau_2-\epsilon)} \rightarrow 0$ for some $\epsilon > 0$.

Then for each $c \in (0, 1)$, each $w_X = w_X^{(k_X)} \in \mathcal{W}_{[\beta_1^*/2], c}^{(k_X)}$ with $\|w_X\|_1 \leq 1/c$, and each $w_Y = w_Y^{(k_Y)} \in \mathcal{W}_{[\beta_2^*/2], c}^{(k_Y)}$ with $\|w_Y\|_1 \leq 1/c$, we have

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

as $m, n \rightarrow \infty$.

Asymptotic risk

Theorem. Fix $d \in \mathbb{N}$, $\vartheta = (\alpha, \beta, \lambda_1, \lambda_2, \gamma, C) \in \Theta$, $\xi = (\kappa_1, \kappa_2, \beta_1^*, \beta_2^*, L) \in \Xi$.

Assume that $\zeta < 1/2$, $\tau_1 > 1/\beta_1^*$, $\tau_2 > 1/\beta_2^*$ and that $\gamma > \gamma_* > 0$.

Let (k_X^L) , (k_Y^L) , (k_X^U) and (k_Y^U) satisfy $k_X^L m^{-1/\beta_1^*} \wedge k_Y^L n^{-1/\beta_2^*} \rightarrow \infty$ and $k_X^U m^{-(\tau_1-\epsilon)} \vee k_Y^U n^{-(\tau_2-\epsilon)} \rightarrow 0$ for some $\epsilon > 0$.

Then for each $c \in (0, 1)$, each $w_X = w_X^{(k_X)} \in \mathcal{W}_{[\beta_1^*/2], c}^{(k_X)}$ with $\|w_X\|_1 \leq 1/c$, and each $w_Y = w_Y^{(k_Y)} \in \mathcal{W}_{[\beta_2^*/2], c}^{(k_Y)}$ with $\|w_Y\|_1 \leq 1/c$, we have

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

as $m, n \rightarrow \infty$. We can take, e.g., $k_X = m^{1/\beta_1^*} \log m$, $k_Y = n^{1/\beta_2^*} \log n$.

Let

$$\tilde{\mathcal{F}}_{d,\vartheta} := \left\{ (f, g) \in \mathcal{F}_{d,\vartheta} : \int_{\mathcal{X}} f(x)^4 g(x)^{-\frac{4\alpha+3d}{2\alpha+d}\gamma} dx \leq C, \min(v_1, v_2) \geq C^{-1} \right\}.$$

Then, under slightly stronger conditions on γ , (k_X^U) and (k_Y^U) ,

$$\sup_{\phi \in \Phi(\xi)} \sup_{(f,g) \in \tilde{\mathcal{F}}_{d,\vartheta}} \max_{\substack{k_X \in \{k_X^L, \dots, k_X^U\} \\ k_Y \in \{k_Y^L, \dots, k_Y^U\}}} d_K \left(\mathcal{L} \left(\frac{\hat{T}_{m,n} - T}{\{m^{-1}v_1 + n^{-1}v_2\}^{1/2}}, N(0, 1) \right) \right) \rightarrow 0.$$

This result can be used to obtain asymptotically valid confidence intervals for T .



Writing

$$\tilde{\phi}(u, v, x) := \phi(u, v, x) - \phi(u, g(x), x) - \phi(f(x), v, x) + \phi(f(x), g(x), x)$$

we have that $\tilde{\phi}(f(x), g(x), x) \equiv \tilde{\phi}_{10}(f(x), g(x), x) \equiv \tilde{\phi}_{01}(f(x), g(x), x) \equiv 0$.

We can therefore linearise the unweighted estimator $\tilde{T}_{m,n}$, and write

$$\tilde{T}_{m,n} - \mathbb{E}\tilde{T}_{m,n} = \tilde{T}_m^{(1)} - \mathbb{E}\tilde{T}_m^{(1)} + \tilde{T}_{m,n}^{(2)} - \mathbb{E}\tilde{T}_{m,n}^{(2)} + o_p(m^{-1/2} + n^{-1/2}),$$

where

$$\begin{aligned}\tilde{T}_m^{(1)} &:= \frac{1}{m} \sum_{i=1}^m \phi\left(\frac{k_X}{mV_d\rho_{(k_X),i,X}^d}, g(X_i), X_i\right) \\ \tilde{T}_{m,n}^{(2)} &:= \frac{1}{m} \sum_{i=1}^m \left\{ \phi\left(f(X_i), \frac{k_Y}{nV_d\rho_{(k_Y),i,Y}^d}, X_i\right) - \phi\left(f(X_i), g(X_i), X_i\right) \right\}.\end{aligned}$$



Moreover, $\tilde{T}_{m,n}^{(2)} - \mathbb{E}\tilde{T}_{m,n}^{(2)} = \tilde{T}_n^{(2)} - \mathbb{E}\tilde{T}_n^{(2)} + o_p(m^{-1/2})$, where

$$\tilde{T}_n^{(2)} := \mathbb{E} \left[\phi \left(f(X_1), \frac{k_Y}{nV_d\rho_{(k_Y),1,Y}^d}, X_1 \right) \mid Y_1, \dots, Y_n \right] - T(f, g).$$

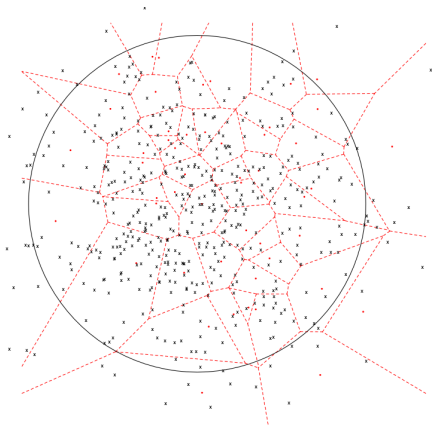
Hence

$$\tilde{T}_{m,n} - \mathbb{E}\tilde{T}_{m,n} = \tilde{T}_m^{(1)} - \mathbb{E}\tilde{T}_m^{(1)} + \tilde{T}_n^{(2)} - \mathbb{E}\tilde{T}_n^{(2)} + o_p(m^{-1/2} + n^{-1/2}),$$

and we have approximated $\tilde{T}_{m,n}$ by a sum of a random variable that depends only on $X_1, \dots, X_m$ and another that depends only on $Y_1, \dots, Y_n$.

Sketch of asymptotic normality proof

For $\tilde{T}_m^{(1)}$, we can partition $\{x : f(x)M_{f,\beta}(x)^{-d} \geq kn^{-1} \log^2 n\}$ using the Voronoi cells associated to a Poisson process with intensity $\frac{m}{k_X} f(\cdot)$. Distant cells are roughly independent and the CLT of [Baldi and Rinott \(1989\)](#) can be applied.



The super-oracle phenomenon

Given access to the unknown $\{f(X_1), \dots, f(X_m), g(X_1), \dots, g(X_m)\}$, the natural oracle estimator of T is

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

This is unbiased, and $m^{1/2}(T_m^* - T) \xrightarrow{d} N(0, \sigma^2)$, where $\sigma^2 := \text{Var}_f \phi_{X_1}$.

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

Under the conditions of our asymptotic risk result,

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

as $m, n \rightarrow \infty$.

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

Suppose that $A := \limsup_{n \rightarrow \infty} m/n \in [0, \infty)$. If $\sigma^2/(v_1 + Av_2) > 1$, then the worst-case risk of $\hat{T}_{m,n}$ is asymptotically better than that of T_m^* , and we have an illustration of the super-oracle phenomenon.

One-sample functionals correspond to $A = 0$, and the Rényi-type functional $\int_{\mathbb{R}^d} f(x)^\kappa dx$ with $\kappa \in (1/2, 1)$ provides such an example, since $v_1/\sigma^2 = \kappa^2$.

In general, the phenomenon occurs if and only if

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



Fix $d \in \mathbb{N}$, $\vartheta = (\alpha, \beta, \lambda_1, \lambda_2, \gamma, C) \in \Theta$.

Let $m = m_n$ be any sequence of positive integers such that $m/n \rightarrow A$ for some $A \in [0, \infty]$, let $(f, g) \in \mathcal{F}_{d, \vartheta}$ and let $(T_{m, n})$ be any estimator sequence.

Write $\mathcal{I}$ for the set of finite subsets of $\mathbb{R}^2$.

Then, for ϕ belonging to a slightly restricted class of functionals $\tilde{\Phi}$, there exists a least favourable sub-model $\{P_{n, t} : t \in \mathbb{R}^2\}$ such that

$$\sup_{I \in \mathcal{I}} \liminf_{n \rightarrow \infty} \max_{t \in I} \mathbb{E}_{P_{n, t}} \left[\frac{\{T_{m, n} - T(f, g)\}^2}{m^{-1}v_1(f, g) + n^{-1}v_2(f, g)} \right] \geq 1.$$



Fix $d \in \mathbb{N}$, $\vartheta = (\alpha, \beta, \lambda_1, \lambda_2, \gamma, C) \in \Theta$.

Let $m = m_n$ be any sequence of positive integers such that $m/n \rightarrow A$ for some $A \in [0, \infty]$, let $(f, g) \in \mathcal{F}_{d, \vartheta}$ and let $(T_{m, n})$ be any estimator sequence.

Write $\mathcal{I}$ for the set of finite subsets of $\mathbb{R}^2$.

Then, for ϕ belonging to a slightly restricted class of functionals $\tilde{\Phi}$, there exists a least favourable sub-model $\{P_{n, t} : t \in \mathbb{R}^2\}$ such that

$$\sup_{I \in \mathcal{I}} \liminf_{n \rightarrow \infty} \max_{t \in I} \mathbb{E}_{P_{n, t}} \left[\frac{\{T_{m, n} - T(f, g)\}^2}{m^{-1}v_1(f, g) + n^{-1}v_2(f, g)} \right] \geq 1.$$

Moreover, there exists $t_0 = t_0(d, \vartheta, \tilde{\Phi}) > 0$ such that, for any $t \in (-t_0, t_0)^2$, the density of $P_{n, t}$ belongs to $\mathcal{F}_{d, \tilde{\vartheta}}$, where $\tilde{\vartheta} = (\alpha, \tilde{\beta}, \lambda_1, \lambda_2, \gamma, C/t_0)$ for some suitable $\tilde{\beta} > 0$.



- ▶ Nearest neighbour methods provide a conceptually simple and computationally efficient way of estimating functionals.



- ▶ Nearest neighbour methods provide a conceptually simple and computationally efficient way of estimating functionals.
- ▶ When $d \geq 4$, weighted versions are required for efficiency.



- ▶ Nearest neighbour methods provide a conceptually simple and computationally efficient way of estimating functionals.
- ▶ When $d \geq 4$, weighted versions are required for efficiency.
- ▶ It may be the case that the worst-case risk is better than that of the natural oracle estimator.



- ▶ Nearest neighbour methods provide a conceptually simple and computationally efficient way of estimating functionals.
- ▶ When $d \geq 4$, weighted versions are required for efficiency.
- ▶ It may be the case that the worst-case risk is better than that of the natural oracle estimator.

Main reference:

- ▶ Berrett, T. B. and Samworth, R. J. (2019) Efficient functional estimation and the super-oracle phenomenon. <https://arxiv.org/abs/1904.09347>.

Happy Birthday Dude, Where is the Party? :D



Other references

- Baldi, P. and Rinott, Y. (1989) On normal approximations of distributions in terms of dependency graphs. *Ann. Prob.*, **17**, 1646–1650.
- Cai, T. T. and Low, M. G. (2011) Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. **39**, 1012–1041.
- Carter, K. M., Raich, R. and Hero, A. O. (2010) On local intrinsic dimension estimation and its applications. *IEEE Trans. Signal Process.*, **58**, 650–663.
- Comon, P. (1994). Independent component analysis, a new concept?. *Signal Process.*, **36**, 287–314.
- Donoho, D. L., Johnstone, I. M., Kerkycharian, G. and Picard, D. (1996) Density estimation by wavelet thresholding. *Ann. Statist.*, **24**, 508–539.
- El Haje Hussein, F. and Golubev, Yu. (2009) On entropy estimation by m -spacing method, *J. Math. Sci.*, **163**, 290–309.
- Goldenshluger, A. and Lepski, O. (2014) On adaptive minimax density estimation on R^d . *Probab. Th. Rel. Fields*, **159**, 479–543.
- Goldfeld, Z., Greenewald, K., Polyanskiy, Y. and Weed, J. (2019). Convergence of smoothed empirical measures with applications to entropy estimation. <https://arxiv.org/abs/1905.13576>.
- Hall, P. and Morton, S. C. (1993) On the estimation of entropy. *Ann. Inst. Statist. Math.*, **45**, 69–88.
- Han, Y., Jiao, J., Weissman, T. and Wu, Y. (2018) Optimal rates of entropy estimation over Lipschitz balls. <https://arxiv.org/abs/1711.02141>.
- Huber, P. (1985) Projection pursuit. *Ann. Statist.*, **13**, 435–475.
- Ibragimov, I. A. and Khas'minskii (1991) Asymptotically normal families of distributions and efficient estimation. *Ann. Statist.*, **19**, 1681–1724.
- Juditsky, A. and Lambert-Lacroix, S. (2004) On minimax density estimation on $\mathbb{R}$. *Bernoulli*, **10**, 187–220.
- Kanamori, T., Suzuki, T. and Sugiyama, M. (2012) f -divergence estimation and two-sample homogeneity test under semiparametric density-ratio models. *IEEE. Trans. Inf. Theory*, **58**, 708–720.
- Laurent, B. (1996) Efficient estimation of integral functionals of a density. *Ann. Statist.*, **24**, 659–681.
- Paninski, L. and Yajima, M. (2008) Undersmoothed kernel entropy estimators. *IEEE Trans. Inf. Th.*, **54**, 4384–4388.
- Robins, J. Li, L., Tchetgen, E. and van der Vaart, A. (2008) Higher order influence functions and minimax estimation of nonlinear functionals. *IMS Collections Probability and Statistics: Essays in Honor of David A. Freedman*, **2**, 335–421.
- Robins, J. M., Li, L., Mukherjee, R., Tchetgen, E. and van der Vaart, A (2017) Minimax estimation of a functional on a structured high-dimensional model. *Ann. Statist.*, **45**, 1951–1987.
- Sricharan, K. Wei, D. and Hero, A. O. (2013) Ensemble Estimators for Multivariate Entropy Estimation. *IEEE Trans. Inf. Th.*, **59**, 4374–4388.
- van der Vaart, A. W. (1998) *Asymptotic Statistics*, Cambridge University Press, Cambridge.
- Vasicek, O. (1976) A test for normality based on sample entropy. *J. Roy. Statist. Soc., Ser. B.*, **38**, 54–59.
- Viola, P. and Wells, W. M. (1997) Alignment by maximization of mutual information (1997). *Int. J. Comput. Vis.*, **24**, 137–154.