A. The von Mises Expansion

Before diving into the auxiliary results of Section 5, let us first derive some properties of the von Mises expansion. It is a simple calculation to verify that the Gateaux derivative is simply the functional derivative of ϕ in the event that $T(F) = \int \phi(f)$.

Lemma 8. *Let $T(F) = \int \phi(f)d\mu$ where $f = dF/d\mu$ is the Radon-Nikodym derivative, ϕ is differentiable and let G be some other distribution with density $g = dG/d\mu$. Then:*

$$dT(G; F - G) = \int \frac{\partial \phi(g(x))}{\partial g(x)}(f(x) - g(x))d\mu(x). \tag{11}$$

Proof.

$$\begin{aligned} dT(G; F-G) &= \lim_{\tau \to 0} \frac{T(G + \tau(F-G)) - T(G)}{\tau} = \lim_{\tau \to 0} \int \frac{1}{\tau} \left[\phi(g(x) + \tau(f(x) - g(x))) - \phi(g(x))\right] d\mu(x) \\ &= \int \lim_{\tau \to 0} \frac{1}{\tau} \left[\phi(g(x) + \tau(f(x) - g(x))) - \phi(g(x))\right] d\mu(x) \\ &= \int \frac{\partial \phi(g(x))}{\partial g(x)}(f(x) - g(x))d\mu(x) \end{aligned}$$

□

We now demonstrate that the remainder for the tth order von Mises expansion is $O(\|p - \hat{p}\|_{t+1}^{t+1} + \|q - \hat{q}\|_{t+1}^{t+1})$ under the assumption that $p, \hat{p}, q, \hat{q}$ are all bounded above and below.

Lemma 9. *Let $T(p, q) = \int p^\alpha q^\beta d\mu$ and uppose that $p, \hat{p}, q, \hat{q}$ are all bounded from above and below. Then R_t, the remainder of the tth order von Mises expansion of $T(p, q)$ around $T(\hat{p}, \hat{q})$ satisfies:*

$$R_t = O\left(\|\hat{p} - p\|_t^t + \|\hat{q} - q\|_t^t\right) \tag{12}$$

Proof. The tth order term in the von Mises expansion is:

$$\frac{1}{t!}\sum_{a=0}^{t} \binom{t}{a} \int \frac{\partial^t \hat{p}^\alpha(x)\hat{q}^\beta}{\partial \hat{p}(x)^a \partial \hat{q}(x)^{t-a}}(p(x) - \hat{p}(x))^a (q(x) - \hat{q}(x))^{t-a} dx =$$

$$\frac{1}{t!}\sum_{a=0}^{t} \binom{t}{a} \int \prod_{i=0}^{a}(\alpha - i) \prod_{i=0}^{t-a}(\beta - i)\hat{p}^{\alpha-a}(x)\hat{q}^{\beta-(t-a)}(x)(p(x) - \hat{p}(x))^a (q(x) - \hat{q}(x))^{t-a} dx,$$

where $\prod_{i=0}^{0} a_i = 1$. If we are to take a $t-1$st order expansion, the remainder is of the same form as the tth term, except that the terms $\hat{p}^{\alpha-a}(x), \hat{q}^{\beta-(t-a)}(x)$ are replaced by functions $\xi_1^{\alpha-a}(x), \xi_2^{\beta-(t-a)}(x)$ for some functions ξ_1, ξ_2 that are bounded between $p, \hat{p}$ and $q, \hat{q}$ respectively. In our setting, $p, q \in [\kappa_l, \kappa_u]$ and $\hat{p}, \hat{q} \in [\kappa_l - \epsilon, \kappa_u + \epsilon]$ so ξ_1, ξ_2 are bounded functions. With this bound, we can simplify the remainder term R_{t-1} to:

$$R_{t-1} \leq C(\alpha, \beta, \kappa_l, \kappa_u, \epsilon, t)\frac{1}{t!}\sum_{a=0}^{t} \binom{t}{a} \int |p(x) - \hat{p}(x)|^a |q(x) - \hat{q}(x)|^{t-a} dx.$$

Looking at the integral pointwise, either $|p(x) - \hat{p}(x)| \leq |q(x) - \hat{q}(x)|$ in which case the expression is upper bounded by $|q(x) - \hat{q}(x)|^t$ or the opposite is true in which case it is bounded by $|p(x) - \hat{p}(x)|^t$. Either way, we can upper bound the integral by the sum. This gives:

$$R_{t-1} \leq C(\alpha, \beta, \kappa_l, \kappa_u, \epsilon, t)\frac{2^t}{t!}\left(\|p - \hat{p}\|_t^t + \|q - \hat{q}\|_t^t\right).$$

□

In many cases, the constant can be worked out:

1. If $\alpha = \beta = 1$, then $R_1 = \alpha\beta$ while $R_2, \ldots, = 0$.
2. If $\alpha, \beta > 0, \alpha+\beta = 1$ as in the Rényi Divergence, $R_2 = 1$ while $R_3 = \frac{5}{6}\kappa_\epsilon^{-2}\alpha\beta$ where $\kappa_\epsilon = \min\{\kappa_l - \epsilon, (\kappa_u + \epsilon)^{-1}\}$.

The first order von Mises expansion is:

$$\begin{aligned}
T(p,q) &= T(\hat{p},\hat{q}) + \int \frac{\partial \hat{p}^\alpha(x)\hat{q}^\beta(x)}{\partial \hat{p}(x)}(p(x)-\hat{p}(x)) + \int \frac{\partial \hat{p}^\alpha(x)\hat{q}^\beta(x)}{\partial \hat{q}(x)}(q(x)-\hat{q}(x)) + O(\|p-\hat{p}\|_2^2 + \|q-\hat{q}\|_2^2) \\
&= T(\hat{p},\hat{q}) + \int \alpha\hat{p}^{\alpha-1}(x)\hat{q}^\beta(x)(p(x)-\hat{p}(x)) + \int \beta\hat{p}^\alpha(x)\hat{q}^{\beta-1}(x)(q(x)-\hat{q}(x)) + O(\|p-\hat{p}\|_2^2 + \|q-\hat{q}\|_2^2) \\
&= (1-\alpha-\beta)T(\hat{p},\hat{q}) + \int \alpha\hat{p}^{\alpha-1}(x)\hat{q}^\beta(x)p(x) + \int \beta\hat{p}^\alpha(x)\hat{q}^{\beta-1}(x)q(x) + O(\|p-\hat{p}\|_2^2 + \|q-\hat{q}\|_2^2) \\
&= C_1 T(\hat{p},\hat{q}) + \theta^p_{1,1} + \theta^q_{1,1} + R_2.
\end{aligned}$$

The second order expansion is computed similarly. The three second order terms are:

$$\begin{aligned}
&\frac{1}{2}\int \alpha(\alpha-1)\hat{p}^{\alpha-2}(x)\hat{q}^\beta(x)(p(x)-\hat{p}(x))^2 \\
&\int \alpha\beta\hat{p}^{\alpha-1}(x)\hat{q}^{\beta-1}(x)(p(x)-\hat{p}(x))(q(x)-\hat{q}(x)) \\
&\frac{1}{2}\int \beta(\beta-1)\hat{p}^\alpha(x)\hat{q}^{\beta-1}(x)(q(x)-\hat{q}(x))^2.
\end{aligned}$$

Adding these together along with the linear terms, expanding and regrouping terms we get:

$$T_2(p,q) = C_2 T(\hat{p},\hat{q}) + \sum_{\substack{i=1,2 \\ f=p,q}} \theta^f_{2,i} + \theta^{p,q}_{2,2} + R_3.$$

B. Full Specification of the Estimators

Here we write out the complete expressions for the estimators $\widehat{T}_{pl}, \widehat{T}_{lin}, \widehat{T}_{quad}$. Recall that we have samples $X_1^n \sim p, Y_1^n \sim q$ and our goal is to estimate $T(p,q) = \int p^\alpha q^\beta$. Define:

$$\begin{aligned}
\hat{p}(x) &= \frac{1}{n}\sum_{i=1}^n K_h(X_i - x) \qquad \hat{q}(x) = \frac{1}{n}\sum_{j=1}^n K_h(Y_j - x) \\
\hat{p}_{DS}(x) &= \frac{2}{n}\sum_{i=1}^{n/2} K_h(X_i - x) \qquad \hat{q}_{DS}(x) = \frac{2}{n}\sum_{j=1}^{n/2} K_h(Y_j - x),
\end{aligned}$$

where DS is used to denote that we are data splitting, and K_h is a kernel with bandwidth h meeting Assumption 3. The estimator $\widehat{T}_{pl}$ is formed by simply plugging in $\hat{p}, \hat{q}$ into the function T. Formally:

$$\widehat{T}_{pl} = \int \hat{p}^\alpha(x)\hat{q}^\beta(x)d\mu(x) \tag{13}$$

The estimator $\widehat{T}_{lin}$ is formed by a first order correction but we must used the data split KDEs to ensure independence between the multiple terms in the estimator.

$$\widehat{T}_{lin} = (1-\alpha-\beta)\int \hat{p}^\alpha_{DS}(x)\hat{q}^\beta_{DS}(x)d\mu(x) + \frac{2}{n}\sum_{i=n/2+1}^n \alpha\hat{p}^{\alpha-1}_{DS}(X_i)\hat{q}^\beta_{DS}(X_i) + \frac{2}{n}\sum_{j=n/2+1}^n \alpha\hat{p}^\alpha_{DS}(Y_j)\hat{q}^\beta_{DS}(Y_j). \tag{14}$$

For the quadratic term we perform an additional correction:

$$\begin{aligned}
\widehat{T}_{quad} &= (1-3\alpha/2-3\beta/2+1/2(\alpha+\beta)^2)\int \hat{p}_{DS}^{\alpha}(x)\hat{q}_{DS}^{\beta}(x)d\mu(x)+ \\
&+\frac{2}{n}\sum_{i=n/2+1}^{n}\alpha(2-\alpha-\beta)\hat{p}_{DS}^{\alpha-1}(X_i)\hat{q}_{DS}^{\beta}(X_i)+\frac{2}{n}\sum_{j=n/2+1}^{n}\beta(2-\alpha-\beta)\hat{p}_{DS}^{\alpha}(Y_j)\hat{q}_{DS}^{\beta-2}(Y_j) \\
&+\frac{4}{n(n/2-1)}\sum_{k\in M}\sum_{i_1\neq i_2=n/2+1}^{n}\phi_k(X_{i_1})\phi_k(X_{i_2})\left[\frac{1}{2}\alpha(\alpha-1)\hat{p}_{DS}^{\alpha-2}(X_{i_2})\hat{q}_{DS}^{\beta}(X_{i_2})\right] \\
&-\frac{2}{n(n/2-1)}\sum_{k,k'\in M}\sum_{i_1\neq i_2=n/2+1}^{n}\phi_k(X_{i_1})\phi_{k'}(X_{i_2})\left[\frac{1}{2}\alpha(\alpha-1)\int\phi_k(x)\phi_{k'}(x)\hat{p}_{DS}^{\alpha-2}(x)\hat{q}_{DS}^{\beta}(x)d\mu(x)\right] \\
&+\frac{4}{n(n/2-1)}\sum_{k\in M}\sum_{j_1\neq j_2=n/2+1}^{n}\phi_k(Y_{j_1})\phi_k(Y_{j_2})\left[\frac{1}{2}\beta(\beta-1)\hat{p}_{DS}^{\alpha}(Y_{j_2})\hat{q}_{DS}^{\beta-2}(Y_{i_2})\right] \\
&-\frac{2}{n(n/2-1)}\sum_{k,k'\in M}\sum_{j_1\neq j_2=n/2+1}^{n}\phi_k(Y_{j_1})\phi_{k'}(Y_{j_2})\left[\frac{1}{2}\beta(\beta-1)\int\phi_k(y)\phi_{k'}(y)\hat{p}_{DS}^{\alpha}(y)\hat{q}_{DS}^{\beta-2}(y)d\mu(y)\right] \\
&+\frac{2}{n}\sum_{j=n/2+1}^{n}\sum_{k\in M}\left(\frac{2}{n}\sum_{i=n/2+1}^{n}\phi_k(X_i)\right)\phi_k(Y_j)\left(\alpha\beta\hat{p}_{DS}^{\alpha-1}(Y_j)\hat{q}_{DS}^{\beta-1}(Y_j)\right).
\end{aligned}$$

Recall that $\{\phi_k\}_{k\in D}$ is an orthonormal basis for $L_2([0,1]^d)$, and M is an appropriately chosen subset of D. The first line of the estimator is simply the plugin term, while the second lines makes up the two linear terms. The third through sixth lines are the two quadratic terms, one involving the data from p and the other involving the data from q. Finally the last line is the bilinear term.

C. Detailed Proofs of Upper Bound

Let us now prove the the auxiliary results stated in Section 5

C.1. Proof of Theorem 5

The truncated kernel density estimator takes the following form: We select a parameter $\epsilon > 0$. If $\tilde{f}$ is the usual kernel density estimator for f, we set $\hat{f}(x) = \tilde{f}(x)$ if $\tilde{f}(x) \in [\kappa_l - \epsilon, \kappa_u + \epsilon]$ and otherwise we set $\hat{f}(x) = f_0(x)$ for some fixed function bounded between κ_l, κ_u.

Recall Assumption 3 ensures that the kernel $K : \mathbb{R}^d \to \mathbb{R}$ satisfies:

1. supp$K \in (-1,1)^d$
2. $\int K(x)dx = 1$ and $\int \prod_i x_i^{p_i} K(x)dx = 0$ for all tuples $p = (p_1, \ldots, p_d)$ with $\sum p_i \leq \lfloor s \rfloor$.

Note that we can use the Legendre polynomials to construct kernels meeting these properties (Tsybakov, 2009).

Let us first establish the rate of convergence of $\tilde{f}$ the regular kernel density estimator in ℓ_p^p, which is $\tilde{f}(x) = \frac{1}{nh^d}\sum_{i=1}^{n} K\left(\frac{x-X_i}{h}\right)$. Denote by $\bar{f}(x) = \mathbb{E}[\tilde{f}(x)] = \mathbb{E}_{X\sim f}[\frac{1}{h^d}K\left(\frac{x-X}{h}\right)]$. Then:

$$\mathbb{E}[\|\tilde{f}-f\|_p^p] \leq 2^p\left(\mathbb{E}[\|\tilde{f}-\bar{f}\|_p^p] + \|\bar{f}-f\|_p^p\right).$$

To bound the first term, let us write $\eta_i(x) = \frac{1}{h^d}K\left(\frac{x-X_i}{h}\right) - \mathbb{E}_{X\sim f}[\frac{1}{h^d}K\left(\frac{x-X}{h}\right)]$. Exchanging integrals, we can look at fixed x and we have:

$$\mathbb{E}|\tilde{f}(x)-\bar{f}(x)|^p = \mathbb{E}|\frac{1}{n}\sum_{i=1}^{n}\eta_i(x)|^p \leq \left(\frac{1}{n^{2p}}\mathbb{E}[(\sum_{i=1}^{n}\eta_i(x))^{2p}]\right)^{1/2}. \tag{15}$$

If we expand the expectation and drop the terms that vanish we get all terms of the form:

$$\sum_{i_1 \neq i_2 \ldots \neq i_t = 1}^{n} \prod_{j=1}^{t} \binom{p - \sum_{k=1}^{j-1} p_k}{p_j} \eta_{i_j}(x)^{p_j} = \frac{n!}{(n-t)!} \prod_{j=1}^{t} \binom{p - \sum_{k=1}^{j-1} p_k}{p_j} \eta_{i_j}(x)^{p_j},$$

where $1 \leq t < p$, $\sum p_j = p$ and $p_j \neq 1 \forall j$. That is, we pick a term in the polynomial with t unique variables, then assign powers p_j to each of the terms, then count the number of ways to assign those powers to those terms (which results in the binomial coefficients). Since $\mathbb{E}[\eta_j(x)] = 0$, the terms where there is some $p_j = 1$ are all zero.

By linearity of expectation and independence, we therefore need to control $\mathbb{E}[|\eta_i(x)|^q]$ for $2 \leq q \leq p$. Applying Jensen's inequality, we get:

$$\mathbb{E}[|\eta_i(x)|^q] \leq 2^q \mathbb{E}[|\frac{1}{h^d} K\left(\frac{X-x}{h}\right)|^q] \leq 2^q \kappa_u h^{-(q-1)d} \int |K(u)^q| du,$$

where the last expression comes from expanding the integral, performing a substitution and bounding $f(x) \leq \kappa_u$. So we can bound by $C(q, \kappa_u, K) h^{-(q-1)d}$. Plugging this into the expression above, we get:

$$\frac{n!}{(n-t)!} C(p_1^j, \kappa_u, K) h^{-pd+td} \leq n^t C'(p_1^j, \kappa_u, K) h^{-pd+td} \leq C'(p_1^j, \kappa_u, K) \frac{n^p}{h^{pd}}.$$

The second inequality holds for n sufficiently large. The third inequality holds whenever $nh^d \geq 1$ which will be true for n sufficiently large, given our setting of h. To summarize, all of the terms can be upper bounded by $c(n^p/h^{pd})$ and there are a constant-in-p number of terms. Plugging this into Equation 15 we get

$$\mathbb{E}[\|\tilde{f} - \bar{f}\|_p^p] \leq C(nh^d)^{-p/2}. \tag{16}$$

Remark 1. *The constant here has exponential dependence on p but we are only concerned with cases where p is a small constant (at most 4).*

As for the bias (note that x, u, t are all d-dimensional vectors here):

$$|\bar{f}(x) - f(x)| = \int \frac{1}{h^d} K\left(\frac{x-t}{h}\right) f(t) dt - f(x) = \int (f(x-uh) + f(x)) K(u) du.$$

Let us define $m = \lfloor s \rfloor$. Taking the $(m-1)$st order von Mises expansion of $f(x+uh)$ about $f(x)$ we get terms of the form:

$$\sum_{r_1, \ldots, r_d | \sum r_i \leq m-1} \frac{1}{|r|!} D^r f(x) h^{|r|} \int \prod_i u_i^{r_i} K(u) du$$

which are all zero by our assumption on K. The remainder term, gives us:

$$\sum_{r_1, \ldots, r_d | \sum r_i = m} \frac{h^m}{m!} \int \xi(r, x, uh) \prod_i u_i^{r_i} K(u) du \leq \sum_{r_1, \ldots, r_d | \sum r_i = m} \frac{Lh^s}{m!} \int \|u\|^{s-m} \prod_i u_i^{r_i} K(u) du,$$

which we will denote $C(m, K, d) L h^s$. Here the function ξ is between $D^r f(x)$ and $D^r f(x-uh)$ and to reach the last expression, we use the fact that $|D^r f(x) - D^r f(x-uh)| \leq L \|uh\|^{s-r}$, i.e. the Hölderian assumption on f. In applying the Hölderian assumption, there is another term of the form $D^r f(x) \int \prod_i u_i^{r_i} K(u) du$ which is zero by the assumption on K. Equipped with this bound, we can bound the bias:

$$\|\bar{f} - f\|_p^p \leq C(m, K, d) L^p h^{ps}. \tag{17}$$

In trading off the bias and the variance, we set $h \asymp n^{\frac{-1}{2s+d}}$ and see that the rate of convergence is $\mathbb{E}[\|\tilde{f} - f\|_p^p = O(n^{\frac{-ps}{2s+d}})$.

To prove Theorem 5, we just have to show that truncation does not significantly affect the rate. Fix $\epsilon > 0$ and define $S_\epsilon = \{x : \kappa_l - \epsilon \leq \tilde{f}(x) \leq \kappa_u + \epsilon\}$. We have:

$$\mathbb{E}[\|\hat{f} - f\|_p^p] \quad = \quad \mathbb{E}\left[\int_{S_\epsilon} |\tilde{f}(x) - f(x)|^p dx + \int_{S_\epsilon^C} |f_0(x) - f(x)| dx\right]$$

$$\leq \mathbb{E}\left[\|\tilde{f} - f\|_p^p\right] + \|f_0 - f\|_\infty^p \mathbb{E}\left[\int \mathbf{1}[x \notin S_\epsilon]dx\right]$$
$$= \mathbb{E}\left[\|\tilde{f} - f\|_p^p\right] + \|f_0 - f\|_\infty^p \int \mathbb{P}_{X_1^n}[x \notin S_\epsilon]dx,$$

so we must control the probability that $x \notin S_\epsilon$. This can be done via Bernstein's inequality. First observe that the bias $|\bar{f} - f| \to 0$ with our choice of h so that for sufficiently large n, $\sup_x \bar{f}(x) - f(x) \leq \epsilon/2$. Once this happens, it is clear that $x \notin S_\epsilon$ implies that $\tilde{f}(x) - \bar{f}(x) \geq \epsilon/2$. Therefore:

$$\mathbb{P}[x \notin S_\epsilon] \leq \mathbb{P}[|\tilde{f}(x) - \bar{f}(x)| \geq \epsilon/2] = \mathbb{P}[|\frac{1}{n}\sum_i \eta_i(x)| \geq \epsilon/2] \leq 2\exp\left(\frac{-nh^d\epsilon^2/4}{\kappa_u\|K\|_2^2 + \frac{1}{3}\|K\|_\infty\epsilon}\right).$$

This last inequality is an application of Bernstein's inequality noting that $|\eta_i(x)| \leq \frac{2}{h^d}\|K\|_\infty$ and $\mathrm{Var}(\eta_i(x)) \leq h^{-d}\kappa_u\|K\|_2^2$ since:

$$\mathrm{Var}(\eta_i(x)) \leq \mathbb{E}_{X_i \sim f}\left[\frac{1}{h^{2d}}K^2(\frac{X_i - x}{h})\right] = \frac{1}{h^d}\int K^2(u)f(x + hu)du \leq h^{-d}\kappa_u\|K\|_2^2.$$

Using our definition $h \asymp n^{\frac{-1}{2s+d}}$ and using the fact that ϵ is some constant $\mathbb{P}[x \notin S_\epsilon] \leq 2\exp(-Cn^{\frac{2s}{2s+d}})$. Plugging this bound in above, we have:

$$\mathbb{E}\left[\|\hat{f} - f\|_p^p\right] \leq \mathbb{E}\left[\|\tilde{f} - f\|_p^p\right] + 2\|f_0 - f\|_\infty^p \exp\left(-Cn^{\frac{2s}{2s+d}}\right)\mathrm{vol}([0,1]^d) = O(n^{\frac{-ps}{2s+d}}),$$

since the second term goes to zero exponentially quickly in n. This proves the theorem.

C.2. Convergence Rate for Estimating Linear Functionals

It is trivial to derive the convergence rate for estimating linear functionals:

$$\mathbb{E}[(\hat{\theta} - \theta)^2] = \frac{1}{n}(\mathbb{E}[\psi^2(X)] - \mathbb{E}[\psi(X)]^2) \leq 2\|\psi\|_\infty^2/n,$$

And by Jensen's inequality, we have $\mathbb{E}[|\hat{\theta} - \theta|] \leq \sqrt{\mathbb{E}[(\hat{\theta} - \theta)^2]}$, so the rate of convergence is $\sqrt{2}\|\psi\|_\infty/\sqrt{n}$.

C.3. Proof of Theorem 6

For the quadratic terms, we use a result of Laurent (1996):

Theorem 10 ((Laurent, 1996)). *Let X_1^n be i.i.d random variables with common density f that belongs to some Hilbert Space $L^2(d\mu)$. Let $\{\phi_i\}_{i\in D}$ be an orthonormal basis of $L^2(d\mu)$. Assume that f is uniformly bounded and belongs to the ellipsoid $\mathcal{E} = \{\sum_{i\in D} a_i\phi_i : \sum_{i\in D} |a_i^2/c_i^2| \leq 1\}$. Let ψ be bounded function and define $\theta = \int \psi(x)f(x)\mu(dx)$ and $\hat{\theta}$ as in Equation 2 where the set $M = M_n \subset D$ has size m. Then whenever $n \geq n_0$ (some absolute constant), we have:*

$$\mathbb{E}[(\hat{\theta} - \theta)^2] = \mathit{Bias}^2(\hat{\theta}) + \mathrm{Var}(\hat{\theta}) \leq \|\psi\|_\infty^2 \sup_{i\notin M_n} |c_i|^4 + 72\|\psi\|_\infty^2\|f\|_\infty^2\left(\frac{2}{n} + \frac{m}{n^2}\right). \tag{18}$$

For the bi-linear term $\theta_{2,2}^{p,q}$ we have the following theorem:

Theorem 11. *Let X_1^n be i.i.d random variables with common density f and Y_1^n be i.i.d. with common density g. Let f, g belong to some Hilbert space $L^2(d\mu)$ and let $\{\phi_i\}_{i\in D}$ be an orthonormal basis for $L^2(d\mu)$. Assume that f, g are uniformly bounded and both belong to the ellipsoid $\mathcal{E} = \{\sum_{i\in D} a_i\phi_i : \sum_{i\in D} |a_i^2/c_i^2| \leq 1\}$. Let $\theta = \int \psi(x)f(x)g(x)\mu(dx)$ and $\hat{\theta}$ be defined by Equation 1 where the set $M = M_n \subset D$ has size m. Then whenever $n \geq n_0$ (some absolute constant), we have:*

$$\mathbb{E}[(\hat{\theta} - \theta)^2] = \mathit{Bias}^2(\hat{\theta}) + \mathrm{Var}(\hat{\theta}) \leq \|\psi\|_\infty^2 \sup_{i\notin M_n} |c_i|^4 + \|\psi\|_\infty^2\|f\|_\infty\|g\|_\infty\left(\frac{2}{n} + \frac{m+1}{n^2}\right). \tag{19}$$

Proof. The bias is:

$$\mathbb{E}[\hat{\theta}] - \theta = \int \sum_{i \in M} \alpha_i \phi_i(x) \psi(x) g(x) dx - \int \psi(x) f(x) g(x) = \int \psi(x) \left(\mathcal{P}_M f(x) - f(x) \right) g(x) dx,$$

where $\alpha_i = \int \phi_i(x) f(x)$ and $\mathcal{P}_M f$ is the projection of f onto the subspace defined by M. Define $\beta_i = \int \phi_i(x) g(x)$. If f, g live in the ellipsoid $\mathcal{E} = \{\sum a_i \phi_i | \sum |a_i|^2 / |c_i|^2 \leq L\}$ then:

$$\text{Bias}^2(\hat{\theta}) = \left(\sum_{i \notin M} \alpha_i \int \psi(x) g(x) \phi_i(x) dx \right)^2 \leq \|\psi\|_\infty^2 \left(\sum_{i \notin M} \alpha_i \beta_i \right)^2.$$

The term inside the parenthesis can be bounded as:

$$\sum_{i \notin M} \alpha_i \beta_i \leq \frac{1}{2} \sup_{i \notin M} |c_i|^2 \sum_{i \notin M} \frac{|\alpha_i|^2 + |\beta_i|^2}{|c_i|^2} \leq L \sup_{i \notin M} |c_i|^2,$$

so the bias is $\text{Bias}^2(\hat{\theta}) \leq \|\psi\|_\infty^2 L^2 \sup_{i \notin M} |c_i|^4$.

As for the variance, let us define $Q(x)$ to be the m-dimensional vector of functions $\phi_i(x) - \alpha_i$ and $R(x)$ to be the m-dimensional vector of functions $\phi_i(x)\psi(x) - \int \psi \phi_i g$. Further define A, B to be the m-dimensional vectors with ith components $\alpha_i = \int \phi_i f$ and $\beta_i = \int \psi \phi_i g$ respectively. Then our estimator can alternatively be written as:

$$\hat{\theta} = \underbrace{\frac{1}{n^2} \sum_{j,k} Q(X_j)^T R(Y_k)}_{T_1} + \underbrace{\frac{1}{n} \sum_j Q(X_j)^T B}_{T_2} + \underbrace{\frac{1}{n} \sum_k A^T R(Y_k)}_{T_3} - A^T B.$$

Notice that Q, R are centered functions. Since Xs are independent of the Ys, $\text{Cov}(T_2, T_3) = 0$. Since T_2 is independent of Y and $\mathbb{E}[R(Y_k)] = 0$, we see that $\text{Cov}(T_1, T_2) = 0$. Similarly, $\text{Cov}(T_1, T_3) = 0$.

Therefore,

$$\text{Var}(\hat{\theta}) = \text{Var}(T_1) + \text{Var}(T_2) + \text{Var}(T_3).$$

Let us analyze T_1. By independence,

$$\begin{aligned}
\text{Var}(T_1) &= \frac{1}{n^2} \text{Var}(Q(X_1)^T R(Y_1)) = \frac{1}{n^2} \sum_{i,i' \in M} \int \phi_i(x) \phi_{i'}(x) \phi_i(y) \phi_{i'}(y) \psi(y)^2 f(x) g(y) dx dy \\
&- \int \alpha_i \alpha_{i'} \phi_i(y) \phi_{i'}(y) \psi(y)^2 g(y) dy - \int \beta_i \beta_{i'} \phi_i(x) \phi_{i'}(x) f(x) + \alpha_i \alpha_{i'} \beta_i \beta_{i'} \\
&\leq \frac{1}{n^2} \sum_{i,i' \in M} \int \phi_i(x) \phi_{i'}(x) \phi_i(y) \phi_{i'}(y) \psi(y)^2 f(x) g(y) dx dy + \frac{1}{n^2} (\sum_i \alpha_i \beta_i)^2 \\
&= \frac{1}{n^2} \int \left(\sum_{i \in M} \phi_i(x) \phi_i(y) \right)^2 \psi(y)^2 f(x) g(y) dx dy + \frac{1}{n^2} (\sum_i \alpha_i \beta_i)^2 \\
&\leq \frac{\|\psi\|_\infty^2 \|f\|_\infty \|g\|_\infty}{n^2} \int \left(\sum_{i \in M} \phi_i(x) \phi_i(y) \right)^2 dx dy + \frac{1}{n^2} \left(\sum_i \alpha_i^2 \right) \left(\sum_i \beta_i^2 \right) \\
&\leq \frac{\|\psi\|_\infty^2 \|f\|_\infty \|g\|_\infty m}{n^2} + \frac{1}{n^2} (\int f^2)(\int g^2 \psi^2) \leq \frac{\|\psi\|_\infty^2 \|f\|_\infty \|g\|_\infty (m+1)}{n^2}.
\end{aligned}$$

To arrive at the third line, notice that the cross terms are non-negative, since $\sum_{i,i'} \alpha_i \alpha_{i'} \phi_i(y) \phi_{i'}(y) = \left(\sum_i \alpha_i \phi_i(y) \right)^2$ (and analogously for the other cross term). Therefore we can simply omit them and provide an upper bound. To go from the fourth to fifth lines, we use Hölder's inequality on the first term and Cauchy-Schwarz on the second term. Notice that the expression involving $\phi_i(x)\phi_i(y)$ is positive, so we can drop the absolute values in the ℓ_1 norm term of Hölder's inequality. To arrive at the fifth line, we expand out the square and use the fact that ϕ_is are orthornormal.

For T_2 again by independence we have:

$$\begin{aligned}\mathrm{Var}(T_2) &= \frac{1}{n}\mathrm{Var}(Q(X_1)^T B) = \mathbb{E}[\left(\sum_{i\in M}(\phi_i(X_1)-\alpha_i)\int\psi\phi_i g\right)^2] \\ &= \sum_{i,i'\in M}\int\phi_i(x)\phi_{i'}(x)f(x)\int\psi\phi_i g\int\psi\phi_{i'}g - \int\alpha_i\psi\phi_i g\int\alpha_{i'}\psi\phi_{i'}g \\ &= \int(\sum_{i\in M}\beta_i\phi_i(x))^2 f(x) - \left(\int(\mathcal{P}_M)\psi g\right)^2 \le \int(\mathcal{P}_M(\psi g))^2 f.\end{aligned}$$

Here the last inequality follows from the fact that $\beta_i = \int\psi\phi_i g$ is the ith fourier coefficient of ψg so $\sum_i \beta_i\phi_i$ is the projection onto M. Of course this quantity is bounded by:

$$\mathrm{Var}(T_2) \le \frac{1}{n}\|f\|_\infty\int\psi^2(x)g^2(x)dx \le \frac{\|\psi\|_\infty^2\|f\|_\infty\|g\|_\infty}{n}.$$

Essentially the same argument reveals that T_3 is bounded in the same way.

$$\begin{aligned}\mathrm{Var}(T_3) &= \frac{1}{n}\mathrm{Var}(A^T R(Y_1)) \le \frac{\|\psi\|_\infty^2}{n}\sum_{i,i'}\alpha_i\alpha_{i'}\left[\int\phi_i(y)\phi_{i'}(y)g(y)dy - \int\phi_i g\int\phi_{i'}g\right] \\ &= \frac{1}{n}\|\psi\|_\infty^2\left[\int(\mathcal{P}_M f)^2 g - (\int(\mathcal{P}_M f)g)^2\right] \le \frac{\|\psi\|_\infty^2\|f\|_\infty\|g\|_\infty}{n},\end{aligned}$$

so the variance of the estimator is:

$$\mathrm{Var}(\hat{\theta}) \le \|\psi\|_\infty^2\|f\|_\infty\|g\|_\infty\left(\frac{m+1}{n^2}+\frac{2}{n}\right).$$

□

Both the quadratic and bilinear terms exhibit the same dependence on $\sup_{i\notin M_n}|c_i|, m, n$ so choosing M_n appropriately will give the rate of convergence for both terms. To establish Theorem 6 we work with the fourier basis $\{\phi_k\}_{k\in\mathbb{Z}^d}$ where $\phi_k(x) = e^{2\pi i k^T x}$ and the Sobolev class $\mathcal{W}(s,L)$ defined by:

$$\mathcal{W}(s,L) = \left\{ f = \sum_{k\in\mathbb{Z}^d} a_k\phi_k \;\middle|\; \sum_{k\in\mathbb{Z}^d}(\sum_{j=1}^d |k_j|^{2s})|a_k|^2 \le L \right\} \tag{20}$$

In Lemma 14 we show that the class $\mathcal{W}(s',L')$ contains $\Sigma(s,L)$ as long as $s' < s$ and with appropriate choice of L'. For now let us work in $\mathcal{W}(s',L')$.

Let us choose:

$$M_n = \{k\in\mathbb{Z}^d \,||k_j| \le \frac{1}{2}m^{1/d}\},\;\; m_0 = \left(18\frac{d}{s'}2^{4s'/d}n^{-2}\right)^{\frac{-d}{4s'+d}} \asymp n^{\frac{2d}{4s'+d}}.$$

Thinking of M_n as an integer lattice with side lengths $m_0 = m^{1/d}$ we see that $|M_n| = m$. Moreover $\sup_{i\notin M_n}|c_i|^4 = L^2(2/m)^{4s'/d}$. For the quadratic terms, this results in the bound:

$$\begin{aligned}\mathbb{E}[(\hat{\theta}-\theta)^2] &\le \|\psi\|_\infty^2\left(L^2(2/m)^{4s'/d} + 72\|f\|_\infty^2 m/n^2 + 144\|f\|_\infty^2/n\right) \\ &\le \|\psi\|_\infty^2\max\{1,\|f\|_\infty^2\}\max\{L^2,1\}\left((2/m)^{4s'/d} + 72m/n^2 + 144/n\right),\end{aligned}$$

and plugging in our definition of m followed by some algebraic simplifications, we get

$$\mathbb{E}[(\hat{\theta}-\theta)^2] \le 18\|f\|_\infty^2\max\{1,\|p\|_\infty^2\}\max\{L^2,1\}\left(\frac{8}{n} + n^{\frac{-8s'}{4s'+d}}\left[2^{\frac{8s'}{d}}d/s'+3\right]\right).$$

For the bilinear terms, plugging into Theorem 11, we get

$$\mathbb{E}[(\hat{\theta}-\theta)^2] \leq \|\psi\|_\infty^2 \max\{1, \|f\|_\infty \|g\|_\infty\} \max\{L^2, 1\} \left((2/m)^{4s'/d} + m/n^2 + 3/n \right),$$

which when we plug in for m we get:

$$\mathbb{E}[(\hat{\theta}-\theta)^2] \leq \|\psi\|_\infty^2 \max\{1, \|f\|_\infty \|g\|_\infty\} \max\{L^2, 1\} \left(3/n + n^{\frac{-8s'}{4s'+d}} \left[18 \times 2^{8s'/d} d/s' + 1 \right] \right).$$

D. Proofs of Corollaries 4 and 3

The proof of Corollary 4 is immediate given the decomposition $\|p-q\|_2^2 = \int p^2 + \int q^2 - 2\int pq$ and the Theorem 6.

For Corollary 3, if we use our estimator $\hat{T}$ for $T(p,q) = \int p^\alpha q^{1-\alpha}$ we can plug $\hat{T}$ into the definition of Rényi divergence to obtain an estimator $\hat{D}_\alpha$. The rate of convergence is:

$$\mathbb{E}[|\hat{D}_\alpha - D_\alpha|] = \frac{1}{\alpha-1}\mathbb{E}\left[\log\left(\hat{T}/T\right)\right] \leq \frac{1}{\alpha-1}\mathbb{E}\left[\log(1+|T-\hat{T}|/T)\right] \leq \frac{1}{\alpha-1} c n^{-\gamma}/T(p,q)$$

where γ is the rate of convergence of our estimator. This is $O(n^{-\gamma})$ as long as $T(p,q) \geq c > 0$.

E. Detailed Proofs for Lower Bound

To prove the main part of the theorem, the $\Omega(n^{\frac{-4s}{4s+d}})$ rate, we use Le Cam's method. We decompose the proof into three parts. In the first part, we adapt Le Cam's method to our setting. In the second part, we show how the properties established on the functions u_j, $j \in [p]$ allow us to apply the technique and establish the theorem. In the third part, we prove the existence of such functions u_j. We conclude this section with a proof of the $\Omega(n^{-1/2})$ when $s > d/4$.

E.1. Proof of Lemma 7

Proof. Define $\Theta_0 = \{g \in \Theta | T(g,q) \geq T(p,q)\}$ and $\Theta_1 = \{g \in \Theta | T(g,q) \leq T(p,q) - 2\beta\}$ so that all $g_\lambda \in \Theta_1$ while $p \in \Theta_0$. Let $\tilde{\Theta}_i = \text{conv}(\{G^n \times Q^n | g \in \Theta_i\})$ and consider the simple versus simple testing problem between $P \in \Theta_0$ and $G_\lambda \in \Theta_1$. The minimax probability of error p_e of such a test is lower bounded by $\frac{1}{2}(1 - \sqrt{h^2(P,G_\lambda)(1-h^2(P,G_\lambda))/4})$ by Theorem 2.2. of Tsybakov (2009). So for any test statistic ψ, taking supremum over $P \in \Theta_0, G \in \Theta_1$ we have:

$$\sup_{\theta_{0,1} \in \tilde{\Theta}_{0,1}} p_e(\psi; \theta_0, \theta_1) \geq \frac{1}{2}\left[1 - \sqrt{\gamma(1-\gamma/4)}\right],$$

where $\gamma \geq h^2(P^n \times Q^n, \bar{G}^n \times Q^n \in \tilde{\Theta}_1)$, which holds since $P^n \times Q^n \in \tilde{\Theta}_0$ and $\bar{G}^n \times Q^n \in \tilde{\Theta}_1$ by convexity. The same bound holds for after taking infimum over ψ. Finally, if we make an error in the testing problem, we suffer loss at least β which results in the statement in the Lemma. □

E.2. The properties of u_j

Recall that in our proof we partition $[0,1]^d$ into m cubes $R_1, \ldots, R_m$ of side length $m^{-1/d}$. On each bin we require a function u_j such that:

$$\text{supp}(u_j) \subset \{x | B(x,\epsilon) \in R_j\},\ \|u_j\|_2^2 = \Theta(m^{-1}),\ \int_{R_j} u_j = 0,\ \int_{R_j} p^{\alpha-1} q^\beta u_j = 0, \|D^r u_j\|_\infty \leq m^{r/d},$$

where the last inequality needs to hold for all tuples r with $\sum_j r_j \leq s+1$. Using these functions u_j, we construct the alternatives $g_\lambda = p + K\sum_{\lambda \in \Lambda} \lambda_j u_j \mathbf{1}_{R_j}$ for all $\lambda \in \Lambda = \{-1,1\}^m$. The third property above ensures that g_λ is a valid density.

Properties 2, 4, and 5 ensure that $T(p,q) - T(g_\lambda, q)$ is sufficiently large. Indeed, by the von Mises expansion:

$$T(p,q) - T(g_\lambda, q) \quad = \quad K\alpha \sum_{j=1}^m \lambda_j \int_{R_j} p^{\alpha-1} q^\beta u_j + K^2 \alpha(\alpha-1) \sum_{j=1}^m \int_{R_j} \xi_p^{\alpha-2}(x) q^\beta(x) u_j^2(x) dx$$

$$\geq \quad c_0 K^2 \sum_{j=1}^{m} \|u_j\|_2^2 \geq c_1 K^2.$$

Here ξ is the function in the Taylor's remainder theorem, bounded between p and g_λ, both of which are bounded above and below. g_λ is bounded above and below by property 5 since $\|D_0 u_j\|_\infty = \|u_j\|_\infty \leq 1$ which means that $g_\lambda \in [1-K, 1+K]$. K will be decreasing with n, so this quantity will certainly be bounded for n large enough. Property 2 allows us to arrive at the last line since each u_j is orthogonal to the derivative of T, so the first term in the expansion is zero. Finally property 4 allows us to lower bound $\|u_j\|_2^2$.

Property 2 is also critical in ensuring that $h^2(P^n \times Q^n, \bar{G}^n \times Q^n)$ is small through the following Theorem of Birge and Massart (1995).

Theorem 12 ((Birgé & Massart, 1995)). *Consider a set of densities p and $p_\lambda = p[1+\sum_j \lambda_j v_j(x)]$ for $\lambda \in \Lambda = \{-1, 1\}^m$. Suppose that (i) $\|v_j\|_\infty \leq 1$ (ii) $\|1_{R_j^C} v_j\|_1 = 0$, (iii) $\int v_j p = 0$ and (iv) $\int v_j^2 p = \alpha_j > 0$ all hold with:*

$$\alpha = \sup_j \|v_j\|_\infty, \; s = n\alpha^2 \sup_j P(R_j), \; c = n \sup_j \alpha_j.$$

Define $\bar{P}_\Lambda^n = \frac{1}{|\Lambda|} \sum_{\lambda \in \Lambda} P_\lambda^n$. Then:

$$h^2(P^n, \bar{P}_\Lambda^n) \leq C(\alpha, s, c) n^2 \sum_{j=1}^{m} \alpha_j^2,$$

where $C < 1/3$ is continuous and non-decreasing with respect to each argument and $C(0,0,0) = 1/16$.

In bounding the Hellinger distance $h^2(P^n \times Q^n, \bar{G}^n \times Q^n)$ we first use the property that hellinger distance decomposes across product measures:

$$h^2(P^n \times Q^n, \bar{G}^n \times Q^n) = 2\left(1 - (1 - h^2(P^n, \bar{G}^n)/2)(1 - h^2(Q^n, Q^n)/2)\right) = h^2(P^n, \bar{G}^n).$$

If we define $v_j(x) = K u_j(x)/p(x)$ then we have $g_\lambda = p[1 + \sum_j \lambda_j v_j]$ as needed by Theorem 12. We immediately satisfy requirements 1, 2, and 3 and we have $\int v_j^2 p = K^2 \int u_j^2/p \leq K^2 \kappa_l / m = \alpha_j$. Thus in applying the theorem we have:

$$h^2(P^n \times Q^m, \bar{G}^n \times Q^m) \leq (1/3) n^2 \sum_{j=1}^{m} \alpha_j^2 \leq \frac{C n^2 K^4}{m}.$$

Property 1 and 5 ensure that $g_\lambda \in \Sigma(s, L)$ via the following argument. Defining $u_\lambda = K \sum_j \lambda_j u_j$, we will first show that u_λ is holder smooth and g_λ will be holder by a final application of the triangle inequality. For u_λ, fix r with $\sum_j r_j = s$ and fix x, y. Let x_1 be the boundary point of R_j, the bin containing x along the line between x and y and let y_1 be the analogous boundary point for y.

$$\begin{aligned}
|D^r u_\lambda(x) - D^r u_\lambda(y)| &\leq |D^r u_\lambda(x) - D^r u_\lambda(x_1)| + |D^r u_\lambda(x_1) - D^r u_\lambda(y_1)| + |D^r u_\lambda(y_1) - D^r u_\lambda(y)| \\
&= |D^r u_\lambda(x) - D^r u_\lambda(x_1)| + |D^r u_\lambda(y_1) - D^r u_\lambda(y)| \\
&= \int_{\gamma(x,x_1)} \nabla D^r u_\lambda(z) dz + \int_{\gamma(y,y_1)} \nabla D^r u_\lambda(z) dz \\
&\leq K \|D^{r+1} u_j\|_\infty (\|x - x_1\|_2 + \|y - y_1\|_2) \\
&\leq K m^{(r+1)/d} \left(\|x - x_1\|_2^{s-r} \|x - x_1\|_2^{1-(s-r)} + \|y - y_1\|_2^{s-r} \|y - y_1\|_2^{1-(s-r)} \right) \\
&\leq K m^{(r+1)/d} \sqrt{d} m^{-\frac{1-(s-r)}{d}} \left(\|x - x_1\|_2^{s-r} + \|y - y_1\|_2^{s-r} \right) \\
&\leq K m^{s/d} \sqrt{d} \|x - y\|_2^{s-r} \leq L \|x - y\|_2^{s-r}
\end{aligned}$$

The first line is an application of the triangle inequality. In the second line we use that u_λ is zero and has all derivatives equal to zero on the boundaries of the cubes R_j. This follows from the fact that u_j is not supported in the band around the border of R_j. The third line is an application of the fundamental theorem of calculus, $\gamma(x, x_1)$ is the path between x and x_1. The fourth line follows from Hölder's inequality, we replace each derivative with its supremum and are left with

just the path integral, which simplifies to the length of the path, i.e. $\|x - x_1\|_2$. In the fifth line we use the assumption $\|D^r u_j\|_\infty \leq m^{r/d}$ for any derivative operator with $\sum_j r_j \leq s+1$. To arrive at the sixth line, notice that since x, x_1 are in the same box R_j, we have $\|x - x_1\|_2 \leq \sqrt{d} m^{-1/d}$ (there are m boxes and each one has length $m^{-1/d}$ on each side). The last line is true since x_1, y_1 are on the line segment between x, y.

In other words, g_λ is holder smooth as long as $K m^{s/d}\sqrt{d} \asymp L$, imposing the requirement that $K = O(m^{-s/d})$. So if we pick $m = n^{\frac{2d}{4s+d}}$ and $K = m^{-s/d} = n^{\frac{-2s}{4s+d}}$ we get that $g_\lambda \in \Sigma(s, L)$ as long as there is some wiggle room around p. We also get that the Hellinger distance is bounded by $O(n^2 n^{\frac{-8s}{4s+d}} n^{\frac{-2d}{4s+d}}) = O(1)$ and the distance in our metric is $n^{\frac{-4s}{4s+d}}$ as we desired. We can apply Theorem 7 and arrive at the result.

E.3. Existence of u_j

To wrap up, we need to show that we can in fact find the functions u_j. We can do this by mapping R_j to $[0,1]^d$ and using an orthonormal system $\{\phi_j\}_{j=1}^q$ for $L^2([0,1]^d)$ with $q \geq 3$. Suppose that ϕ_j satisfy (i) $\phi_1 = 1$, $\phi_j(x) = 0$ for $x \notin [\epsilon, 1-\epsilon]^d$ and (iii) $\|D^r \phi_j\|_\infty \leq K < \infty$ for all j. Certainly we can find such an orthonormal system.

Now for any function $f \in L^2([0,1]^d)$, we can easily find a unit-normed function $\tilde{v} \in \text{span}(\{\phi_j\})$ such that $\tilde{v} \perp \phi_1$, and $\tilde{v} \perp f$. If we write $\tilde{v} = \sum_i c_i \phi_i$ we have that $D^r v = c_i D^r \phi_i$ so that $\|D^r v\|_\infty \leq K \sum_i |c_i| \leq K\sqrt{q}$ since $\tilde{v}$ is unit-normed. Notice that the vector $v = \tilde{v}(K\sqrt{q})^{-1}$ has upper and lower-bounded ℓ_2^2-norm while having all $\|D^r v\|_\infty \leq 1$.

To construct the functions u_j, map the $R_j = \Pi_{i=1}^d [j_i m^{-1/d}, (j_i+1)m^{-1/d}]$ to $[0,1]^d$ and let the function $f = p^{\alpha-1}(x) q^\beta(x)$ mapped appropriately to $[0,1]^d$. Use the function v_j constructed in the previous paragraph. In mapping back to R_j, let $u_j(x) = v_j(m^{1/d}(x - (j_1, \ldots, j_d))^T)$ so that $\int_{R_j} u_j^2(x)dx = m^{-1}\int v_j^2(x)dx = \Omega(1/m)$ and $\|D^r u_j\|_\infty \leq m^{r/d}$. These functions u_j meet the requirements 1-5 outlined above, allowing us to apply Le Cam's method.

E.4. An $n^{-1/2}$ Lower Bound when $s > d/4$

To obtain the $n^{-1/2}$ lower bound for the highly-smooth setting, we will reduce the problem of estimating $T(p,q)$ to that of estimating a quadratic functional of the two densities:

$$\theta(p,q) = \int a_1(x)p(x) + a_2(x)q(x) + a_3(x)p(x)q(x) + a_4(x)p^2(x) + a_5(x)q(x)d\mu(x) \tag{21}$$

for some known functions $a_i : [0,1]^d \to \mathbb{R}$, $i \in \{1, \ldots, 5\}$. We will then use the following lower bound on the rate of estimating these functionals to establish a lower bound in our problem:

Theorem 13. *Let $a_i : [0,1]^d \to \mathbb{R}, i \in \{1, \ldots, 5\}$ be continuous, bounded, non-constant functions and let $\theta(p,q)$ be as in Equation 21. Then:*

$$\liminf_{n\to\infty} \inf_{\hat{\theta}_n} \sup_{p,q \in \Sigma(s,L)} \mathbb{P}_{X_1^n \sim p, Y_1^n \sim q}[|\hat{\theta}_n - \theta(p,q)| \geq \epsilon n^{-1/2}] \geq c > 0 \tag{22}$$

For some constants $\epsilon, c > 0$.

Proof. We will use Le Cam's Method to establish the lower bound. Let us fix q once and for all. We will only vary p. Let $p_0(x) = 1$ and $p_1(x) = 1 + u(x)$ for some function $u(x)$ that we will select later. By Theorem 2.2 of (Tsybakov, 2009) (essentially the Neyman-Pearson Lemma) if we can upper bound $KL(p_1^n \times q^n, p_0^n \times q^n)$ we have a lower bound on the probability of making an error in the simple versus simple hypothesis test between the two possible distributions when $X_1^n, \sim p_1$ and $Y_1^n \sim q$. Mathematically, define $p_{e,1}(\psi) = \mathbb{P}_{X_1^n \sim p_1, Y_1^n \sim q}[\psi(X_1^n, Y_1^n) \neq 1]$ for a test statistic ψ taking values in $\{0,1\}$. Also define $p_{e,1} = \inf_\psi p_{e,1}(\psi)$. Then Theorem 2.2 of (Tsybakov, 2009) says that if $KL(p_1^n \times q^n, p_0^n \times q^n) \leq \alpha < \infty$ then

$$p_{e,1} \geq \max\left(\frac{1}{4}\exp(-\alpha), \frac{1-\sqrt{\alpha/2}}{2}\right)$$

So let us bound the KL-divergence:

$$KL(p_1^n \times q^n, p_0^n \times q^n) = nKL(p_1, p_0) = n\int (1+u(x))\log(1+u(x))dx \leq n\int u(x) + u^2(x)dx = n\|u\|_2^2$$

Here we used that $\int u(x) = 0$ if p_1 is to remain a density. This is one of the requirements on the function u that we will pick. If the KL-divergence is to remain bounded, we will also require that $\|u\|_2^2 \leq c/n$ for some constant.

If we make a mistake in the testing problem, we suffer at least $1/2|\theta(p_0, q) - \theta(p_1, q)|$ loss in the estimation problem. So we must lower bound the absolute difference between the two functional values.

$$\begin{aligned} |\theta(p_0, q) - \theta(p_1, q)| &= |\int a_1(x)u(x) + a_3(x)q(x)u(x) + 2a_4(x)u(x) + a_4(x)u^2(x)d\mu(x)| \\ &= |\int f(x)u(x) + a_4(x)u^2(x)d\mu(x)| \end{aligned}$$

where $f(x) = a_1(x) + a_3(x)q(x) + 2a_4(x)$. Suppose we had a function v such that:

$$\int v(x) = 0,\ \|v(x)\|_2^2 = O(1),\ p_1 = 1 + 1/\sqrt{n}v(x) \in \Sigma(s, L),\ \int f(x)v(x) = \Omega(1)$$

Then if we use $u(x) = n^{-1/2}v(x)$ the loss we suffer is at least $c_1/\sqrt{n} - c_2/n \geq \epsilon n^{-1/2}$ for some $\epsilon > 0$ for n sufficiently large. At the same time, the KL-divergence between the two hypothesis is also $O(1)$. So we would be able to apply Le Cam's inequality.

So, we just need to find a sufficiently smooth function v with constant ℓ_2^2 norm and constant inner product with f. To do this, consider an orthonormal system $\phi_1, \ldots, \phi_q$ with $q \geq 3$ of $L^2([0,1]^d)$ such that (i) $\phi_j(x) = 1$, (ii) $f \in \text{span}(\{\phi_j\}_{j=1}^q)$ and (iii) $\|D^r\phi_j\|_\infty \leq K < \infty$ for all j and all tuples r with $\sum_j r_j \leq s + 1$. It is always possible to construct such a system as long as f itself has bounded r-th derivatives, which is true since f itself is a continuous, bounded function over a compact domain. Let L denote the linear space spanned by $\{\phi_j\}$. Earlier we showed that if $v \in L$, then $v \in \Sigma(s, A)$ for sufficiently large constant A. So we can let v be any unit-normed function in $L' = \{v \in L | \langle v, f\rangle = c, \langle v, \phi_1\rangle = 0\}$, which is an affine space of dimension at least 1 (since $f \neq c\phi_1$).

Then $u(x) = v(x)/\sqrt{n}$ meets all of the requirements. Notice that since $v \in \Sigma(s, A)$, we have that $u \in \Sigma(s, A/\sqrt{n}) \subset \Sigma(s, L)$ for n sufficiently large. □

In what follows, the functional θ that we are trying to estimate will actually be a random quantity. However, since Theorem 13 applies to any set of five bounded continuous function $a_1, \ldots, a_5$, it actually applies to any distribution over this space of five bounded continuous functions. Mathematically, for any distribution $\mathcal{D}$ over this space of bounded continuous functions:

$$\liminf_{n\to\infty} \inf_{\hat{\theta}_n} \sup_{p,q\in\Sigma(s,L)} \mathbb{P}_{X_1^n\sim p, Y_1^n\sim q, (a_1,\ldots,a_5)\sim\mathcal{D}} \left[|\hat{\theta}_n(a_1^5) - \theta(a_1^5, p, q)| \geq \epsilon n^{-1/2}\right] \geq c > 0$$

where $\theta(a_1^5, p, q)$ is given in Equation 21.

Let us use Theorem 13 to prove a lower bound for estimating $T(p, q) = \int p^\alpha q^\beta$. Suppose we had an estimator $\widehat{T}_n$ for $T(p, q)$ that converges at rate $o(n^{-1/2})$, say $\forall p, q, n, \mathbb{E}[|\widehat{T}_n - T(p, q)|] \leq c_1 n^{-1/2-\epsilon}$ for some constants $c_1, \epsilon > 0$. We will use it to construct an estimator for a quadratic functional of p, q with better-than-$\sqrt{n}$ rate, which will contradict Theorem 13.

The quadratic functional of p, q will be the terms in the second order expansion of $T(p, q)$ about $T(\hat{p}_n, \hat{q}_n)$.

Given $2n$ samples, as in our upper bound, we use the first n to construct estimators $\hat{p}_n, \hat{q}_n$ for p, q respectively. We use the second n samples to compute $\widehat{T}_n$. The estimator for θ will be $\hat{\theta}_{2n} = \widehat{T}_n - C_2T(\hat{p}_n, \hat{q}_n)$. Where we are collecting all of the terms of the form $T(\hat{p}_n, \hat{q}_n)$ together. Recall that C_2 is the coefficient for all of these terms.

The risk of the estimator is:

$$\begin{aligned} \mathbb{E}_{X_1^{2n}}[|\hat{\theta}_n - \theta|] &\leq \mathbb{E}_{X_{n+1}^{2n}}[|\widehat{T}_n - T|] + \mathbb{E}_{X_1^{2n}}[|T - C_2T(\hat{p}, \hat{q}) - \theta|] \\ &\leq c_1 n^{-1/2-\epsilon} + O(\mathbb{E}_{X_1^n}[\|p - \hat{p}\|_3^3 + \|q - \hat{q}\|_3^3]) \\ &\leq c_1 n^{-1/2-\epsilon} + c_2 n^{\frac{-3s}{2s+d}} \end{aligned}$$

for constants $c_1, c_2 > 0$. Now if $s > d/4$, both terms are $o(n^{-1/2})$, so we have $\mathbb{E}[|\hat{\theta}_n - \theta|] = o(n^{-1/2})$. The functions $\hat{p}_n, \hat{q}_n$ are deterministic functions of X_1^n, Y_1^n, so we can think of X_1^n as encoding a distribution over functions $\hat{p}_n, \hat{q}_n$.

More formally, let $\mathcal{D}$ encode the following distribution: We drawn X_1^n, Y_1^n from p, q respectively and compute $\hat{p}_n, \hat{q}_n$. With these, the five functions $a_1, \ldots, a_5$ are:

$$\begin{aligned} a_1 &= \alpha(2-\alpha-\beta)\hat{p}_n^{\alpha-1}\hat{q}_n^{\beta} \\ a_2 &= \beta(2-\alpha-\beta)\hat{p}_n^{\alpha}\hat{q}_n^{\beta-1} \\ a_3 &= \alpha\beta\hat{p}_n^{\alpha-1}\hat{q}_n^{\beta-1} \\ a_4 &= 1/2\alpha(\alpha-1)\hat{p}_n^{\alpha-2}\hat{q}_n^{\beta} \\ a_5 &= 1/2\beta(\beta-1)\hat{p}_n^{\alpha}\hat{q}_n^{\beta-2} \end{aligned}$$

Notice that all of these functions are continuous and they can be bounded from above and below if we use the truncated kernel density estimators. Now whenever $s > d/4$:

$$\mathbb{E}_{(a_1,\ldots,a_5)\sim\mathcal{D}}\mathbb{E}_{X_1^n\sim p, Y_1^n\sim q}\left[|\hat{\theta}-\theta|\right] = \mathbb{E}_{X_1^{2n}\sim p, Y_1^{2n}\sim q}\left[|\hat{\theta}-\theta|\right] \leq cn^{-1/2-\epsilon}$$

which contradicts the lower bound. Via Markov's inequality, $\mathbb{P}_{X_1^{2n}}[|\hat{\theta}_n-\theta| \geq c_4 n^{-1/2}] \leq o(n^{-1/2})/n^{-1/2} \to 0$ which contradicts our discussion following Theorem 13. This shows that when $s > d/4$, one cannot estimate $T(p,q)$ are faster than $\sqrt{n}$ rate.

E.5. Translating to T_α and D_α

Suppose we have an estimator $\hat{S}_\alpha$ for the Tsallis-α divergence, such that for all $p, q \in \Sigma(s,l)\mathbb{E}[|\hat{S}_\alpha - S_\alpha|] \leq \epsilon_n$. We can define an estimator $\hat{T}$ for $T(p,q) = \int p^\alpha q^{1-\alpha}$ as $\hat{T} = (\alpha-1)\hat{S}_\alpha + 1$. The error between $\hat{T}$ and T is:

$$\mathbb{E}[|\hat{T}-T|] = |\alpha-1|\mathbb{E}[|\hat{S}_\alpha - S_\alpha|] \leq |\alpha-1|\epsilon_n$$

We therefore know that $\epsilon_n = \Omega(n^{-\gamma})$ where $\gamma = \min\{\frac{4s}{4s+d}, 1/2\}$ since otherwise we would have an estimator $\hat{T}$ for $T(p,q)$ with rate $o(n^{-\gamma})$, which contradicts Theorem 2.

For D_α, we use the same proof structure, but computing the error for $\hat{T}$ is more involved. The estimator $\hat{T} = \exp\{(\alpha-1)\hat{D}\}$ has error:

$$\mathbb{E}[|\hat{T}-T|] = \mathbb{E}\left[|\exp\{(\alpha-1)\hat{D}\} - \exp\{(\alpha-1)D_\alpha\}|\right]$$

We would like to eliminate the absolute value, so we will have to consider all of the cases. If $\alpha < 1$ and $D > \hat{D}$ then the first term dominates the second so we can simply drop the absolute value sign. In this case we can use convexity of e^x to upper bound by:

$$\leq (\alpha-1)\mathbb{E}[e^{(\alpha-1)\hat{D}}(\hat{D}-D_\alpha)] = (1-\alpha)\mathbb{E}[e^{(\alpha-1)\hat{D}}(D_\alpha-\hat{D})] \leq C\epsilon_n$$

as long as D_α is bounded from below, which implies that for n large enough, $e^{(\alpha-1)\hat{D}} = O(1)$. Actually the other cases are analogous, for example if $\hat{D} > D$, then to remove the absolute value, we must swap the two terms, after which we can use convexity to arrive at the same upper bound. Thus we have shown that $\mathbb{E}[|\hat{T}-T|] = O(\epsilon_n)$ which implies that $\mathbb{E}[|\hat{D}-D|] = \Omega(n^{-\gamma})$ as claimed.

F. More Auxiliary Results

Lemma 14 (Hölder is contained in Sobolev)**.** *Let $f \in \Sigma(s,L)$ belong to the periodic holder class with smoothness s. Then f belongs to the sobolev ellipsoid $\mathcal{W}(s',L')$ where $\phi_k(x) = e^{2i\pi k^T x}$ is the fourier basis, $k \in \mathbb{Z}^d$, $s' < s$ and:*

$$L' = \frac{dCL^2}{(2\pi)^{2\lfloor s\rfloor}}$$

with $C = \sum_{l=0}^{\infty} 4^{l(s'-s)}$.

Proof. Let us decompose $s = r + \alpha$ where $r = \lfloor s\rfloor$ and $\alpha \in (0,1]$. We need to bound:

$$\sum_{(k_1,\ldots,k_d)\in\mathbb{Z}^d}(\sum_{j=1}^{d}|k_j|^{2s'})|\alpha_k|^2$$

where $\alpha_k = \int f(x)\phi_k(x)dx$. This is equivalent to bounding, for each $j = [d]$, $\sum_{k\in\mathbb{Z}^d} |k_j|^{2s'}|\alpha_k|^2$ so let us fix a dimension j for now. Using repeated integration by parts and the fact that $D^{\vec{r}}f$ is period for all $\vec{r}$ with $\sum_j r_j \leq r$. we get

$$\left|\int \frac{\partial^r}{\partial x_j^r} f(x)\phi_k(x)dx\right| = |2\pi i k_j|^r| \int f(x)\phi_k(x)dx| = |2\pi i k_j|^r|\alpha_k|$$

Let us write $g(x) = \frac{\partial^r}{\partial x_j^r} f(x)$. Then since $f \in \Sigma(s, L)$, we know that g satisfies:

$$|g(x) - g(y)| \leq L\|x - h\|^\alpha$$

for all x, y. We will use this fact to bound $\sum_{k\in\mathbb{Z}^d} |k_j|^{2\alpha'}|b_k|^2$ where $b_k = \int g(x)\phi_k(x)$ and $\alpha' < \alpha$ which will give us a bound on $\sum_{k\in\mathbb{Z}^d} |k_j|^{2s'}|\alpha_k|$ via the above calculation. In particular, suppose that $\sum_{k\in ZZ^d} |k_j|^{2\alpha'}|b_k|^2 \leq \gamma_j$, then:

$$\sum_{k\in\mathbb{Z}^d} |k_j|^{2s'}|\alpha_k|^2 = \sum_{k\in\mathbb{Z}^d} |k_j|^{2r+2\alpha'}|\alpha_k^2| = |2\pi i|^{-2r} \sum_{k\in\mathbb{Z}^d} |k_j|^{2\alpha'}|b_k|^2 \leq (2\pi)^{-2r}\gamma_j$$

Notice that:

$$g(x_1, \ldots, x_j - h, \ldots, x_d) - g(x_1, \ldots, x_j + h, \ldots, x_d) = \sum_{k\in\mathbb{Z}^d} b_k e^{2i\pi k^T x} 2i \sin(2\pi k_j h)$$

This means that:

$$4\sum_{k\in\mathbb{Z}^d} |b_k|^2 \sin^2(2\pi k_j h) = \int (g(x_1, \ldots, x_j - h, \ldots, x_d) - g(x_1, \ldots, x_j + h, \ldots, x_d))^2 dx \leq L^2|h|^{2\alpha}$$

Notice that $\sin^2(\pi/2) > \sin^2(\pi/4) \geq 1/2$ so if we pick $h = 1/(8q)$ and $k_j \in \{q, \ldots, 2q-1\} \cup \{-q, \ldots, -2q+1\}$ we can lower bound the left hand side. To be concrete, letting $S_q = \{k \in \mathbb{Z}^d | k_j \in \{q, \ldots, 2q-1\} \cup \{-q, \ldots, -2q+1\}\}$:

$$\sum_{k\in\mathbb{Z}^d} |b_k|^2|k_j|^{2\alpha'} = \sum_{l=0}^{\infty} \sum_{k\in S_{2^l}} |b_k|^2|k_j|^{2\alpha'} \leq \sum_{l=0}^{\infty} (2^{l+1})^{2\alpha'} \sum_{k\in S_{2^l}} |b_k|^2$$

But:

$$\sum_{k\in S_{2^l}} |b_k|^2 \leq 2\sum_{k\in S_{2^l}} |b_k|^2 \sin^2(2\pi k_j(1/2^{l+3})) \leq 2\sum_{k\in\mathbb{Z}^d} |b_k|^2 \sin^2(2\pi k_j(1/2^{l+3})) \leq \frac{L^2}{2} 2^{-2\alpha(l+3)}$$

Using this bound above, we get:

$$\sum_{k\in\mathbb{Z}^d} |b_k|^2|k_j|^{2\alpha'} \leq \frac{L^2}{2}\frac{4^{2\alpha'}}{8^{2\alpha}} \sum_{l=0}^{\infty} 4^{l(\alpha'-\alpha)} \leq CL^2$$

whenever the series converges (as long as $\alpha' < \alpha$).

Using this as our value for γ_j and summing over the d dimensions, we get:

$$\sum_{j=1}^{d} \sum_{k\in\mathbb{Z}^d} |k_j|^{2s'}|\alpha_k| \leq d(2\pi)^{-2r}\gamma_j \leq \frac{dCL^2}{(2\pi)^{2r}}$$

□