Harvard University

Harvard University Biostatistics Working Paper Series

Lepski's Method and Adaptive Estimation of Nonlinear Integral Functionals of Density

Rajarshi Mukherjee[*]   Eric J. Tchetgen Tchetgen[†]

James M. Robins[‡]

[*]Stanford University, rmukherj@stanford.edu

[†]Harvard University, etchetge@hsph.harvard.edu

[‡]Harvard University, robins@hsph.harvard.edu

LEPSKI'S METHOD AND ADAPTIVE ESTIMATION OF NONLINEAR INTEGRAL FUNCTIONALS OF DENSITY

By Rajarshi Mukherjee[*,??] Eric Tchetgen Tchetgen[†,??] and James Robins[‡,??]

We study the adaptive minimax estimation of non-linear integral functionals of a density and extend the results obtained for linear and quadratic functionals to general functionals. The typical rate optimal non-adaptive minimax estimators of "smooth" non-linear functionals are higher order U-statistics. Since Lepski's method requires tight control of tails of such estimators, we bypass such calculations by a modification of Lepski's method which is applicable in such situations. As a necessary ingredient, we also provide a method to control higher order moments of minimax estimator of cubic integral functionals. Following a standard constrained risk inequality method, we also show the optimality of our adaptation rates.

Introduction. Estimation of statistical functionals in nonparametric problems has received considerable attention over the last few decades. Of specific interest have been both linear and non-linear integral functionals of an underlying density. For example, a large body of work has focused on estimating the entropy of an underlying distribution. Beirlant et al. (1997) provides an overview of results and related techniques. More recent works include estimation of Rényi and Tsallis entropies (Leonenko and Seleznjev, 2010; Pál, Póczos and Szepesvári, 2010). For more references and examples one can refer to Kandasamy et al. (2014). We consider a framework for estimating such integral functionals of a density. In particular, suppose $X_1, \ldots, X_n$ are i.i.d on $[0,1]$ with density $f(x)$ with respect to Lebesgue measure. We take $f \in H(\beta, C)$ where $H(\beta, C)$ is a Hölder ball of smoothness β and radius $C > 0$. We are interested in estimation of $\phi(f)$ where $\phi : \mathcal{F} :\to \mathbb{R}$ is a non-linear functional of density and $\mathcal{F}$ refers to the class of all densities on $[0,1]$. It is well known (Birgé and Massart, 1995) that if $\phi(f) = \int T(f(x))d\mu(x)$, and T is sufficiently smooth, then the minimax rate of estimation over $f \in H(\beta, C)$ in squared error norm is $n^{-\frac{8\beta}{1+4\beta}}$ when $\beta < \frac{1}{4}$

---

[*]Stein Fellow, Department of Statistics, Stanford University
[†]Professor, Department of Biostatistics, Harvard University
[‡]Professor, Department of Epidemiology, Harvard University

1

and n^{-1} for $\beta \geq \frac{1}{4}$.

There exits extensive literature addressing minimax estimation of linear and quadratic functionals in density, white noise or nonparametric regression models. Although definitely not exhaustive, a comprehensive snapshot of this immense body of work can be found in Bickel and Ritov (1988); Cai and Low (2003, 2004, 2005); Donoho, Liu and MacGibbon (1990); Donoho and Nussbaum (1990); Fan (1991); Hall and Marron (1987); Kerkyacharian et al. (1996); Laurent et al. (1996) and other references therein. Most estimators proposed in the literature above, which attain the minimax rate of convergence over certain smoothness classes of the underlying function, depend explicitly on the knowledge of the smoothness index of the class. In particular, a standard technique in estimation of these functionals is expanding the infinite dimensional function of interest in an suitable orthonormal basis of $L_2[0,1]$ and estimate an approximate functional created by truncating the basis expansion at certain point. The point of truncation decides the approximation error of the truncated functional as an surrogate for the actual functional and depends on the smoothness of the function of interest and approximation properties of the orthonormal basis used. This point of truncation is then delicately balanced with the bias and variance of the resulting estimator and therefore directly depends on the smoothness of the function. Thus, it becomes of interest to understand the question of adaptive estimation i.e. the construction and analysis of estimators without prior knowledge of the smoothness.

The question of adaptation of linear and quadratic functionals has been studied in detail as well (Cai et al., 2005a,b; Cai and Low, 2006; Cai et al., 2006, 2008; Efromovich and Low, 1994; Efromovich et al., 1996b; Efromovich and Samarov, 2000; Giné and Nickl, 2008; Klemela and Tsybakov, 2001; Laurent and Massart, 2000; Low, 1992) and references therein. However, adaptive estimation of general non-linear functionals has not been addressed in complete generality. In this paper we address the problem of adaptive estimation of general smooth non-linear functionals of a density.

Robins et al. (2008) have developed a concrete theory for addressing minimax estimation of a class of non-linear functionals in certain non-parametric and semi-parametric problems under low regularity smoothness conditions. Following similar logic of construction Tchetgen et al. (2008) constructs a minimax estimator of $\int f^3 d\mu$ for $\beta < \frac{1}{4}$. The specific minimax estimator is a third order U-statistics and the construction of the kernel depends explicitly on the knowledge of the underlying smoothness. Based on the technique of Birgé and Massart (1995) one can show that the minimax estimator of a general non-linear functional $\phi(f) = \int T(f) d\mu$ for smooth T can be constructed

by using ideas from linear, quadratic and cubic functionals of the density and appealing to a standard Taylor expansion argument. While producing adaptive estimators of non-linear functionals as well, a similar strategy can be followed and it becomes crucial to understand the adaptive estimation of $\int f^3 d\mu$.

In our search for answers regarding $\int f^3 d\mu$, we start with looking at the general idea driving estimation of quadratic functionals of the density i.e. $\phi(f) = \int f^2 d\mu$. Giné and Nickl (2008) provide adaptive estimators of quadratic functional $\phi(f)$ where the estimators are based on certain types of second order U-statistics with specific kernels. Our results provide a richer class of estimators based on a compactly supported wavelet basis and is in line with the theory established by Efromovich et al. (1996b); Giné and Nickl (2008).

Unlike a quadratic functional, the minimax estimator for $\phi(f) = \int f^r d\mu$ is in general a rth order U-statistic. In order to apply regular Lepski's Method we will need to obtain suitable exponential deviation inequalities for higher order U-statistics. Although, such moment inequalities do exist (Adamczak et al., 2006), the bounds include complicated quantities which needs to be controlled in a problem specific manner. To bypass such complications, we employ a modification of Lepski's method where we test for the smoothness using our previously obtained second order U-statistics and use the selected smoothness to estimate the required functional.

In this paper we focus on adaptation over the non$-\sqrt{n}$ regime i.e. $\beta < \frac{1}{4}$; although our proofs carries over easily for $-\sqrt{n}$ range i.e. $\beta \geq \frac{1}{4}$ as well where adaptation is possible without paying a price and it is possible to achieve asymptotic efficiency for $\beta > \frac{1}{4}$. Moreover, the case of higher dimensions of X i.e. $d > 1$ can also be achieved by arguments similar to those in this paper. A brief discussion of these possible extensions is given in Section 8.

The main contributions of this paper are as followed. This work extends previous results obtained for linear and quadratic functionals to new adaptation theory for non-linear integral functionals of the density. In order to do so, we develop a suitable variant of Lepski's method which bypasses establishing exponential tail bounds for estimators of general non-linear integral functionals. In applying this modified Lepski's method, the main challenge lies in obtaining suitable bounds on higher order moments of suitable estimators of general non-linear integral functionals. This requires control of moments of higher order U-statistics based on orthogonal projection kernels. We crucially use the structure of the projection kernels based on compactly supported wavelet basis. Following ideas from Robins et al. (2015), we implement a binning argument to keep track of membership of sampled obser-

vations in partitions of the sample space created by a particular resolution level of the wavelet expansion, and this turns out to be crucial in controlling the higher order moments of our estimator at the right level.

We would like to mention that our proofs for analyzing adaptive estimator of general nonlinear integral functional of the density uses projection kernels based on Haar Basis expansion. We provide proof for general compactly supported wavelet based procedure when analyzing the quadratic functional. However, for analysis regarding general nonlinear integral functional of the density, the requirement of Haar Basis arises at one specific instance in the proof. We expect that similar results continue to hold for more general compactly supported wavelets using arguments developed herein. However we do not further consider other wavelet bases here.

The paper is organized as follows. In Section 1, we discuss Lepski's method the main idea of the paper. In Section 2, we introduce basic notations, terminology and some theory about compactly supported wavelets we will need in the sequel. In Section 3, we study the estimation of the quadratic functional and provide a general class of adaptive estimators when the kernels of the U-statistics are based on a class of compactly supported wavelet bases. Section 4 is devoted to the development of a modified Lepski's method suitable for analyzing higher order moments of the density. We then use this results to develop an adaptive estimator of $\int f^3 d\mu$ in section 5. Section 6 is used to understand how construction of an adaptive estimator of a general nonlinear functional $\int T(f)d\mu$ for smooth T can be derived from our analyses in Section 5. A lower bound on the required price of adaptation is provided in Section 7. Section 8 contains some discussions and future work. Finally all proofs are collected in the Appendix.

1. Lepski's Method and Heuristics of the Main Idea.

The purpose of this section is to heuristically explain the main idea behind modifying Lepski's Method suitably in our context. We do this at level of abstraction and do not provide formal results. Sections 4 and 5 are devoted to making the heuristics of this section more precise. We begin with a discussion of Lepski's Method as a recipe for producing adaptive estimators from a sequence of candidate estimators. Our discussion is inspired by the wonderful article of Birgé (2001). As astutely observed by Birgé (2001), Lepski's Method can be succinctly described in a relatively abstract set up as follows.

Using notations in essence similar to Birgé (2001), consider a family of experiment spaces $\{(\chi, \mathcal{B}(\chi), P), P \in S_\theta\}_{\theta \in \Theta}$ for measurable space χ, sigma field $\mathcal{B}(\chi)$ and probability measure P in one of the parameter sets $\{S_\theta\}_{\theta \in \Theta}$.

With an i.i.d sample of size n from such an experiment, we consider uniform rates of convergence for estimation of certain objects of interest over these parameter sets. In particular, suppose $s : P \to \mathcal{Y}$ is an object of interest for the experiment of interest with $\mathcal{Y}$ being a pseudo-metric space equipped with pseudo-metric $d(\cdot, \cdot)$. For a given estimator $\hat{s}$ based on sample of size n of a required object of interest s, define its rate of convergence over S_θ as $r(\theta, \hat{s}) = \sup_{s \in S_\theta} \mathbb{E}_0 (d^q(s, \hat{s}))$, where $q \geq 1$. An estimator $\hat{s}$ is often called (minimax) rate optimal over S_θ if $r(\theta, \hat{s}) \asymp r(\theta) := \inf_{\tilde{s}} r(\theta, \tilde{s})$.

Now, starting from a family of rate optimal estimators $\{\hat{s}_\theta\}_{\theta \in \Theta}$, Lepski's Method provides a strategy for building a new estimator $\hat{\theta}$ which has "good" performance simultaneously over all sets $S_\theta, \theta \in \Theta$. Working under the regime of estimating a whole function in the context of nonparametric regression, Lepskii (1991, 1992, 1993) develop such a strategy. Later these methods have been further used by Efromovich and Low (1994); Efromovich et al. (1996b); Giné and Nickl (2008); Klemela and Tsybakov (2001) and others for studying adaptation theory of linear and quadratic functionals in nonparametric problems. The method can be summarized roughly in brief as follows.

Suppose that $\Theta \subset \mathbb{R}$ is a bounded subset such that S_θ is non-decreasing with respect to θ, the risks and minimax rates $r(\theta, \hat{s}_\theta), r(\theta)$ are continuous with respect to θ, and for each $\theta \in \Theta$, $\exists$ a rate optimal estimator $\hat{s}_\theta$ where for large enough n, $d^q(s, \hat{s}_\theta)$ is suitably concentrated around its expectation. One then chooses, for each n, a suitable fine discretization $\theta_1 < \ldots < \theta_{K(n)}$ of Θ and finally for some large enough constant C define the candidate estimator to be $\hat{s}_{\theta_{\hat{j}}}$ where

$$\hat{j} := \inf \left\{ j \leq K(n) : d^q(\hat{s}_{\theta_j}, \hat{s}_{\theta_l}) \leq Cr(\theta_l, \hat{s}_{\theta_l}), \ \forall l \in (j, K(n)) \right\}.$$

Let us try to intuitively elaborate on the method described above. In its heart, the definition of $\hat{j}$ is devoted to choosing the "best" θ from a a point of view of the risk of the estimator of interest. Often, the "best" θ is the one corresponding to the unknown data generating mechanism and if $\hat{j}$ selects a value "close enough" to this θ with high probability, then owing to the continuity property of the risk, one obtains desired adaptive performance of the final estimator $\hat{s}_{\theta_{\hat{j}}}$. The required high probability selection of the desired θ is driven by the concentration of $d^q(s, \hat{s}_\theta)$ around its expectation.

Consider now a situation where suitable concentration of $d^q(s, \hat{s}_\theta)$ around its expectation is not easily achieved. This is often the case when the sequence of estimators $\{\hat{s}_\theta\}_{\theta \in \Theta}$ is not sufficiently "nice" for application of standard concentration inequalities. In such cases there can be two possible ways out. The first entails obtaining a sufficiently sharp concentration

inequality for the sequence of estimators $\{\hat{s}_\theta\}_{\theta \in \Theta}$ at hand. This is often an important and difficult question in its own right, even outside the context of the problem of adaptive estimation at hand. In certain cases however, an alternative strategy might be available. This paper pursues the second approach in the context of estimating a non-linear integral functional of a density.

At a high level of abstraction, our method can be described as follows. As mentioned earlier, the main challenge in applying Lepski's Method is obtaining a suitably sharp concentration of $d^q(s, \hat{s}_\theta)$ around its expectation. However, suppose that there exists another sequence of estimators of a different object of interest $\widetilde{s} : P \to \widetilde{\mathcal{Y}}$ with $\widetilde{\mathcal{Y}}$ being a pseudo-metric space equipped with pseudo-metric $\widetilde{d}(\cdot, \cdot)$ and corresponding sequence of estimators $\widetilde{\hat{s}}_\theta$ with concentration of $\widetilde{d}^q(\widetilde{s}, \widetilde{\hat{s}}_\theta)$ around its expectation and a further property that $\widetilde{d}^q(\widetilde{\hat{s}}_\theta, \widetilde{\hat{s}}_{\theta'})$ and $\widetilde{r}(\theta, \widetilde{\hat{s}})$ equals $d^q(\hat{s}_\theta, \hat{s}_{\theta'})$ and $r(\theta, \hat{s})$ up to multiplicative constants uniformly in $(\theta, \theta') \in \Theta \times \Theta$. Our idea is to define

$$\widetilde{\hat{j}} := \inf \left\{ j \le K(n) : \widetilde{d}^q(\widetilde{\hat{s}}_{\theta_j}, \widetilde{\hat{s}}_{\theta_l}) \le C\widetilde{r}(\theta_l, \widetilde{\hat{s}}_{\theta_l}), \ \forall l \in (j, K(n)) \right\},$$

and using $\hat{s}_{\theta_{\hat{j}}}$ as our candidate estimator. Under some conditions on the actual sequence of estimators $\hat{s}_\theta$ one then obtains desired adaptation properties of the final estimator $\hat{s}_{\theta_{\hat{j}}}$. These conditions on $\hat{s}_\theta$ needs to be less demanding than sharp concentration of $d^q(s, \hat{s}_\theta)$ around its expectation, and this turns out to be the case in our context.

To fix ideas, for our case the experiment space will be $(\chi, \mathcal{B}(\chi), P : P \in S_\theta)_\theta = ([0, 1], \mathcal{B}([0, 1]), P = f d\mu : f \in H(\beta, C))_{\beta \in (0, \frac{1}{4})}$; i.e. parameter spaces of interest are $H(\beta, C)$ where β is identified with θ in the discussion above and the object of interest is $s = \int T(f) d\mu$ based on i.i.d. observations $X_1, \ldots, X_n$ from density $f \in H(\beta, C)$. As discussed earlier, the case of quadratic functionals., i.e. $T(x) = x^2$ has been studied in detail in the literature. In particular, one has suitable concentration of $\widetilde{d}^2(\int f^2 d\mu, \widetilde{\hat{s}}_\beta)$ around its expectation, where $\widetilde{d}(z_1, z_2) = |z_1 - z_2|$ for $z_1, z_2 \in \mathbb{R}$ and $\widetilde{\hat{s}}_\beta$ is rate optimal estimator of the quadratic functional. Such an estimator is often a second order U-statistic, for which the desired concentration type results can be obtained in reference to Giné, Latała and Zinn (2000); Houdré and Reynaud-Bouret (2003). However, for estimating $s = \int f^3 d\mu$, which is crucial for understanding the general smooth non-linear functional problem, the estimator sequence $\{s_\beta\}$ is a third order U-statistic. In general, for estimating $s = \int f^r d\mu$, the estimator sequence $\{s_\beta\}$ is a rth order U-statistic. In order to apply the standard Lepski's Method we will need to obtain suitable exponential deviation inequalities for higher order U-statistics. Although,

such moment inequalities exist (Adamczak et al., 2006), the bounds include complicated quantities that must be controlled in a problem specific manner. However, owing to Tchetgen et al. (2008), we observe that the biases of the estimator sequence $\{s_\beta\}$ are within multiplicative constants of that of the estimator for quadratic functional i.e. $\{\widetilde{s}_\beta\}$ and their variances remain of same order as well. Since we are interested in estimation in squared error norm, this motivates us to use $\widetilde{\widehat{j}}$ constructed using the estimator for quadratic functional and analyzing $s_{\beta_{\widetilde{\widehat{j}}}}$ as our final estimator.

In Sections 3 and 4 below, we elaborate on the ideas laid down above. In particular, we provide a concrete example of the application of standard Lepski's Method while estimating quadratic functional of the density. Subsequently, we employ the modification of Lepski's method as discussed above to study adaptive estimation of more general non-linear integral functionals of the density.

2. Notation. It is well known that without imposing restrictions on function classes, consistent inference in non-parametric problems is not feasible. Our results are stated in terms of certain regularity conditions on the class of densities. We will place the following kind of bounds on their roughness or complexity.

DEFINITION 2.1. *A function $h(\cdot)$ with domain $[0,1]^d$ is said to belong to a Hölder ball $H(\beta, C)$, with Hölder exponent $\beta > 0$ and radius $C > 0$, if and only if $h(\cdot)$ is uniformly bounded by C, all partial derivatives of $h(\cdot)$ up to order $\lfloor \beta \rfloor$ exist and are bounded, and all partial derivatives $\nabla^{\lfloor \beta \rfloor}$ of order $\lfloor \beta \rfloor$ satisfy*

$$\sup_{x, x+\delta x \in [0,1]^d} \left| \nabla^{\lfloor \beta \rfloor} h(x + \delta x) - \nabla^{\lfloor \beta \rfloor} h(x) \right| \leq C ||\delta x||^{\beta - \lfloor \beta \rfloor}.$$

All our results are in terms of Hölder balls. However, they are easily extendable to Sobolev balls . With some abuse of notation we will denote by $H(\beta)$ the class of functions in $H(\beta, C)$ such that $\|f\|_\infty \leq M^*$ and $f_{\min} \geq M_*$ for known constants $M^* > M_* > 0$. Our results will be based on known C, M^*, M_* only to the extent that these quantities will determine multiplicative constants of convergence rates obtained throughout. Adaptation w.r.t. these parameters may be of interest, however beyond the scope of this paper.

A crucial ingredient for constructing our estimators is the use of orthogonal projection kernels onto increasing finite dimensional subspaces of $L_2[0,1]$.

The construction of such kernels is based on a suitable choice of an orthonormal basis of $L_2[0,1]$. With this in mind we provide a brief discussion on orthogonal projection kernels and compactly supported wavelet bases.

The kernels (possibly dependent on n) are assumed to be measurable maps $K_n : [0,1] \times [0,1] \to \mathbb{R}$ that are symmetric in their arguments and satisfy $\int \int K_n^2(x_1, x_2) dx_1 dx_2 < \infty$ for all n. The corresponding kernel operator (which we denote by the same symbol with an abuse of notation)

$$K_n h(x) = \int h(v) K_n(x, v) dv$$

are continuous linear operators $K_n : L_2[0,1] \to L_2[0,1]$. Throughout we will work with kernels whose operator norms $\|K_n\| = \sup\{\|K_n f\|_2 : \|f\|_2 \le 1\}$ are uniformly bounded, i.e., $\sup_n \|K_n\| < \infty$. The operator norm $\|K_n\|$ is however typically much smaller than the $L_2[0,1] \times L_2[0,1]$ norm of the kernel. In our case this will typically be of the order of k given by the dimension of the projected space.

Kernels K_n most commonly used in statistical applications are usually projection kernels. A projection kernel operator satisfies $K_n h = h$ for all h in its range space; that is for any function f in the range of the kernel, $f(x) = \int f(v) K_n(x, v) dv$ for a.e. x. For a given orthonormal basis $e_1, e_2, \ldots$ of $L_2[0,1]$, the orthogonal projection onto $lin(e_1, \ldots, e_k)$ is the map $K_k : f \to \sum_{j=1}^k \langle f, e_j \rangle e_j$. It is given by the kernel operator $K_k : L_2[0,1] \to L_2[0,1]$ with kernel

$$K_k(x_1, x_2) = \sum_{j=1}^k e_j(x_1) e_j(x_2).$$

It is easy to show by orthonormality properties that $\|K_k\| = 1$ and $\int \int K_k^2 = k$.

We now provide examples of orthogonal projection kernels that will be used extensively throughout.

2.1. *Haar Basis.* For "father" and "mother" functions $\phi(x) = \mathcal{I}(0,1](x)$ and $\psi(x) = \mathcal{I}(0,1/2](x) - \mathcal{I}(1/2,1](x)$, the Haar basis is the set of functions $\{\phi, \psi_{i,j} : i = 0, 1, \ldots, j = 0, 1, \ldots, 2^i - 1\}$, for $\psi_{i,j}(x) = 2^{i/2}\psi(2^i x - j)$, $i = 0, 1, \ldots, j = 0, 1, \ldots, 2^i - 1$. The Haar basis is a complete orthonormal basis of $L_2[0,1]$. The linear span of the first $k = 2^I$ basis elements $\{\phi, \psi_{i,j} : i = 0, \ldots, I - 1, j = 0, \ldots, 2^i - 1\}$ is equal to the linear span of the scaled and shifted father functions $\{\psi_{I,j} : j = 0, \ldots, 2^I - 1\}$ given by the kernel

$$K_k(x_1, x_2) = \sum_{j=0}^{2^I - 1} \phi_{I,j}(x_1) \phi_{I,j}(x_2)$$

$$= k \sum_{j=1}^{k} I\left(x_1 \in \left(\frac{j-1}{k}, \frac{j}{k}\right]\right) I\left(x_2 \in \left(\frac{j-1}{k}, \frac{j}{k}\right]\right).$$

2.2. *Wavelets.* Consider expansions of functions $h \in L_2(\mathbb{R})$ on an orthonormal basis of compactly supported bounded wavelets of the form

$$h(x) = \sum_{j\in\mathbb{Z}^d} \sum_{v\in\{0,1\}^d} \langle h, \psi_{0,j}^v \rangle \psi_{0,j}^v(x) + \sum_{i=0}^{\infty} \sum_{j\in\mathbb{Z}^d} \sum_{v\in\{0,1\}^d-\{0\}} \langle h, \psi_{i,j}^v \rangle \psi_{i,j}^v(x),$$

(2.1)

where the base functions $\psi_{i,j}^v$ are orthogonal for different indices (i,j,v) and are scaled and translated versions of the 2^d base functions $\psi_{0,0}^v$, i.e., $\psi_{i,j}^v(x) = 2^{id/2}\psi_{0,0}^v(2^i x - j)$. Such a wavelet basis can be obtained as tensor products $\psi_{0,0}^v = \phi^{v_1} \times \ldots \times \phi^{v_d}$ of a given father wavelet ϕ^0 and mother wavelet ϕ^1 in one dimension.

We are interested in functions f with support $[0,1]$. In view of the compact support of the wavelets, for each resolution level i and index v, only 2^i base elements $\psi_{i,j}^v$ are non-zero on $[0,1]$; let us denote the corresponding set of indices j by j_i. Truncating the expansion at resolution level $i = I$ then gives an orthogonal projection on a subspace of dimension k of order 2^I. Let

$$K_k(x_1, x_2) = \sum_{j\in J_0} \sum_{v\in\{0,1\}} \psi_{0,j}^v(x_1)\psi_{0,j}^v(x_2) + \sum_{i=0}^{I} \sum_{j\in J_i} \psi_{i,j}^1(x_1)\psi_{i,j}^1(x_2).$$

It is worth noting that we can re-express the wavelet expansion (2.1) to start from a level I as

$$h(x) = \sum_{j\in\mathbb{Z}} \sum_{v\in\{0,1\}} \langle h, \psi_{I,j}^v \rangle \psi_{I,j}^v(x) + \sum_{i=I+1}^{\infty} \sum_{j\in\mathbb{Z}} \langle h, \psi_{i,j}^1 \rangle \psi_{i,j}^1(x).$$

The projection kernel K_k sets the coefficients in the second sum equal to zero and hence can also be expressed as

$$K_k(x_1, x_2) = \sum_{j\in J_I} \sum_{v\in\{0,1\}} \psi_{I,j}^v(x_1)\psi_{I,j}^v(x_2).$$

Owing to the discussion above, with some abuse of notation, we will work with the definition of projection kernel as

$$K_k(x_1, x_2) = \sum_{j=1}^{k} \psi_{k,j}(x_1)\psi_{k,j}(x_2),$$

which will correspond to the orthogonal projection operator onto the first $O(\log_2 k)$ wavelet basis. Owing to compactness of the support of wavelet bases only a fixed $O(k)$ many $\psi_{k,j}$'s will intersect $[0,1]$. We have then conveniently renamed them as $1, \ldots, k$. This simplifies notation without loss of generality. In particular, this is exact for Haar kernel since as previously noted, we then have that

$$K_k(x,y) = \sum_{l=1}^{k} \psi_{k,j}(x)\psi_{k,j}(y)$$

where $\psi_{k,j}(x) = \sqrt{k}\mathcal{I}(x \in (\frac{j-1}{k}, \frac{j}{k})]$, $j = 1, \ldots, k$ is the $\log_2 k$ level Haar basis of $L_2[0,1]$. Also, for larger expressions involved in some proofs, we often write $\psi_{k,l} = \psi_l^k$ for convenience.

With these conventions, let us now discuss some approximation properties of these projection kernels. Letting $\overline{\psi} := lin\{\psi_{k,j} : j = 1, \ldots, k\}$, define

$$f_k(x) := \Pi_{\text{Lebesgue}}\left(f|\overline{\psi}\right) := \int f(y)K_k(x,y)dy.$$

For convenience of notation, define

$$\alpha_{k,j} = \int \psi_{k,j} f.$$

It is well known that (Härdle et al., 1998), choosing $\psi_{k,l}$'s to be the $\log_2 k$ level compactly supported wavelet basis with suitable vanishing moment conditions on the mother wavelet, one has that

$$\sup_{h \in H(\beta,C)} \|h - h_k\|_2 \lesssim k^{-\beta}.$$

The results in this paper are mostly asymptotic in nature and thus requires some standard asymptotic notations. If a_n and b_n are two sequences of real numbers then $a_n \gg b_n$ (and $a_n \ll b_n$) implies that $a_n/b_n \to \infty$ (and $a_n/b_n \to 0$) as $n \to \infty$, respectively. Similarly $a_n \gtrsim b_n$ (and $a_n \lesssim b_n$) implies that $\liminf a_n/b_n = C$ for some $C \in (0, \infty]$ (and $\limsup a_n/b_n = C$ for some $C \in [0, \infty)$). Alternatively, $a_n = o(b_n)$ will also imply $a_n \ll b_n$ and $a_n = O(b_n)$ will imply that $\limsup a_n/b_n = C$ for some $C \in [0, \infty)$). Finally we comment briefly on the various constants appearing throughout the text and proofs. Given that our primary results concern convergence rates of various estimators, we will not emphasize the role of constants throughout and rely on fairly generic notation throughout for such constants. Some conventions we follow whenever required is a follows. Throughout the paper

we denote by $C(\psi_0, \|f\|_\infty)$ a non-negative constant that depends on a fixed ψ_0 (a function that can be taken as a majorant of both father and mother wavelets in absolute value) and $\|f\|_\infty$. Often, C_{ψ_0} will denote a number which depends only on ψ_0. Hence if $\|f\|_\infty$ is known to us, such constants are deterministic and eventually can all be replaced by an universal constant by taking suitable care of all possible constants appearing in this paper. Finally, we will also use $C(\psi_0, \|f\|_\infty, f_{\min})$ to denote numbers which depends on $\psi_0, \|f\|_\infty, f_{\min}$ only.

3. Lepski's Method and Minimax Adaptive Estimation of $\phi(f) = \int f^2 d\mu$.

By way of introduction, this section we will provide a concrete example where Lepski's Method applies in its standard form. For the sake of simplicity, we will restrict ourselves to the adaptation over two parameter spaces indexed by two smoothness classes $\beta_1 < \beta_0 < \frac{1}{4}$. The more general case can be addressed from our results in Section 3. However, before going into further details, we need additional notation and a basic result which drives the main idea behind Lepski's Method. For our candidate sequence of estimators, define

$$U_n^{(k)} = \frac{1}{n(n-1)} \sum_{i_1 \neq i_2} K_k(X_{i_1}, X_{i_2}),$$

$$k(\beta) = \lceil n^{\frac{2}{1+4\beta}} \rceil,$$

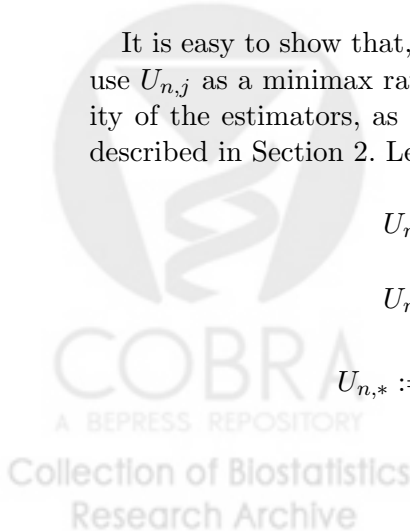$$U_{n,j} := U_n^{(k_j)}, \ k_j = k(\beta_j) := \lceil n^{\frac{2}{1+4\beta_j}} \rceil,$$

$$U_n^{(k_j^*)}, \ k_j^* = k^*(\beta_j) = \lceil \left(\frac{n^2}{\log n}\right)^{\frac{1}{1+4\beta_j}} \rceil.$$

It is easy to show that, knowing the exact smoothness β_j one can readily use $U_{n,j}$ as a minimax rate optimal estimator of $\phi(f)$. Asymptotic normality of the estimators, as stated by Lemma C.3, drives Lepski's Method as described in Section 2. Let,

$$U_{n,0} := U_n^{(k_0)}, \ k_0 = \lceil n^{\frac{2}{1+4\beta_0}} \rceil,$$

$$U_{n,1} := U_n^{(k_1)}, \ k_1 = \lceil n^{\frac{2}{1+4\beta_1}} \rceil,$$

$$U_{n,*} := U_n^{(k_*)}, \ k_* = \lceil \left(\frac{n^2}{\log n}\right)^{\frac{1}{1+4\beta_1}} \rceil.$$

The method we now describe, is along the lines of Efromovich et al. (1996b) suitably adapted to deal with our situation. Our goal is to choose a data dependent $\hat{k} \in \{k_0, k_1\}$ so that if we base our estimation on $U_n^{(\hat{k})}$, the mean squared error of estimation will be suitably controlled. In particular,

$$\sup_f E \left[U_{n,\hat{j}} - \psi(f) \right]^2$$

$$\leq \sum_{k=0}^{1} \sup_f E \left\{ \mathcal{I} \left\{ \hat{j} = k \right\} \left[U_{n,\hat{j}} - \psi(f) \right]^2 \right\}$$

$$\equiv \sup_f R_0(\beta_f) + \sup_f R_1(\beta_f)$$

We divide our study into two cases having two sub-cases each.

First, consider the case $\beta = \beta_0$. Then, $R_0(\beta_0)$ achieves the minimax rate since the truth is β_0 and we are choosing β_0. Next we need to bound $R_1(\beta_0)$ by $n^{-\frac{8\beta_0}{1+4\beta_0}}$ up to a potentially logarithmic factor. It is worth noting, that along the lines of Efromovich et al. (1996b), since β_0 corresponds to higher smoothness regime, we do not need to pay the logarithmic price. The term $R_1(\beta_0)$ corresponds to the scenario where the truth is β_0 but the procedure wrongly chooses β_1 as the underlying smoothness. If this happens too often, then the mean squared error will be sub-optimal. Therefore, our testing procedure should guarantee that type of error, i.e., choosing a lower smoothness when the truth is a higher smoothness, does not happen too often. Since, the probability of making such an error must reduce the mean squared error $n^{-\frac{8\beta_1}{1+4\beta_1}}$ down to $n^{-\frac{8\beta_0}{1+4\beta_0}}$ for any β_1, it must be of $O(\frac{1}{n})$. So the problem boils down to designing a selection procedure so that probability of selecting the lower smoothness is $O(\frac{1}{n})$. If one can find a sequence of statistics T_n such that T_n converges weakly to N(0,1) "fast enough" under β_0 and $|T_n|$ diverges polynomially in n under β_1, then a simple test for selecting the lower smoothness with required error rate is given by $\mathcal{I}(T_n > \sqrt{2 \log n})$. To see this, note that, the probability of selecting the lower smoothness under β_0 is approximately $Pr(|N(0,1)| > \sqrt{2 \log n}) = O(\frac{1}{n})$. Provided the approximation in the CLT is also of the order of $\frac{1}{n}$ we have achieved our goal for this first case.

Considering the complementary case of $\beta = \beta_1$, we will need the error probability of selecting the wrong smoothness $Pr(|T_n| \leq \sqrt{2 \log n})$ to converge to 0 "polynomially fast". As we establish below, this can be achieved with the right choice of test statistic. Finally, by Lemma C.3, a candidate

T_n which satisfies the above mentioned properties is given by

$$\frac{U_{n,*} - U_{n,0}}{\sqrt{Var_f(U_{n,*})}}.$$

We now provide the exact technical details that justify the argument laid down so far.

We start with a few further notations. Define

$$I(k_1, k_2) := \phi(f_{k_2}) - \phi(f_{k_1}) \ k_1 \leq k_2,$$

corresponding to the difference in truncation bias at k_1 versus k_2. With this notation, it is easy to see that $I(k_0, k_*) = O(k_0^{-2\beta_f})$ when the truth is β_f. A suitable estimator of $I(k_0, k_*)$ is indeed

$$\hat{I}(k_0, k_*) = U_{n,*} - U_{n,0}.$$

We then have the following theorem.

THEOREM 3.1. *For a deterministic constant C depending on the parent wavelets, we have the following result. Let*

$$\hat{j} = \mathcal{I}\left(\hat{I}(k_0, k_*) > C\sqrt{\log n}\frac{\sqrt{k_*}}{n}\right)$$

and let

$$\hat{\hat{\phi}} = U_{n,\hat{j}}.$$

Then

$$\sup_{f \in H(\beta_0)} \mathbb{E}_f\left[\hat{\phi} - \phi(f)\right]^2 \lesssim n^{-\frac{8\beta_0}{1+4\beta_0}},$$

$$\sup_{f \in H(\beta_1)} \mathbb{E}_f\left[\hat{\phi} - \phi(f)\right]^2 \lesssim (\log n)^{\frac{4\beta_1}{1+4\beta_1}} n^{-\frac{8\beta_1}{1+4\beta_1}}.$$

The proof can be found in the appendix. However, it is worth mentioning that the two crucial ingredients of the proof is control of the error of testing at a suitably vanishing level and bounding higher order central moments of the non-adaptive minimax estimators dependent on non-random truncation level k. These properties are both derived from exponential tail inequalities for second order U-statistics using results of Giné, Latała and Zinn (2000); Houdré and Reynaud-Bouret (2003). In particular, this is a typical structure for most applications of Lepski's method i.e. the treatment crucially relies on some exponential inequalities for the deviations of the sequence of non-adaptive minimax estimators (Birgé, 2001). For estimation of general non-linear functionals this poses a challenge. The next section is devoted in understanding a possible way out.

4. A General Result by a Modification of Lepski's Method.

The results of the previous section strictly hold only for the quadratic functional $\phi(f) = \int f^2 d\mu$ so far. This is because, we have crucially used a deviation inequality which only holds for degenerate second order U-statistics. The minimax estimator for $\phi(f) = \int f^r d\mu$ is in general a rth order U-statistic. In order to apply standard Lepski's Method we will need to obtain suitable exponential deviation inequalities for higher order U-statistics. Although, such moment inequalities do exist (Adamczak et al., 2006), the bounds include complicated quantities which needs to be controlled in a problem specific manner. However, owing to Tchetgen et al. (2008), we observe that the biases of candidate estimators are within multiplicative constants of that of the estimator for quadratic functional and the variances remain of same order as well. Since we are interested in estimation in squared error norm, this motivates us to use the idea discussed in Section 2 to construct our final estimator. We now explain this below.

For a given choice of $d > 1$, let N be the largest integer such that $d^{N-1} \le n^{1-\frac{2}{\log \log n}}$. Set $k_{j-1} = d^{j-1}n$ for $j = 1, \ldots, N$. This holds if and only if, $(N-1) \le \frac{\log n}{\log d}\left(1 - \frac{2}{\log \log n}\right)$ and also that $\frac{k_N}{n^2} = o(\frac{1}{n})$. Now, for $j = 0, \ldots, N-1$, define β_j to be the solution of $k_j = n^{\frac{2}{1+4\beta_j}}$. This implies that $1 + 4\beta_j = \frac{2\log n}{(j-1)\log d + \log n}$, $j = 0, \ldots, N-1$. Also, note that by construction, $k_0 \le k_1 \le \ldots \le k_{N-1}$ and therefore, $\beta_0 \ge \beta_1 \ge \ldots \ge \beta_{N-1}$. Also, a simple calculation shows that, $\beta_{N-1} \ge \frac{1}{4\log \log n - 1}$. Therefore, $\frac{1}{4} \ge \beta_0 \ge \beta_1 \ge \ldots \ge \beta_{N-1} \ge \frac{1}{4\log \log n - 1}$. Further, note that for all $0 \le l_1 \le l_2 \le N-1$, there exists constants c_1, c_2 such that $\beta_{l_1} - \beta_{l_2} \in \left[c_1 \frac{l_2 - l_1}{\log n}, c_2 \frac{l_2 - l_1}{\log n}\right]$. Further, define $k_j^* = \frac{k_j}{\log n^{\frac{1}{1+4\beta_j}}} = \left(\frac{n^2}{\log n}\right)^{\frac{1}{1+4\beta_j}}$ and $R(k_j^*) = \frac{k_j^*}{n^2}$. This definition implies that for $l_1 > l_2$, $\frac{k_{l_1}^*}{k_{l_2}^*} = \left(\frac{n^2}{\log n}\right)^{\frac{4\beta_{l_2} - 4\beta_{l_1}}{(1+4\beta_{l_1})(1+4\beta_{l_2})}} \to \infty$. However, note that $\frac{k_{l_1}^*}{k_{l_2}}$ might not enjoy similar properties for certain ranges of $l_1 > l_2$. Especially, for β values within $\frac{1}{\log n}$ rate of each other, the corresponding k values do not enjoy the above mentioned property. Let $s^* = s^*(n)$ be the smallest integer such that, $k(\beta_{s^*})^* \ge n$. Finally, define

$$\hat{j} := \min\left\{j : \hat{I}^2(k_j^*, k_l^*) \le C_{\text{opt}}^2 \log n R(k_l^*) \ \forall \ l \ge j, s^* \ge j \le N-1\right\}.$$

where C_{opt} is a deterministic constant to be specified later and similar to Section 3, $\hat{I}(k_j, k_l) = U_{k_l} - U_{k_j}$ for $k_j < k_l$.

THEOREM 4.1. *Suppose an estimator $U_{n,k}$ of $\phi(f)$ satisfies the following properties.*

1. *For sufficiently large k, n and fixed choice of $q > 1$,*

$$\max\left\{\sup_{f \in H(\beta)} \mathbb{E}_f^{\frac{1}{q}}\left[(U_{n,k} - \mathbb{E}_f(U_{n,k}))^{2q}\right], \sup_{f \in H(\beta)} |\mathbb{E}_f(U_{n,k}) - \phi(f)|^2\right\} \leq C_q \frac{k}{n^2}$$

2. *There exists fixed constants C_1, C_2 such that for any $f \in H(\beta, C)$,*

$$C_1 I(k_1, k_2) \leq I_c(k_1, k_2) \leq C_2 I(k_1, k_2),$$

for sufficiently large $k_1 < k_2$ where $I_c(k_1, k_2) = \mathbb{E}_f(U_{n,k_2}) - \mathbb{E}_f(U_{n,k_1})$ and by our previous convention $I(k_1, k_2) = \int f_{k_2}^2 - \int f_{k_1}^2$.

Then for any $\beta \in (0, \frac{1}{4})$, we have that

$$\sup_{f \in H(\beta, C)} \mathbb{E}_f\left[\left(U_{n,k_{\hat{j}}^*} - \phi(f)\right)^2\right] \lesssim n^{-\frac{8\beta}{1+4\beta}} \log n^{\frac{4\beta}{1+4\beta}}.$$

Some remarks are in order about the implications of the theorem above. It says that if a sequence of estimators, having common index with the minimax estimator of the quadratic functional, has bias and higher order moments of the same order of magnitude as the quadratic functional estimator, then following the discussion in Section 1, the new sequence of estimators also shares the same rate of convergence towards its corresponding functional simultaneously over the functional spaces which dictates the adaptation of the quadratic functional. Establishing an equivalence in the order of the bias for these functionals requires fairly standard arguments. The main challenge lies in obtaining suitable bounds on higher order moments of these estimators. This requires control of moments of higher order U-statistics based on orthogonal projection kernels. We crucially use the structure of the projection kernels based on compactly supported wavelet basis. Following ideas from Robins et al. (2015), we implement a binning argument to keep track of membership of sampled observations in partitions of the sample space created by a particular resolution level of the wavelet expansion, and this turns out to be crucial in controlling the higher order moments of our estimator at the right level.

5. Minimax Adaptive Estimation of $\int f^3 d\mu$.

As mentioned earlier, based on the technique of Birgé and Massart (1995) one can show that the minimax estimator of a general non-linear functional $\phi(f) = \int T(f) d\mu$ for smooth T can be constructed by using ideas from linear,

quadratic and cubic functionals of the density and appealing to a standard Taylor expansion argument. While producing adaptive estimators of non-linear functionals as well, a similar strategy can be followed and it becomes crucial to understand the adaptive estimation of $\int f^3 d\mu$. In particular, a non-adaptive minimax rate optimal estimator of $\phi(f) = \int f^3 d\mu$ is given by

$$
\begin{aligned}
\widehat{\phi}_{k_1,k_2,k_3} &= \mathbb{V}_n \left[\widetilde{K}_{k_1,k_2,k_3} \left(X_{i_1}, X_{i_2}, X_{i_3} \right) \right] \\
&= \mathbb{V}_n \left[\int K_{k_1} \left(x, X_{i_1} \right) K_{k_1} \left(x, X_{i_2} \right) K_{k_1} \left(x, X_{i_3} \right) dx \right] \\
&+ 3\mathbb{V}_n \left[\int K_{k_1} \left(x, X_{i_1} \right) \left(K_{k_3} \left(x, X_{i_2} \right) - K_{k_1} \left(x, X_{i_2} \right) \right) K_{k_1} \left(x, X_{i_3} \right) dx \right] \\
&+ 3\mathbb{V}_n \left[\int K_{k_1} \left(x, X_{i_1} \right) \left(K_{k_3} \left(x, X_{i_2} \right) - K_{k_1} \left(x, X_{i_2} \right) \right) \left(K_{k_3} \left(x, X_{i_3} \right) - K_{k_1} \left(x, X_{i_3} \right) \right) dx \right] \\
&+ 3\mathbb{V}_n \left[\int \left(\begin{array}{c} \left(K_{k_2} \left(x, X_{i_1} \right) - K_{k_1} \left(x, X_{i_1} \right) \right) \left(K_{k_2} \left(x, X_{i_2} \right) - K_{k_1} \left(x, X_{i_2} \right) \right) \\ \left(K_{k_3} \left(x, X_{i_3} \right) - K_{k_2} \left(x, X_{i_3} \right) \right) \end{array} \right) dx \right] \\
&+ \mathbb{V}_n \left[\int \left(\begin{array}{c} \left(K_{k_2} \left(x, X_{i_1} \right) - K_{k_1} \left(x, X_{i_1} \right) \right) \left(K_{k_2} \left(x, X_{i_2} \right) - K_{k_1} \left(x, X_{i_2} \right) \right) \\ \left(K_{k_2} \left(x, X_{i_3} \right) - K_{k_1} \left(x, X_{i_3} \right) \right) \end{array} \right) dx \right],
\end{aligned}
$$

where $k_1 = k_1(\beta) \sim n, n^{(3/2-2\beta)/(1+4\beta)} \leq k_2 = k_2(\beta) \leq n^{(3/2+2\beta)/(1+4\beta)}$ and $k_3 = k_3(\beta) \sim n^{2/(1+4\beta)}$ and $\mathbb{V}_n(h(X_{i_1}, X_{i_2}, X_{i_3}))$ corresponds to to the U-statistic based on h i.e. $\mathbb{V}_n(h(X_{i_1}, X_{i_2}, X_{i_3})) = \frac{1}{n(n-1)(n-2)} \sum_{i_1 \neq i_2 \neq i_3} h(X_{i_1}, X_{i_2}, X_{i_3});$ see Tchetgen et al. (2008) for more details. By Tchetgen et al. (2008), the bias of this estimator is given by

$$
\begin{aligned}
&\mathbb{E}_f \left(\widehat{\phi}_{k_1(\beta),k_2(\beta),k_3(\beta)} \right) - \phi(f) \\
&= \int \left(f_{k_3} - f_{k_2} \right)^3 + 3 \int \left(f_{k_3} - f_{k_2} \right)^2 \left(f_{k_2} - f_{k_1} \right) + \int \left(f^3 - f_{k_3}^3 \right)
\end{aligned}
\tag{5.1}
$$

which is dominated by

$$
\int \left(f^3 - f_{k_3}^3 \right) \sim k_3^{-2\beta} \leq C \frac{k_3}{n^2}.
$$

Now note that,

$$
\sup_{f \in H(\beta,C)} |\mathbb{E}_f(\widehat{\phi}_{k_1^{(2)},k_2^{(2)},k_3^{(2)}}) - \mathbb{E}_f(\widehat{\phi}_{k_1^{(1)},k_2^{(1)},k_3^{(1)}}| \sim \left(k_3^{(1)} \right)^{-2\beta}, \tag{5.2}
$$

whenever $k^{(1)} = \left(k_1^{(1)}, k_2^{(1)}, k_3^{(1)}\right) \le k^{(2)} = \left(k_1^{(2)}, k_2^{(2)}, k_3^{(2)}\right)$ where we say $v_1 \le v_2$ for two vectors $v_1, v_2 \in \mathbb{R}_+^s$ whenever $\max v_1 \le \max v_2$. As shown by Tchetgen et al. (2008), the variance is also dominated by $\frac{k_3}{n^2}$. Therefore, similar to the discussion in Section 2, the bias and variance of the candidate estimator is similar to that of the estimator of the quadratic functional of the density. We will use this fact to produce an adaptive estimator. The next result is the main result of this section and helps us apply Theorem 4.1 to construct an adaptive estimator of $\phi(f) = \int f^3 d\mu$.

THEOREM 5.1. *For any $k_1 \le k_2 \le k_3$ as above, one has for all sufficiently large n,*

$$\mathbb{E}_f \left(\widehat{\phi}_{k_1, k_2, k_3} - \mathbb{E}_f \left(\widehat{\phi}_{k_1, k_2, k_3}\right)\right)^{2q} \le C(\|\psi_0\|_\infty, \|f\|_\infty, f_{\min}) \left(\frac{k_3}{n^2}\right)^q,$$

whenever the projection kernels in the construction are based on the Haar Basis.

Now combining (5.1), (5.2), Theorem 5.1 with Theorem 4.1, we can construct an adaptive estimator of cubic integral functional as follows.

We use the ideas laid down heuristically in Section 2 and detailed in Section 4. In particular, fix a $d > 1$. Let N be the largest integer such that $d^{N-1} \le n^{1 - \frac{2}{\log\log n}}$. Set $l_{j-1} = d^{j-1}n$ for $j = 1, \ldots, N$. Now, for $j = 0, \ldots, N - 1$, define β_j to be the solution of $l_j = n^{\frac{2}{1+4\beta_j}}$. Further, define $l_j^* = \frac{l_j}{\log n^{\frac{1}{1+4\beta_j}}} = \left(\frac{n^2}{\log n}\right)^{\frac{1}{1+4\beta_j}}$ and $R(l_j^*) = \frac{l_j^*}{n^2}$. Let $s^* = s^*(n)$ be the smallest integer such that, $l(\beta_{s^*})^* \ge n$. Finally, define

$$\hat{j} := \min\left\{j : \hat{I}^2(l_j^*, l_m^*) \le C_{\text{opt}}^2 \log n R(k_m^*) \ \forall \ m \ge j, s^* \ge j \le N - 1\right\}.$$

where C_{opt} will be a deterministic constant and similar to Section 3, $\hat{I}(l_j, l_m) = U_{l_m} - U_{l_j}$ for $l_j < l_m$. Now let

$$\hat{\phi} = \hat{\phi}_{\hat{k_1}, \hat{k_2}, \hat{k_3}},$$

where $\hat{k_1} \sim n$, $\hat{k_2} \sim n^{\frac{\frac{3}{2} + 2\beta_{\hat{j}}}{1 + 4\beta_{\hat{j}}}}$ and $\hat{k_3} \sim l_{\hat{j}}^*$.

THEOREM 5.2. *Consider projection kernels based on Haar wavelets. The estimator $\hat{\phi}$ of $\phi(f) = \int f^3$ satisfies for any $\beta \in (0, \frac{1}{4})$,*

$$\sup_{f \in H(\beta, C)} \mathbb{E}_f \left[\left(\hat{\phi} - \phi(f) \right)^2 \right] \lesssim n^{-\frac{8\beta}{1+4\beta}} \log n^{\frac{4\beta}{1+4\beta}},$$

for some deterministic C_{opt} (depending only on parent wavelets and $\|f\|_\infty$).

REMARK 5.3. *For the proof of Theorem 5.1, the requirement of Haar wavelet arises at one specific instance in the proof. We expect that similar results continue to hold for more general compactly supported wavelets using arguments developed herein. However we do not further consider other wavelet bases.*

6. Minimax Adaptive Estimation of $\int T(f) d\mu$.

We now discuss adaptive estimation of a more general integral functional of the density. As mentioned earlier that such functionals arise naturally in information theory. In particular, a large body of work has focused on estimating the Shannon entropy (Beirlant et al., 1997) and more recent works include estimation of Renyi and Tsallis entropies (Leonenko and Seleznjev, 2010; Pál, Póczos and Szepesvári, 2010). As mentioned, we will use ideas developed so far for adaptive estimation of up to cubic functionals to come up with an adaptive estimator of smooth non-linear functionals. We only sketch the idea here and omit technical details. In particular, suppose that $T(f)$ admits an expansion around f_0 such that

$$T(f)(x) = T(f_0)(x) + T^{(1)}(f_0)(x)(f(x) - f_0(x)) + T^{(2)}(f_0)(f(x) - f_0(x))^2$$
$$+ T^{(3)}(f_0)(f(x) - f_0(x))^3 + O\left((f(x) - f_0(x))^4\right)$$

as $f(x) \to f_0(x)$. The idea now is to sample split and obtain an adaptive estimator $\widehat{f}(x)$ of $f(x)$ so that $\left(\widehat{f}(x) - f(x)\right) = O_p\left(n^{-2\beta/(1+2\beta)}\right)$. Next,

$$\phi(f) = \int T\left(\widehat{f}\right)(x) + \int T^{(1)}\left(\widehat{f}\right)(x)\left(f(x) - \widehat{f}(x)\right) + \int T^{(2)}\left(\widehat{f}\right)\left(f(x) - \widehat{f}(x)\right)^2$$
$$+ \int T^{(3)}\left(\widehat{f}\right)\left(f(x) - \widehat{f}(x)\right)^3 + O\left(\int \left(f(x) - \widehat{f}(x)\right)^4\right).$$

Note that, $\int \left(f(x) - \widehat{f}(x)\right)^4 = O_P\left(n^{-8\beta/(1+2\beta)}\right) = o_p\left(n^{-8\beta/(1+4\beta)}\right)$. Therefore, we need only learn how to adapt to functionals of the form $\int g_1\left(\widehat{f}\right) f(x) + \int g_2\left(\widehat{f}\right) f(x)^2 + \int g_3\left(\widehat{f}\right) f(x)^3$. The linear functional estimation theory is

well understood and the quadratic or cubic terms are rather straightforward generalizations of our statistics. For the quadratic term for instance, use $\mathbb{V}_n \left[\int g_2 \left(\widehat{f} \right) (x) \, K_k \left(x, X_{i_1} \right) K_k \left(x, X_{i_2} \right) dx \right]$ instead of the one used earlier.

7. Adaptation Lower bound for Estimation of $\int T(f) d\mu$.

In this section, we implement ideas similar to Efromovich et al. (1996a) to provide a lower bound on the required price to be paid for adaptation over $\beta < \frac{1}{4}$. Efromovich et al. (1996a) proved that, while estimating a quadratic functional of the density, if an estimator achieves a parametric rate for $\beta \geq \frac{1}{4}$ then it must incur a heavier penalty than $(\log n)^{\frac{4\beta}{1+4\beta}}$ for $\beta < \frac{1}{4}$. Here we provide a result of similar flavor regarding constraints on estimation rates at two points $\beta_1, \beta_2 < \frac{1}{4}$.

We begin by describing the main tool in our proof, which is a general version of constrained risk inequality due to Cai et al. (2011), obtained as an extension of Brown et al. (1996). For the sake of completeness, begin with a Summary of these results. Suppose Z has distribution $\mathbb{P}_\theta$ where θ belongs to some parameter space Θ. Let $\hat{Q} = \hat{Q}(Z)$ be an estimator of a function $Q(\theta)$ based on Z with bias $B(\theta) := \mathbb{E}_\theta(\hat{Q}) - Q(\theta)$. Now suppose that Θ_0 and Θ_1 form a disjoint partition of Θ with priors π_0 and π_1 supported on them respectively. Also, let $\mu_i = \int Q(\theta) d\pi_i$ and $\sigma_i^2 = \int (Q(\theta) - \mu_i)^2 d\pi_i$, $i = 0, 1$ be the mean and variance of $Q(\theta)$ under the two priors π_0 and π_1. Letting γ_i be the marginal density with respect to some common dominating measure of Z under π_i, $i = 0, 1$, let us denote by $\mathbb{E}_{\gamma_0}(g(Z))$ the expectation of $g(Z)$ with respect to the marginal density of Z under prior π_0 and distinguish it from $\mathbb{E}_\theta(g(Z))$, which is the expectation under $\mathbb{P}_\theta$. Lastly, denote the chi-square divergence between γ_0 and γ_1 by $\chi = \left\{ \mathbb{E}_{\gamma_0} \left(\frac{\gamma_1}{\gamma_0} - 1 \right)^2 \right\}^{\frac{1}{2}}$. Then we have the following result.

PROPOSITION 7.1 (Cai et al. (2011)). *If $\int \mathbb{E}_\theta \left(\hat{Q}(Z) - Q(\theta) \right)^2 d\pi_0(\theta) \leq \epsilon^2$, then*

$$| \int B(\theta) d\pi_1(\theta) - \int B(\theta) d\pi_0(\theta) | \geq |\mu_1 - \mu_0| - (\epsilon + \sigma_0)\chi.$$

Since the maximum risk is always at least as large as the average risk, this immediately yields a lower bound on the minimax risk. In order to use this result, we will need to produce suitable priors on appropriate parameter spaces, so that we capture the price needed for adaptation. Since the uniform density $f_0 \equiv \mathcal{I}(0, 1)$ is in all the smoothness classes we consider here, we take

$\Theta_0 = \{f_0^{(n)}\}$ and the prior $\pi_0 = \delta_{f_0^{(n)}}$, the dirac mass at the uniform density. Here and below $f^{(n)}$ refers to the joint density of $X_1, \ldots, X_n$ which are i.i.d with density f. Now let us construct the alternative parameter space along the lines of Efromovich et al. (1996a), which in turn relies on Ingster (1987). Take h to be a function supported on $[0,1]$ such that $\int h = 0$, $\int h^2 = c$ for some $c > 0$ to be specified later, $h \in H(1, C)$, and $1 + h \geq 0$. Let v_n be an increasing sequence of positive integers, to be specified later up to multiplicative constants, and denote by $\bar{a}$ the vector $(a_0, \ldots, a_{v_n-1}) \in \{-1, +1\}^{v_n}$. Now define

$$f_{\bar{a}} = f_0 + \sum_{i=0}^{v_n-1} a_i v_n^{-\beta} h(v_n x - i),$$

and let $\Theta_1 = \{f_{\bar{a}}^{(n)} : \bar{a} \in \{-1, +1\}^{v_n}\}$. First,note that by construction one always has $f_{\bar{a}} \in H(\beta, C)$. Finally let π_1 be the uniform prior putting $\frac{1}{2^{v_n}}$ mass at each point of Θ_1. In order to apply Proposition 7.1, we evaluate the following quantities. Using the notation introduced earlier,

$$\mu_0 = \int T(f_0(x))dx = T(1)$$

and $\sigma_0 = 0$. Also,

$$\mu_1 = \frac{1}{2^{v_n}} \sum_{\bar{a} \in \{-1, +1\}^{v_n}} \int T\left(f_0(x) + \sum_{i=0}^{v_n-1} a_i v_n^{-\beta} h(v_n x - i)\right) dx.$$

Therefore

$$|\mu_1 - \mu_0| = \frac{1}{2^{v_n}} \sum_{\bar{a} \in \{-1, +1\}^{v_n}} \int \left[T\left(1 + \sum_{i=0}^{v_n-1} a_i v_n^{-\beta} h(v_n x - i)\right) - T(1)\right] dx.$$

Now by Taylor expansion, for each $\bar{a}$,

$$\int \left[T\left(1 + \sum_{i=0}^{v_n-1} a_i v_n^{-\beta} h(v_n x - i)\right) - T(1)\right]$$

$$= T'(1) \sum_{i=0}^{v_n-1} a_i v_n^{-\beta} \int h(v_n x - i)dx + T''(1) \int \left[\sum_{i=0}^{v_n-1} a_i v_n^{-\beta} h(v_n x - i)\right]^2 dx$$

$$+ \int T'''(\xi(x)) \left[\sum_{i=0}^{v_n-1} a_i v_n^{-\beta} h(v_n x - i)\right]^3 dx \text{ where } \|\xi - 1\|_\infty \leq 1$$

$$\geq C_T v_n^{-2\beta}$$

for some constant $C_T > 0$ depending only on T. The above calculations follow by the inclusion of the support of h in $[0,1]$, boundedness of h and standard change of variables. Therefore, $|\mu_1 - \mu_0| \geq C v_n^{-2\beta}$ as well, where we have dropped the subscript T for notational convenience. Now suppose that the bias at f_0 is smaller than $\epsilon_n \lesssim \left(\frac{n}{\sqrt{\log n}}\right)^{-\frac{4\beta^*}{4\beta^*+1}}$ for some $\beta^* < \frac{1}{4}$. Then for any for any $\beta < \beta^*$, we have by proportion 7.1 that $\int B(\theta) d\pi_1(\theta) \geq C v_n^{-2\beta} - \epsilon_n \chi$. Therefore, if we can show that for $v_n = \left(\frac{n}{\sqrt{c \log n}}\right)^{\frac{2}{1+4\beta}}$ one has $\chi \ll n^c$, then one gets the desired rate for the incurred penalty by choosing c small enough. This can be derived following the arguments in Ingster (1987). Therefore, we have sketched the proof of the following theorem.

THEOREM 7.2. *Suppose one has*

$$\sup_{f \in H(\beta^*)} \mathbb{E}_f \left(\hat{\phi} - \phi(f)\right)^2 \lesssim \left(\frac{n^2}{\log n}\right)^{-\frac{4\beta^*}{1+4\beta^*}},$$

for an estimator $\hat{\phi}$ of $\phi(f) = \int T(f) d\mu$ with T having bounded third derivative. Then for any $\beta > \beta^*$,

$$\sup_{f \in H(\beta)} \mathbb{E}_f \left(\hat{\phi} - \phi(f)\right)^2 \gtrsim \left(\frac{n^2}{\log n}\right)^{-\frac{4\beta}{1+4\beta}}.$$

8. Discussions. Our results are all provided for one dimension $d = 1$ corresponding to a low smoothness regime $\beta < \frac{d}{4}$. Although we do not provide explicit details, our proofs can be easily extended to incorporate the case of $d > 1$ and to consider $\sqrt{n}$ rate of convergence of the constructed estimators for $\beta > \frac{d}{4}$. The case of $\beta = \frac{d}{4}$ is a little more subtle; however this also can be done by adding one more level of discretization near the truncation level $k = n$. Since, the crux of the arguments remain the same, we do not elaborate on such proofs here.

Following the idea of possibly modifying Lepski's method to include a larger class of mechanisms to choose a tuning parameter for adaptive estimation, another potentially interesting question is to investigate which method provides a better performance- either in a finite sample sense or regarding constants of asymptotic mean squared error. Finally, the study of adaptive non-parametric divergence and entropy based on more than one

densities is also of interest. Since, the quadratic functional estimator is a U-statistics is based on a bounded kernel, the results of Houdré and Reynaud-Bouret (2003) were readily applicable for deriving a suitable exponential tail bound. It is worth exploring how to modify such arguments for second order U-statistics estimators that arise in context of non-parametric regression problems and fail to be bounded in case of unbounded error distributions. We keep the study of such questions for possible future research.

APPENDIX A: PROOF OF THEOREMS

Proof of Theorem 3.1.

PROOF. Recall that,

$$
\sup_{f} E\left[U_{n,\widehat{j}} - \phi(f)\right]^2
$$

$$
\leq \sum_{k=0}^{1} \sup_{f} E\left\{\mathcal{I}\left\{\widehat{j} = k\right\}\left[U_{n,\widehat{j}} - \phi(f)\right]^2\right\}
$$

$$
\equiv \sup_{f} R_0(\beta_f) + \sup_{f} R_1(\beta_f)
$$

Higher Smoothness $(\beta_f = \beta_0)$.

(i) First let us consider,

$$
\sup_{f} R_0(\beta_0) = \sup_{f \in H(\beta_0)} \mathbb{E}_f\left[\mathcal{I}(\hat{j} = 0)\left(U_{n,0} - \phi(f)\right)^2\right]
$$

$$
\leq \sup_{f \in H(\beta_0)} \mathbb{E}_f\left[\left(U_{n,0} - \phi(f)\right)^2\right] \lesssim n^{-\frac{8\beta_0}{1+4\beta_0}}
$$

by our choice of $U_{n,0}$.

(ii) Next we consider $\sup_{f \in H(\beta_0)} R_1(\beta_0)$. For this part, we will bound $R_1(\beta_0)$ from above uniformly in all $f \in H(\beta_0)$. First note that by Holder's Inequality, for any conjugate pair (p, q) we have

$$
R_1(\beta_0) = \mathbb{E}_f\left[\mathcal{I}(\hat{j} = 1)\left(U_{n,1} - \phi(f)\right)^2\right]
$$

$$
\leq \left\{\mathbb{P}_f\left(\mathcal{I}(\hat{j} = 1)\right)\right\}^{\frac{1}{p}}\left\{\mathbb{E}_f\left[\left(U_{n,1} - \phi(f)\right)^{2q}\right]\right\}^{\frac{1}{q}}
$$

The proof of $\sup_{f \in H(\beta_0)} R_1(\beta_0) \lesssim n^{-\frac{8\beta_0}{1+4\beta_0}}$ now follows by the following two lemmas.

LEMMA A.1. *For kernels based on compactly supported wavelet bases,*

$$\sup_{f \in H(\beta_0)} \mathbb{P}_f \left(\mathcal{I}(\hat{j} = 1) \right) \lesssim \frac{1}{n}$$

.

LEMMA A.2. *For kernels based on compactly supported wavelet bases,*

$$\sup_{f \in H(\beta_0)} \left\{ \mathbb{E}_f \left[(U_{n,1} - \phi(f))^{2q} \right] \right\}^{\frac{1}{q}} \lesssim \frac{k_1}{n^2}$$

.

Lower Smoothness $(\beta_f = \beta_1)$.

(i) First note that,

$$\sup_f R_1(\beta_1) = \sup_{f \in H(\beta_1)} \mathbb{E}_f \left[\mathcal{I}(\hat{j} = 1) \left(U_{n,1} - \phi(f) \right)^2 \right]$$
$$\leq \sup_{f \in H(\beta_1)} \mathbb{E}_f \left[(U_{n,1} - \phi(f))^2 \right] \lesssim n^{-\frac{8\beta_1}{1+4\beta_1}}$$

by our choice of $U_{n,1}$.

(ii) Now,

$$R_0(\beta_1) = \mathbb{E}_f \left(\mathbf{1} \left\{ \widehat{I}^2 \left(k_0, k_1^* \right) < C \ln (n) \frac{k_1^*}{n^2} \right\} \left[U_{n,0} - \phi \left(f \right) \right]^2 \right).$$

We consider three cases as below with $\nu_n = \ln(n)^{-1/(8\beta_1+2)}$. In the calculations below C, C', C'', C''' are arbitrary constants which can be chosen for the calculations to be valid and can change from place to place. Also for controlling higher order moments of related second order U-statistics below, we use results along the lines of Lemma A.2 with obvious modifications to the proof.

Case 1: $\mathbf{I^2 \left(k_0, k_1^* \right) > C' \left(1 + \nu_n \right) \ln (n) \frac{k_1^*}{n^2}}$

In this case we have,

$$R_0(\beta)$$
$$= \mathbb{E}_f \left(\mathbf{1} \left\{ \widehat{I}^2 \left(k_0, k_1^* \right) < C \ln (n) \frac{k_1^*}{n^2} \right\} \mathbf{1} \left\{ I^2 \left(k_0, k_1^* \right) > C' \left(1 + \nu_n \right) \ln (n) \frac{k_1^*}{n^2} \right\} \left[U_{n,0} - \phi \left(f \right) \right]^2 \right)$$

$$\leq \mathbb{E}_f \left(\left(\mathbf{1} \left\{ I\left(k_0, k_1^*\right) - \widehat{I}\left(k_0, k_1^*\right) > \left(C'' \nu_n\right) I\left(k_0, k_1^*\right) \right\} \left[U_{n,0} - \phi\left(f\right)\right]^2 \right) \right)$$

$$\leq C''' \left(\nu_n I\left(k_0, k_1^*\right)\right)^{-2} \mathbb{E}_f \left(\left(I\left(k_0, k_1^*\right) - \widehat{I}\left(k_0, k_1^*\right) \right)^2 \left[U_{n,0} - \phi\left(f\right)\right]^2 \right)$$

$$\leq C''' \left(\nu_n I\left(k_0, k_1^*\right)\right)^{-2} \mathbb{E}_f^{1/2} \left(\left(I\left(k_0, k_1^*\right) - \widehat{I}\left(k_0, k_1^*\right) \right)^4 \right) \mathbb{E}_f^{1/2} \left(\left[U_{n,0} - \phi\left(f\right)\right]^4 \right)$$

$$\leq C''' \left(\nu_n I\left(k_0, k_1^*\right)\right)^{-2} \left(n^{-2} k_1^*\right) \left(I\left(k_0, k_1^*\right) + \left(k_1^*\right)^{-2\beta_1} \right)^2 \leq C''' \left(\nu_n\right)^{-2} \left(n^{-2} k_1^*\right)$$

$$= C''' \ln(n)^{1/(4\beta_1 + 1)} \left(\frac{n^2}{\ln\left(n\right)} \right)^{1/(1+4\beta_1)} n^{-2} = C''' n^{2/(1+4\beta_1)} n^{-2} = C''' n^{-\frac{8\beta_1}{1+4\beta_1}}$$

Case 2: $\mathbf{I^2\left(k_0, k_1^*\right) \leq C\frac{k_1}{n^2}}$

In this case,

$$\mathbb{E}_f \left(\left[U_{n,0} - \phi\left(f\right)\right]^2 \right) = k_0^{-4\beta_1} + \frac{k_0}{n^2} \asymp I^2\left(k_0, k_1\right) + \frac{k_0}{n^2} \leq C' n^{-\frac{8\beta_1}{1+4\beta_1}}.$$

Case 3: $\mathbf{I^2\left(k_0, k_1^*\right) \leq C\left(1 + \nu_n\right)\ln\left(n\right)\frac{k_1^*}{n^2}}$ and $\mathbf{I^2\left(k_0, k_1^*\right) \leq C\frac{k_1}{n^2}}$

In this case, $I^2\left(k_0, k_1^*\right) < \left(1 + \nu_n\right)\ln\left(n\right)\frac{k_1^*}{n^2}$ and $I^2\left(k_0, k_1\right) > \frac{k_1^*}{n^2}$, so that

$$\mathbb{E}_f \left(\left[U_{n,0} - \phi\left(f\right)\right]^2 \right) \leq C' \left(I^2\left(k_0, k_1^*\right) + \left(k_1^*\right)^{-4\beta_1} + \frac{k_0}{n^2} \right) C'' \left(I^2\left(k_0, k_1^*\right) + \left(k_1^*\right)^{-4\beta_1} + \frac{k_0}{n^2} \right)$$

$$\leq C''' \ln\left(n\right)^{1 - 1/(1+4\beta_1)} n^{-8\beta_1/(1+4\beta_1)} = C''' n^{-8\beta_1/(1+4\beta_1)} \ln\left(n\right)^{4\beta_1/(1+4\beta_1)}.$$

$\square$

Proof of Theorem 4.1.

PROOF. Suppose $\beta_f \in \left(\beta_{j+1}, \beta_j\right]$. Let $l_c(C^*)$ be the largest l such that $I^2(k_l^*, k_j^*) > C^* \log nR(k_j^*)$. Let us explain the reason for existence of such an $l_c(C^*)$. Note that it is enough to show that the ratio $\frac{I^2(k_l^*, k_j^*)}{C^* \log nR(k_j^*)}$ is upper bounded by a quantity that l increases through positive integers. By our choice of $\beta_f \in \left(\beta_{j+1}, \beta_j\right]$, we have that $\frac{I^2(k_l^*, k_j^*)}{C^* \log nR(k_j^*)}$ is at most a constant times $\frac{(k_l^*)^{-4\beta_{j+1}}}{\frac{k_j^*}{n^2}}$. The power of n in this ration is $\frac{8\beta_j}{1+4\beta_j} - \frac{8\beta_{j+1}}{1+4\beta_l}$ which indeed decreases as l increases. Also, note that trivially $l_c(C^*) < j$. Also, by definition of $l_c(C^*)$, for any $l \leq l_c$, $I^2(k_l^*, k_j^*) > C^* \log nR(k_j^*)$ and for $l > l_c(C^*)$, $I^2(k_l^*, k_j^*) \leq C^* \log nR(k_j^*)$. Our proof relies on the fact that the testing procedure does not select an index $l \leq l_c(C^*)$ or $l > j+1$ with high probability. This is captured by the following two lemmas.

LEMMA A.3. *For kernels based on compactly supported wavelet bases,*

$$\sup_{f \in H(\beta)} \mathbb{P}_f \left(\hat{j} = l \right) \leq \frac{C_{\psi_0}}{n},$$

for any $l \leq l_c(C^*)$ whenever $C^* > 4C_{\text{opt}}^2$.

LEMMA A.4. *For kernels based on compactly supported wavelet bases,*

$$\sup_{f \in H(\beta)} \mathbb{P}_f \left(\hat{j} \geq j + 2 \right) \leq \frac{C_{\psi_0}}{n},$$

whenever $C^* > 4C_{\text{opt}}^2$.

Let us first complete the proof of Theorem 4.1 assuming the validity of Lemmas A.3 and A.4. Note that,

$$\mathbb{E}_f \left[\left(U_{n,k_{\hat{j}}^*} - \phi(f) \right)^2 \right] = \sum_{l=s^*}^{N-1} \mathbb{E}_f \left[\mathcal{I}(\hat{j} = l) \left(U_{n,k_l^*} - \phi(f) \right)^2 \right]$$
$$= T_1 + T_2 + T_3, \qquad (A.1)$$

where $T_1 = \sum_{l=s^*}^{l_c(C^*)} \mathbb{E}_f \left[\mathcal{I}(\hat{j} = l) \left(U_{n,k_l^*} - \phi(f) \right)^2 \right]$, $T_2 = \sum_{l=l_c(C^*)+1}^{j+1} \mathbb{E}_f \left[\mathcal{I}(\hat{j} = l) \left(U_{n,k_l^*} - \phi(f) \right)^2 \right]$

and $T_3 = \sum_{l=j+2}^{N-1} \mathbb{E}_f \left[\mathcal{I}(\hat{j} = l) \left(U_{n,k_l^*} - \phi(f) \right)^2 \right]$. In the following we control

the terms T_h, $h = 1, 2, 3$ individually at $n^{-\frac{8\beta_f}{1+4\beta_f}} \log n^{\frac{4\beta_f}{1+4\beta_f}}$ to show the desired result modulo the proofs of the above two lemmas. We then finish the proof by proving the lemmas. For the following let (p, q) denote a pair of positive real numbers such that $\frac{1}{p} + \frac{1}{q} = 1$. The specific choices of the pair will be clear from the proof. In particular, we will always choose q to be an integer and p will be sufficiently close to 1.

Control of T_1.

$$T_1 = \sum_{l=s^*}^{l_c(C^*)} \mathbb{E}_f \left[\mathcal{I}(\hat{j} = l) \left(U_{n,k_l^*} - \phi(f) \right)^2 \right]$$
$$\leq \sum_{l=s^*}^{l_c(C^*)} \mathbb{P}_f^{\frac{1}{p}} \left(\hat{j} = l \right) \mathbb{E}_f^{\frac{1}{q}} \left[\left(U_{n,k_l^*} - \phi(f) \right)^{2q} \right]$$

$$\leq \sum_{l=s^*}^{l_c(C^*)} \left(\frac{C_{\psi_0}}{n} \right)^{\frac{1}{p}} \left[\frac{k_l^*}{n^2} + (k_l^*)^{-2\beta_f} \right] \leq \log n \left(\frac{C_{\psi_0}}{n} \right)^{\frac{1}{p}}$$

where the second last inequality above follows from Lemma A.3 and condition 2 of the theorem, and the last inequality follows from the choice of $N \lesssim \log n$. Therefore for p sufficiently close to 1, we have desired control over T_1.

Control of T_2.

$$T_2 = \sum_{l=l_c(C^*)+1}^{j+1} \mathbb{E}_f \left[\mathcal{I}(\hat{j}=l) \left(U_{n,k_l^*} - \phi(f) \right)^2 \right] \leq \sum_{l=l_c(C^*)+1}^{j+1} \mathbb{P}_f^{\frac{1}{p}} \left(\hat{j}=l \right) \mathbb{E}_f^{\frac{1}{q}} \left[\left(U_{n,k_l^*} - \phi(f) \right)^{2q} \right]$$

$$\leq C_q \sum_{l=l_c(C^*)+1}^{j+1} \mathbb{P}_f^{\frac{1}{p}} \left(\hat{j}=l \right) \left\{ \begin{array}{c} \mathbb{E}_f^{\frac{1}{q}} \left[\left(U_{n,k_l^*} - \mathbb{E}_f \left(U_{n,k_l^*} \right) \right)^{2q} \right] \\ + \mathbb{E}_f^{\frac{1}{q}} \left[\left(\mathbb{E}_f \left(U_{n,k_l^*} \right) - \mathbb{E}_f \left(U_{n,k_j^*} \right) \right)^{2q} \right] + \mathbb{E}_f^{\frac{1}{q}} \left[\left(U_{n,k_j^*} - \phi(f) \right)^{2q} \right] \end{array} \right\}$$

$$\leq C_q \sum_{l=l_c(C^*)+1}^{j+1} \mathbb{P}_f^{\frac{1}{p}} \left(\hat{j}=l \right) \left\{ \frac{k_l^*}{n^2} + I_c^2(k_j^*, k_l^*) + \frac{k_j^*}{n^2} + (k_j^*)^{-2\beta_f} \right\}$$

$$\leq C_q \sum_{l=l_c(C^*)+1}^{j+1} \mathbb{P}_f^{\frac{1}{p}} \left(\hat{j}=l \right) \left\{ \frac{k_l^*}{n^2} + I^2(k_j^*, k_l^*) + \frac{k_j^*}{n^2} \right\}$$

$$\leq C_q C^* \log n R(k_j^*) \sum_{l=l_c(C^*)+1}^{j+1} \mathbb{P}_f^{\frac{1}{p}} \left(\hat{j}=l \right) \leq C_q C^* \log n R(k_j^*).$$

Above, the fourth last inequality follows from condition 1 of the theorem, the third last inequality follows condition 2 of the theorem and using the fact that $\beta_{j+1} < \beta_f \leq \beta_j$ (which implies that $(k_j^*)^{-2\beta_f} \leq (k_j^*)^{-2\beta_j} \left(\frac{n^2}{\log n} \right)^{\frac{2\beta_j - 2\beta_{j+1}}{1+4\beta_j}} \leq C \frac{k_j^*}{n^2} n^{\frac{C}{\log n}} \leq C' \frac{k_j^*}{n^2}$). The second last inequality follows from our choice of $l_c(C^*)$. The last inequality follows since $\sum_{l=l_c(C^*)+1}^{j+1} \mathbb{P}_f \left(\hat{j}=l \right) \leq 1$, we have that for p sufficiently close to 1, one also must have that $\sum_{l=l_c(C^*)+1}^{j+1} \mathbb{P}_f^{\frac{1}{p}} \left(\hat{j}=l \right) \leq$ 1. Finally noting that $C_q C^* \log n R(k_j^*) \lesssim n^{-\frac{8\beta_f}{1+4\beta_f}} \log n^{\frac{4\beta_f}{1+4\beta_f}}$ completes the required control T_2.

Control of T_3.

$$T_3 = \sum_{l=j+2}^{N-1} \mathbb{E}_f \left[\mathcal{I}(\hat{j} = l) \left(U_{n,k_l^*} - \phi(f) \right)^2 \right]$$

$$\leq \sum_{l=j+2}^{N-1} \mathbb{P}_f^{\frac{1}{p}} \left(\hat{j} = l \right) \mathbb{E}_f^{\frac{1}{q}} \left[\left(U_{n,k_l^*} - \phi(f) \right)^{2q} \right]$$

$$\leq \sum_{l=j+2}^{N-1} \left(\frac{C_{\psi_0}}{n} \right)^{\frac{1}{p}} \left[\frac{k_l^*}{n^2} + (k_l^*)^{-2\beta_f} \right] \leq \log n \left(\frac{C_{\psi_0}}{n} \right)^{\frac{1}{p}}$$

where the second last inequality above follows from Lemma A.4 and condition 2 of the theorem, and the last inequality follows from the choice of $N \lesssim \log n$. Therefore for p sufficiently close to 1, we have desired control over T_1. $\qquad\square$

Proof of Theorem 5.1.

PROOF. The idea of proof is similar to the second proof of A.2. However, the computations are much more cumbersome since ow we control moments of a third order U-statistics.

We decompose $U_{n,1}$ as,

$$\widehat{\psi}_{k_1,k_2,k_3} = V_{n,1} + R_{n,1},$$

where

$$V_{n,1} = \frac{1}{n(n-1)(n-2)} \sum_{i \neq j \neq s} \left\{ \widetilde{K}(X_i, X_j, X_s) \mathcal{I}((X_i, X_j, X_s) \in \bigcup_{m=1}^{M_n} \chi_{n,m} \times \chi_{n,m} \times \chi_{n,m}) \right\}$$

and

$$R_{n,1} = \widehat{\psi}_{k_1,k_2,k_3} - V_{n,1}.$$

The sets $\chi_{n,m}, m = 1, \ldots, M_n$ are constructed according to Robins et al. (2015). Therefore, for Haar basis in one dimension, we can take $M_n = n$ and $\chi_{n,m} = [\frac{m-1}{n}, \frac{m}{n})$ for $m = 1, \ldots, M_n$. Therefore,

$$\mathbb{E}_f \left[\left(\widehat{\psi}_{k_1,k_2,k_3} - \mathbb{E}_f(\widehat{\psi}_{k_1,k_2,k_3}) \right)^{2q} \right] \leq C_q \left\{ \mathbb{E}_f \left[(V_{n,1} - \mathbb{E}_f(V_{n,1}))^{2q} \right] + \mathbb{E}_f \left[(R_{n,1} - \mathbb{E}_f(R_{n,1}))^{2q} \right] \right\}$$

Below, we only control the first summand i.e. $\mathbb{E}_f \left[(V_{n,1} - \mathbb{E}_f(V_{n,1}))^{2q} \right]$ which suffices for Haar basis. However, for non-Haar bases this is not sufficient. We believe, that for non-Haar bases one can show that the second summand has sub-optimal rate.

Control of $\mathbb{E}_{\mathbf{f}}\left[(\mathbf{V_{n,1}} - \mathbb{E}_{\mathbf{f}}(\mathbf{V_{n,1}}))^{\mathbf{2q}}\right]$.

Write $V_{n,1} = \sum_{m=1}^{M_n} V_{n,m}$ with

$$V_{n,m} = \frac{1}{n(n-1)(n-2)} \sum_{i \neq j \neq s} \widetilde{K}(X_i, X_j, X_s)\mathcal{I}((X_i, X_j, X_s) \in \chi_{n,m} \times \chi_{n,m} \times \chi_{n,m}).$$

Following Robins et al. (2015), we define the following membership based quantities. Let,

$$I_{n,r} := m \text{ if } X_r \in \chi_{n,m}, \ r = 1, \ldots, n, \ m = 1, \ldots, M_n,$$

$$N_{n,m} := \# \left(r : I_{n,r} = m \right).$$

Then, $(N_{n,m}, 1 \leq m \leq M_n) \sim \text{Multinomial}(p_{nm}, 1 \leq m \leq M_n)$, where

$$p_{n,m} = \int_{\chi_{n,m}} f(x)dx.$$

Note that $p_{n,m} \in [\frac{f_{\min}}{n}, \frac{\|f\|_\infty}{n}]$ where $f_{\min} = \min f(x)$. Given the vector $\mathbf{I}_n = (I_{n,1}, \ldots, I_{n,M_n})$, the observations $X_1, \ldots, X_n$ are independent with distribution of $X_r | I_{n,r} = m$ being $\frac{I_{\chi_{n,m}}dF}{p_{n,m}}$. Now,

$$\mathbb{E}_f\left[(V_{n,1} - \mathbb{E}_f(V_{n,1}))^{2q}\right]$$
$$\leq C_q \left[\begin{array}{c} \mathbb{E}_f\left[(V_{n,1} - \mathbb{E}_f(V_{n,1}|\mathbf{I}_n))^{2q}\right] \\ +\mathbb{E}_f\left[(\mathbb{E}_f(V_{n,1}|\mathbf{I}_n) - \mathbb{E}_f(V_{n,1}))^{2q}\right] \end{array} \right]$$

Control of $\mathbb{E}_{\mathbf{f}}\left[(\mathbb{E}_{\mathbf{f}}(\mathbf{V_{n,1}}|\mathbf{I_n}) - \mathbb{E}_{\mathbf{f}}(\mathbf{V_{n,1}}))^{\mathbf{2q}}\right]$.

Now, note that,

$$\mathbb{E}_f(V_{n,1}|\mathbf{I}_n) - \mathbb{E}_f(V_{n,1}) = \sum_{m=1}^{M_n} \left(\frac{N_{n,m}(N_{n,m}-1)(N_{n,m}-2)}{n(n-1)(n-2)p_{n,m}^3} - 1 \right) \alpha_{n,m},$$

where $\alpha_{n,m} = \int \int \widetilde{K}(x_1, x_2, x_3)\mathcal{I}((x_1, x_2, x_3) \in \chi_{n,m} \times \chi_{n,m} \times \chi_{n,m})d(F(x_1))d(F(x_2))d(F(x_3))$. The next lemma provides control of the terms $\alpha_{n,m}$.

LEMMA A.5. *For compactly supported wavelet bases one has*

$$\max_m |\alpha_{n,m}| \leq \frac{C(\psi_0, \|f\|_\infty)}{M_n}.$$

Therefore, we have $\sum_{m=1}^{M_n} |\alpha_{n,m}| \leq C_{\psi_0}$. Now, since $N_{n,m}$'s are negatively associated (see definition in Joag-Dev and Proschan (1983)), we have by Theorem 2 of Shao (2000) followed by Marcinkiewicz Zygmund inequality that,

$$
\mathbb{E}_f \left[\left(\mathbb{E}_f(V_{n,1}|\mathbf{I}_n) - \mathbb{E}_f(V_{n,1}) \right)^{2q} \right]
$$

$$
= \mathbb{E}_f \left[\left(\frac{1}{M_n} \sum_{m=1}^{M_n} \left\{ \frac{N_{n,m}(N_{n,m}-1)(N_{n,m}-2)}{n(n-1)(n-2)p_{n,m}^3} - 1 \right\} \alpha_{n,m} \right)^{2q} \right] M_n^{2q}
$$

$$
\leq C_q M_n^{2q} M_n^{-\frac{2q}{2}} \frac{1}{M_n} \sum_{m=1}^{M_n} \mathbb{E}_f \left[\left(\frac{N_m(N_m-1)(N_{n,m}-2)}{n(n-1)(n-2)p_{n,m}^3} - 1 \right)^{2q} \right] \alpha_{n,m}^{2q}
$$

$$
\leq C_q^* M_n^{q-1} \sum_{m=1}^{M_n} |\alpha_{n,m}|^{2q} \tag{A.2}
$$

$$
\leq C_q^* M_n^{q-1} \{\max(\alpha_{n,m})\}^{2q-1} \sum_{m=1}^{M_n} |\alpha_{n,m}| \leq C_{\psi_0}^q M_n^{q-1} \{\max(\alpha_{n,m})\}^{2q-1}
$$

In the above display, equation A.2 follows from Lemma 5.6 of Robins et al. (2015). Now, by Lemma A.5, $\{\max(\alpha_{n,m})\}^{2q-1} \leq (C(\psi_0, \|f\|_\infty))^{2q-1} \left(\frac{1}{M_n}\right)^{2q-1}$. Therefore, $\mathbb{E}_f \left[\left(\mathbb{E}_f(V_{n,1}|\mathbf{I}_n) - \mathbb{E}_f(V_{n,1}) \right)^{2q} \right] \leq (C_{\psi_0}\|f\|_\infty)^{2q-1} \left(\frac{1}{M_n}\right)^q \leq (C_{\psi_0}\|f\|_\infty)^{2q-1} \left(\frac{k_3}{n^2}\right)^q$ for our choice of $M_n \asymp n$.

Control of $\mathbb{E}_\mathbf{f} \left[(\mathbf{V_{n,1}} - \mathbb{E}_\mathbf{f}(\mathbf{V_{n,1}}|\mathbf{I_n}))^{2\mathbf{q}} \right]$.

Suppose, $V_{n,m} = V_{n,m}^{(1)} + V_{n,m}^{(2)} + V_{n,m}^{(3)} \mathbb{E}_f(V_{n,m}|\mathbf{I}_n)$ denote the Hoeffding decomposition of $V_{n,m}$ w.r.t the conditional distribution given $\mathbf{I}_n$. Therefore,

$$
\mathbb{E}_f \left[(V_{n,1} - \mathbb{E}_f(V_{n,1}|\mathbf{I}_n))^{2q} \right] = \mathbb{E}_f \left[\left(\sum_{m=1}^{M_n} \left\{ V_{n,m}^{(1)} + V_{n,m}^{(2)} + V_{n,m}^{(3)} \right\} \right)^{2q} \right]
$$

Note that $\mathbb{E}_f \left(V_{n,m}^{(1)}|\mathbf{I}_n \right) = 0$, $\mathbb{E}_f \left(V_{n,m}^{(2)}|\mathbf{I}_n \right) = 0$, $\mathbb{E}_f \left(V_{n,m}^{(3)}|\mathbf{I}_n \right) = 0$ and they are independent over m conditional on $\mathbf{I}_n$. Hence by a conditional version of Rosenthal's Inequality (noting that the constants of the inequality does not depend on the underlying distribution), we have

$$
\mathbb{E}_f \left[\left(\sum_{m=1}^{M_n} \left\{ V_{n,m}^{(1)} + V_{n,m}^{(2)} + V_{n,m}^{(3)} \right\} \right)^{2q} \right]
$$

$$\leq C_q \mathbb{E}_f \left[\sum_{m=1}^{M_n} \mathbb{E}_f \left(|V_{n,m}^{(1)} + V_{n,m}^{(2)} + V_{n,m}^{(3)}|^{2q} | \mathbf{I}_n \right) + \left\{ \sum_{m=1}^{M_n} \mathbb{E}_f \left(|V_{n,m}^{(1)} + V_{n,m}^{(2)} + V_{n,m}^{(3)}|^2 \mathbf{I}_n \right) \right\}^{\frac{2q}{2}} \right]$$

$$= C_q \left[\sum_{m=1}^{M_n} \mathbb{E}_f \left(|V_{n,m}^{(1)} + V_{n,m}^{(2)} + V_{n,m}^{(3)}|^{2q} \right) + \mathbb{E}_f \left\{ \sum_{m=1}^{M_n} \mathbb{E}_f \left(|V_{n,m}^{(1)} + V_{n,m}^{(2)} + V_{n,m}^{(3)}|^2 \mathbf{I}_n \right) \right\}^{\frac{2q}{2}} \right]$$

$$\text{(A.3)}$$

Below we control

$$\sum_{m=1}^{M_n} \mathbb{E}_f \left(|V_{n,m}^{(1)} + V_{n,m}^{(2)} + V_{n,m}^{(3)}|^{2q} \right) \text{ and } \mathbb{E}_f \left\{ \sum_{m=1}^{M_n} \mathbb{E}_f \left(|V_{n,m}^{(1)} + V_{n,m}^{(2)} + V_{n,m}^{(3)}|^2 \mathbf{I}_n \right) \right\}^{\frac{2q}{2}}$$

separately, at the required rate.

Control of $\sum_{m=1}^{M_n} \mathbb{E}_f \left(|V_{n,m}^{(1)} + V_{n,m}^{(2)} + V_{n,m}^{(3)}|^{2q} \right)$.

We now provide a general control over terms like $\mathbb{E}_f \left(|V_{n,m}^{(1)} + V_{n,m}^{(2)} + V_{n,m}^{(3)}|^{2q} \right)$.
To this end, note that, it is enough to control $\mathbb{E}_f \left((V_{n,m}^{(1)})^{2q} \right)$, $\mathbb{E}_f \left((V_{n,m}^{(2)})^{2q} \right)$,
and $\mathbb{E}_f \left((V_{n,m}^{(3)})^{2q} \right)$ separately. We have dropped the absolute value sign assuming without loss of generality that q is a sufficiently large integer.

Control of $\mathbb{E}_f \left((V_{n,m}^{(1)})^{2q} \right)$.

Note that, by Hoeffding decomposition of U-statistics with asymmetric kernel, we have

$$V_{n,m}^{(1)} = \frac{N_{n,m}(N_{n,m}-1)(N_{n,m}-2)}{n(n-1)(n-2)} \times$$

$$\frac{1}{N_{n,m}} \sum_{i=1}^{N_{n,m}} \left\{ \begin{array}{l} \mathbb{E}_{f_{X_j}, f_{X_s}} \left[\widetilde{K}\left(X_i, X_j, X_s\right) \mathcal{I}((X_i, X_j, X_s) \in \chi_{n,m} \times \chi_{n,m} \times \chi_{n,m}) | \mathbf{I}_n \right] \\ + \mathbb{E}_{f_{X_j}, f_{X_s}} \left[\widetilde{K}\left(X_i, X_s, X_j\right) \mathcal{I}((X_i, X_j, X_s) \in \chi_{n,m} \times \chi_{n,m} \times \chi_{n,m}) | \mathbf{I}_n \right] \\ + \mathbb{E}_{f_{X_j}, f_{X_s}} \left[\widetilde{K}\left(X_j, X_i, X_s\right) \mathcal{I}((X_i, X_j, X_s) \in \chi_{n,m} \times \chi_{n,m} \times \chi_{n,m}) | \mathbf{I}_n \right] \\ + \mathbb{E}_{f_{X_j}, f_{X_s}} \left[\widetilde{K}\left(X_s, X_i, X_j\right) \mathcal{I}((X_i, X_j, X_s) \in \chi_{n,m} \times \chi_{n,m} \times \chi_{n,m}) | \mathbf{I}_n \right] \\ + \mathbb{E}_{f_{X_j}, f_{X_s}} \left[\widetilde{K}\left(X_j, X_s, X_i\right) \mathcal{I}((X_i, X_j, X_s) \in \chi_{n,m} \times \chi_{n,m} \times \chi_{n,m}) | \mathbf{I}_n \right] \\ + \mathbb{E}_{f_{X_j}, f_{X_s}} \left[\widetilde{K}\left(X_s, X_j, X_i\right) \mathcal{I}((X_i, X_j, X_s) \in \chi_{n,m} \times \chi_{n,m} \times \chi_{n,m}) | \mathbf{I}_n \right] \\ - 6 \mathbb{E}_f \left[\widetilde{K}\left(X_i, X_j, X_s\right) \mathcal{I}((X_i, X_j, X_s) \in \chi_{n,m} \times \chi_{n,m} \times \chi_{n,m}) | \mathbf{I}_n \right] \end{array} \right\}$$

By an application of conditional version of Marcinkiewicz Zygmund Inequality (noting that the constants of the inequality does not depend on the underlying distribution) we have that

$$\mathbb{E}_f\left((V_{n,m}^{(1)})^{2q}|\mathbf{I}_n\right)$$

$$\leq C_q \frac{N_{n,m}^{2q}(N_{n,m}-1)^{2q}(N_{n,m}-2)^{2q}}{n^{2q}(n-1)^{2q}(n-2)^{2q}}\frac{N_{n,m}^{-\frac{2q}{2}}}{N_{n,m}}\times$$

$$\sum_{i=1}^{N_{n,m}}\mathbb{E}_f\left(\left\{\begin{array}{c}\mathbb{E}_{f_{X_j},f_{X_s}}\left[\widetilde{K}\left(X_i,X_j,X_s\right)\mathcal{I}((X_i,X_j,X_s)\in\chi_{n,m}\times\chi_{n,m}\times\chi_{n,m})|\mathbf{I}_n\right]\\ +\mathbb{E}_{f_{X_j},f_{X_s}}\left[\widetilde{K}\left(X_i,X_s,X_j\right)\mathcal{I}((X_i,X_j,X_s)\in\chi_{n,m}\times\chi_{n,m}\times\chi_{n,m})|\mathbf{I}_n\right]\\ +\mathbb{E}_{f_{X_j},f_{X_s}}\left[\widetilde{K}\left(X_j,X_i,X_s\right)\mathcal{I}((X_i,X_j,X_s)\in\chi_{n,m}\times\chi_{n,m}\times\chi_{n,m})|\mathbf{I}_n\right]\\ +\mathbb{E}_{f_{X_j},f_{X_s}}\left[\widetilde{K}\left(X_s,X_i,X_j\right)\mathcal{I}((X_i,X_j,X_s)\in\chi_{n,m}\times\chi_{n,m}\times\chi_{n,m})|\mathbf{I}_n\right]\\ +\mathbb{E}_{f_{X_j},f_{X_s}}\left[\widetilde{K}\left(X_j,X_s,X_i\right)\mathcal{I}((X_i,X_j,X_s)\in\chi_{n,m}\times\chi_{n,m}\times\chi_{n,m})|\mathbf{I}_n\right]\\ +\mathbb{E}_{f_{X_j},f_{X_s}}\left[\widetilde{K}\left(X_s,X_j,X_i\right)\mathcal{I}((X_i,X_j,X_s)\in\chi_{n,m}\times\chi_{n,m}\times\chi_{n,m})|\mathbf{I}_n\right]\\ -6\mathbb{E}_f\left[\widetilde{K}\left(X_i,X_j,X_s\right)\mathcal{I}((X_i,X_j,X_s)\in\chi_{n,m}\times\chi_{n,m}\times\chi_{n,m})|\mathbf{I}_n\right]\end{array}\right\}^{2q}\Bigg|\mathbf{I}_n\right)$$

$$\leq C_q \frac{N_{n,m}^{2q}(N_{n,m}-1)^{2q}(N_{n,m}-2)^{2q}}{n^{2q}(n-1)^{2q}(n-2)^{2q}}\frac{N_{n,m}^{-\frac{2q}{2}}}{N_{n,m}}\times$$

$$\sum_{i=1}^{N_{n,m}}\left(\begin{array}{c}\mathbb{E}_f\left(\left\{\mathbb{E}_{f_{X_j},f_{X_s}}\left[\widetilde{K}\left(X_i,X_j,X_s\right)\mathcal{I}((X_i,X_j,X_s)\in\chi_{n,m}\times\chi_{n,m}\times\chi_{n,m})|\mathbf{I}_n\right]\right\}^{2q}|\mathbf{I}_n\right)\\ +\mathbb{E}_f\left(\left\{\mathbb{E}_{f_{X_j},f_{X_s}}\left[\widetilde{K}\left(X_i,X_s,X_j\right)\mathcal{I}((X_i,X_j,X_s)\in\chi_{n,m}\times\chi_{n,m}\times\chi_{n,m})|\mathbf{I}_n\right]\right\}^{2q}|\mathbf{I}_n\right)\\ +\mathbb{E}_f\left(\left\{\mathbb{E}_{f_{X_j},f_{X_s}}\left[\widetilde{K}\left(X_j,X_i,X_s\right)\mathcal{I}((X_i,X_j,X_s)\in\chi_{n,m}\times\chi_{n,m}\times\chi_{n,m})|\mathbf{I}_n\right]\right\}^{2q}|\mathbf{I}_n\right)\\ +\mathbb{E}_f\left(\left\{\mathbb{E}_{f_{X_j},f_{X_s}}\left[\widetilde{K}\left(X_s,X_i,X_j\right)\mathcal{I}((X_i,X_j,X_s)\in\chi_{n,m}\times\chi_{n,m}\times\chi_{n,m})|\mathbf{I}_n\right]\right\}^{2q}|\mathbf{I}_n\right)\\ +\mathbb{E}_f\left(\left\{\mathbb{E}_{f_{X_j},f_{X_s}}\left[\widetilde{K}\left(X_j,X_s,X_i\right)\mathcal{I}((X_i,X_j,X_s)\in\chi_{n,m}\times\chi_{n,m}\times\chi_{n,m})|\mathbf{I}_n\right]\right\}^{2q}|\mathbf{I}_n\right)\\ +\mathbb{E}_f\left(\left\{\mathbb{E}_{f_{X_j},f_{X_s}}\left[\widetilde{K}\left(X_s,X_j,X_i\right)\mathcal{I}((X_i,X_j,X_s)\in\chi_{n,m}\times\chi_{n,m}\times\chi_{n,m})|\mathbf{I}_n\right]\right\}^{2q}|\mathbf{I}_n\right)\end{array}\right)$$

$$\text{(A.4)}$$

Above, the last inequality follows from conditional Jensen's Inequality. A typical term in the above summand looks like

$$\frac{1}{p_{n,m}^{4q+1}}\int_{\chi_{n,m}}\left[\int_{\chi_{n,m}}\int_{\chi_{n,m}}\int\sum_{l_1}\sum_{l_2}\sum_{l_3}\left(\begin{array}{c}\psi_{l_1}^{k_1}(x)\psi_{l_1}^{k_1}(x_1)\psi_{l_2}^{k_2}(x)\psi_{l_2}^{k_2}(x_2)\\ \psi_{l_3}^{k_3}(x)\psi_{l_3}^{k_3}(x_3)f(x_2)f(x_3)\end{array}\right)dxdx_2dx_3\right]^{2q}f(x_1)dx_1$$

$$(A.5)$$

for some tuple k_1, k_2, k_3. To evaluate each of the above, we consider three different ordering of k_1, k_2, k_3. Let us consider the the integrals in the square bracket first. Since here x_1 is fixed, we look for the subinterval of resolution k_1 where x_1 lies. In each of the cases we look at the specific subinterval containing x_1 for every fixed x_1. Further this subinterval needs to intersect $\chi_{n,m}$. Say, we index that by k_1. Call this subinterval $S_{k_1}(x_1)$. Taking the integral inside the sum over location parameters, we have that each summand is bounded by $C(\psi_0, \|f\|_\infty) \frac{k_1 k_2 k_3}{(\max(k_1, k_2, k_3))^3}$. For each fixed x_1, the number of summands contributing will be bounded by $C_{\psi_0} \frac{\max(k_1, k_2, k_3)}{k_1}$. Therefore, whichever interval x must lie it should intersect that containing x_i. Therefore, the term in the square bracket is bounded by C_{ψ_0} for each fixed x_1. Raising to the power of $2q$ and integrating with respect to f over $\chi_{n,m}$ yields the outer integral to be bounded by $C(\psi_0, \|f\|_\infty) p_{n,m}$. Therefore, (A.5) is always bounded by $\frac{C(\psi_0, \|f\|_\infty)}{p_{n,m}^{4q}}$. Therefore, by (A.4) we have that almost surely,

$$\mathbb{E}_f \left((V_{n,m}^{(1)})^{2q} | \mathbf{I}_n \right) \leq C_q C(\psi_0, \|f\|_\infty) \frac{N_{n,m}^{2q}(N_{n,m} - 1)^{2q}(N_{n,m} - 2)^{2q}}{n^{2q}(n-1)^{2q}(n-2)^{2q}} \frac{N_{n,m}^{-\frac{2q}{2}}}{N_{n,m}} \times \frac{N_{n,m}}{p_{n,m}^{2q}}$$

$$(A.6)$$

Now, noting that $p_{n,m} \geq \frac{f_{\min}}{n}$ and using Lemma 5.6 of Robins et al. (2015), we have that

$$\mathbb{E}_f \mathbb{E}_f \left((V_{n,m}^{(1)})^{2q} | \mathbf{I}_n \right) \leq \frac{C_q(\|\psi_0\|_\infty, \|f\|_\infty, f_{\min})}{n^{2q}}, \qquad (A.7)$$

which is at the desired rate of control.

Control of $\mathbb{E}_\mathbf{f} \left((\mathbf{V}_{\mathbf{n,m}}^{(2)})^{2\mathbf{q}} \right)$.

For the second term of the U-statistics it is enough to control for the expectation of

$$I_{21}^{(m)} = \frac{(N_{n,m}(N_{n,m} - 1)(N_{n,m} - 1))^{2q}}{n^{2q}(n-1)^{2q}(n-2)^{2q}}$$

$$\times \left[\frac{1}{N_{n,m}} \left(\int_{\chi_{n,m}} \int_{\chi_{n,m}} (\mathbb{E}_{f_{X_i}}(\widetilde{K}(X_i, X_j, X_s) \mathcal{I}((X_i, X_j, X_s) \in \chi_{n,m} \times \chi_{n,m} \times \chi_{n,m}) | \mathbf{I}_n))^2 \frac{f(x_j) f(x_s)}{p_{n,m}^2} dx_j dx_s \right)^{\frac{2q}{2}} \right]$$

and

$$I_{22}^{(m)} = \frac{(N_{n,m}(N_{n,m}-1)(N_{n,m}-1))^{2q}}{n^{2q}(n-1)^{2q}(n-2)^{2q}}$$

$$\times \left[\frac{1}{N_{n,m}} \left(\int_{\chi_{n,m}} \int_{\chi_{n,m}} (\mathbb{E}_{f_{X_i}}(\widetilde{K}(X_i,X_j,X_s)\mathcal{I}((X_i,X_j,X_s) \in \chi_{n,m} \times \chi_{n,m} \times \chi_{n,m})|\mathbf{I}_n))^{2q} \frac{f(x_j)f(x_s)}{p_{n,m}^2} dx_j dx_s \right) \right]$$

Control of $I_{21}^{(m)}$.

First note that one can argue by simply looking at orders of truncation that,

$$\mathbb{E}_{f_{X_i}}(\widetilde{K}(X_i,X_j,X_s)\mathcal{I}((X_i,X_j,X_s) \in \chi_{n,m} \times \chi_{n,m} \times \chi_{n,m})|\mathbf{I}_n)$$

$$\asymp \frac{1}{p_{n,m}} \left[\begin{array}{l} \sum_{l_1}\sum_{l_2}\sum_{l_3} \int_{\chi_{n,m}} \int \psi_{l_1}^{k_1}(x)\psi_{l_1}^{k_1}(x_i)\psi_{l_2}^{k_3}(x)\psi_{l_2}^{k_3}(x_j)\psi_{l_3}^{k_3}(x)\psi_{l_3}^{k_3}(x_s)f(x_1)dxdx_1 \\ + \sum_{l_1}\sum_{l_2}\sum_{l_3} \int_{\chi_{n,m}} \int \psi_{l_1}^{k_3}(x)\psi_{l_1}^{k_3}(x_i)\psi_{l_2}^{k_1}(x)\psi_{l_2}^{k_1}(x_j)\psi_{l_3}^{k_3}(x)\psi_{l_3}^{k_3}(x_s)f(x_1)dxdx_1 \end{array} \right]$$

Let us control the first term in the square bracket first. For each fixed value of (x_j, x_s), find the boxes of length $\frac{1}{k_3}$ which contains x_j and x_s. If these boxes are disjoint then they contribute to the summand of the first term. Therefore, only way of getting a contribution in the first term is when x_j and x_s are in the same box of resolution k_3. Therefore, x must also belong to the same box of resolution k_3 to contribute to the summand. Therefore, the first term inside the square bracket is always bounded by $C(\psi_0, \|f\|_\infty)\frac{k_1 k_3^2}{k_3 M_n}$ for every choice (x_j, x_s) lying in the same k_3 resolution box and 0 otherwise. Therefore, the integral of the square of the first term over $(x_j, x_s) \in \chi_{n,m} \times \chi_{n,m}$ is bounded by $C(\psi_0, \|f\|_\infty)\left(\frac{k_1 k_3^2}{k_3 M_n}\right)^2 \frac{1}{k_3^2}\frac{k_3}{M_n} = C(\psi_0, \|f\|_\infty)\frac{k_1^2 k_3}{M_n^3}$. Now, let us look at the second term of the square bracket. Once again fix a $(x_j, x_s) \in \chi_{n,m} \times \chi_{n,m}$ and find the k_3 resolution box that x_s belongs to. Now note that x_i, x must belong to the same k_3 resolution box and hence the number of summands contributing to the sum in second term of the square bracket is again bounded by C_{ψ_0}. Therefore, the second term inside the square bracket is always bounded by $C(\psi_0, \|f\|_\infty)\frac{k_1 k_3^2}{k_3^2}$ for every choice (x_j, x_s) with x_s lying in the some k_3 resolution box and 0 otherwise. Therefore, the integral of the square of the second term over $(x_j, x_s) \in \chi_{n,m} \times \chi_{n,m}$ is bounded by $C(\psi_0, \|f\|_\infty)\left(\frac{k_1 k_3^2}{k_3^2}\right)^2 \frac{k_3}{M_n}\frac{1}{k_3}\frac{1}{M_n} = C(\psi_0, \|f\|_\infty)\frac{k_1^2}{M_n^2}$. Since, by our choice, $k_3 \gg M_n$, the first term of the square bracket dominates after squaring and integrating over over $(x_j, x_s) \in \chi_{n,m} \times \chi_{n,m}$. Taking into account the division by $p_{n,m}$'s, the final contribution to $I_{21}^{(m)}$ is bounded

$\left(C(\psi_0, \|f\|_\infty) \frac{k_1^2 k_3}{M_n^3} \right)^q \frac{1}{p_{n,m}^{4q}}$. Therefore,

$$I_{21}^{(m)} \leq C(\|\psi_0\|_\infty, \|f\|_\infty, f_{\min}) \, (N_{n,m}(N_{n,m}-1)(N_{n,m}-1))^{2q} \left(\frac{k_1^2 k_3}{M_n^3} \right)^q \frac{1}{n^{2q}}$$

$$\leq C(\|\psi_0\|_\infty, \|f\|_\infty, f_{\min}) \, (N_{n,m}(N_{n,m}-1)(N_{n,m}-1))^{2q} \left(\frac{k_3}{n^3} \right)^q.$$

$$(A.8)$$

Taking expectations by using Lemma 5.6 of Robins et al. (2015) yields the desired control.

Control of $I_{22}^{(m)}$.

The calculation technique of this term is similar to that of $I_{21}^{(m)}$. The only difference is that instead of taking square of the square bracket term we take the $2q^{\text{th}}$ power. Hence, the integral over $(x_j, x_s) \in \chi_{n,m} \times \chi_{n,m}$ is bounded by

$$C(\psi_0, \|f\|_\infty) \left[\left(\frac{k_1 k_3^2}{k_3 M_n} \right)^{2q} \frac{1}{k_3^2} \frac{k_3}{M_n} + \left(\frac{k_1 k_3^2}{k_3^2} \right)^{2q} \frac{k_3}{M_n} \frac{1}{k_3} \frac{1}{M_n} \right].$$

Taking into account the division by $p_{n,m}$'s, the final contribution to $I_{22}^{(m)}$ is bounded

$$C(\psi_0, \|f\|_\infty) \left[\left(\frac{k_1 k_3^2}{k_3 M_n} \right)^{2q} \frac{1}{k_3^2} \frac{k_3}{M_n} + \left(\frac{k_1 k_3^2}{k_3^2} \right)^{2q} \frac{k_3}{M_n} \frac{1}{k_3} \frac{1}{M_n} \right] \frac{1}{p_{n,m}^{2q+2}}.$$

Therefore,

$$I_{21}^{(m)} \leq C(\|\psi_0\|_\infty, \|f\|_\infty, f_{\min}) \, (N_{n,m}(N_{n,m}-1)(N_{n,m}-2))^{2q} \left[\left(\frac{k_1 k_3^2}{k_3 M_n} \right)^{2q} \frac{1}{k_3^2} \frac{k_3}{M_n} + \left(\frac{k_1 k_3^2}{k_3^2} \right)^{2q} \frac{k_3}{M_n} \frac{1}{k_3} \frac{1}{M_n} \right] \frac{1}{n^{4q-2}}$$

$$\leq C(\|\psi_0\|_\infty, \|f\|_\infty, f_{\min}) \, (N_{n,m}(N_{n,m}-1)(N_{n,m}-2))^{2q} \frac{1}{M_n} \left(\frac{k_3}{n^2} \right)^{2q-1}.$$

$$(A.9)$$

Taking expectations by using Lemma 5.6 of Robins et al. (2015) yields the desired control.

Combining Controls of $I_{21}^{(m)}$ and $I_{22}^{(m)}$ for control of $\mathbb{E}_{\mathbf{f}}\left((\mathbf{V_{n,m}^{(2)}})^{\mathbf{2q}} \right)$.

We get

$$\mathbb{E}_{\mathbf{f}}\left((\mathbf{V_{n,m}^{(2)}})^{\mathbf{2q}} \right) \leq C(\|\psi_0\|_\infty, \|f\|_\infty, f_{\min}) \left[\frac{1}{M_n} \left(\frac{k_3}{n^2} \right)^{2q-1} + \left(\frac{k_3}{n^3} \right)^q \right] \quad (A.10)$$

Control of $\mathbb{E}_{\mathbf{f}}\left((\mathbf{V}_{\mathbf{n,m}}^{\mathbf{(3)}})^{\mathbf{2q}}\right).$

For this part we will need a moment bound for third order U-statistics. We do that in the following Lemma.

LEMMA A.6. *For any $q > 2$, there exists a constant C_q such that for any i.i.d random variables $X_1, X_2, ..., X_m$ and degenerate symmetric kernel K.*

$$E\left|\frac{1}{m(m-1)(m-2)}\sum_{i_1\neq i_2\neq i_3} K(X_{i_1}, X_{i_2}, X_{i_3})\right|^q$$
$$\leq C_q m^{-q} E\left|K(X_{i_1}, X_{i_2}, X_{i_3})\right|^q$$

We will now apply Lemma A.6 to control moments of the degenerate part of the kernel $K(X_i, X_j, X_s) = \int K_{k_1}(x, X_i)K_{k_3}(x, X_j)K_{k_3}(x, X_s)I((X_i, X_j, X_s) \in \chi_{n,m} \times \chi_{n,m} \times \chi_{n,m})dx$. Standard arguments show that this is the term with the highest contribution among all the different terms of $V_{n,m}^{(3)}$. Denoting the degenerate part of this kernel by $K_d^{(m)}$ we can show by Lemma A.6 and a standard contraction of norm under conditional expectation argument that

$$\mathbb{E}_f\left((V_{n,m}^{(3)})^{2q} \mid \mathbf{I}_n\right)$$

$$\leq \frac{C(\psi_0, \|f\|_\infty)^q (N_{n,m}(N_{n,m}-1)(N_{n,m}-2))^{2q}}{n^{2q}(n-1)^{2q}(n-2)^{2q}}$$

$$\times \frac{E\left|K_d^{(m)}(X_1, X_2, X_3)\right|^{2q}}{N_{n,m}^{2q}}$$

$$\leq \frac{C(\psi_0, \|f\|_\infty)^q (N_{n,m}(N_{n,m}-1)(N_{n,m}-2))^{2q}}{n^{2q}(n-1)^{2q}(n-2)^{2q}}$$

$$\times \frac{E\left|\int K_{k_1}(x, X_1)K_{k_3}(x, X_2)K_{k_3}(x, X_3)I((X_i, X_j, X_s) \in \chi_{n,m} \times \chi_{n,m} \times \chi_{n,m})dx\right|^{2q}}{N_{n,m}^{2q}}$$

$$= \frac{C(\psi_0, \|f\|_\infty)^q ((N_{n,m}-1)(N_{n,m}-2))^{2q}}{n^{2q}(n-1)^{2q}(n-2)^{2q}}$$

$$\times E\left|\int K_{k_1}(x, X_1)K_{k_3}(x, X_2)K_{k_3}(x, X_3)I((X_i, X_j, X_s) \in \chi_{n,m} \times \chi_{n,m} \times \chi_{n,m})dx\right|^{2q}$$

$$\text{(A.11)}$$

Therefore it is enough to control

$$E\left|\int K_{k_1}(x, X_1)K_{k_3}(x, X_2)K_{k_3}(x, X_3)I((X_i, X_j, X_s) \in \chi_{n,m} \times \chi_{n,m} \times \chi_{n,m})dx\right|^{2q}$$

at a level so that the expectation of the right hand side of (A.11) is controlled at the desired level of $\left(\frac{k_3}{n^2}\right)^q$. We do this below.

Control of $\mathbf{E}\left|\int \mathbf{K_{k_1}}(\mathbf{x},\mathbf{X_1})\mathbf{K_{k_3}}(\mathbf{x},\mathbf{X_2})\mathbf{K_{k_3}}(\mathbf{x},\mathbf{X_3})\mathbf{I}((\mathbf{X_i},\mathbf{X_j},\mathbf{X_s})\in\chi_{\mathbf{n,m}}\times\chi_{\mathbf{n,m}}\times\chi_{\mathbf{n,m}})\mathbf{dx}\right|^{\mathbf{2q}}.$

$$E\left|\int K_{k_1}(x,X_1)K_{k_3}(x,X_2)K_{k_3}(x,X_3)I((X_i,X_j,X_s)\in\chi_{n,m}\times\chi_{n,m}\times\chi_{n,m})dx\right|^{2q}$$

$$\leq\int_{\chi_{n,m}}\int_{\chi_{n,m}}\int_{\chi_{n,m}}\left[\sum_{l_1}\sum_{l_2}\sum_{l_3}\int\left|\begin{array}{c}\psi_{l_1}^{k_1}(x)\psi_{l_1}^{k_1}(x_1)\psi_{l_2}^{k_3}(x)\\\psi_{l_2}^{k_3}(x_2)\psi_{l_3}^{k_3}(x)\psi_{l_3}^{k_3}(x_3)\end{array}\right|dx\right]^{2q}\frac{f(x)f(x_1)f(x_2)f(x_3)}{p_{n,m}^3}dx_1dx_2dx_3$$

Now, for each fixed value of (x_1,x_2,x_3), find the boxes of length $\frac{1}{k_3}$ which contains x_2 and x_3. Therefore, only way of getting a contribution is when x_2 and x_3 are in the same box of resolution k_3. Therefore, x must also belong to the same box of resolution k_3 to contribute to the summand. Therefore, the term inside the square bracket is always bounded by $\left(C_{\psi_0}\frac{k_1k_3^2}{k_3}\right)^{2q}$ for every choice (x_2,x_3) lying in the same k_3 resolution box and 0 otherwise. Therefore, the final integral with respect to (x_1,x_2,x_3) is bounded by

$$C(\psi_0,\|f\|_\infty)^q\frac{\left(\frac{k_1k_3^2}{k_3}\right)^{2q}}{p_{n,m}^3}\frac{1}{M_n}\frac{1}{k_3^2}\frac{k_3}{n}=C_q(\|\psi_0\|_\infty,\|f\|_\infty,f_{\min})k_3^{2q-1}n^{2q+1}.$$

Combining with (A.11). Therefore, from (A.11), we have that almost surely,

$$\mathbb{E}_f\left((V_{n,m}^{(3)})^{2q}\mid\mathbf{I}_n\right)$$

$$\leq\frac{C(\psi_0,\|f\|_\infty)^q\left((N_{n,m}-1)(N_{n,m}-2)\right)^{2q}}{n^{2q}(n-1)^{2q}(n-2)^{2q}}$$

$$\times E\left|\int K_{k_1}(x,X_1)K_{k_3}(x,X_2)K_{k_3}(x,X_3)I((X_i,X_j,X_s)\in\chi_{n,m}\times\chi_{n,m}\times\chi_{n,m})dx\right|^{2q}$$

$$\leq C_q(\|\psi_0\|_\infty,\|f\|_\infty,f_{\min})\frac{\left((N_{n,m}-1)(N_{n,m}-2)\right)^{2q}}{n^{2q}(n-1)^{2q}(n-2)^{2q}}k_3^{2q-1}n^{2q+1}$$

$$\leq C_q(\|\psi_0\|_\infty,\|f\|_\infty,f_{\min})\left((N_{n,m}-1)(N_{n,m}-2)\right)^{2q}\frac{k_3^{2q-1}}{n^{4q-1}}$$

$$(A.12)$$

Therefore, by Lemma 5.6 of Robins et al. (2015) we have that,

$$\mathbb{E}_f\left((V_{n,m}^{(3)})^{2q}\right) \le C_q(\|\psi_0\|_\infty, \|f\|_\infty, f_{\min})\frac{k_3^{2q-1}}{n^{4q-1}} = C_q(\|\psi_0\|_\infty, \|f\|_\infty, f_{\min})\left(\frac{k_3}{n^2}\right)^{2q-1}\frac{1}{n}$$
(A.13)

Combining (A.7), (A.10) and (A.13) and summing over $M_n = O(n)$ partitions we get the desired bound that

$$\sum_{m=1}^{M_n}\mathbb{E}_f\left(|V_{n,m}^{(1)} + V_{n,m}^{(2)} + V_{n,m}^{(3)}|^{2q}\right) \le C_q(\|\psi_0\|_\infty, \|f\|_\infty, f_{\min})\left(\frac{k_3}{n^2}\right)^{2q-1}.$$

Control of $\mathbb{E}_\mathbf{f}\left\{\sum_{\mathbf{m=1}}^{\mathbf{M_n}}\mathbb{E}_\mathbf{f}\left(|\mathbf{V}_{\mathbf{n,m}}^{(1)} + \mathbf{V}_{\mathbf{n,m}}^{(2)} + \mathbf{V}_{\mathbf{n,m}}^{(3)}|^\mathbf{2}|\mathbf{I_n}\right)\right\}^{\frac{\mathbf{2q}}{\mathbf{2}}}$.

Unlike the previous case, here we will will employ a conditional version of Gine-Latala inequality for moments of U-statistics along with the observation that the constants of the inequality does not depend on the underlying distribution. The only difference will be that we will have to do an extra step of taking the expectation of the moment of sum of functions of multinomial coordinate statistics. To deal with this we again use the Marcinkiewicz Zygmund Inequality following an use of Theorem 2 of Shao (2000). First, note that it is enough to control $\mathbb{E}_f\left\{\sum_{m=1}^{M_n}\mathbb{E}_f\left((V_{n,m}^{(1)})^2|\mathbf{I}_n\right)\right\}^{\frac{2q}{2}}$, $\mathbb{E}_f\left\{\sum_{m=1}^{M_n}\mathbb{E}_f\left((V_{n,m}^{(2)})^2|\mathbf{I}_n\right)\right\}^{\frac{2q}{2}}$, $\mathbb{E}_f\left\{\sum_{m=1}^{M_n}\mathbb{E}_f\left((V_{n,m}^{(3)})^2|\mathbf{I}_n\right)\right\}^{\frac{2q}{2}}$ separately at the desired rate.

Control of $\mathbb{E}_\mathbf{f}\left\{\sum_{\mathbf{m=1}}^{\mathbf{M_n}}\mathbb{E}_\mathbf{f}\left((\mathbf{V}_{\mathbf{n,m}}^{(1)})^\mathbf{2}|\mathbf{I_n}\right)\right\}^{\frac{\mathbf{2q}}{\mathbf{2}}}$.

Note that by (A.6),

$$\mathbb{E}_f\left((V_{n,m}^{(1)})^{2q}|\mathbf{I}_n\right) \le C(\psi_0, \|f\|_\infty)^q\frac{N_{n,m}^q(N_{n,m}-1)^{2q}(N_{n,m}-2)^{2q}}{p_{n,m}^{2q}n^{2q}(n-1)^{2q}(n-2)^{2q}}$$

$$\le C_q(\|\psi_0\|_\infty, \|f\|_\infty, f_{\min})\frac{N_{n,m}^q(N_{n,m}-1)^{2q}(N_{n,m}-2)^{2q}}{(n-1)^{2q}(n-2)^{2q}}$$

Therefore,

$$\mathbb{E}_f\left((V_{n,m}^{(1)})^2|\mathbf{I}_n\right) \le C_2(\|\psi_0\|_\infty, \|f\|_\infty, f_{\min})\frac{N_{n,m}(N_{n,m}-1)^2(N_{n,m}-2)^2}{(n-1)^2(n-2)^2}$$

Therefore, by Theorem 2 of Shao (2000),

$$\mathbb{E}_f\left\{\sum_{m=1}^{M_n}\mathbb{E}_f\left((V_{n,m}^{(1)})^2|\mathbf{I}_n\right)\right\}^{\frac{2q}{2}} \le C_2^*\mathbb{E}_f\left\{\sum_{m=1}^{M_n}\frac{N_{n,m}'(N_{n,m}'-1)^2(N_{n,m}'-2)^2}{n^4}\right\}^{\frac{2q}{2}}$$

where $N'_{n,m}$, $m = 1, \ldots, M_n$ are independent with the same marginals as $N_{n,m}$, $m = 1, \ldots, M_n$. Now,

$$
\mathbb{E}_f \left\{ \sum_{m=1}^{M_n} \frac{N'_{n,m}(N'_{n,m}-1)^2(N'_{n,m}-2)^2}{n^4} \right\}^{\frac{2q}{2}}
$$
$$
\leq C_q \left[\begin{array}{c} \mathbb{E}_f \left\{ \sum_{m=1}^{M_n} \left(\frac{N'_{n,m}(N'_{n,m}-1)^2(N'_{n,m}-2)^2}{n^4} - \mathbb{E}_f \left(\frac{N'_{n,m}(N'_{n,m}-1)^2(N'_{n,m}-2)^2}{n^4} \right) \right) \right\}^q \\ + \left\{ \sum_{m=1}^{M_n} \mathbb{E}_f \left(\frac{N'_{n,m}(N'_{n,m}-1)^2(N'_{n,m}-2)^2}{n^4} \right) \right\}^q \end{array} \right]
$$
(A.14)

The first term in the above summand can be controlled by Marcinkiewicz Zygmund Inequality as follows,

$$
\mathbb{E}_f \left\{ \sum_{m=1}^{M_n} \left(\frac{N'_{n,m}(N'_{n,m}-1)^2(N'_{n,m}-2)^2}{n^4} - \mathbb{E}_f \left(\frac{N'_{n,m}(N'_{n,m}-1)^2(N'_{n,m}-2)^2}{n^4} \right) \right) \right\}^q
$$
$$
\leq C_q M_n^q M_n^{-\frac{q}{2}} \frac{1}{M_n} \sum_{m=1}^{M_n} \mathbb{E}_f \left(\frac{N'_{n,m}(N'_{n,m}-1)^2(N'_{n,m}-2)^2}{n^4} - \mathbb{E}_f \left(\frac{N'_{n,m}(N'_{n,m}-1)^2(N'_{n,m}-2)^2}{n^4} \right) \right)^q
$$
$$
\leq C'_q M_n^{\frac{q}{2}-1} \sum_{m=1}^{M_n} \mathbb{E}_f \left\{ \left(\frac{N'_{n,m}(N'_{n,m}-1)^2(N'_{n,m}-2)^2}{n^4} \right) \right\}^q
$$
(A.15)

Therefore, it is enough to control $\mathbb{E}_f \left\{ \left(\frac{N'_{n,m}(N'_{n,m}-1)^2(N'_{n,m}-2)^2}{n^4} \right) \right\}^q$ for this part of the proof. But, by Lemma 5.6 of Robins et al. (2015), we have that there exists a constant $C_q(\|f\|_\infty, \psi_0)$ such that

$$
\mathbb{E}_f \left\{ \left(\frac{N'_{n,m}(N'_{n,m}-1)^2(N'_{n,m}-2)^2}{n^4} \right) \right\}^q \leq \frac{C_q(\|f\|_\infty, \psi_0)}{n^{4q}}. \quad (A.16)
$$

By a similar argument,

$$
\mathbb{E}_f \left(\frac{N'_{n,m}(N'_{n,m}-1)^2(N'_{n,m}-2)^2}{n^4} \right) \leq \frac{C_q(\|f\|_\infty, \psi_0)}{n^4} \quad (A.17)
$$

Combining (A.14) with (A.15), (A.16) and (A.17), we have,

$$
\mathbb{E}_f \left\{ \sum_{m=1}^{M_n} \frac{N'_{n,m}(N'_{n,m}-1)^2(N'_{n,m}-2)^2}{n^4} \right\}^{\frac{2q}{2}}
$$
$$
\leq C_q \left[M_n^{\frac{q}{2}-1} M_n \frac{C_q(\|f\|_\infty, \psi_0)}{n^{4q}} + \left(M_n \frac{C_q(\|f\|_\infty, \psi_0)}{n^4} \right)^q \right] \leq C'_q(\|f\|_\infty, \psi_0) \frac{k_3}{n^2}.
$$

This completes the control of $\mathbb{E}_f \left\{ \sum_{m=1}^{M_n} \mathbb{E}_f \left((V_{n,m}^{(1)})^2 | \mathbf{I}_n \right) \right\}^{\frac{2q}{2}}$.

Control of $\mathbb{E}_{\mathbf{f}}\left\{\sum_{\mathbf{m}=\mathbf{1}}^{\mathbf{M_n}} \mathbb{E}_{\mathbf{f}}\left((\mathbf{V}_{\mathbf{n},\mathbf{m}}^{(\mathbf{2})})^{\mathbf{2}}|\mathbf{I_n}\right)\right\}^{\frac{\mathbf{2q}}{\mathbf{2}}}.$

Note that by (A.8) and (A.9),

$$
\mathbb{E}_f\left\{\sum_{m=1}^{M_n} \mathbb{E}_f\left((V_{n,m}^{(2)})^2|I_n\right)\right\}^{\frac{2q}{2}}
$$

$$
\leq C_q(\|\psi_0\|_\infty, \|f\|_\infty, f_{\min})\mathbb{E}_f\left[
\begin{array}{c}
\left(\frac{k_3}{n^3}\right)\sum_{m=1}^{M_n}\left(N_{n,m}(N_{n,m}-1)(N_{n,m}-2)\right)^2 \\
+\left(\frac{k_3}{n^2}\right)\frac{1}{M_n}\sum_{m=1}^{M_n}\left(N_{n,m}(N_{n,m}-1)(N_{n,m}-2)\right)^2
\end{array}
\right]^q
$$

$$
\leq C_q(\|\psi_0\|_\infty, \|f\|_\infty, f_{\min})\mathbb{E}_f\left[
\begin{array}{c}
\left(\frac{k_3}{n^3}\right)\sum_{m=1}^{M_n}\left(N'_{n,m}(N'_{n,m}-1)(N'_{n,m}-2)\right)^2 \\
+\left(\frac{k_3}{n^2}\right)\frac{1}{M_n}\sum_{m=1}^{M_n}\left(N'_{n,m}(N'_{n,m}-1)(N'_{n,m}-2)\right)^2
\end{array}
\right]^q
$$

$$(A.18)$$

where the last inequality in the above display is again by Theorem 2 of Shao (2000), with $N'_{n,m}$, $m = 1,\ldots,M_n$ are independent with the same marginals as $N_{n,m}$, $m = 1,\ldots,M_n$. Proceeding as the last section, we can apply the Marcinkiewicz Zygmund Inequality to a centered version of the right hand side of above and obtain that

$$
\mathbb{E}_f\left\{\sum_{m=1}^{M_n} \mathbb{E}_f\left((V_{n,m}^{(2)})^2|I_n\right)\right\}^{\frac{2q}{2}}
$$

$$
\leq C_q(\|\psi_0\|_\infty, \|f\|_\infty, f_{\min})\mathbb{E}_f\left[
\begin{array}{c}
\left(\frac{k_3}{n^3}\right)\sum_{m=1}^{M_n}\left(N'_{n,m}(N'_{n,m}-1)(N'_{n,m}-2)\right)^2 \\
+\left(\frac{k_3}{n^2}\right)\frac{1}{M_n}\sum_{m=1}^{M_n}\left(N'_{n,m}(N'_{n,m}-1)(N'_{n,m}-2)\right)^2
\end{array}
\right]^q
$$

$$
\leq C_q(\|\psi_0\|_\infty, \|f\|_\infty, f_{\min})\left[
\begin{array}{c}
\left(\frac{k_3}{n^3}\right)^q\left\{M_n^q M_n^{-\frac{q}{2}-1} + M_n^q\right\} \\
+\left(\frac{k_3}{n^2}\right)^q\left\{M_n^{-\frac{q}{2}-1} + 1\right\}
\end{array}
\right]
$$

$$
\leq C_q(\|\psi_0\|_\infty, \|f\|_\infty, f_{\min})\left(\frac{k_3}{n^2}\right)^q
$$

Control of $\mathbb{E}_{\mathbf{f}}\left\{\sum_{\mathbf{m}=\mathbf{1}}^{\mathbf{M_n}} \mathbb{E}_{\mathbf{f}}\left((\mathbf{V}_{\mathbf{n},\mathbf{m}}^{(\mathbf{3})})^{\mathbf{2}}|\mathbf{I_n}\right)\right\}^{\frac{\mathbf{2q}}{\mathbf{2}}}.$

Note that by (A.12),

$$
\mathbb{E}_f\left((V_{n,m}^{(3)})^2 \mid I_n\right) \leq C_2(\|\psi_0\|_\infty, \|f\|_\infty, f_{\min})\left((N_{n,m}-1)(N_{n,m}-2)\right)^{2q}\frac{k_3}{n^3}
$$

which is similar to the first term of (A.18) which we controlled earlier.

$\square$

Proof of Theorem 5.2.

PROOF. The proof is a simple application of Theorem 4.1 with slight modification to account for the fact that the estimator depends on three truncation points (k_1, k_2, k_3) instead of single level of truncation parametrization considered in Theorem 4.1. However, as noted in equations (5.1), 5.2, and Theorem 5.1, the required problem can be equivalently parametrized by the highest order of truncation i.e. k_3. Since, k_3 decides a corresponding smoothness index β as the solution of $k_3 = n^{\frac{1}{1+4\beta}}$, this also decides the other two levels of truncation. As a result, the proof of Theorem 4.1 goes through in a similar fashion. We do not provide the details here for the sake of brevity. $\square$

APPENDIX B: PROOF OF LEMMAS

Proof of Lemma A.1.

PROOF.

$$\mathbb{P}_f\left(\mathcal{I}(\hat{j} = 1)\right) \leq \mathbb{P}_f\left[|\hat{I}(k_0, k_*) - I(k_0, k_*)| \geq C\sqrt{\log n}\frac{\sqrt{k_*}}{n}\right]$$

since there exists a constant C_ψ such that $I(k_0, k_*) \leq C_\psi k^{-2\beta_0} \ll \frac{\sqrt{k_*}}{n}$. Now, by Hoeffding decomposition, we have

$$
\begin{aligned}
&\hat{I}(k_0, k_*) - I(k_0, k_*) \\
&= U_{n,*} - U_{n,0} - \mathbb{E}_f\left(U_{n,*} - U_{n,0}\right) \\
&= \frac{1}{n(n-1)} \sum_{i \neq j} [K_*(X_i, X_j) - K_0(X_i, X_j) - \mathbb{E}_f\left(K_*(X_i, X_j) - K_0(X_i, X_j)\right)] \\
&= \frac{2}{n} \sum_{i=1}^n \left[\mathbb{E}_{X_j}\left(K_*(X_i, X_j) - K_0(X_i, X_j)\right) - \mathbb{E}_f\left(K_*(X_i, X_j) - K_0(X_i, X_j)\right)\right] \\
&+ \frac{1}{n(n-1)} \sum_{i \neq j} \left[\begin{array}{c} K_*(X_i, X_j) - K_0(X_i, X_j) - \mathbb{E}_{X_j}\left(K_*(X_i, X_j) - K_0(X_i, X_j)\right) \\ -\mathbb{E}_{X_i}\left(K_*(X_i, X_j) - K_0(X_i, X_j)\right) + \mathbb{E}_f\left(K_*(X_i, X_j) - K_0(X_i, X_j)\right) \end{array}\right] \\
&:= T_1 + T_2
\end{aligned}
$$

Therefore,

$$
\begin{aligned}
&\mathbb{P}_f\left[|\hat{I}(k_0, k_*) - I(k_0, k_*)| \geq C\sqrt{\log n}\frac{\sqrt{k_*}}{n}\right] \\
&= \mathbb{P}_f\left[|T_1 + T_2| \geq C\sqrt{\log n}\frac{\sqrt{k_*}}{n}\right]
\end{aligned}
$$

$$\leq \mathbb{P}_f\left[|T_1| \geq C\delta_n\sqrt{\log n}\frac{\sqrt{k_*}}{n}\right] + \mathbb{P}_f\left[|T_2| \geq C(1-\delta_n)\sqrt{\log n}\frac{\sqrt{k_*}}{n}\right]$$

for some sequence $\delta_n \geq 0$ to be specified later.

Control of $\mathbb{P}_f\left[|T_1| \geq C\delta_n\sqrt{\log n}\frac{\sqrt{k_*}}{n}\right]$.

$$\mathbb{P}_f\left[|T_1| \geq C\delta_n\sqrt{\log n}\frac{\sqrt{k_*}}{n}\right]$$
$$= \mathbb{P}_f\left[|\frac{2}{n}\sum_{i=1}^n\left[\mathbb{E}_{X_j}(K_*(X_i,X_j) - K_0(X_i,X_j)) - \mathbb{E}_f(K_*(X_i,X_j) - K_0(X_i,X_j))\right]| \geq C\delta_n\sqrt{\log n}\frac{\sqrt{k_*}}{n}\right]$$

Let $R_i = \mathbb{E}_{X_j}(K_*(X_i,X_j) - K_0(X_i,X_j))$. Then the above display can be bounded from above by Markov's Inequality as follows,

$$\mathbb{P}_f\left[\frac{2}{n}|\sum_{i=1}^n(R_i - \mathbb{E}_f(R_i))| \geq c\delta_n\sqrt{\log n}\frac{\sqrt{k_*}}{n}\right]$$
$$\leq \frac{\mathbb{E}_f\left[\sum_{i=1}^n|R_i - \mathbb{E}_f(R_i)|\right]^{2r}}{\left(c\delta_n\sqrt{\log n}\frac{\sqrt{k_*}}{n}n\right)^{2r}}$$
$$\leq C_r n^{2r}\frac{1}{n^r}\sum_{i=1}^n\mathbb{E}_f\left[|R_i - \mathbb{E}_f(R_i)|^{2r}\right]\frac{1}{\left(c\delta_n\sqrt{\log n}\frac{\sqrt{k_*}}{n}n\right)^{2r}}$$
$$\leq C_r n^{2r}\frac{n}{n^r}\mathbb{E}_f\left[|R_i|^{2r}\right]\frac{1}{\left(c\delta_n\sqrt{\log n}\frac{\sqrt{k_*}}{n}n\right)^{2r}}$$

(B.1)

where the last two lines is by applications of Marcinkiewicz Zygmund Inequality followed by Jensen's Inequality. Next, we evaluate $\mathbb{E}_f(|R|^{2r})$.

$$\mathbb{E}_f(|R|^{2r}) = \mathbb{E}_{f_{X_1}}\left[\left\{\mathbb{E}_{f_{X_2}}(K_*(X_1,X_2) - K_0(X_1,X_2))\right\}^{2r}\right]$$
$$\leq 2^{2r-1}\left[\mathbb{E}_{f_{X_1}}\left\{\mathbb{E}_{f_{X_2}}(K_*(X_1,X_2))\right\}^{2r} + \mathbb{E}_{f_{X_1}}\left\{\mathbb{E}_{f_{X_2}}(K_0(X_1,X_2))\right\}^{2r}\right]$$

Let us evaluate $\mathbb{E}_{f_{X_1}}\left\{\mathbb{E}_{f_{X_2}}(K_k(X_1,X_2))\right\}^{2r}$ for a general k. We first consider,

$$\mathbb{E}_{f_{X_2}}(K_k(X_1,X_2)) = \sum_{l=1}^k\psi_{k,l}(X_1)\mathbb{E}_{f_{X_2}}(\psi_{k,l}(X_2))$$

$$= \sum_{l=1}^{k} \psi_{k,l}(X_1)\alpha_{k,l} \leq (\sup_l \alpha_{k,l}) \sum_{l=1}^{k} |\psi_{k,l}(X_1)|$$

$$\leq C_{\psi_0} \|f\|_\infty^2$$

where the last line follows by Lemma C.2. Therefore, for some constant $C_{\psi_0} > 0$,

$$\mathbb{E}_{f_{X_1}} \left\{ \mathbb{E}_{f_{X_2}} \left(K_k(X_1, X_2) \right) \right\}^{2r} \leq C_{\psi_0}^r \|f\|_\infty^2.$$

Therefore from equation B.1, we have

$$\mathbb{P}_f \left[\frac{2}{n} | \sum_{i=1}^{n} (R_i - \mathbb{E}_f(R_i)) | \geq c\delta_n \sqrt{\log n} \frac{\sqrt{k_*}}{n} \right]$$

$$\leq C_r n^{2r} \frac{n}{n^r} \mathbb{E}_f \left[|R|^{2r} \right] \frac{1}{\left(c\delta_n \sqrt{\log n} \frac{\sqrt{k_*}}{n} n \right)^{2r}}$$

$$\leq C_{\psi_0}^r \|f\|_\infty^2 \frac{n^{r+1}}{\left(c\delta_n \sqrt{\log n} \sqrt{k_*} \right)^{2r}} \leq \frac{C_{\psi_0}^r \|f\|_\infty^2}{n}$$

by choosing r large enough and $\delta_n > 0$ at most a sub-algebraic sequence going to 0. To see this note that $\frac{n^{r+1}}{\left(\delta_n \sqrt{\log n} \sqrt{k_*} \right)^{2r}} = \frac{1}{\left(\delta_n \sqrt{\log n} \right)^{2r}} \frac{(\log n)^{\frac{r}{1+4\beta_1}}}{n^{-(r+1)+\frac{2r}{1+4\beta_1}}}$. Choosing $r \geq \frac{2}{1-4\beta_1}$ yields the desire result.

Control of $\mathbb{P}_{\mathbf{f}} \left[|\mathbf{T_2}| \geq \mathbf{C}(\mathbf{1} - \delta_{\mathbf{n}})\sqrt{\log \mathbf{n}} \frac{\sqrt{\mathbf{k_*}}}{\mathbf{n}} \right]$. Lemma C.4 and Lemma C.5 imply that for some deterministic constant C_{ψ_0} and all $t > 0$,

$$\mathbb{P}_f \left[|\frac{1}{n(n-1)} \sum_{i \neq j} H(X_i, X_j)| > \frac{C_{\psi_0}}{n-1} \left(\sqrt{k_* t} + t + \sqrt{\frac{k_*}{n}} t^{\frac{3}{2}} + \frac{k_*}{n} t^2 \right) \right] \leq 6e^{-t}$$

Using, $2t^{\frac{3}{2}} \leq t + t^2$ we have,

$$\mathbb{P}_f \left[|T_2| > \frac{C_{\psi_0}}{n-1} \left(\sqrt{k_* t} + t + \sqrt{\frac{k_*}{4n}} (t + t^2) + \frac{k_*}{n} t^2 \right) \right] \leq 6e^{-t}$$

Setting $t = \log n$ we have,

$$\mathbb{P}_f \left[|T_2| > \frac{C_{\psi_0}}{n-1} \left(\sqrt{k_* \log n} + \log n + \sqrt{\frac{k_*}{4n}} (\log n + (\log n)^2) + \frac{k_*}{n} (\log n)^2 \right) \right] \leq \frac{6}{n}$$

Now, since $\frac{\sqrt{k_* \log n}}{n-1} \geq \max \left\{ \frac{\log n}{n-1}, \frac{1}{n-1}\sqrt{\frac{k_*}{4n}}(\log n + (\log n)^2), \frac{k_*}{n(n-1)}(\log n)^2 \right\}$ and $\frac{1}{n-1} > \frac{1}{2n}$ for sufficiently large n, we have,

$$\mathbb{P}_f \left[|T_2| > 2C_{\psi_0} \frac{\sqrt{k_* \log n}}{n} \right] \leq \frac{6}{n}$$

This is enough to prove the desired result and ends the proof of Lemma A.1. $\qquad \square$

Proof of Lemma A.2.

$$
\begin{aligned}
\mathbb{E}_f \left[(U_{n,1} - \phi(f))^{2q} \right] &\leq 2^{2q-1}\mathbb{E}_f \left[(U_{n,1} - \mathbb{E}_f(U_{n,1}))^{2q} \right] + 2^{2q-1}\mathbb{E}_f \left[(\mathbb{E}_f(U_{n,1} - \phi(f)))^{2q} \right] \\
&\leq 2^{2q-1}\mathbb{E}_f \left[(U_{n,1} - \mathbb{E}_f(U_{n,1}))^{2q} \right] + C_{\psi_0}^{2q}\|f\|_\infty^{2q} k_1^{-2q\beta_0} \\
&\leq 2^{2q-1}\mathbb{E}_f \left[(U_{n,1} - \mathbb{E}_f(U_{n,1}))^{2q} \right] + C_{\psi_0}^{2q}\|f\|_\infty^{2q} \left(\frac{k_1}{n^2} \right)^q
\end{aligned}
$$

where the second last inequality follows from Lemma C.1. Therefore, it is enough to show that

$$\sup_{f \in H(\beta_0)} \mathbb{E}_f \left[(U_{n,1} - \mathbb{E}_f(U_{n,1}))^{2q} \right] \leq C_{\psi_0}^q \|f\|_\infty^{Cq} \left(\frac{k_1}{n^2} \right)^q \qquad \text{(B.2)}$$

We will give two proofs of inequality B.2. The first will be valid for any kernels based on compactly supported wavelet bases. The second will be valid only for Haar wavelets.

First Proof of Inequality B.2.

$$
\begin{aligned}
&\mathbb{E}_f \left[\left(U_{n,1} - \mathbb{E}_f(U_{n,1})^{2q} \right) \right] \\
&\leq C_q \left\{ \begin{array}{c} \mathbb{E}_f \left[\left(\frac{2}{n}\sum_{i=1}^n \left\{ \mathbb{E}_{f_{X_j}}(K_1(X_i, X_j)) - \mathbb{E}_f(K_1(X_i, X_j)) \right\} \right)^{2q} \right] \\ + \mathbb{E}_f \left[\left(\frac{1}{n(n-1)}\sum_{i \neq j} H_1(X_i, X_j) \right)^{2q} \right] \end{array} \right\} \\
&= C_q(I_1 + I_2)
\end{aligned}
$$

where $H_1(X_i, X_j) = K_1(X_i, X_j) - \mathbb{E}_{f_{X_i}}(K_1(X_i, X_j)) - \mathbb{E}_{f_{X_j}}(K_1(X_i, X_j)) + \mathbb{E}_f(K_1(X_i, X_j))$. We first bound I_1 by Marcinkiewicz Zygmund Inequality.

44

To this end, first write $R_i = \mathbb{E}_{f_{X_j}}(K_1(X_i, X_j))$. Therefore, by Marcinkiewicz
Zygmund Inequality and Jensen's Inequality, we have,

$$I_1 = 2^{2q}\mathbb{E}_f\left[\left(\frac{1}{n}\sum_{i=1}^{n}\{R_i - \mathbb{E}_f(R_i)\}\right)^{2q}\right]$$

$$\leq C_q n^{-2q/2}\frac{1}{n}\sum_{i=1}^{n}\mathbb{E}_f\left[(R - i - \mathbb{E}_f(R_i))^{2q}\right]$$

$$\leq 2C_q n^{-q}\mathbb{E}_f(|R_1|^{2q})$$

Therefore, it is enough to control $\mathbb{E}_f(|R_i|^{2q}) = \mathbb{E}_{f_{X_i}}\left[|\ Ej\ \{K_1(X_i, X_j)\}\ |^{2q}\right]$.

$$|\ Ej\ \{K_1(X_i, X_j)\}\ | = |\sum_{l=1}^{k}\psi_{l,k}(X_i)\mathbb{E}_{f_{X_j}}(\psi_{l,k}(X - j))\ |$$

$$= |\sum_{l=1}^{k}\alpha_{l,k}\psi_{l,k}(X_i)\ | \leq \sup(|\alpha_{l,k}|)\sum_{l=1}^{k}|\psi_{l,k}(X_i)|$$

$$\leq C_{\psi_0}\|f\|_\infty\frac{1}{\sqrt{k_1}}C_{\psi_0}\|f\|_\infty\sqrt{k_1} = C_{\psi_0}^2\|f\|_\infty^2$$

by Lemma C.2. Therefore,

$$\mathbb{E}_f(|R_i|^{2q}) \leq C_{\psi_0}^2\|f\|_\infty^2$$

as well. This in turn implies

$$I_1 \leq 2C_q n^{-q}\mathbb{E}_f(|R_1|^{2q}) \leq 2C_q C_{\psi_0}^2\|f\|_\infty^2 n^{-q} \leq 2C_q C_{\psi_0}^2\|f\|_\infty^2\left(\frac{k_1}{n^2}\right)^q$$

as required. Therefore, it is enough to show that

$$I_2 \leq C_{\psi_0}^q\|f\|_\infty^{Cq}\left(\frac{k_1}{n^2}\right)^q$$

i.e. we need to control $\mathbb{E}_f\left(\frac{1}{n(n-1)}\sum_{i\neq j}H_1(X_i, X_j)\right)^{2q}$ where $H_1(X_i, X_j) = K_1(X_i, X_j) - \mathbb{E}_{f_{X_j}}(K_1(X_i, X_j)) - \mathbb{E}_{f_{X_i}}(K_1(X_i, X_j)) + \mathbb{E}_f(K_1(X_i, X_j))$. Following the arguments of Lemma C.4 and Lemma C.5, we can also show that for all $t > 0$,

$$\mathbb{P}_f\left[|\frac{1}{n(n-1)}\sum_{i\neq j}H_1(X_i, X_j)| > \frac{C_{\psi_0}}{n-1}\left(\sqrt{k_1 t} + t + \sqrt{\frac{k_1}{n}}t^{\frac{3}{2}} + \frac{k_1}{n^2}t^2\right)\right] \leq 6e^{-t}.$$

Using $2t^{\frac{3}{2}} \le t^2 + t$, we have,

$$\mathbb{P}_f \left[|\frac{1}{n(n-1)} \sum_{i \ne j} H_1(X_i, X_j)| > \frac{C_{\psi_0}}{n-1} \left(\sqrt{k_1 t} + t \left(\sqrt{\frac{k_1}{4n}} + 1 \right) + \left(\sqrt{\frac{k_1}{4n}} + \frac{k_1}{n^2} \right) t^2 \right) \right] \le 6e^{-t}.$$

Calling, $\frac{1}{n(n-1)} \sum_{i \ne j} H_1(X_i, X_j) = Z$, we have an exponential inequality of the form,

$$\mathbb{P}_f \left[|Z| > a_1 \sqrt{t} + a_2 t + a_3 t^2 \right] \le 6e^{-t},$$

with $a_1 = \frac{C_{\psi_0}}{n-1} \sqrt{k_1}, a_2 = \frac{C_{\psi_0}}{n-1} \left(\sqrt{\frac{k_1}{4n}} + 1 \right), a_3 = \frac{C_{\psi_0}}{n-1} \left(\sqrt{\frac{k_1}{4n}} + \frac{k_1}{n^2} \right)$. From this we need to estimate,

$$\mathbb{E}_f(|Z|^{2q}) = 2q \int_0^\infty x^{2q-1} \mathbb{P}_f(|Z| \ge x) dx.$$

We do this as follows. Suppose, $t(x)$ solve for $a_1 \sqrt{t(x)} + a_2 t(x) + a_3 t^2(x) = x$. Now, trivially $a_1 \sqrt{t} + a_2 t + a_3 t^2$ is an increasing function of $t \in \mathbb{R}^+$. Therefore, if we can find a function $h(x) \ge 0$ such that $a_1 \sqrt{h(x)} + a_2 h(x) + a_3 h^2(x) \le x$ for all $x \ge 0$, then $h(x) \le t(x)$ for all $x \ge 0$. It is not too difficult to see that one such $h(x)$ is given by $h(x) = b_1 x^2 \wedge b_2 x \wedge b_3 \sqrt{x}$ where for $b_1 = \frac{c_1}{a_1^2}, b_2 = \frac{c_2}{a_2}, b_3 = \frac{c_3}{\sqrt{a_3}}$ for fixed constants $c_1, c_2, c_3 > 0$. Therefore,

$$\begin{aligned}
\mathbb{E}_f(|Z|^{2q}) &= 2q \int_0^\infty x^{2q-1} \mathbb{P}_f(|Z| \ge x) dx \\
&\le 2q \int_0^\infty x^{2q-1} \mathbb{P}_f(|Z| \ge a_1 \sqrt{h(x)} + a_2 h(x) + a_3 h^2(x)) dx \\
&\le 12q \int_0^\infty x^{2q-1} e^{-h(x)} dx \\
&= 12q \int_0^\infty x^{2q-1} e^{-\{b_1 x^2 \wedge b_2 x \wedge b_3 \sqrt{x}\}} dx \\
&\le 12q \left[\int_0^\infty x^{2q-1} e^{-b_1 x^2} dx + \int_0^\infty x^{2q-1} e^{-b_2 x} dx + \int_0^\infty x^{2q-1} e^{-b_3 \sqrt{x}} dx \right] \\
&= 12q \left(\frac{\Gamma(q)}{2b_1^q} + \frac{\Gamma(2q)}{b_2^{2q}} + \frac{2\Gamma(4q)}{b_3^{4q}} \right) \le C_{\psi_0}^q \left(\frac{k_1}{n^2} \right)^q
\end{aligned}$$

by our choices of b_1, b_2, b_3. This completes the first proof.

Proof of Lemma A.3. Pick any $l \le l_c(C^*)$. Then

$$\mathbb{P}_f(\hat{j} = l) \le \mathbb{P}_f(\hat{I}^2(k_l^*, k_j^*) \le C_{\text{opt}}^2 \log n R(k_j^*))$$

$$= \mathbb{P}_f \left(\hat{I}(k_l^*, k_j^*) - I(k_l^*, k_j^*) \leq C_{\text{opt}} \sqrt{\log n} \sqrt{R(k_j^*)} - I(k_l^*, k_j^*) \right)$$

$$- \mathbb{P}_f \left(\hat{I}(k_l^*, k_j^*) - I(k_l^*, k_j^*) < -C_{\text{opt}} \sqrt{\log n} \sqrt{R(k_j^*)} - I(k_l^*, k_j^*) \right)$$

$$\tag{B.3}$$

Recall that by $l \leq l_c(C^*)$, we have that $I^2(k_l^*, k_j^*) > C^* \log n R(k_j^*)$. We consider two cases.

Case 1: $\mathbf{I(k_l^*, k_j^*) > \sqrt{C^* \log n R(k_j^*)}}$.

In this case we have that the right hand side of (B.3) is bounded by

$$\mathbb{P}_f \left(\hat{I}(k_l^*, k_j^*) - I(k_l^*, k_j^*) \leq C_{\text{opt}} \sqrt{\log n} \sqrt{R(k_j^*)} - \sqrt{C^* \log n R(k_j^*)} \right)$$

$$+ \mathbb{P}_f \left(\hat{I}(k_l^*, k_j^*) - I(k_l^*, k_j^*) < -C_{\text{opt}} \sqrt{\log n} \sqrt{R(k_j^*)} - \sqrt{C^* \log n R(k_j^*)} \right)$$

$$\leq \frac{C_{\psi_0}}{n},$$

where the last line follows from the proof of Lemma A.1 along the lines of Lemma C.4 and C.5, provided C_{opt} is according to the constant specified in Theorem 3.1 and $C^* > 4C_{\text{opt}}^2$.

Case 2: $\mathbf{I(k_l^*, k_j^*) < -\sqrt{C^* \log n R(k_j^*)}}$.

In this case we have that the right hand side of (B.3) equals,

$$1 - \mathbb{P}_f \left(\hat{I}(k_l^*, k_j^*) - I(k_l^*, k_j^*) > C_{\text{opt}} \sqrt{\log n} \sqrt{R(k_j^*)} - \sqrt{C^* \log n R(k_j^*)} \right)$$

$$- 1 + \mathbb{P}_f \left(\hat{I}(k_l^*, k_j^*) - I(k_l^*, k_j^*) \geq -C_{\text{opt}} \sqrt{\log n} \sqrt{R(k_j^*)} - \sqrt{C^* \log n R(k_j^*)} \right)$$

$$\leq \frac{C_{\psi_0}}{n},$$

where the last line once again follows from the proof of Lemma A.1 along the lines of Lemma C.4 and C.5, provided C_{opt} is according to the constant specified in Theorem 3.1 and $C^* > 4C_{\text{opt}}^2$.

Proof of Lemma A.4.

PROOF.

$$\mathbb{P}_f(\hat{j} \geq j + 2) \leq \mathbb{P}_f \left(\exists l > j : \hat{I}^2(k_{j+1}^*, k_l^*) > C_{\text{opt}}^2 \log n R(k_l^*) \right)$$

$$\leq \sum_{l=j+2}^{N-1} \mathbb{P}_f \left(\hat{I}^2(k_{j+1}^*, k_l^*) > C_{\text{opt}}^2 \log n R(k_l^*) \right)$$

Now, for any $l \geq j + 2$,

$$\mathbb{P}_f \left(\hat{I}^2(k_{j+1}^*, k_l^*) > C_{\text{opt}}^2 \log n R(k_l^*) \right)$$

$$= \mathbb{P}_f \left(\hat{I}(k_{j+1}^*, k_l^*) - I(k_{j+1}^*, k_l^*) > C_{\text{opt}} \sqrt{\log n} \sqrt{R(k_l^*)} - I(k_{j+1}^*, k_l^*) \right)$$

$$+ \mathbb{P}_f \left(\hat{I}(k_{j+1}^*, k_l^*) - I(k_{j+1}^*, k_l^*) < -C_{\text{opt}} \sqrt{\log n} \sqrt{R(k_l^*)} - I(k_{j+1}^*, k_l^*) \right).$$

Thereafter, note that $\frac{I^2(k_{j+1}^*, k_l^*)}{R(k_l^*)} \leq C_{\psi_0} \frac{(k_{j+1}^*)^{-4\beta_f}}{\frac{k_l^*}{n^2}}$, which has a power of n equal to $\frac{8\beta_l}{1+4\beta_l} - \frac{8\beta_f}{1+4\beta_{j+1}} \leq 0$ since $\beta_l < \beta_{j+1} \leq \beta_f$. Hence arguing as in proof of Lemma A.1 along the lines of Lemma C.4 and C.5, we have the desired result provided C_{opt} is according to the constant specified in Theorem 3.1. $\qquad \square$

Proof of Lemma A.5.

PROOF. We begin by noting that it is enough to control

$$\int_{\chi_{n,m}} \int_{\chi_{n,m}} \int_{\chi_{n,m}} \int K_{k_1}(x, x_1) K_{k_2}(x, x_2) K_{k_3}(x, x_3) f(x_1) f(x_2) f(x_3) dx dx_1 dx_2 dx_3$$

for arbitrary $M_n \leq k_1 \leq k_2 \leq k_3$. We assume that k_1 divides k_2, k_2 divides k_3 and M_n divides k_1. The proof for a general tuple follows by similar arguments with suitable obvious modifications. The above integral equals,

$$\int_{\chi_{n,m}} \int_{\chi_{n,m}} \int_{\chi_{n,m}} \int \sum_{l_1} \sum_{l_2} \sum_{l_3} \left\{ \begin{array}{c} \psi_{l_1}^{k_1}(x) \psi_{l_1}^{k_1}(x_1) \psi_{l_2}^{k_2}(x) \\ \psi_{l_2}^{k_2}(x_2) \psi_{l_3}^{k_3}(x) \psi_{l_3}^{k_3}(x_3) \end{array} \right\} \prod_{i=1}^{3} f(x_i) dx_i$$

where $\psi_l^k(x)$ is k dilated an l shifted wavelet bases. Taking the integral inside the summation, any of the summand equals at most $C(\psi_0, \|f\|_\infty) \frac{1}{k_3} k_1 k_2 k_3 \frac{1}{k_1 k_2 k_3} = \frac{C(\psi_0, \|f\|_\infty)}{k_3}$. The reason being, each of the summand corresponds to an integral of x_1, x_2, x_3 over intervals of length $\frac{1}{k_1}$, $\frac{1}{k_2}$, and $\frac{1}{k_3}$ respectively. Moreover, since k_3 is the finest refinement, the integral over x is simply over an interval of length $\frac{1}{k_3}$. The bound on the summand then follows. Therefore, we now need to count the number of summands that contribute to the sum above. This can be argued as follows. For every given subinterval $\chi_{n,m}$ there are less than $C_{\psi_0} \frac{k_1}{M_n}$ subintervals of length $\frac{C_{\psi_0}}{k_1}$ with support intersecting that of some $\psi_{l_1}^{k_1}$. For each given subinterval of length $\frac{C_{\psi_0}}{k_1}$ with support intersecting that of some $\psi_{l_1}^{k_1}$, there are less than $C_{\psi_0} \frac{k_2}{k_1}$ subintervals of length

$\frac{C_{\psi_0}}{k_2}$ with support intersecting that of some $\psi_{l_2}^{k_2}$. Finally, for each given subinterval of length $\frac{C_{\psi_0}}{k_2}$ with support intersecting that of some $\psi_{l_2}^{k_2}$, there are less than $C_{\psi_0}\frac{k_3}{k_2}$ subintervals of length $\frac{C_{\psi_0}}{k_3}$ with support intersecting that of some $\psi_{l_3}^{k_3}$. Therefore, total number of terms contributing to the sum above is at most $C_{\psi_0}\frac{k_1}{M_n}\frac{k_2}{k_1}\frac{k_3}{k_2} = C_{\psi_0}\frac{k_3}{M_n}$. This implies that in absolute value, each $\alpha_{n,m}$ is at most $C(\psi_0, \|f\|_\infty)\frac{1}{k_3}\frac{k_3}{M_n} = C(\psi_0, \|f\|_\infty)\frac{1}{M_n}$, as promised. $\qquad\square$

Proof of Lemma A.6.

PROOF. We evoke Proposition 2.4 of Giné, Latała and Zinn (2000) (page 9) which implies that for a universal constant depending on q

$$
E\left|\frac{1}{m(m-1)(m-2)}\sum_{i_1 \neq i_2 \neq i_3} K(X_{i_1}, X_{i_2}, X_{i_3})\right|^q
$$

$$
\lesssim (m(m-1)(m-2))^{-q} C_q \left\{ \begin{array}{c} m^{3q/2}E\left[\left|E\left(K^2(X_{i_1}, X_{i_2}, X_{i_3})|X_{i_3}\right)\right|^{q/2}\right] \\ +m^q E\left[\max_{i_3}\left|E\left(K^2(X_{i_1}, X_{i_2}, X_{i_3})|X_{i_3}\right)\right|^{q/2}\right] \\ +m^{q/2}E\left[\max_{i_3,i_2}\left|E\left(K^2(X_{i_1}, X_{i_2}, X_{i_3})|X_{i_3}, X_{i_2}\right)\right|^{q/2}\right] \\ +E\left[\max_{i_3,i_2,i_1}\left|(K(X_{i_1}, X_{i_2}, X_{i_3}))\right|^q\right] \end{array} \right\}
$$

$$(B.4)$$

Now, we have that for $Z_1, ..., Z_m > 0$ identically distributed, possibly dependent with $E(Z_{(m)}) \leq mE(Z_1) = E\left(\sum_i Z_i\right)$ where $Z_{(m)} = \max_{i \in \{1,...,m\}} Z_i$. So that B.4 is bounded above by

$$
C_q(m(m-1)(m-2))^{-q} C_q \left\{ \begin{array}{c} m^{3q/2}E\left[|K(X_{i_1}, X_{i_2}, X_{i_3})|^q\right] \\ +m^{q+1}E\left[|K(X_{i_1}, X_{i_2}, X_{i_3})|^q\right] \\ +m^{q/2+2}E\left[|K(X_{i_1}, X_{i_2}, X_{i_3})|^q\right] \\ +m^3 E\left[|K(X_{i_1}, X_{i_2}, X_{i_3})|^q\right] \end{array} \right\}
$$

$$
\lesssim C_q m^{-q} E\left[|K(X_{i_1}, X_{i_2}, X_{i_3})|^q\right]
$$

$\qquad\square$

APPENDIX C: TECHNICAL LEMMAS

The following lemma regarding the bias of U_n^k will be used regularly.

LEMMA C.1. *Suppose $f \in H(\beta)$. Then for compactly supported wavelet bases*

$$\sup_f \mid \mathbb{E}_f(U_n^k) - \phi(f) \mid \leq C_{\psi_0} \|f\|_\infty k^{-2\beta_f}.$$

PROOF. The proof involves simple algebra and properties of compactly supported wavelet bases (Härdle et al., 1998) and hence is omitted.

We will be using the following lemma about properties of compactly supported wavelets.

LEMMA C.2. *For kernels based on compactly supported wavelets, as defined by Equation (2.1), the following hold.*

1. $\|f_k\|_\infty \leq C_{\psi_0} \|f\|_\infty$.
2. $\sup_l |\alpha_{k,l}| \leq \frac{C_{\psi_0} \|f\|_\infty}{k}$.
3. $\sup_x \sum_{l=1}^k |\psi_{k,l}(x)| \leq C_{\psi_0}.k$

PROOF. Once again, the proofs follow from simple algebra and properties of compactly supported wavelet bases (Härdle et al., 1998). □

□

We will also need the following asymptotic normality of the quadratic estimators (Robins et al., 2015).

LEMMA C.3. *For $f \in H(\beta, c)$ we have $\frac{U_n^{k(\beta)} - \mathbb{E}_f\left(U_n^{k(\beta)}\right)}{\sqrt{Var_f\left(U_n^{k(\beta)}\right)}} \Rightarrow N(0,1)$ where* $Var_f\left(U_n^{k(\beta)}\right) \sim \frac{k(\beta)}{n^2}$.

We will need the following lemma about tails of second order U-statistics.

LEMMA C.4. *Let,*

$$H(X,Y) = K_*(X,Y) - K_0(X,Y) - \mathbb{E}_Y(K_*(X,Y) - K_0(X,Y))$$
$$- \mathbb{E}_X(K_*(X,Y) - K_0(X,Y)) + \mathbb{E}_f(K_*(X,Y) - K_0(X,Y))$$

. Then the following exponential inequality holds for every $t > 0$ and a deterministic constant $C > 0$.

$$\mathbb{P}_f\left[|\sum_{i \neq j} H(X_i, X_j)| \geq C(A_1\sqrt{t} + A_2 t^2 + A_3 t^{\frac{3}{2}} + A_4 t^2)\right] \leq 5.6e^{-t},$$

where

$$A_1^2 = n(n-1)\mathbb{E}_f[H^2(X_1, X_2)],$$

$$A_2 = \sup \left\{ \begin{array}{l} \left| \mathbb{E}_f \left[\sum_{i=1}^n \sum_{j=1}^{i-1} H(X_1, X_2) a_i(X_1) b_j(X_2) \right] \right| : \\ \mathbb{E}_f \left(\sum_{i=1}^n a_i^2(X_1) \right) \leq 1, \mathbb{E}_f \left(\sum_{j=1}^n b_j^2(X_2) \right) \leq 1 \end{array} \right\},$$

$$A_3^2 = n \sup_x \{ \mathbb{E}_{f_{X_2}}(H^2(x, X_2)) \},$$

$$A_4 = \sup_{x,y} |H(x, y)|.$$

PROOF. This follows directly from Theorem 3.4 of Houdré and Reynaud-Bouret (2003). □

We now estimate the terms A_1, A_2, A_3, A_4 in the following lemma.

LEMMA C.5. *For kernels based on compactly supported wavelet bases, there exists deterministic constant C_{ψ_0} (depending on ψ_0 and $\|f\|_\infty$) such that,*

$$A_1^2 \leq C_{\psi_0}^2 n(n-1)(k_* + k_0), A_2 \leq C_{\psi_0} n, A_3^2 \leq C_{\psi_0} n(k_* + k_0), A_4 \leq C_{\psi_0}(k_* + k_0).$$

PROOF. The proof follows along the same line of arguments as those laid down in Proposition 2 of Bull and Nickl (2013). □

REFERENCES

ADAMCZAK, R. et al. (2006). Moment inequalities for U-statistics. *The Annals of Probability* **34** 2288–2314.

BEIRLANT, J., DUDEWICZ, E. J., GYÖRFI, L. and VAN DER MEULEN, E. C. (1997). Nonparametric entropy estimation: An overview. *International Journal of Mathematical and Statistical Sciences* **6** 17–39.

BICKEL, P. J. and RITOV, Y. (1988). Estimating integrated squared density derivatives: sharp best order of convergence estimates. *Sankhyā: The Indian Journal of Statistics, Series A* 381–393.

BIRGÉ, L. (2001). An alternative point of view on Lepski's method. *Lecture Notes-Monograph Series* 113–133.

BIRGÉ, L. and MASSART, P. (1995). Estimation of integral functionals of a density. *The Annals of Statistics* 11–29.

BROWN, L. D., LOW, M. G. et al. (1996). A constrained risk inequality with applications to nonparametric functional estimation. *The annals of Statistics* **24** 2524–2535.

BULL, A. D. and NICKL, R. (2013). Adaptive confidence sets in L^ 2. *Probability Theory and Related Fields* **156** 889–919.

CAI, T. T. and LOW, M. G. (2003). A note on nonparametric estimation of linear functionals. *Annals of statistics* 1140–1153.

CAI, T. T. and LOW, M. G. (2004). Minimax estimation of linear functionals over non-convex parameter spaces. *Annals of statistics* 552–576.

CAI, T. T. and LOW, M. G. (2005). Nonquadratic estimators of a quadratic functional. *The Annals of Statistics* 2930–2956.

CAI, T. T., LOW, M. G. et al. (2005a). Adaptive estimation of linear functionals under different performance measures. *Bernoulli* **11** 341–358.

CAI, T. T., LOW, M. G. et al. (2005b). On adaptive estimation of linear functionals. *The Annals of Statistics* **33** 2311–2343.

CAI, T. T. and LOW, M. G. (2006). Adaptation under probabilistic error for estimating linear functionals. *Journal of multivariate analysis* **97** 231–245.

CAI, T. T., LOW, M. G. et al. (2006). Optimal adaptive estimation of a quadratic functional. *The Annals of Statistics* **34** 2298–2325.

CAI, T. T., LOW, M. G. et al. (2011). Testing composite hypotheses, Hermite polynomials and optimal estimation of a nonsmooth functional. *The Annals of Statistics* **39** 1012–1041.

CAI, T. T., WANG, L. et al. (2008). Adaptive variance function estimation in heteroscedastic nonparametric regression. *The Annals of Statistics* **36** 2025–2054.

DONOHO, D. L., LIU, R. C. and MACGIBBON, B. (1990). Minimax risk over hyperrectangles, and implications. *The Annals of Statistics* 1416–1437.

DONOHO, D. L. and NUSSBAUM, M. (1990). Minimax quadratic estimation of a quadratic functional. *Journal of Complexity* **6** 290–323.

EFROMOVICH, S. and LOW, M. G. (1994). Adaptive estimates of linear functionals. *Probability theory and related fields* **98** 261–275.

EFROMOVICH, S., LOW, M. et al. (1996a). On Bickel and Ritov's conjecture about adaptive estimation of the integral of the square of density derivative. *The Annals of Statistics* **24** 682–686.

EFROMOVICH, S., LOW, M. et al. (1996b). On optimal adaptive estimation of a quadratic functional. *The Annals of Statistics* **24** 1106–1125.

EFROMOVICH, S. and SAMAROV, A. (2000). Adaptive estimation of the integral of squared regression derivatives. *Scandinavian journal of statistics* **27** 335–351.

FAN, J. (1991). On the estimation of quadratic functionals. *The Annals of Statistics* 1273–1294.

GINÉ, E., LATAŁA, R. and ZINN, J. (2000). Exponential and moment inequalities for U-statistics. In *High Dimensional Probability II* 13–38. Springer.

GINÉ, E. and NICKL, R. (2008). A simple adaptive estimator of the integrated square of a density. *Bernoulli* 47–61.

HALL, P. and MARRON, J. S. (1987). Estimation of integrated squared density derivatives. *Statistics & Probability Letters* **6** 109–115.

HÄRDLE, W., KERKYACHARIAN, G., TSYBAKOV, A. and PICARD, D. (1998). *Wavelets, approximation, and statistical applications*. Springer.

HOUDRÉ, C. and REYNAUD-BOURET, P. (2003). Exponential inequalities, with constants, for U-statistics of order two. *Stochastic inequalities and applications* 55–69.

INGSTER, Y. I. (1987). Minimax testing of nonparametric hypotheses on a distribution density in the L_p metrics. *Theory of Probability & Its Applications* **31** 333–337.

JOAG-DEV, K. and PROSCHAN, F. (1983). Negative association of random variables with applications. *The Annals of Statistics* 286–295.

KANDASAMY, K., KRISHNAMURTHY, A., POCZOS, B., WASSERMAN, L. and ROBINS, J. M. (2014). Influence Functions for Machine Learning: Nonparametric Estimators for Entropies, Divergences and Mutual Informations. *arXiv preprint arXiv:1411.4342*.

KERKYACHARIAN, G., PICARD, D. et al. (1996). Estimating nonquadratic functionals of

a density using Haar wavelets. *The Annals of Statistics* **24** 485–507.

Klemela, J. and Tsybakov, A. B. (2001). Sharp adaptive estimation of linear functionals. *Annals of statistics* 1567–1600.

Laurent, B. et al. (1996). Efficient estimation of integral functionals of a density. *The Annals of Statistics* **24** 659–681.

Laurent, B. and Massart, P. (2000). Adaptive estimation of a quadratic functional by model selection. *Annals of Statistics* 1302–1338.

Leonenko, N. and Seleznjev, O. (2010). Statistical inference for the -entropy and the quadratic Rényi entropy. *Journal of Multivariate Analysis* **101** 1981–1994.

Lepskii, O. (1991). On a problem of adaptive estimation in Gaussian white noise. *Theory of Probability & Its Applications* **35** 454–466.

Lepskii, O. (1992). Asymptotically minimax adaptive estimation. I: Upper bounds. Optimally adaptive estimates. *Theory of Probability & Its Applications* **36** 682–697.

Lepskii, O. (1993). Asymptotically minimax adaptive estimation. II. Schemes without optimal adaptation: Adaptive estimators. *Theory of Probability & Its Applications* **37** 433–448.

Low, M. G. (1992). Renormalization and white noise approximation for nonparametric functional estimation problems. *The Annals of Statistics* 545–554.

Pál, D., Póczos, B. and Szepesvári, C. (2010). Estimation of Rényi entropy and mutual information based on generalized nearest-neighbor graphs. In *Advances in Neural Information Processing Systems* 1849–1857.

Robins, J., Li, L., Tchetgen, E., van der Vaart, A. et al. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics: Essays in Honor of David A. Freedman* 335–421. Institute of Mathematical Statistics.

Robins, J., Li, L., Tchetgen, E. and Van der Vaart, A. W. (2015). Asymptotic Normality of Quadratic Estimators. submitted to *Festschrift for Everist Gine, Stochastic Processes and Its Applications*.

Shao, Q.-M. (2000). A comparison theorem on moment inequalities between negatively associated and independent random variables. *Journal of Theoretical Probability* **13** 343–356.

Tchetgen, E., Li, L., Robins, J. and van der Vaart, A. (2008). Minimax estimation of the integral of a power of a density. *Statistics & Probability Letters* **78** 3307–3311.