A Proof of Theorems 3 and 4

We analyze the two estimators separately and Theorem 3 follows immediately from Theorems 10 and 11 below. For the estimator without data splitting, the result follows from below and the inequality:

$$\begin{aligned}&\mathbb{E}\left[\left(\hat{\theta}_p + \hat{\theta}_q - 2\hat{\theta}_{pq} - \theta_p - \theta_q + 2\theta_{pq}\right)^2\right]\\&\le \mathbb{E}[(\hat{\theta}_p - \theta_p)^2] + \mathbb{E}[(\hat{\theta}_q - \theta_q)^2] + \mathbb{E}[(\hat{\theta}_{pq} - \theta_{pq})^2]\\&+ 2\left|\mathbb{E}\hat{\theta}_p - \theta_p\right|\left|\mathbb{E}\hat{\theta}_q - \theta_q\right| + 4\sqrt{\mathbb{E}[(\hat{\theta}_{pq} - \theta_{pq})^2]}\left(\sqrt{\mathbb{E}[(\hat{\theta}_p - \theta_p)^2]} + \sqrt{\mathbb{E}[(\hat{\theta}_q - \theta_q)^2]}\right)\end{aligned}$$

Plugging in the bounds from Theorems 10 and 11 immediately establish the rate of convergence for the estimator without data splitting.

For the quadratic term estimators, we make a slight modification to a theorem from Gine and Nickl [5]. The only difference between our proof and theirs is in controlling the bias, where we use the bounded-variation assumption while they use a Sobolev assumption. However this has little bearing, as the bias is still of the same order, and we have the following theorem characterizing the behavior of the quadratic estimator:

Theorem 10 (Adapted from [5])**.** *Under Assumption 2, we have:*

$$\left|\mathbb{E}[\hat{\theta}_p] - \theta_p\right| \le c_b h^{2\beta} \qquad \mathbb{E}\left[(\hat{\theta}_p - \mathbb{E}[\theta_p])^2\right] \le c_v\left(\frac{1}{n} + \frac{1}{n^2h^d}\right), \tag{17}$$

and when $\beta > d/4$:

$$\sqrt{n}(\hat{\theta}_p - \theta_p) \rightsquigarrow \mathcal{N}(0, 4\operatorname*{Var}_{x\sim p}(p(x))). \tag{18}$$

While we are not aware of any analyses of the bilinear term, it is not particularly different from the quadratic term, and we have the following theorem:

Theorem 11. *Under Assumption 2, we have:*

$$\left|\mathbb{E}[\hat{\theta}_{pq}] - \theta_{pq}\right| \le c_b h^{2\beta} \qquad \mathbb{E}\left[(\hat{\theta}_{pq} - \mathbb{E}[\theta_{pq}])^2\right] \le c_v\left(\frac{1}{n} + \frac{1}{n^2h^d}\right), \tag{19}$$

and when $\beta > d/4$:

$$\sqrt{n}(\hat{\theta}_{pq} - \theta_{pq}) \rightsquigarrow \mathcal{N}(0, \operatorname*{Var}_{x\sim p}(q(x)) + \operatorname*{Var}_{y\sim q}(p(y))). \tag{20}$$

Proof of Theorem 10. We reproduce the proof of Gine and Nickl for completeness. The bias can be bounded by:

$$\begin{aligned}\mathbb{E}[\hat{\theta}_p] - \theta_p &= \int\int K_h(x,y)p(y)dyp(x)dx - \int p(x)p(x)dx = \int\int K_h(x,y)[p(y)-p(x)]p(x)dydx\\&= \int\int K(u)[p(x-uh) - p(x)]p(x)dudx = \int K(u)\left[(p_0 \star p)(uh) - (p_0\star p)(0)\right]du,\end{aligned}$$

where $p_0(x) = p(-x)$ and $\star$ denotes convolution. Now by Lemma 14 below, we know that $p_0 \star p \in \mathcal{W}_1^{2\beta}(C^2)$ and can take a Taylor expansion of order $2\beta - 1$. When we take such an expansion, by the properties of the kernel, all but the remainder term is annihilated and we are left with:

$$\frac{h^{2\beta}}{(2\beta)!}\sum_{r_1,\ldots,r_d|\sum_i r_i = 2s}\int K(u)\Pi_i u_i^{r_i}\xi(r,uh)du \le c_b h^{2\beta},$$

where we used the fact the function is integrable by the fact that $\xi \in L^1$, which in turn follows from the fact that $p_0 \star p \in \mathcal{W}_1^{2\beta}(C^2)$ and by Taylor's remainder theorem. We are also using the compactness of K here so that we only have to integrate over $(-1,1)^d$ in which case all polynomial functions are also L_1 integrable. This shows that the bias is $O(h^{2\beta})$.

Note that the main difference between our proof and that of Gine and Nickl is in the smoothness assumption, which comes into play here. Under the bounded variation assumption, we were able to argue that smoothness is additive under convolution. The same is true under the Sobolev assumption, and this property is exploited by Gine and Nickl in exactly the same way as we do here. Unfortunately, Hölder smoothness is not additive under convolution, so the more standard assumption does not provide the semiparametric rate of convergence.

As for the variance, we may write:

$$\mathbb{E}[\hat{\theta}_p^2] - (\mathbb{E}\hat{\theta}_p)^2 = \mathbb{E}\left[\frac{1}{n^2(n-1)^2}\sum_{i \neq j, s \neq t} K_h(X_i, X_j)K_h(X_s, X_t)\right] - (\mathbb{E}\hat{\theta}_p)^2,$$

which we can split into three cases. When $i \neq j \neq s \neq t$, each term in the sum is exactly $(\mathbb{E}\hat{\theta})^2$, and this happens for $n(n-1)(n-2)(n-3)$ terms in the sum. When one of the first indices is equal to one of the second indices we get:

$$\begin{aligned}\mathbb{E}K_h(X_i, X_j)K_h(X_i, X_t) &= \int\int\int K_h(X_i, X_j)K_h(X_i, X_t)p(X_i)p(X_j)p(X_t)dX_i dX_j dX_t \\ &= \int\int\int K(u_j)K(u_t)p(X_i - u_j h)p(X_i - u_t h)p(X_i)du_j du_t dX_i \\ &\leq ||K||_2^2 ||p||_2^2,\end{aligned}$$

where we performed a substitution to annihilate the dependence on h. There are $4n(n-1)(n-2)$ expressions of this form, so in total, these terms contribute:

$$\frac{1}{n}||K||_2^2||p||_2^2.$$

Finally, the $2n(n-1)$ terms where $i = s, j = t$ or vice versa in total contribute:

$$\begin{aligned}\frac{2}{n(n-1)}\mathbb{E}K_h^2(X_i, X_j) &= \frac{2}{h^{2d}n(n-1)}\int K^2(\frac{X_i - X_j}{h})p(X_i)p(X_j)dX_i dX_j \\ &= \frac{2}{h^d n(n-1)}\int K^2(u_j)p(X_i)p(X_i - u_j h)du_j dX_i \\ &\leq \frac{2||K||_2^2||p||_2^2}{h^d n^2}.\end{aligned}$$

Adding together these terms, establishes the variance bound in the theorem. The rate of convergence in Theorem 3 follows from plugging the definition of h, which was selected to optimize the tradeoff between bias and variance.

As for asymptotic normality, we decompose the proof into several steps.

1. Control the bias.
2. Apply Hoeffding's decomposition.
3. Control the second order term, which will be lower order.
4. Show that the first order term is close to $P_n p - \theta$ (here P_n is the empirical measure).
5. Apply the Lindberg-Levy central limit theorem to $P_n p - \theta$.

As usual we have the decomposition:

$$\hat{\theta}_p - \theta_p = \underbrace{\hat{\theta}_p - \mathbb{E}\hat{\theta}_p}_{\text{Variance}} + \underbrace{\mathbb{E}\hat{\theta}_p - \theta}_{\text{Bias}}.$$

We already controlled the bias above. Specifically we know that $\sqrt{n}(\mathbb{E}\hat{\theta}_p - \theta_p) \leq \sqrt{n}h^{2\beta} \to 0$ with our setting of h and under the assumption that $\beta > d/4$.

As is common in the analysis of U-statistics, we apply Hoeffding's decomposition before proceeding. That is, we write:

$$\hat{\theta}_p - \mathbb{E}\hat{\theta}_p = U_n(\pi_2 K_h) + 2P_n(\pi_1 K_h),$$

where $U_n f = \frac{1}{n(n-1)}\sum_{i\neq j} f(X_i, X_j)$ is the U-process and $P_n f = \frac{1}{n}\sum_i f(X_i)$ is the empirical process and:

$$\begin{aligned}(\pi_1 K_h)(X) &= \mathbb{E}_{x\sim p} K_h(x, X) - \mathbb{E}_{x,y\sim p} K_h(x,y)\\ (\pi_2 K_h)(X,Y) &= K_h(X,Y) - \mathbb{E}_{x\sim p}K_h(x,Y) - \mathbb{E}_{y\sim p}K_h(X,y) + \mathbb{E}_{x,y\sim p}K_h(x,y).\end{aligned}$$

It is easy to very that our estimator can be decomposed in this manner. Moreover, since everything is centered, the two terms also have zero covariance. Also notice that $\mathbb{E}_{x\sim p}K_h(x,Y) = \bar{p}(Y)$ and $\mathbb{E}_{x,y\sim p}K_h = \int \bar{p}(x)p(x)$ where $\bar{p}$ is the expectation of the density estimate.

We now control the second order term $U_n(\pi_2 K_h)$ by showing convergence in quadratic mean.

$$\mathbb{E}[(U_n(\pi_2 K_h))^2] = \frac{1}{n(n-1)}\mathbb{E}[(\pi_2 K_h(X_1, X_2))^2] \leq \frac{c}{n^2 h^d}||K||_2^2 ||p||_2^2.$$

The first equality follows from the fact that each term is conditionally centered, so all cross terms are zero, while the inequality is the result of performing a substitution as we have seen before. Thus $\sqrt{n}U_n(\pi_2 K_h) \to 0$ since $\frac{1}{nh^d} \to 0$ when $\beta > d/4$.

For the first order term $P_n(\pi_1 K_h)$, we now show that it is close to $P_n p - \theta_p$.

$$\mathbb{E}[(P_n(\pi_1 K_h) - (P_n p - \int p^2))^2] \leq \frac{1}{n}\mathbb{E}[(\bar{p}(X) - p(x))^2] \leq \frac{||\bar{p} - p||_\infty^2}{n} = \frac{ch^{2\beta}}{n},$$

so that $\sqrt{n}P_n(\pi_1 K_h) \to^{q.m.} \sqrt{n}(P_n p - \theta_p)$ since $h^{2\beta} \to 0$.

Now by the Lindberg-Levy CLT, we know that:

$$\sqrt{n}(2P_n p - 2\theta_p) \rightsquigarrow \mathcal{N}(0, 4 \underset{x\sim p}{\mathrm{Var}}(p(X))),$$

which concludes the proof of the theorem. □

We now prove Theorem 11, although the arguments are fairly similar.

Proof of Theorem 11. The bias is:

$$\begin{aligned}\mathbb{E}[\hat{\theta}_{pq}] - \theta_{pq} &= \int\int K_h(x,y)p(y)dyq(x)dx - \int p(x)q(x)dx\\ &= \int\int K_h(x,y)[p(y) - p(x)]q(x)dydx\\ &= \int\int K(u)[p(x-uh) - p(x)]q(x)dudx = \int K(u)\left[(p_0 \star q)(uh) - (p_0\star q)(0)\right]du,\end{aligned}$$

where, as before, $p_0(x) = p(-x)$ and $\star$ denotes convolution. So we can proceed as in the quadratic setting. Specifically, by Lemma 14, we can take a Taylor expansion of order $2\beta + 1$, annihilate all but the remainder term, which we know is bounded by the fact that $p_0 \star q \in \mathcal{W}_1^{2\beta}(C^2)$. Formally, the remainder term is:

$$\frac{h^{2\beta}}{(2\beta)!}\sum_{r_1,\ldots,r_d | \sum_i r_i = 2s}\int K(u)\Pi_i u_i^{r_i}\xi(r, uh)du \leq c_b h^{2\beta},$$

where we used the fact the function is integrable by the fact that $\xi \in L^1$, since $p_0 \star q \in \mathcal{W}_1^{2\beta}(C^2)$. Thus the bias is $O(h^{2\beta})$.

The variance can be bounded in a similar way to the quadratic estimator:

$$\mathbb{E}[\hat{\theta}_{pq}^2] - \mathbb{E}[\hat{\theta}_{pq}]^2 = \frac{1}{n^4}\sum_{i,j,s,t}\mathbb{E}[K_h(X_i, Y_j)K_h(X_s, Y_t)] - \mathbb{E}[\hat{\theta}_{pq}]^2.$$

Whenever $i \neq s$ and $j \neq t$ all of the terms are independent so they cancel out with the $\mathbb{E}[\hat{\theta}_{pq}]^2$ term. This happens for $n^2(n-1)^2$ terms.

When $i = s, j \neq t$, we substitute $u_j = h^{-1}(X_i - Y_j)$ and $u_t = h^{-1}(X_i - Y_t)$ for Y_j, Y_t to see that:

$$\begin{aligned}\frac{1}{n^4}\sum_{i,j\neq t}\mathbb{E}[K_h(X_i,Y_j)K_h(X_i,Y_t)] &= \frac{n-1}{h^{2d}n^2}\int\int\int K(\frac{X_i-Y_j}{h})K(\frac{X_i-Y_t}{h})p(X_i)q(Y_j)q(Y_t)\\ &= \frac{n-1}{n^2}\int\int\int K(u_j)K(u_t)p(X_i)q(X_i-u_jh)q(X_i-u_th)\\ &\leq \frac{1}{n}||K||_2^2||q||_2^2.\end{aligned}$$

Thus, the total contribution from the terms where $j = t, i \neq s$ is bounded by $\frac{1}{n}||K||_2^2||p||_2^2$.

When $j = t, i = s$, we can only perform one substitution so a factor of h^d will remain. Formally:

$$\begin{aligned}\frac{1}{n^2h^{2d}}\int\int K^2(\frac{X_i-Y_j}{h})p(X_i)q(Y_j) &= \frac{1}{n^2h^d}\int\int K^2(u_j)p(X_i)q(X_i-u_jh)\\ &\leq \frac{1}{n^2h^d}||K||_2^2||p||_2||q||_2.\end{aligned}$$

Therefore, the total variance is $O(n^{-1}+n^{-2}h^{-d})$ as in the theorem statement.

The proof of asymptotic normality of the bilinear estimator is not too different from the proof for the quadratic estimator. We can start by ignoring the bias, as when $b \geq d/4$, we know that $\sqrt{n}(\mathbb{E}\hat{\theta}_{pq} - \theta_{pq}) \to 0$. To analyze the variance term we make use of the following decomposition:

$$\begin{aligned}\hat{\theta}_{pq} - \mathbb{E}\hat{\theta}_{pq} &= \frac{1}{n^2}\sum_{ij}K_h(X_i,Y_j) + \frac{1}{n}\sum_i \bar{q}(X_i) - \frac{1}{n}\sum_i \bar{q}(X_i) + \frac{1}{n}\sum_j \bar{p}(Y_i) - \frac{1}{n}\sum_j \bar{p}(Y_i) - \mathbb{E}\hat{\theta}_{pq}\\ &= V_n(\pi_2K_h) + P_n(\pi_{11}K_h) + Q_n(\pi_{12}K_h),\end{aligned}$$

Where:

$$\begin{aligned}(\pi_2K_h)(X,Y) &= K_h(X,Y) - \bar{q}(X) - \bar{p}(Y) + \mathbb{E}K_h(x,y)\\ (\pi_{11}(K_h))(X) &= \bar{q}(X) - \mathbb{E}K_h(x,y)\\ (\pi_{12}(K_h))(Y) &= \bar{p}(Y) - \mathbb{E}K_h(x,y).\end{aligned}$$

Here P_n, Q_n are the empirical processes associated with the samples X, Y respectively and $\bar{p}, \bar{q}$ are the expectations of the kernel density estimators. Also, V_n is the V-process, that is $V_nf = \frac{1}{n^2}\sum_{i,j} f(X_i, Y_j)$. Notice that each term is conditionally centered, which implies that each pair of terms has zero covariance. Thus we only have to look at the variances.

As before, the goal is to show that the V-process term is lower order and then to apply the Lindeberg-Levy CLT to the other two terms. Since each term is conditionally centered:

$$\mathbb{E}[(V_n(\pi_2K_h))^2] = \frac{1}{n^2}\mathbb{E}[(\pi_2K_h(X,Y))^2] \leq \frac{c}{n^2h^d}||K||_2^2||p||_2||q||_2,$$

where the last step follows by performing the substitution $u = \frac{X-Y}{h}$ in each term of the integral.

For the first order terms, we first show that they are close to $q(X) - \mathbb{E}[q(x)]$ and $p(Y) - \mathbb{E}[p(y)]$ so that we can apply the CLT to the latter. We will show convergence in quadratic mean.

$$\mathbb{E}\left[(P_n(\pi_{11}K_h) - (P_nq - \int pq))^2\right] \leq \frac{1}{n}\mathbb{E}\left[(\bar{q}(x) - q(x))^2\right] \leq \frac{\|\bar{q}-q\|_\infty^2}{n} \leq \frac{h^{2\beta}}{n},$$

which means that, under our choice of h and with $\beta > d/4$, $\sqrt{n}P_n(\pi_{11}K_h) \to^{q.m.} \sqrt{n}P_n(q - \theta_{pq})$. Exactly the same argument shows that $\sqrt{n}Q_n(\pi_{12}K_h) \to^{q.m.} \sqrt{n}Q_n(p - \theta_{pq})$.

Finally, by the Lindeberg-Levy CLT, we know that:

$$\sqrt{n}(P_nq - \theta_{pq}) \rightsquigarrow \mathcal{N}(0, \underset{x\sim p}{\mathrm{Var}}(q(x))), \qquad \sqrt{n}(Q_np - \theta_{pq}) \rightsquigarrow \mathcal{N}(0, \underset{y\sim q}{\mathrm{Var}}(p(y))),$$

and since x and y are independent, both of these central limit theorems hold jointly. Since in our estimate for $\hat{D}$ we have a term of the form $2\hat{\theta}_{pq}$, the contribution of this term to the total variance is $4\operatorname{Var}(\hat{\theta}_{pq})$. This concludes the proof. □

B Proof of Theorem 6

In this section we fill in the missing details in the proof of Theorem 6. We will apply the Berry-Esséen inequality for multi-sample U-statistics from Chen, Goldstein and Shao [4], which we reproduce below.

In order to state the theorem we need to make several definitions. We make some simplifications to their result for ease of notation. Consider k independent sequences $X_{j1}, \dots, X_{jn}$ $j = 1, \dots, k$ of i.i.d. random variables, all of length n (this can be relaxed). Let $m_j \geq 1$ for each j and let $\omega(x_{jl}, l \in [m_j], j \in [k])$ be a function that is symmetric with respect to the m_j arguments. In other words, ω is invariant under permutation of two arguments from the same sequence. Let $\theta = \mathbb{E}\omega(X_{jl})$.

The multi-sample U-statistic is defined as:

$$U_n = \left\{ \prod_{j=1}^{k} \binom{n}{m_j}^{-1} \right\} \sum \omega(X_{jl}, j \in [k], l = i_{j1}, \dots, i_{jm_j}), \tag{21}$$

where the sum is carried out over all indices satisfying $1 \leq i_{j1} < \ldots < i_{jm} \leq n$.

Let:

$$\sigma^2 = \mathbb{E}\omega^2(X_{jl}),$$

and for each $j \in [k]$ define:

$$\omega_j(x) = \mathbb{E}[\omega(X_{jl})|X_{j1} = x],$$

with:

$$\sigma_j^2 = \mathbb{E}\omega_j^2(X_{j1}).$$

Lastly, define:

$$\sigma_n^2 = \sum_{j=1}^{k} \frac{m_j^2}{n} \sigma_j^2.$$

We are finally ready to state the theorem:

Theorem 12 (Theorem 10.4 of [4]). *Assume that* $\theta = 0$, $\sigma^2 < \infty$, $\max_{j \in [k]} \sigma_j^2 > 0$. *Then for* $2 < p \leq 3$*:*

$$\sup_{z \in \mathbb{R}} \left| \mathbb{P}\left(\sigma_n^{-1} U_n \leq z\right) - \Phi(z) \right| \leq \frac{6.1}{\sigma_n^p} \sum_{j=1}^{k} \frac{m_j^p}{n^{p-1}} \mathbb{E}[|\omega_j(X_{j1})|^p] + \frac{(1+\sqrt{2})\sigma}{\sigma_n} \sum_{j=1}^{k} \frac{m_j^2}{n}. \tag{22}$$

As we did in our estimator, we split the data into four groups, two samples of size n from each distribution, which we will denote with superscripts, i.e. $X_i^{(1)}$ will be the ith sample from the first group of the data from p. We can write $\hat{D} - \mathbb{E}\hat{D}$ as a zero-mean multi-sample U-statistic with four groups where the first X and Y groups are used for $\hat{\theta}_p - \theta_p$ and $\hat{\theta}_q - \theta_q$ respectively, while the second two groups are used for the cross term $\hat{\theta}_{pq} - \theta_{pq}$.

In other words, ω will be a function that takes 6 variables, two from the $X^{(1)}$ group, two from the $Y^{(1)}$ group and one each from the $X^{(2)}$ and $Y^{(2)}$ groups. Formally, we define:

$$\begin{aligned} &\omega(x_{11}, x_{12}, y_{11}, y_{12}, x_{21}, y_{21}) \\ &= K_h(x_{11}, x_{12}) - \mathbb{E}\hat{\theta}_p + K_h(y_{11}, y_{12}) - \mathbb{E}\hat{\theta}_q - 2K_h(x_{21}, y_{21}) + 2\mathbb{E}\hat{\theta}_{pq}. \end{aligned}$$

With this definition, it is clear that $U_n = \hat{D} - \mathbb{E}\hat{D}$.

To apply Theorem 12 on the appropriate term in the proof, we just have to bound a number of quantities involving ω. As we will see, we will not achieve the $n^{-1/2}$ rate because the function ω depends on the bandwidth h, which is decreasing, so the variance σ is increasing. Specifically:

$$\sigma^2 = \mathbb{E}[\omega^2(X_{11}, X_{12}, Y_{11}, Y_{12}, X_{21}, Y_{21})]$$

$$= \mathbb{E}(K_h(X_{11}, X_{12}) - \mathbb{E}\hat{\theta}_p)^2 + \mathbb{E}(K_h(Y_{11}, Y_{12}) - \mathbb{E}\hat{\theta}_q)^2 + 4\mathbb{E}(K_h(X_{21}, Y_{21}) - \mathbb{E}\hat{\theta}_{pq})^2.$$

Each of the three terms can be analyzed in exactly the same way so we focus on the first term:

$$\begin{aligned}\mathbb{E}(K_h(X_{11}, X_{12}) - \mathbb{E}\hat{\theta}_p)^2 &\leq \int\int K_h^2(X_1, X_2)p(X_1)p(X_2) = \frac{1}{h^d}\int\int K(u)p(X_1 + uh)p(X_1)\\ &\leq \frac{1}{h^d}\|K\|_2^2\|p\|_2^2,\end{aligned}$$

and the same substitution on the other two terms shows that the variance is:

$$\sigma^2 \leq \frac{1}{h^d}\|K\|_2^2\left(\|p\|_2^2 + \|q\|_2^2 + 4\|p\|_2\|q\|_2\right). \tag{23}$$

A similar argument gives us a bound on σ_j^2 $j = 1, \dots 4$. First, since the other terms are centered, we can write $\omega_1(x) = \mathbb{E}(K_h(x, X_2)) - \mathbb{E}\hat{\theta}_p$ with similar expressions for the other terms. Then, σ_1^2 can be simplified to:

$$\begin{aligned}\mathbb{E}\omega_1^2(X) &= \mathbb{E}(\mathbb{E}K_h(X_1, X_2))^2 - (\mathbb{E}\hat{\theta}_p)^2\\ &\leq \int\left(\int K_h(X_1, X_2)p(X_2)\right)^2 p(X_1) = \int\left(\int K(u)p(X_1 - uh)\right)^2 p(X_1) \leq \|K\|_\infty^2.\end{aligned}$$

With exactly the same argument for the other three. Thus:

$$\sigma_n^2 = \frac{1}{n}\sum_{j=1}^{4} m_j^2\sigma_j^2 \leq \frac{10}{n}\|K\|_\infty^2. \tag{24}$$

The last thing we need is the third moments of the linearizations $\mathbb{E}[|\omega_j(x)|^3]$.

$$\begin{aligned}\mathbb{E}\left[\left|\mathbb{E}K_h(X_1, X_2) - \mathbb{E}\hat{\theta}_p\right|^3\right] &= \int\left|\int K_h(x, X_2)p(X_2) - \int\int K_h(X_1, X_2)p(X_1)p(X_2)\right|^3 p(x)\\ &= \int\left|\int K(u)p(x - uh) - \int\int K(u)p(X_1 - uh)p(X_1)\right|^3 p(x)\\ &= \int\left|\int K(u)\left(p(x - uh) - \int p(X_1 - uh)p(X_1)\right)\right|^3 p(x)\\ &\leq 8\|K\|_\infty^3\|p\|_\infty^3\end{aligned}$$

It is easy to verify that each of the third moments are bounded by:

$$\mathbb{E}|\omega_j|^3 \leq 8\|K\|_\infty^3(\|p\|_\infty^3 + \|q\|_\infty^3), \forall j. \tag{25}$$

And plugging in all of these calculations into Theorem 12 shows that:

$$\begin{aligned}&\sup_z\left|\mathbb{P}\left(\sqrt{n}\tilde{\sigma}_n^{-1}U_n \leq z\right) - \Phi(z)\right| \leq\\ &\leq \frac{n^{3/2}(6.1)(18)}{n^2 10^{3/2}\|K\|_\infty^3}8\|K\|_\infty^3(\|p\|_\infty^3 + \|q\|_\infty^3) + \frac{\sqrt{n}(1+\sqrt{2})}{n\sqrt{h^d}\sqrt{10}\|K\|_\infty}\|K\|_2\sqrt{\|p\|_2^2 + \|q\|_2^2 + 4\|p\|_2\|q\|_2}\\ &\leq \frac{27}{\sqrt{n}}\left(\|p\|_\infty^3 + \|q\|_\infty^3\right) + \frac{8}{\sqrt{nh^d}}\frac{\|K\|_2}{\|K\|_\infty}\sqrt{\|p\|_2^2 + \|q\|_2^2 + 4\|p\|_2\|q\|_2}.\end{aligned}$$

This gives the bound in Equation 15.

C Proof of Theorem 7

For completeness we introduce the construction used by Krishnamurthy et al [9]. For the remainder of the proof, we will work of $[0, 1]^d$ and assume that p is pointwise lower bounded by $1/\kappa_l$, noting that a lower bound here applies to the more general setting. For the construction, suppose we have a disjoint collection of subset $A_1, \ldots, A_m \subset [0,1]^d$ for some parameter m with associated functions u_j that are compactly supported on A_j. Specifically assume that we have u_j satisfying:

$$\mathrm{supp}(u_j) \subset \{x | B(x, \epsilon \subset A_j\}, \|u_j\|_2^2 = \Omega(m^{-1}), \int_{A_j} u_j = \int_{A_j} p_0(x) u_j(x) = \int_{A_j} q_0(x) u_j(x) = 0, \|D^r u_j\|_1 \asymp m^{r/d-1}$$

The first condition ensure that the u_js are orthogonal to each other, while the second and third will ensure separation in terms of L_2^2 divergence. The last condition holds for all derivative operators with $r \leq \beta$ and it will ensure that the densities we construct belong to the bounded variation class. The only difference between these requirements and those from [9] are the orthogonality to p, q, and the bounded-variation condition, which replaces a point-wise analog.

Deferring the question of existence of these functions, we can proceed to construct p_λ. Let the index set $\Lambda = \{-1, +1\}^m$ and define the functions $p_\lambda = p_0 + K \sum_{j=1}^m \lambda_j u_j$, where K will be defined subsequently. A simple computation then reveals that:

$$\begin{aligned} T(p_0, q_0) - T(p_\lambda, q_0) &= \int p_0^2 - p_\lambda^2 + 2\left[\int p_\lambda q_0 - \int p_0 q_0\right] \\ &= \int (p_0 - p_\lambda)(p_0 + p_\lambda) + 2\left[\int p_\lambda q_0 - \int p_0 q_0\right] \\ &= K^2 \sum_{j=1}^m \|u_j\|_2^2 = \Theta(K^2) \end{aligned}$$

where we expand p_λ and use the orthogonality properties extensively. This gives us the desired separation.

To bound the hellinger distance, we use Theorem 1 of Birge and Massart [3] and the argument following Theorem 12 of Krishnamurthy et al [9].

Theorem 13. *[3] Consider a set of densities p_0 and $p_\lambda = p[1 + \sum_j \lambda_j v_j(x)]$ for $\lambda \in \Lambda = \{-1, 1\}^m$ with partition $A_1, \ldots, A_m \subset [0,1]^d$. Suppose that (i) $\|v_j\|_\infty \leq 1$, (ii) $\|\mathbf{1}_{A_j^C} v_j\|_1 = 0$, (iii) $\int v_j p_0 = 0$ and (iv) $\int v_j^2 p_0 = \alpha_j > 0$ all hold with:*

$$\alpha = \sup_j \|v_j\|_\infty, s = n\alpha^2 \sup_j P_0(A_j), c = n \sup_j \alpha_j$$

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

$$h^2(P_0^n, \overline{P^n}) \leq C(\alpha, s, c) n^2 \sum_{j=1}^m \alpha_j^2 \tag{26}$$

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

The exact same bound on the hellinger distances holds for the measures $P_0^n \times Q_0^n$ against $\overline{P^n} \times Q^n$. Defining $v_j = K u_j / p_0$ then the densities we used in our construction meet the specification in the above theorem. We immediately satisfy the first three requirements and we have $\int v_j^2 p = K^2 \int u_j^2 / p \leq K^2 \kappa_l / m \triangleq \alpha_j$. Thus we have the hellinger bound of:

$$h^2(P_0^n \times Q_0^n, \overline{P^n} \times Q^n) \leq (1/3) n^2 \sum_{j=1}^m \alpha_j^2 \leq \frac{C n^2 K^4}{m}$$

We lastly have to make sure that the p_λ functions satisfy the bounded variation assumption. This follows from an application of the triangle inequality provided that $\|D^r u_j\|_1 \leq O(m^{r/d-1})$.

$$\|D^r p_\lambda\|_1 = \|D^r p + K \sum_{j=1}^m \lambda_j D^r u_j\|_1 \leq \|D^r p\| + K \sum_{j=1}^m \|D^r u_j\|_1 \leq \|D^r p\| + K \sum_{j=1}^m \|D^r u_j\|_1 \leq \|D^r p\| + O(K m^{r/d})$$

So as long as $K \asymp m^{-r/d}$ and there is some wiggle room around the bounded variation assumption for p, p_λ will meet the bounded variation assumption.

Before we construct the u_js, we put everything together. We must select $K \asymp m^{-\beta/d}$ so that $p_\lambda \in \mathcal{W}_1^\beta(C)$, and then to make the hellinger distance $O(1)$, we must set $m \asymp n^{\frac{2d}{4\beta+d}}$. This makes $K^2 \asymp n^{\frac{-4\beta}{4\beta+d}}$ which is precisely the lower bound on the convergence rate in absolute error.

Lastly we present the construction of the u_j functions. The construction is identical to the one used by Krishnamurthy et al [9], but we must make some modifications to ensure that bounded variation condition is satisfied. We reproduce the details here for completeness.

Let $\{\phi_j\}_{j=1}^q$ be an orthonormal collection of functions for $L^2([0,1]^d)$ with $q \geq 4$. We can choose ϕ_j to satisfy (i) $\phi_1 = 1$, (ii) $\phi_j(x) = 0$ for $x | B(x,\epsilon) \not\subset [0,1]^d$ and (iii) $\|D^r \phi_j\|_\infty \leq \kappa < \infty$ for all j. Certainly we can find such an orthonormal system.

Now for any pair of function $f, g \in L^2([0,1]^d)$, we can find a unit-normed function in $\tilde{w} \in \mathrm{span}(\phi_j)$ such that $\tilde{w} \perp \phi_1, \tilde{w} \perp f, \tilde{w} \perp g$. If we write $\tilde{w} = \sum_j c_j \phi_j$, we have $D^r \tilde{w} = \sum_j c_i D^r \phi_j$ so that $\|D^r \tilde{w}\|_\infty \leq \kappa \sum |c_i| \leq \kappa \sqrt{q}$ since $\tilde{w}$ is unit normed. Thus the vector $w = \tilde{w}/(K\sqrt{q})$ has ℓ_2 norm equal to $(K\sqrt{q})^{-1}$ while have $\|D^r w\|_\infty \leq 1$ for all tuples r.

For the u_j functions, we use the partition $A_j = \prod_{i=1}^d [j_i m^{-1/d}, (j_i+1)m^{-1/d}]$ where $j = (j_1, \ldots, j_d)$ and $j_i \in [m^{1/d}]$ for each i. Map A_j to $[0,1]^d$ and appropriately map the densities p, q from A_j to $[0,1]^d$. We construct u_j by using the construction for w above on the segment of the density corresponding to A_j. In particular, let w_j be the function from above and let $u_j = w_j(m^{1/d}(x - (j_1, \ldots, j_d)))$. With this rescaling and shift, $u_j \in A_j$, $\mathrm{supp}(u_j) \subset \{x | B(x,\epsilon) \in A_j\}$, and $\int u_j^2(x) = m^{-1} \int w_j^2(x) = \Theta(1/m)$. For the last property, by a change of variables and Hólder's inequality, we have:

$$\|D^r u_j\|_1 = \int |D^r w_j(m^{1/d}(x - (j_1, \ldots, j_d)))| d\mu(x) = \frac{1}{m} \int \|m^{r/d} D^r w_j(y)\| dA_j(y) \leq m^{r/d-1}.$$

Thus these function u_j meet all of the requirements.

D Proof of Lemma 8

Recall that the asymptotic variance of the estimator is:

$$\sigma^2 = 4\left(\underset{X\sim p}{\mathrm{Var}}(p(X)) + \underset{Y\sim q}{\mathrm{Var}}(q(Y)) + \underset{X\sim p}{\mathrm{Var}}(q(X)) + \underset{Y\sim q}{\mathrm{Var}}(p(X)\right),$$

and our estimator $\hat{\sigma}^2$ is formed by simply plugging in kernel density estimates $\hat{p}, \hat{q}$ for all occurences of the densities. We will first bound:

$$\mathbb{E}_{X_1^n, Y_1^n}\left[|\sigma^2 - \hat{\sigma}^2|\right] = O(n^{\frac{-\beta}{2\beta+d}}),$$

and our high probability bound will follow from Markov's inequality. We will show the following bounds, and the expected ℓ_1 bound will follow by application of the triangle inequality. Below, let $f, g \in \mathcal{W}_1^\beta(C)$ be any two densities; we will interchangeably substitute p, q for f, g.

$$\mathbb{E}\left[\left|\int \hat{f}^3 - \int f^3\right|\right] \leq O\left(h^\beta + \frac{1}{(nh^d)^{1/2}}\right) \tag{27}$$

$$\mathbb{E}\left[\left|\left(\int \hat{f}^2\right)^2 - \left(\int f^2\right)^2\right|\right] \leq O\left(h^{2\beta} + \frac{1}{\sqrt{n}} + \frac{1}{nh^{d/2}}\right) \tag{28}$$

$$\mathbb{E}\left[\left|\int \hat{f}^2\hat{g} - \int f^2 g\right|\right] \leq O\left(h^\beta + \frac{1}{\sqrt{nh^d}}\right) \tag{29}$$

$$\mathbb{E}\left[\left|\left(\int \hat{f}\hat{g}\right)^2 - \left(\int fg\right)^2\right|\right] \leq O\left(h^{2\beta} + \frac{1}{\sqrt{n}} + \frac{1}{nh^{d/2}}\right) \tag{30}$$

Before establishing the above inequalities, let us conclude the proof. The overall rate of convergence in absolute loss is $O(h^\beta + \frac{1}{\sqrt{nh^d}})$. TBy choosing $h \asymp n^{\frac{-1}{2\beta+d}}$, the rate of convergence is $O(n^{\frac{-\beta}{2\beta+d}})$. Finally we wrap up with an application of Markov's Inequality.

Now we turn to establishing the bounds. For Equation 27, we can write:

$$\begin{aligned}\mathbb{E}\left[\left|\int \hat{f}^3 - \int f^3\right|\right] &\leq \mathbb{E}\|\hat{f} - f\|_3^3 + 3\mathbb{E}\left[\int |f(x)\hat{f}(x)(f(x) - \hat{f}(x))| d\mu(x)\right] \\ &\leq \mathbb{E}\|f - \hat{f}\|_3^3 + 3\mathbb{E}\|f - \hat{f}\|_\infty \|f\hat{f}\|_1 \\ &\leq O\left(h^{3\beta} + \frac{1}{(nh^d)^{3/2}} + h^\beta + \frac{1}{(nh^d)^{1/2}}\right).\end{aligned}$$

The first step is a fairly straightforward expansion followed by the triangle inequality while in the second step we apply Hölder's inequality. The last step follows from well known analysis on the rate of convergence of the kernel density estimator.

For Equation 28 we should actually use the U-statistic estimator for θ_p that we have been analyzing all along. The bound above follows from Theorem 10 and the following chain of inequalities:

$$\begin{aligned}\mathbb{E}\left[\left|\left(\int \hat{f}^2\right)^2 - \left(\int f^2\right)^2\right|\right] &\leq \mathbb{E}\left[\left(\int \hat{f}^2 - f^2\right)^2\right] + 2\|f\|_2^2 \mathbb{E}\left[\left|\int \hat{f}^2 - f^2\right|\right] \\ &\leq \mathbb{E}\left[\left(\int \hat{f}^2 - f^2\right)^2\right] + C\sqrt{\mathbb{E}\left[\left(\int \hat{f}^2 - f^2\right)^2\right]} \\ &\leq O\left(h^{4\beta} + \frac{1}{n} + \frac{1}{n^2h^d} + h^{2\beta} + \frac{1}{\sqrt{n}} + \frac{1}{nh^{d/2}}\right).\end{aligned}$$

The first inequality is a result of some simple manipulations followed by the triangle inequality and the second step is Jensen's inequality. We already have a bound on the MSE of the estimator $\hat{\theta}_p - \theta_p$ which gives us the inequality in Equation 28. Applying that bound leads to the last inequality.

The bound for Equation 30 follows from exactly the same argument with an application Theorem 11 instead of Theorem 10 in the last step. So we simply need to establish Equation 29.

$$\begin{aligned}\mathbb{E}\left[\left|\int \hat{f}^2\hat{g} - \int f^2 g\right|\right] &= \mathbb{E}\left[\left|\int (\hat{f}^2 - f^2)\hat{g}\right|\right] + \mathbb{E}\left[\left|\int f^2(\hat{g} - g)\right|\right] \\ &\leq \mathbb{E}\|\hat{f}^2 - f^2\|_2 \|\hat{g}\|_2 + \|f^2\|_2 \|\hat{g} - g\|_2 \\ &\leq \mathbb{E}\|\hat{f}^2 - f^2\|_2 (\|\hat{g} - g\|_2 + \|g\|_2) + \|f^2\|_2 \|\hat{g} - g\|_2 \\ &\leq O\left(h^{2\beta} + \frac{1}{nh^{d/2}} + \frac{1}{\sqrt{n}} + h^\beta + \frac{1}{\sqrt{nh^d}}\right).\end{aligned}$$

Here we use that $\|\hat{g}\|_1 = 1$ and that $\|f^2\|_2$ and $\|g\|_2$ are both bounded. We use the standard rate of convergence analysis of the kernel density estimator to bound $\mathbb{E}\|\hat{g} - g\|_2 \leq O(h^\beta + (nh^d)^{-1})$. We finally use Theorem 10 to bound $\|\hat{f}^2 - f^2\|_2$. Note that we are exploiting independence between the samples for $\hat{f}$ and $\hat{g}$ to push the expectation inside of the product in the first term. In the last line we omitted the term $\mathbb{E}\|\hat{f}^2 - f^2\|_2 \|\hat{g} - g\|_2$ since it converges much faster than the other two terms.

To prove the second bound, we show that $\bar{\sigma}^2$ is close to σ^2. We just have to look at two forms:

$$T_1 = \int \bar{p}^2(x)p(x) - \int p^3(x) \qquad T_2 = \left(\int \bar{p}(x)p(x)\right)^2 - \left(\int p^2(x)\right)^2.$$

For T_1 we can write:

$$T_1 = \int (\bar{p}^2(x) - p^2(x))p(x) = \int (\bar{p}(x) - p(x))(\bar{p}(x) - p(x) + 2p(x))p(x)$$

$$= \int (\bar{p}(x) - p(x))^2 p(x) + 2\int p^2(x)(\bar{p}(x) - p(x))$$
$$\leq \left(\sup_x |\bar{p}(x) - p(x)|\right)^2 + 2\|p\|_2^2 \sup_x |\bar{p}(x) - p(x)| \leq O(h^{2\beta} + h^\beta),$$

since p is L_2-integrable and the kernel density estimator has point-wise bias $O(h^\beta)$.

For T_2 we have:

$$T_2 = \left(\int (\bar{p}(x) - p(x))p(x)\right)^2 + 2\left(\int p^2(x)\right)^2 \left(\int (\bar{p}(x) - p(x))p(x)\right)$$
$$\leq \left(\sup_x |\bar{p}(x) - p(x)|\right)^2 + 2\|p\|_2^4 \sup_x \|\bar{p}(x) - p(x)\| \leq O(h^{2\beta} + h^\beta).$$

Wwith $h \asymp n^{\frac{-1}{2\beta+d}}$ the additional bias incurred is:

$$\mathbb{E}\left|\hat{\sigma}^2 - \bar{\sigma}^2\right| \leq \mathbb{E}\left|\hat{\sigma}^2 - \sigma^2\right| + \left|\sigma^2 - \bar{\sigma}^2\right| \leq O(n^{\frac{-\beta}{2\beta+d}}).$$

and so $\hat{\sigma}^2$ is an equally good estimator of σ^2 and $\bar{\sigma}^2$ (up to constants).

E A Convolution Lemma

In this section we show that bounded-variation smoothness is additive under convolution.

Lemma 14. *If $f, g \in \mathcal{W}_1^\beta(\mathbb{R}^d, C)$, then $h = f \star g \in \mathcal{W}_1^{2\beta}(\mathbb{R}^d, C^2)$.*

Proof. The proof uses the fact that:

$$\frac{\partial h(x)}{\partial x} = \left(\frac{\partial f}{\partial x} \star g\right)(x)$$

which follows by pushing the derivative operator inside of the integral and continuity of f, g and their derivatives. Using the above identity, we have:

$$\frac{\partial^{2\beta} h(x)}{\partial x^{2\beta}} = \left(\frac{\partial^\beta f}{\partial x^\beta} \star \frac{\partial^\beta g}{\partial x^\beta}\right)(x),$$

or more concisely:

$$\|h^{(2\beta)}\|_1 = \|f^{(\beta)} \star g^{(\beta)}\|_1 \leq \|f^{(\beta)}\|_1 \|g^{(\beta)}\|_1 \leq C^2.$$

The first inequality is Young's inequality. This implies that L_1 is closed under convolution.

It is clear, by the fact that derivatives can be distributed across the convolution that for $k < 2\beta$, $D^k h \in L^1$. This proof strategy extends mutatis mutandis to higher dimension. □