Bernoulli **24**(1), 2018, 699–739
DOI: 10.3150/16-BEJ893

Optimal adaptive inference in random design binary regression

RAJARSHI MUKHERJEE and SUBHABRATA SEN

Department of Statistics, Stanford University, Sequoia Hall, 390 Serra Mall, Stanford, CA 94305-4065, USA. E-mail: rmukherj@stanford.edu; ssen90@stanford.edu

We construct confidence sets for the regression function in nonparametric binary regression with an unknown design density – a nuisance parameter in the problem. These confidence sets are adaptive in L^2 loss over a continuous class of Sobolev type spaces. Adaptation holds in the smoothness of the regression function, over the maximal parameter spaces where adaptation is possible, provided the design density is smooth enough. We identify two key regimes – one where adaptation is possible, and one where some critical regions must be removed. We address related questions about goodness of fit testing and adaptive estimation of relevant infinite dimensional parameters.

Keywords: adaptive confidence sets; binary regression; U-statistics

In many epidemiological studies, a binary response variable Y is independently observed on a population of individuals along with multiple covariates $\mathbf{X}$ to explain the variability in the response. In the context of epidemiological studies, the probability of observing a specific outcome conditional on the covariates is often referred to as the propensity score. Estimating propensity score type functions from observed data is often of interest, and these estimates are subsequently used in multiple inferential procedures such as propensity score matching [50], inverse probability weighted inference [49] etc. In the context of semiparametric inference for missing data type problems, a nice exposition to the importance of understanding questions of similar flavor can be found in [56].

Historically, regression models with binary outcomes have been approached through both parametric [40] and nonparametric lenses [2,51]. Although parametric regression has the natural advantage of being simpler in interpretation and implementation, it often lacks the desired complexity required to capture varieties of dependence between covariates and outcomes. Nonparametric binary regression attempts to address this question, but it has its own share of shortcomings – the two major concerns being dependence on a priori knowledge about the true underlying regression function class and ease of implementation. Motivated by these, in this paper we study inference (estimation, testing, and confidence sets) in binary regression problems under nonparametric models having random covariates with unknown design density, with primary focus on adaptation over function classes.

To fix ideas, suppose we observe data $(\mathbf{x}_i, y_i)_{i=1}^n$, where $\mathbf{x}_i \in [0,1]^d$ and $y_i \in \{0,1\}$. Consider the binary regression model

$$\mathbb{E}(y|\mathbf{x}) = \mathbb{P}(y=1|\mathbf{x}) = f(\mathbf{x}), \qquad y \in \{0,1\}, \mathbf{x} \sim g. \tag{0.1}$$

For the rest of the paper, we assume g to be absolutely continuous with respect to the Lebesgue measure on $[0,1]^d$. Owing to the binary nature of the outcomes, the model is completely parametrized by the tuple (f,g) and admits a likelihood representation

$$l(y,\mathbf{x}|f,g) = f(\mathbf{x})^y\big(1-f(\mathbf{x})\big)^{1-y}g(\mathbf{x}). \tag{0.2}$$

We will be interested in making inferences about the regression function f (treating g as an unknown nuisance function), assuming f and g belong to Sobolev type spaces $B^{\beta}_{2,\infty}(M)$ and $B^{\gamma}_{2,\infty}(M')$ respectively – see Section 4.1 for a precise definition.

It is worth noting that, whereas an adaptive inference framework for Gaussian and density settings is well studied ([8,27–29,33,53] for goodness of fit testing, [9,14,15,30,32,35–37] for adaptive estimation, and [5,7,10,21,25,38,39,48,54] for honest adaptive confidence sets), the corresponding inferential questions in binary regression, with design density unknown, have received less attention.

In many instances, results in estimation and hypothesis testing for a non-Gaussian setup might be derived from a related Gaussian setup by appealing to the theory of asymptotic equivalence of experiments. However, it is well known that such equivalence only takes effect above certain threshold of smoothness for the underlying functions of interest. Also, asymptotic equivalence of regression models with multidimensional covariates and random covariate density is a lesser studied subject. Therefore the question of adaptive estimation for binary regression with multivariate random design cannot be addressed by simply invoking results from asymptotic equivalence. Moreover, the theory of asymptotic equivalence of experiments does not throw any light on the construction of adaptive confidence balls – one of the main questions of interest in this paper.

We also note that in contrast to the usual framework for random design Gaussian regression problems, we consider a setup where the design density is unknown – hence a nuisance parameter in the problem. Although [11] comments briefly on the case of nonparametric regression with uniformly random design density, these do not extend to the unknown design density case. Our setup, while being more realistic, makes our proofs technically more involved. The basic heuristic for our analysis is that in case the unknown design density is smooth enough, modulo certain modifications (to be made precise later), the "effect of estimating" the unknown design density is negligible compared to the errors in making inference for the unknown regression function.

In particular, the main results of this paper are summarized below.

(a) We produce estimators of underlying regression and design density which apart from jointly adapting over desired regimes of smoothness in an L_2 sense has the additional property of satisfying suitable boundedness (in both point-wise and Besov type norm sense) properties if the underlying functions are also similarly bounded (see Theorem 1.1).

(b) We provide complete solution (lower and upper bounds) to the problem of asymptotic minimax goodness of fit testing with both simple and composite null hypotheses (see Theorem 1.2) and unknown design density. An analogous result (with sharp asymptotics) for *simple null hypothesis* in Gaussian regression with multi-dimensional covariates with *known* design density and regression function having at least $\frac{d}{4}$ derivatives was developed by [29].

(c) We provide theory for adaptive confidence sets which complements those obtained in density [7] and sequence models [11,48] (see Theorem 1.3). A part of the adaptation theory for

Hölder balls was sketched briefly in [46] using the theory of higher order influence functions, where honest adaptation was possible in parts of the parameter space. Our results are over Besov balls, where following ideas of [7], we identify regions of the parameter space where adaptation is not possible without removing parts of the parameter space. We make this more precise in Section 1.

(d) All of our procedures are based on second order U-statistics constructed from projection kernels of suitable wavelet bases. We therefore extend the exponential inequality obtained in [7] to more general second order U-statistics based on wavelet projection kernels (See Lemma 4.1). For the case of testing of composite alternatives (1.3), this also adds to the chi-square type empirical wavelet coefficient procedure of [12].

Notation

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 $a_n/b_n \to \infty$ (respectively $a_n/b_n \to 0$) as $n \to \infty$. 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 $C \in [0, \infty)$). 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 for such constants. In particular, for any fixed tuple v of real numbers, $C(v)$ will denote a constant depending on elements of v only. Throughout the paper, we shall use $\mathbb{E}_P$ and $\mathbb{P}_P$ to denote expectation and probability under the measure P, and $\mathcal{I}$ will stand for the indicator function. For any linear subspace $L \subseteq L_2[0,1]^d$, let $\Pi(h|L)$ denote the orthogonal projection of h onto L under the Lebesgue measure. Finally, for suitable functions $h : [0,1]^d \to \mathbb{R}$, we let $\|h\|_q := (\int_{[0,1]^d} |h(\mathbf{x})|^q \, d\mathbf{x})^{1/q}$ and $\|h\|_\infty := \sup_{\mathbf{x}\in[0,1]^d} |h(\mathbf{x})|$ denote the usual L_q and L_∞ semi-norm of h, respectively.

Organization

The rest of the paper is organized as follows. In Section 1, we describe the main results along with the definition of honest adaptive confidence sets. Section 2 discusses our choice of model and places it in the broader perspective of heteroscedastic nonparametric regression. We collect the technical details (definition of Besov type spaces along with discussion on compactly supported wavelet bases) and proofs of the main theorems in Section 3. In Section 4, we discuss the assumptions made in the paper and scope of future research. Finally, we collect the proofs of certain technical lemmas in the Appendix.

1. Main results

We outline our main results in this section. We work with certain smoothness classes for both the regression and design density with suitable additional assumptions on the boundedness. For

conciseness of notation, we define,

$$\mathcal{P}\big(\beta,\gamma,M,M',B_L,B_U\big)=\left\{\begin{array}{c}(f,g): f\in B^{\beta}_{2,\infty}(M), g\in B^{\gamma}_{2,\infty}\big(M'\big), 0<f<1,\\ 0<B_L\le g\le B_U, \int g(\mathbf{x})\,d\mathbf{x}=1\end{array}\right\}. \quad (1.1)$$

Above and throughout the paper, by the pair (f,g) we shall refer to the probability measure P generated according to (0.2) by the regression function f and marginal density g respectively. Therefore, by an abuse of notation, we will refer to the elements of $\mathcal{P}$ interchangeably as either the pair (f,g) or the corresponding probability measure P. We will always assume that the radius and boundedness parameters (M, M', B_L, B_U) are known to us. There are indeed some subtleties involved in inference without the knowledge of these parameters. These issues can be dealt with using our arguments adapted to Theorem 4 of [7]. The lower and upper bound requirements on the design density can also be relaxed to a certain extent at the cost of more involved proofs. However, for focused discussion, these will not be addressed in this paper. For notational brevity, we will henceforth denote $\mathcal{P}(\beta,\gamma,M,M',B_L,B_U)$ simply as $\mathcal{P}(\beta,\gamma)$.

1.1. Adaptive estimation of parameters

Our first result establishes the existence of certain rate optimal estimators for the regression and design density in our setup. We further establish that these estimators satisfy certain additional boundedness properties almost surely, which is invaluable for subsequent inference in this setup.

Theorem 1.1.

1. *Let $0<\gamma_{\min}<\gamma_{\max}$ be given. There exists a sequence of estimators $\hat{g}$ of the design density g and constant C, both depending on $(M',\gamma_{\min},\gamma_{\max},B_U)$, such that for each $\gamma\in[\gamma_{\min},\gamma_{\max}]$ and $\beta>0$,*

$$\sup_{P\in\mathcal{P}(\beta,\gamma)}\mathbb{E}_P\big[\|\hat{g}-g\|_2^2\big]\le Cn^{-\frac{2\gamma}{2\gamma+d}},$$

$$\liminf_{n\to\infty}\inf_{P\in\mathcal{P}(\beta,\gamma)}\mathbb{P}_P\big[\hat{g}\in B^{\gamma}_{2,\infty}(C)\big]=1,$$

and there exists constants $0<B'_L\le B'_U$ (depending on B_L, B_U) such that $B'_L\le\hat{g}\le B'_U$ almost surely. Further, there exists a universal constant $c>0$ such that

$$\inf_{\hat{g}}\sup_{P\in\mathcal{P}(\beta,\gamma)}\mathbb{E}_P\big[\|\hat{g}-g\|_2^2\big]\ge cn^{-\frac{2\gamma}{2\gamma+d}}.$$

2. *Let $0<\beta_{\min}\le\beta_{\max}$, $\gamma_{\min}<\gamma_{\max}$ be given. If $\gamma_{\min}>\beta_{\max}$, there exists a sequence of estimators $\hat{f}$ and constant C, both depending on $(M,M',B_U,B_L,\beta_{\min},\beta_{\max},\gamma_{\max})$, such that*

for every $\beta \in [\beta_{\min}, \beta_{\max}]$ *and* $\gamma \in [\gamma_{\min}, \gamma_{\max}]$

$$\sup_{P \in \mathcal{P}(\beta,\gamma)} \mathbb{E}_P\big[\|\hat{f} - f\|_2^2\big] \le C n^{-\frac{2\beta}{2\beta+d}},$$

$$\liminf_{n\to\infty} \inf_{P \in \mathcal{P}(\beta,\gamma)} \mathbb{P}_P\big[\hat{f} \in B^{\beta}_{2,\infty}(C)\big] = 1,$$

and there exist constants $C_L \le C_U$ *(depending on* B_L *and* B_U*) such that* $C_L \le \hat{f} \le C_U$ *almost surely. Further, there exists a constant* $c > 0$*, independent of* n*, such that*

$$\inf_{\hat{f}} \sup_{P \in \mathcal{P}(\beta,\gamma)} \mathbb{E}_P\big[\|\hat{f} - f\|_2^2\big] \ge c n^{-\frac{2\beta}{2\beta+d}}.$$

The proof of Theorem 1.1 is outlined in Section 4.4.

Remark 1. The result of Theorem 1.1 part 1 is similar to that in [7]. It is worth noting that results of the kind stating that $\hat{g} \in B^{\gamma}_{2,\infty}(M^*)$ with high probability uniformly over $\mathcal{P}(\beta,\gamma)$ for a suitably large constant M^* is not too hard to show. However, our proof shows that a suitably bounded estimator $\hat{g}$, which adapts over smoothness and satisfies $\hat{g} \in B^{\gamma}_{2,\infty}(M^*)$ with probability larger than $1 - \frac{1}{n^{\theta}}$ uniformly over $\mathcal{P}(\beta,\gamma)$, for any $\theta > 0$ and correspondingly large enough M^*. Additionally, the results of Theorem 1.1 part 2 are relatively less common in an unknown design density setting. Indeed, adaptive estimation of regression function with random design over Besov type smoothness classes has been obtained by model selection type techniques by [4] for the case of Gaussian errors. Our results in contrast, as remarked in Section 3, hold for any regression model with bounded outcomes and compactly supported covariates having suitable marginal design density.

Remark 2. The dependence of our constants on $\gamma_{\max}$ stems from deciding the regularity of the wavelet basis used. Once we fix a wavelet basis with regularity $S > \gamma_{\max}$, the dependence of our constants on $\gamma_{\max}$ can be reduced to dependence on S.

1.2. Construction of confidence sets

To tackle the question of adaptive confidence sets in our setup, we need to first analyze the goodness of fit problem in this setup. The next theorem characterizes the minimax testing rate for our problem. The proof is deferred to Section 4.2. To this end, we introduce the parameter spaces

$$\mathcal{P}_0(\beta,\gamma) = \left\{(f,g) : f \equiv 1/2, g \in B^{\gamma}_{2,\infty}\big(M'\big), B_L < g < B_U, \int g(\mathbf{x})\,d\mathbf{x} = 1\right\},$$

$$\mathcal{P}\big(\beta,\gamma,\rho_n^2\big) = \left\{\begin{array}{c}(f,g) : f \in B^{\beta}_{2,\infty}(M), \left\|f - \frac{1}{2}\right\|_2^2 > \rho_n^2, g \in B^{\gamma}_{2,\infty}\big(M'\big), \\ B_L < g < B_U, \int g(\mathbf{x})\,d\mathbf{x} = 1\end{array}\right\}.$$

Further, for $\beta_1 > \beta_2$, we define,

$$\mathcal{P}\big(\beta_1, \beta_2, \gamma, \rho_n^2\big) = \left\{ \begin{array}{c} (f,g) : f \in B_{2,\infty}^{\beta_2}(M), \big\| f - B_{2,\infty}^{\beta_1}(M) \big\|_2^2 > \rho_n^2, \\ g \in B_{2,\infty}^{\gamma}\big(M'\big), B_L < g < B_U, \int g(\mathbf{x})\, d\mathbf{x} = 1 \end{array} \right\}.$$

Finally we recall $\mathcal{P}(\beta, \gamma)$ defined in (1.1).

Theorem 1.2.

1. *Consider the testing problem*

$$H_0 : P \in \mathcal{P}_0(\beta, \gamma) \quad vs. \quad H_1 : P \in \mathcal{P}\big(\beta, \gamma, \rho_n^2\big),$$

for $\gamma > \beta$. We have:

- *For any $0 < \alpha < 1$, there exists $D > 0$ sufficiently large (depending on α, M, M') and a test ϕ such that for $\rho_n^2 = Dn^{-\frac{4\beta}{4\beta+d}}$*

$$\limsup_{n\to\infty} \Big(\sup_{P \in \mathcal{P}_0(\beta,\gamma)} \mathbb{P}_P[\phi = 1] + \sup_{P \in \mathcal{P}(\beta,\gamma,\rho_n^2)} \mathbb{P}_P[\phi = 0] \Big) \le \alpha. \tag{1.2}$$

- *For any test ϕ which satisfies (1.2) introduced above, the corresponding sequence ρ_n^2 satisfies*

$$\liminf_{n\to\infty} \rho_n^2 \gtrsim n^{-\frac{4\beta}{4\beta+d}}.$$

2. *Consider the testing problem*

$$H_0 : P \in \mathcal{P}(\beta_1, \gamma) \quad vs. \quad H_1 : P \in \mathcal{P}\big(\beta_1, \beta_2, \gamma, \rho_n^2\big), \tag{1.3}$$

for $\beta_2 < \beta_1$ and $\gamma > 2\beta_2$. Then:

- *For any $0 < \alpha < 1$, there exists $D > 0$ sufficiently large (depending on α, M, M') and a test ϕ such that for $\rho_n^2 = Dn^{-\frac{4\beta_2}{4\beta_2+d}}$*

$$\limsup_{n\to\infty} \Big[\sup_{P \in \mathcal{P}(\beta_1,\gamma)} \mathbb{P}_P[\phi = 1] + \sup_{P \in \mathcal{P}(\beta_1,\beta_2,\gamma,\rho_n^2)} \mathbb{P}_P[\phi = 0] \Big] \le \alpha. \tag{1.4}$$

- *For any test ϕ which satisfies (1.4) introduced above, the corresponding sequence ρ_n^2 satisfies*

$$\liminf_{n\to\infty} \rho_n^2 \gtrsim n^{-\frac{4\beta_2}{4\beta_2+d}}.$$

A few remarks are in order about the results above. First, it is interesting to note whether the complexity of the null hypothesis affects the minimal rate of separation between the null and the

alternative necessary to carry out the test. Our result answers this question in the negative. As mentioned earlier, although the results appear to be of similar flavor to those in [7,12], the rigorous derivations require careful understanding and modifications to accommodate for the effect of estimating an unknown density. A possible approach to the testing problem (1.3) can be the method of [12] without further modification. However, such an approach results in unbiased estimation of $\|\Pi(fg|L)\|_2^2$ for appropriate subspaces $L \subset L_2[0,1]^d$ instead of $\|\Pi(f|L)\|_2^2$ required for understanding the minimum separation $\|f - B_{2,\infty}^{\beta_1}(M)\|_2^2$. Instead, our proof shows that under the alternative, the quantity $\|\Pi(f\frac{g}{\hat{g}}|L)\|_2^2$ is also large enough for suitable subspaces L. This quantity is easier to estimate modulo the availability of a nice estimator $\hat{g}$ – which is in turn guaranteed by Theorem 1.1. However, this also necessitates modifying the testing procedure of [12] suitably to incorporate the effect of estimating g. We make this more clear in the proof of Theorem 2.

Next, we outline the construction of honest adaptive confidence sets in our setup. We briefly introduce the relevant notions for convenience. A confidence set $C_n = C(\mathbf{x}_1, y_1, \ldots, \mathbf{x}_n, y_n)$ is a random measurable subset of L^2. We define the L^2 radius of a set C as

$$|C| = \inf\bigl\{\tau : C \subset \bigl\{\psi : \|\psi - g\|_2 \le \tau\bigr\} \text{ for some } g\bigr\}.$$

We seek to determine the maximal parameter spaces $\mathcal{P}_n$ so that adaptive confidence sets exist. We define a confidence set $C_n = C_n(\mathbf{x}_1, y_1, \ldots, \mathbf{x}_n, y_n)$ to be *honest* over a sequence of models $\mathcal{P}_n$ if

$$\inf_{P \in \mathcal{P}_n} \mathbb{E}_P[f \in C_n] \ge 1 - \alpha - r_n, \tag{1.5}$$

where $r_n \to 0$ as $n \to \infty$ and $\alpha < 1$ is a fixed level of confidence. Further, we call a confidence set C_n adaptive over a sequence of models $\mathcal{P}_n$ if there exists a constant C depending on the known parameters of the model space $\mathcal{P}_n$ such that

$$\sup_{P \in \mathcal{P}_n \cap \mathcal{P}(\beta,\gamma)} \mathbb{P}_P\bigl[|C_n|^2 \ge Cn^{-\frac{2\beta}{2\beta+d}}\bigr] \le \alpha', \tag{1.6}$$

where $0 < \alpha' < 1$ is a fixed constant.

Now we define the parameter spaces over which we will produce honest adaptive confidence sets in L_2. For given interval of smoothness of regression function $[\beta_{\min}, \beta_{\max}]$ such that $\beta_{\max} > 2\beta_{\min}$, we define a grid following ideas from [7]. With $N = \lceil \log_2(\frac{\beta_{\max}}{\beta_{\min}}) \rceil$ define $\beta_j = 2^{j-1}\beta_{\min}$, $j = 1, \ldots, N$. With this notation, we define

$$\mathcal{F}_n\bigl(M^*\bigr) = B_{2,\infty}^{\beta_N}(M) \cup \left(\bigcup_{j=1}^{N-1} \mathcal{B}_{2,\infty}^{\beta_j}\bigl(M, M^*\rho_n(\beta_j)\bigr)\right),$$

$$\mathcal{P}_n\bigl(M^*, M', \gamma\bigr) = \left\{ \begin{array}{c} (f,g) : f \in \mathcal{F}_n\bigl(M^*\bigr), 0 < f < 1, g \in B_{2,\infty}^{\gamma}\bigl(M'\bigr), \\ B_L \le g \le B_U, \int g(\mathbf{x})\,d\mathbf{x} = 1 \end{array} \right\},$$

where $\mathcal{B}^{\beta_j}_{2,\infty}(M, M^*\rho_n(\beta_j)) = \{h \in B^{\beta_j}_{2,\infty}(M) : \|h - B^{\beta_{j+1}}_{2,\infty}(M)\|_2 \geq M^*\rho_n(\beta_j)\}$, $\rho_n(\beta) = n^{-\frac{2\beta/d}{4\beta/d+1}}$, and the choice of M^* is solely guided by M, M', $\beta_{\min}$, $\beta_{\max}$, B_L, B_U and can be read off from the proof of the next theorem.

Theorem 1.3. *Let $0 < \beta_{\min} \leq \beta_{\max}$, $2\beta_{\max} < \gamma_{\min} \leq \gamma_{\max}$ be given. Then there exists a confidence set C_n depending only on the tuple $(M, M', \beta_{\min}, \beta_{\max}, \gamma_{\min}, \gamma_{\max}, B_L, B_U, \alpha, \alpha')$ which is honest and adaptive in the sense of (1.5) and (1.6) over $\bigcup_{\gamma=\gamma_{\min}}^{\gamma_{\max}} \mathcal{P}_n(M^*, M', \gamma)$, whenever M^* is large enough.*

Theorem 1.3 is proved in Section 4.3. It is of interest to determine whether the models $\bigcup_\gamma \mathcal{P}_n$, are in some sense, the maximal spaces over which adaptation is possible. We note that the testing lower bounds established in Theorem 1.2 part 2 above imply that $\bigcup_\gamma \mathcal{P}_n$ is indeed the largest parameter space, up to multiplicative constants of $\rho_n(\beta_j)$, $j = 1, \ldots, N$, over which adaptation is possible. Moreover, results of the flavor of [7], Theorem 3, can be recovered from the proof of Theorem 1.3.

2. Choice of binary regression model

In this section, we comment on our choice of binary regression model.

Regarding the generality of our model choice, there are two main points that need addressing. The first concerns the framing of model (0.1) without going through a link function – as is the general custom for generalized linear models. Indeed for a link function formulation as

$$\mathbb{E}(y|\mathbf{x}) = \theta\big(h(x)\big), \qquad y \in \{0, 1\}, \mathbf{x} \sim g, \tag{2.1}$$

for θ a distribution function of a symmetric random variable (probit, logistic etc.), our results still go through provided θ satisfies some regularity conditions. In general for a smooth function θ, the function $\theta(h(x))$ shares the smoothness index of h and identifying $f := \theta \circ h$ lands us back in model (0.1). To keep things simple, we work with model (0.1) throughout.

The second point to note is that we have not considered the fixed design case in our set up. The fixed design problem can be addressed similarly with more straightforward generalization of ideas from [7,48], and [12] due to lack of extra nuisance parameter g. We omit this for the sake of brevity.

The other point worth discussing concerns the generalizability of the binary regression model to more general nonparametric regression models. In this context note that, additive Gaussian noise is the simplest example of a situation where the regression function can be parametrized separately from other components of the model such as conditional variance given the covariates. This facilitates the development of a satisfying adaptation theory, even over large classes of unknown design densities. In non-Gaussian settings, given a likelihood specification for the conditional distribution of outcomes given covariates, one can attempt to produce a similar theory for adaptive inference.

As a step towards a general theory of adaptive inference in nonparametric regression, we consider the case of binary outcomes. Binary regression automatically belongs to a heteroscedastic

variance regime – a more challenging scenario in general [1,3,16–19]. However, it is different from standard heteroscedastic additive Gaussian noise regression problems in that the mean regression function is intimately tied to the conditional variance and shares the same smoothness. In this case, the simplicity of the conditional distribution of the outcome given regressors allows us to answer the question of adaptive inference to some degree of generality.

Finally, we remark that most of our results actually hold not only for the case of binary regression, but also for any regression model with bounded outcomes and compactly supported covariates having suitable marginal design density. It is for the proof of matching lower bounds to show that our results are asymptotically rate optimal, that we need the binary regression model.

3. Discussion

Although we have tried to describe adaptive confidence sets for binary regression to some degree of generality, it is instructive to discuss some of the assumptions made in the process. Throughout our paper, we assume a lower bound of smoothness on the marginal density g of $\mathbf{x}$. Although our assumption is not sharp, we believe that such an assumption on the marginal density of $\mathbf{x}$ is necessary to a certain extent. This ensures that we learn about g at a rate fast enough so that it does not reflect too adversely on the inference for f. One can also wonder if the requirement $\gamma_{\min} > 2\beta_{\max}$ in Theorem 1.3 can be relaxed. Using results from [46], it is possible to further reduce our lower bound on the smoothness of g in the context of adaptive confidence sets over smoothness of f satisfying $\beta_{\min} < \beta_{\max} < 2\beta_{\min}$. It remains an interesting and challenging question to understand the sharp lower bound on the smoothness for g under which one no longer derive results similar to those obtained here in the other regime that is, $2\beta_{\min} < \beta_{\max}$. In a future project, we plan to investigate this issue with special focus on using higher order influence functions [45,46]. Indeed, even in the case of non-adaptive inference, [45,46] require certain smoothness lower bounds on the unknown design density. A related point of view for constructing honest adaptive confidence sets is often in the context of "self similar" functions – a case where construction of fully adaptive honest confidence sets are possible without further removing parts of the self similar function spaces [6,21,31,41,43,44,54]. Although we do not pursue this in our paper, it is possible to use ideas from our paper to answer similar questions.

4. Technical details

4.1. Wavelets and Besov spaces

In this section, we collect some facts about wavelets and Besov spaces. We also introduce some notation that we use later. For $d > 1$, consider expansions of functions $h \in L_2([0,1]^d)$ on an orthonormal basis of compactly supported bounded wavelets of the form

$$h(\mathbf{x}) = \sum_{k \in \mathbb{Z}^d} \langle h, \boldsymbol{\psi}^0_{0,k} \rangle \boldsymbol{\psi}^0_{0,k}(\mathbf{x}) + \sum_{l=0}^{\infty} \sum_{k \in \mathbb{Z}^d} \sum_{v \in \{0,1\}^d - \{0\}^d} \langle h, \boldsymbol{\psi}^v_{l,k} \rangle \boldsymbol{\psi}^v_{l,k}(\mathbf{x}),$$

where the base functions $\boldsymbol{\psi}^v_{l,k}$ are orthogonal for different indices (l,k,v) and are scaled and translated versions of the 2^d S-regular base functions $\boldsymbol{\psi}^v_{0,0}$ with $S > \beta$, that is, $\boldsymbol{\psi}^v_{l,k}(x) = 2^{ld/2}\boldsymbol{\psi}^v_{0,0}(2^l\mathbf{x}-k) = \prod_{j=1}^d 2^{\frac{l}{2}}\psi^{v_j}_{0,0}(2^l x_j - k_j)$ for $k = (K_1,\ldots,k_d) \in \mathbb{Z}^d$ and $v = (v_1,\ldots,v_d) \in \{0,1\}^d$ with $\psi^0_{0,0} = \phi$ and $\psi^1_{0,0} = \psi$ being the scaling function and mother wavelet of regularity S, respectively as defined in one dimensional case. As our choices of wavelets, we will throughout use compactly supported scaling and wavelet functions of Cohen–Daubechies–Vial type with S first null moments[13]. In view of the compact support of the wavelets, for each resolution level l and index v, only $O(2^{ld})$ base elements $\psi^v_{l,k}$ are non-zero on $[0,1]$; let us denote the corresponding set of indices k by $\mathcal{Z}_l$ obtaining the representation,

$$h(\mathbf{x}) = \sum_{k\in\mathcal{Z}_{J_0}} \langle h, \boldsymbol{\psi}^0_{J_0,k}\rangle \boldsymbol{\psi}^0_{J_0,k}(\mathbf{x}) + \sum_{l=J_0}^{\infty}\sum_{k\in\mathcal{Z}_l}\sum_{v\in\{0,1\}^d-\{0\}^d} \langle h, \boldsymbol{\psi}^v_{l,k}\rangle \boldsymbol{\psi}^v_{l,k}(\mathbf{x}), \tag{4.1}$$

where $J_0 = J_0(S) \geq 1$ is such that $2^{J_0} \geq S$ [13,23]. Thereafter, letting for any $h \in L_2[0,1]^d$, $\|\langle h, \boldsymbol{\psi}_{l',\cdot}\rangle\|_2$ be the vector L_2 norm of the vector $(\langle h, \boldsymbol{\psi}^v_{l',k'}\rangle : k' \in \mathcal{Z}_{l'}, v \in \{0,1\}^d)$, define

$$B^{\beta}_{2,\infty}(M) := \left\{ \begin{array}{l} h \in L_2\big([0,1]^d\big) : \|h\|_{\beta,2} := 2^{J_0\beta}\big\|\langle h, \boldsymbol{\psi}^0_{J_0,\cdot}\rangle\big\|_2 \\ + \sup_{l\geq J_0} 2^{l\beta}\Bigg(\sum_{k\in\mathbb{Z}^d}\sum_{v\in\{0,1\}^d-\{0\}^d} \langle h, \boldsymbol{\psi}^v_{l,k}\rangle^2\Bigg)^{\frac{1}{2}} \leq M \end{array} \right\}. \tag{4.2}$$

We will be working with projections onto subspaces defined by truncating expansions as above at certain resolution levels. For example letting

$$V_j := \operatorname{span}\big\{\boldsymbol{\psi}^v_{l,k}, J_0 \leq l \leq j, k \in \mathcal{Z}_l, v \in \{0,1\}^d\big\}, \qquad j \geq J_0 \tag{4.3}$$

one immediately has the following orthogonal projection kernel onto V_j as

$$K_{V_j}(\mathbf{x}_1,\mathbf{x}_2) = \sum_{k\in\mathcal{Z}_{J_0}} \boldsymbol{\psi}^0_{J_0,k}(\mathbf{x}_1)\boldsymbol{\psi}^0_{J_0,k}(\mathbf{x}_2) + \sum_{l=J_0}^{j}\sum_{k\in\mathcal{Z}_l}\sum_{v\in\{0,1\}^d-\{0\}} \boldsymbol{\psi}^v_{l,k}(\mathbf{x}_1)\boldsymbol{\psi}^v_{l,k}(\mathbf{x}_2). \tag{4.4}$$

Owing to the MRA property of the wavelet basis, it is easy to see that K_{V_j} has the equivalent representation as

$$K_{V_j}(\mathbf{x}_1,\mathbf{x}_2) = \sum_{k\in\mathcal{Z}_j}\sum_{v\in\{0,1\}^d} \psi^v_{jk}(x_1)\psi^v_{jk}(x_2). \tag{4.5}$$

We will also consider,

$$W_j := \operatorname{span}\big\{\boldsymbol{\psi}^v_{j,k}, k \in \mathcal{Z}_j, v \in \{0,1\}^d - \{0\}^d\big\}, \qquad j \geq J_0 \tag{4.6}$$

and the corresponding orthogonal projection kernel onto W_j as

$$K_{W_j}(\mathbf{x}_1,\mathbf{x}_2) = \sum_{k\in\mathcal{Z}_j}\sum_{v\in\{0,1\}^d-\{0\}^d} \boldsymbol{\psi}^v_{j,k}(\mathbf{x}_1)\boldsymbol{\psi}^v_{j,k}(\mathbf{x}_2). \tag{4.7}$$

4.2. Proof of Theorem 1.2

We will describe the proof of Theorem 1.2 in this section. To this end, we will crucially utilize Lemma 4.1. The proof will be deferred to the Appendix B. The U-statistics appearing in this paper are mostly based on projection kernels sandwiched between arbitrary bounded functions. This necessitates generalizing the U-statistics bounds obtained in [7]. In particular, we are interested in tail bounds of U-statistics based on kernel $R(\mathbf{O}_1, \mathbf{O}_2) = L(\mathbf{O}_1)K_{V_j}(\mathbf{X}_1, \mathbf{X}_2)L(\mathbf{O}_2)$ and $R(\mathbf{O}_1, \mathbf{O}_2) = L(\mathbf{O}_1)K_{W_j}(\mathbf{X}_1, \mathbf{X}_2)L(\mathbf{O}_2)$ where $\mathbf{O} = (Y, \mathbf{X})$ and $Y \in \mathbb{R}$, $\mathbf{X} \in [0,1]^d$. Assume that $|L(\mathbf{O})| \leq B$ (which corresponds to our situation).

Lemma 4.1. *There exists constant $C := C(B, B_U, S) > 0$ such that*

$$\mathbb{P}\left(\left|\frac{1}{n(n-1)}\sum_{i_1 \neq i_2} R(\mathbf{O}_{i_1}, \mathbf{O}_{i_2}) - \mathbb{E}\big(R(\mathbf{O}_1, \mathbf{O}_2)\big)\right| \geq t\right)$$
$$\leq e^{-Cnt^2} + e^{-\frac{Ct^2}{a_1^2}} + e^{-\frac{Ct}{a_2}} + e^{-\frac{C\sqrt{t}}{\sqrt{a_3}}},$$

where $a_1 = \frac{1}{n-1}2^{\frac{jd}{2}}$, $a_2 = \frac{1}{n-1}(\sqrt{\frac{2^{jd}}{n}} + 1)$, $a_3 = \frac{1}{n-1}(\sqrt{\frac{2^{jd}}{n}} + \frac{2^{jd}}{n})$, $R(\mathbf{O}_1, \mathbf{O}_2) = L(\mathbf{O}_1)K_{V_j}(\mathbf{X}_1, \mathbf{X}_2)L(\mathbf{O}_2)$ *or* $R(\mathbf{O}_1, \mathbf{O}_2) = L(\mathbf{O}_1)K_{W_j}(\mathbf{X}_1, \mathbf{X}_2)L(\mathbf{O}_2)$ *with* K_{V_j} *and* K_{W_j} *constructed using compactly supported wavelet bases of regularity* S, $\mathbf{O} = (Y, \mathbf{X})$, $|L(\mathbf{O})| \leq B$ *almost surely* $\mathbf{O}$, *and* $\mathbf{X} \in [0,1]^d$ *has density* g *such that* $g(\mathbf{x}) \leq B_U$ *for all* $\mathbf{x} \in [0,1]^d$.

4.2.1. *Proof of part* 1

We will first introduce a test with the desired properties. We use the statistic

$$T = \frac{1}{n(n-1)}\sum_{1 \leq i \neq j \leq n}(y_i - 1/2)K_{V_{j_0}}(\mathbf{x}_i, \mathbf{x}_j)(y_j - 1/2).$$

Here, we choose $j_0 = \lceil \frac{2}{4\beta+d}\log_2 n \rceil$. We reject this test when $|T| > C\frac{2^{j_0 d/2}}{n}$, for some constant C to be chosen appropriately. We first control the Type I error for this test. We have, under $P \in \mathcal{P}_0(\beta, \gamma)$, $\mathbb{E}_P[T] = 0$. Applying Lemma 4.1, we obtain $\mathbb{P}_P[|T| > C\frac{2^{j_0 d/2}}{n}] \leq e^{-C}$. Thus, the Type I error may be controlled at the desired level α by choosing the cut-off C sufficiently large. To control the Type II error, we fix $P \in \mathcal{P}(\beta, \gamma, \rho_n^2)$. In this case, we have,

$$\mathbb{E}_P[T] = \big\|\Pi_{V_{j_0}}\big((f-1/2)g\big)\big\|_2^2 = \big\|(f-1/2)g\big\|_2^2 - \big\|\Pi_{V_{j_0}^{\perp}}\big((f-1/2)g\big)\big\|_2^2$$
$$\geq B_L^2\rho_n^2 - \frac{2C(M, M')}{\sqrt{1-2^{-2\beta}}}2^{-2j_0\beta},$$

where the last line follows since $\gamma > \beta$, using arguments similar to the proof of Lemma 4.5.

Thus we have,

$$\mathbb{P}_P\Big[|T| > C\frac{2^{j_0 d/2}}{n}\Big] = 1 - \mathbb{P}_P\Big[|T| \le C\frac{2^{j_0 d/2}}{n}\Big],$$

$$\mathbb{P}_P\Big[|T| \le C\frac{2^{j_0 d/2}}{n}\Big] \le \mathbb{P}_P\Big[\big|T - E_P[T]\big| \ge E_P[T] - C\frac{2^{j_0 d/2}}{n}\Big]$$

$$\le \mathbb{P}_P\big[\big|T - E_P[T]\big| \ge B_L^2\rho_n^2 - C' n^{-\frac{4\beta}{4\beta+d}}\big],$$

where C' depends on B_L, M, M'. The proof is completed by an application of Lemma 4.1, upon setting $\rho_n^2 = Dn^{-\frac{4\beta}{4\beta+d}}$ for some constant D sufficiently large.

Next, we establish a matching (up to constants) lower bound on the testing rate for this problem. Assume that $\rho_n^2 \ll n^{-\frac{4\beta}{4\beta+d}}$. The proof of the lower bound is then based on Theorem 2.1 of [47]. In particular, let $H : [0,1]^d \to \mathbb{R}$ be a C^∞ function supported on $[0, \frac{1}{2}]^d$ such that $\int H(\mathbf{x})\,d\mathbf{x} = 0$ and $\int H^2(\mathbf{x})\,d\mathbf{x} = 1$ and let $k = \lceil c_0 n^{\frac{2d}{4\beta+d}} \rceil$ for some $c_0 > 0$. Now suppose that $\Omega_1, \ldots, \Omega_k$ be the translates of the cube $k^{-\frac{1}{d}}[0, \frac{1}{2}]^d$ that are disjoint and contained in $[0,1]^d$. Letting $\mathbf{x}_1, \ldots, \mathbf{x}_k$ denote the bottom left corners of these cubes, we set for $\lambda = (\lambda_1, \ldots, \lambda_k) \in \{-1, +1\}^k$,

$$f_\lambda = \frac{1}{2} + \Big(\frac{1}{k}\Big)^{\frac{\beta}{d}} \sum_{j=1}^{k} \lambda_j H\big((\mathbf{x} - \mathbf{x}_j)k^{\frac{1}{d}}\big).$$

The construction ensures that $f_\lambda \in B^\beta_{2,\infty}(M)$ (H can be chosen to guarantee desired M) for every $\lambda = (\lambda_1, \ldots, \lambda_k) \in \{-1, +1\}^k$ and $\|f_\lambda - \frac{1}{2}\|_2^2 = (\frac{1}{k})^{\frac{2\beta}{d}}$. Therefore, by the choice of k, each f_λ corresponds to a measure in the alternative hypothesis. Choose π to be the uniform prior on $\{-1, +1\}^k$. We use the notation of Theorem 2.1 of [47], let us partition the sample space $\chi = \{0,1\} \times [0,1]^d$ into $\chi_j = \{0,1\} \times \Omega_j$, $j = 1, \ldots, k$ and the remaining set. Letting P_λ and Q_λ be the probability measure on $\{0,1\} \times [0,1]^d$ corresponding to likelihood (0.2) for $f = f_\lambda$ and $f \equiv \frac{1}{2}$ respectively, its obvious that $P_\lambda(\chi_j) = Q_\lambda(\chi_j) = p_j$ (say), since $\int H(\mathbf{x})\,d\mathbf{x} = 0$. Also, $p_j \in \frac{1}{k}[\underline{B}, \overline{B}]$ for fixed constants $\underline{B}, \overline{B}$. Moreover, $\delta = \max_j \sup_\lambda \int_{\chi_j} \frac{(q-p)^2}{p_\lambda}\frac{d\mu}{p_j}$ since $p = \int p_\lambda\,d\pi(\lambda) = \int f_\lambda^y (1 - f_\lambda)^{1-y}\,d\pi(\lambda) = \frac{1}{2} = q$. Finally, $p_\lambda - q = p_\lambda - p = (f_\lambda - \frac{1}{2})^y(\frac{1}{2} - f_\lambda)^{1-y}$ implies that $a = b = \max_j \sup_\lambda \int_{\Omega_j} \frac{(f_\lambda - \frac{1}{2})^2}{p_j} \in k^{-\frac{2\beta}{d}}[\underline{B}, \overline{B}]$. Therefore, by Theorem 2.1 of [47] if $\rho(P_1, P_2)$ denotes the Hellinger affinity between two probability measures P_1, P_2 defined on the same probability space

$$\rho\Big(\int P_\lambda d\big(\pi(\lambda)\big), \int Q_\lambda d\big(\pi(\lambda)\big)\Big) \ge 1 - C\frac{n^2}{k}k^{-\frac{4\beta}{d}},$$

which can be made arbitrarily close to one for large enough c_0. This proves the theorem since if the Hellinger affinity is bounded away from 1, then there does any consistent sequence of tests

distinguishing between the null hypothesis and the easier alternative corresponding to the f_λ's constructed above [57].

4.2.2. *Proof of part* 2

We will construct a test with the desired properties below. The proof of the testing lower bound follows from the argument outlined for the previous part of the Theorem. Our proof is similar in spirit to that of [12], though the details are considerably different.

Similar to the argument for the previous part, we set $j_0 = \lceil \frac{2}{4\beta_2+d} \log_2 n \rceil$. We assume that we have data $\{\mathbf{x}_i, y_i\}_{i=1}^{2n}$. We split it into two equal parts and use the second part to construct the estimator $\hat{g}$ of the design density g introduced in Theorem 1.1. Throughout the proof, $\mathbb{E}_{i,P}[\cdot]$ will denote the expectation with respect to the ith half of the sample, with the other half held fixed, under the distribution P. For $J_0 \le l \le j_0$, we construct the test statistics

$$T_n(l) = \frac{1}{n(n-1)} \sum_{1 \le i \ne j \le n} \frac{y_i}{\hat{g}(\mathbf{x}_i)} K_{W_l}(\mathbf{x}_i, \mathbf{x}_j) \frac{y_j}{\hat{g}(\mathbf{x}_j)}.$$

By Markov inequality, there exits a constant C^* such that

$$\mathbb{P}_P\big[\|\hat{g} - g\|_2^2 > C^{*2} n^{-\frac{2\gamma}{2\gamma+d}}\big] < \frac{\alpha}{4}. \tag{4.8}$$

We will condition on this event throughout this proof. The construction of the test depends on the following two lemmas.

Lemma 4.2. *For* $0 < \alpha < 1$, *there exists* ζ *sufficiently large such that*

$$\mathbb{P}_P \left[\begin{array}{c} \forall J_0 \le l \le j_0, \\ \left| T_n(l) - \left\| \Pi_{W_l}\left(f \frac{g}{\hat{g}} \right) \right\|_2^2 \right| \le \zeta \sqrt{\frac{2^{(l+j_0)d/2}}{n^2} + 2^{ld/4} \frac{\|\Pi_{W_l}(f\frac{g}{\hat{g}})\|_2^2}{n}} \end{array} \right] \ge 1 - 3\alpha/4.$$

Lemma 4.3.

- *Under* H_0, $\sup_{J_0 \le l \le j_0} (\|\Pi_{W_l}(f\frac{g}{\hat{g}})\|_2 - (\frac{M}{2^{l\beta_1}} + \frac{C^*}{B'_L} n^{-\frac{\gamma}{2\gamma+d}})) \le 0$ *with probability at least* $(1 - \alpha/4)$.
- *Let* $\{\tau_l : J_0 \le l \le j_0\}$ *be a sequence of numbers satisfying* $\sum_{l=J_0}^{j_0} \tau_l \le \frac{3}{4}\sqrt{D} n^{-\frac{2\beta_2}{4\beta_2+d}}$. *Then under* $(f, g) \in H_1$, *with probability at least* $1 - \alpha/4$, *there exists* $J_0 \le l \le j_0$ *such that* $\|\Pi_{W_l}(f\frac{g}{\hat{g}})\|_2 \ge \frac{M}{2^{l\beta_1}} + \tau_l$.

Before proving these two lemmas, we first complete the proof of the theorem assuming the validity of these two lemmas.

We consider the test Ψ which rejects if at least one of the $T_n(l) > \widetilde{C}_l$, where

$$\widetilde{C}_l = \left(\frac{M}{2^{l\beta_1}} + \frac{C^*}{B'_L} n^{-\frac{\gamma}{2\gamma+d}} + \zeta \frac{2^{(l+j_0)d/8}}{\sqrt{n}} \right)^2 \tag{4.9}$$

for ζ suitably large, to be chosen appropriately. We will use the following deviation bounds to control the Type I and II errors of this testing procedure.

We first control the Type I error of this procedure. Under H_0, with probability at least $1-\alpha$, for all $J_0 \le l \le j_0$,

$$\begin{aligned} T_n(l) &\le \left\| \Pi_{W_l}\left(f\frac{g}{\hat{g}} \right) \right\|_2^2 + \zeta \frac{2^{(l+j_0)d/4}}{n} + \zeta 2^{ld/8} \frac{\|\Pi_{W_l}(f\frac{g}{\hat{g}})\|_2}{\sqrt{n}} \\ &\le \left(\frac{M}{2^{l\beta_1}} + \frac{C^*}{B'_L} n^{-\frac{\gamma}{2\gamma+d}} \right)^2 + \zeta^2 \frac{2^{(l+j_0)d/4}}{n} + 2\zeta \frac{2^{(l+j_0)d/8}}{\sqrt{n}} \left(\frac{M}{2^{l\beta_1}} + \frac{C^*}{B'_L} n^{-\frac{\gamma}{2\gamma+d}} \right) \\ &\le \left(\frac{M}{2^{l\beta_1}} + \frac{C^*}{B'_L} n^{-\frac{\gamma}{2\gamma+d}} + \zeta \frac{2^{(l+j_0)d/8}}{\sqrt{n}} \right)^2, \end{aligned}$$

where we assume that $\zeta > 1$ without loss of generality. This controls the Type I error.

To control the Type II error, we fix $(f,g) \in \mathcal{P}(\beta_1, \beta_2, \gamma, \rho_n^2)$. Using Lemma 4.3, there exits $J_0 \le l \le j_0$ such that

$$\left\| \Pi_{W_l}\left(f\frac{g}{\hat{g}} \right) \right\|_2 \ge \frac{M}{2^{l\beta_1}} + \tau_l,$$

where we choose $\tau_l = C_1(n^{-\frac{\gamma}{2\gamma+d}} + \frac{2^{(l+j_0)d/8}}{\sqrt{n}})$. Thus with probability at least $(1-\alpha)$, we have,

$$\begin{aligned} T_n(l) &\ge \left\| \Pi_{W_l}\left(f\frac{g}{\hat{g}} \right) \right\|_2^2 - \zeta \sqrt{\frac{2^{(l+j_0)d/2}}{n^2} + 2^{ld/4} \frac{\|\Pi_{W_l}(f\frac{g}{\hat{g}})\|_2^2}{n}} \\ &\ge \left\| \Pi_{W_l}\left(f\frac{g}{\hat{g}} \right) \right\|_2 \left(\left\| \Pi_{W_l}\left(f\frac{g}{\hat{g}} \right) \right\|_2 - \zeta \frac{2^{ld/8}}{\sqrt{n}} \right) - \zeta \frac{2^{(l+j_0)d/4}}{n} \\ &\ge \left(\frac{M}{2^{l\beta_1}} + \tau_l \right) \left(\frac{M}{2^{l\beta_1}} + \tau_l - \zeta \frac{2^{ld/8}}{\sqrt{n}} \right) - \zeta \frac{2^{(l+j_0)d/4}}{n} \\ &\ge \left(\frac{M}{2^{l\beta_1}} + \tau_l \right) \left(\frac{M}{2^{l\beta_1}} + \tau_l/2 \right) - \zeta \frac{2^{(l+j_0)d/4}}{n}, \end{aligned}$$

where we choose $C_1 > 2\zeta$. Now, choosing C_1 even larger, specifically $C_1^2 > 4\zeta$, it follows that $\zeta \frac{2^{(l+j_0)d/4}}{n} \le \tau_l^2/4$. Thus for some $J_0 \le l \le j_0$, with probability at least $1-\alpha$,

$$T_n(l) \ge \left(\frac{M}{2^{l\beta_1}} + \frac{\tau_l}{2} \right)^2 \ge \left(\frac{M}{2^{l\beta_1}} + \frac{C^*}{B'_L} n^{-\frac{\gamma}{2\gamma+d}} + \zeta \frac{2^{(l+j_0)d/8}}{\sqrt{n}} \right)^2,$$

provided we choose $C_1 > 2C^*/B'_L$. This controls the Type II error of this test.

The proof of the theorem will now be completed with the proofs of Lemma 4.2 and Lemma 4.3 in the next two subsections.

4.2.3. *Proof of Lemma 4.2*

The proof of this lemma can indeed be completed by invoking Lemma 4.1, which yields a much stronger control of the tail bound than demanded by Lemma 4.2. However, for the sake of simplicity we provide simpler proof by simple union bound and Chebyshev's inequality. The proof follows by an argument similar to [12], Lemma 4.2. We have, for $J_0 \le l \le j_0$,

$$\mathbb{E}_{1,P}\big[T_n(l)\big] = \left\|\Pi_{W_l}\left(f\frac{g}{\hat g}\right)\right\|_2^2 \operatorname{var}_{1,P}\big[T_n(l)\big] \le C(S, B_L, B_U)\left(\frac{\|\Pi_{W_l}(f\frac{g}{\hat g})\|_2^2}{n} + \frac{2^{ld}}{n^2}\right).$$

The validity of the variance bound of the last display above, follows from Hoeffding's decomposition, boundedness of f, g, $\hat g$, and standard properties of compactly supported wavelet bases. Therefore, we have, using union bound and Chebyshev's inequality,

$$\begin{aligned}
&\mathbb{P}_P\left[\exists l, J_0 \le l \le j_0, \left|T_n(l) - \left\|\Pi_{W_l}\left(f\frac{g}{\hat g}\right)\right\|_2^2\right| > \zeta\sqrt{\frac{2^{(l+j_0)d/2}}{n^2} + 2^{ld/4}\frac{\|\Pi_{W_l}(f\frac{g}{\hat g})\|_2^2}{n}}\right] \\
&\le \sum_{l=J_0}^{j_0} \mathbb{E}_P\left[\mathbb{P}_{1,P}\left[\left|T_n(l) - \left\|\Pi_{W_l}\left(f\frac{g}{\hat g}\right)\right\|_2^2\right| > \zeta\sqrt{\frac{2^{(l+j_0)d/2}}{n^2} + 2^{ld/4}\frac{\|\Pi_{W_l}(f\frac{g}{\hat g})\|_2^2}{n}}\right]\right] \\
&\le \sum_{l=J_0}^{j_0} \mathbb{E}_P\left[\frac{\operatorname{var}_{1,P}[T_n(l)]}{\zeta^2[\frac{2^{(l+j_0)d/2}}{n^2} + 2^{ld/4}\frac{\|\Pi_{W_l}(f\frac{g}{\hat g})\|_2^2}{n}]}\right] \\
&\le \frac{C(S, B_L, B_U)}{\zeta^2}\sum_{l=J_0}^{j_0}\big[2^{-(j_0-l)d/2} + 2^{-ld/4}\big].
\end{aligned}$$

The proof follows upon noting that $\sum_{l=J_0}^{j_0} 2^{-(j_0-l)d/2} \le \frac{2^{d/2}}{2^{d/2}-1}$ and $\sum_{l=J_0}^{j_0} 2^{-ld/4} \le \frac{2^{-J_0 d/4}}{1-2^{-d/4}}$.

4.2.4. *Proof of Lemma 4.3*

Let us consider $f \in B^{\beta_1}_{2,\infty}(M)$. Setting $\hat\Delta = \frac{g-\hat g}{\hat g}$, we have, for $J_0 \le l \le j_0$,

$$\left\|\Pi_{W_l}\left(f\frac{g}{\hat g}\right)\right\|_2 \le \big\|\Pi_{W_l}(f)\big\|_2 + \big\|\Pi_{W_l}(f\hat\Delta)\big\|_2.$$

For $f \in B^{\beta_1}_{2,\infty}(M)$, it follows from definition that $\|\Pi_{W_l}(f)\|_2 \le \frac{M}{2^{l\beta_1}}$. Recalling the definition of C^* from (4.8), the property of $\hat g$ from Theorem 1.1, and using the contraction property of the norm under projections, we have with probability at least $(1-\alpha/4)$,

$$\big\|\Pi_{W_l}(f\hat\Delta)\big\|_2^2 \le \|f\hat\Delta\|_2^2 \le \left(\frac{C^*}{B_L'}\right)^2 n^{-\frac{2\gamma}{2\gamma+d}}.$$

Combining, we get the desired result for functions $f \in B^{\beta_1}_{2,\infty}(M)$.

Next, we consider functions $f \in B^{\beta_2}_{2,\infty}(M)$ such that $\|f - B^{\beta_1}_{2,\infty}(M)\|_2 > \rho_n$. We first note that for any $h \in B^{\beta_1}_{2,\infty}(M)$,

$$\left\|\Pi_{V_{j_0}}(f) - h\right\|_2 \geq \rho_n - \left\|f - \Pi_{V_{j_0}}(f)\right\|_2 \geq \frac{5}{6}\rho_n$$

if D is chosen large enough. This implies that with probability at least $(1 - \alpha/4)$, we have,

$$\left\|\Pi_{V_{j_0}}\left(f\frac{g}{\hat{g}}\right) - h\right\|_2 \geq \frac{5}{6}\rho_n - \left\|\Pi_{V_{j_0}}(f\hat{\Delta})\right\|_2 \geq \frac{3}{4}\rho_n$$

for n sufficiently large, if $\gamma > 2\beta_2$. Thus, if $\{\tau_l : J_0 \leq l \leq j_0\}$ is a sequence of numbers such that $\frac{3}{4}\rho_n \geq \sum_{l=J_0}^{j_0} \tau_l$, following the argument of [12], Lemma 4.1, it is easy to see that there exists $J_0 \leq l \leq j_0$ such that $\|\Pi_{W_l}(f\frac{g}{\hat{g}})\|_2 \geq \frac{M}{2^{l\beta_1}} + \tau_l$. We choose

$$\tau_l = C_1\left(n^{-\frac{\gamma}{2\gamma+d}} + \frac{2^{(l+j_0)d/8}}{\sqrt{n}}\right),$$

where C_1 will be chosen suitably. It is easy to see that for any chosen C_1, $\sum_l \tau_l \leq \frac{3}{4}\rho_n$ can be enforced by choosing D sufficiently large.

4.3. Proof of Theorem 1.3

This proof idea is motivated by [7]. For $\beta \in [\beta_{\min}, \beta_{\max}]$ and $2\beta_{\max} < \gamma < \gamma_{\max}$ consider $P = (f, g) \in \mathcal{P}_n(M^*, M', \gamma) \cap \mathcal{P}(\beta, \gamma)$. Recall the finite grid $\{\beta_1, \ldots, \beta_N\}$ used for the construction of the parameter spaces $\mathcal{P}_n(M^*, M', \gamma)$. We define $\mathcal{F}_n(M^*, j) = \mathcal{B}^{\beta_j}_{2,\infty}(M, M^*\rho_n(\beta_j))$, $j = 1, \ldots, N-1$, $\mathcal{F}_n(M^*, N) = B^{\beta_N}_{2,\infty}(M)$. In addition, we set, for $j \in \{1, \ldots, N\}$,

$$\mathcal{P}_n\big(j, M^*, M', \gamma\big) = \left\{ \begin{array}{c} (h, g) : h \in \mathcal{F}_n\big(M^*, j\big), 0 < h < 1, g \in B^{\gamma}_{2,\infty}\big(M'\big), \\ B_L < g < B_U, \int g(\mathbf{x})\, d\mathbf{x} = 1 \end{array} \right\}.$$

Recall further the test Ψ introduced in the proof of Theorem 1.2 part 2. Note that cut-off for the test, as in (4.9), depends on the smoothness of g. However, a close inspection of the proof reveals that the only requirement on the smoothness of g is that of being at least as large as twice the maximum smoothness of f. Since, our estimator $\hat{g}$ is an adaptive estimator of g, and $\gamma_{\min} > 2\beta_{\max}$, we can use $\gamma_{\min}$ in the cut-off (4.9) for the test Ψ, maintaining the validity of the results. The test Ψ with $\beta = \beta_j$ will be referred to as $\Psi(j)$. We first test the hypothesis $H_0 : h \in \mathcal{F}_n(M^*, 2)$ vs. $H_1 : h \in \mathcal{F}_n(M^*, 1)$ at level $\alpha/4N$. If we reject H_0, we set $\hat{\beta} = \beta_1$ and stop. Otherwise we continue. At the jth step, $1 < j < N-1$, we test $H_0 : h \in \mathcal{F}_n(M^*, j+1)$ vs. $H_1 : h \in \mathcal{F}_n(M^*, j)$ using the appropriate test $\Psi(j)$ at level $\alpha/(4N)$. If we reject H_0 at step j, we set $\hat{\beta} = \beta_j$ and stop. Otherwise we continue – if none of the hypotheses are rejected,

we set $\hat{\beta} = \beta_N$. This procedure determines the "shell" in which f belongs. Once this has been accomplished, we construct a confidence set using ideas introduced in [48].

Without loss of generality, we assume we have data $\{\mathbf{x}_i, y_i : 1 \le i \le 3n\}$. We split the data into three equal parts – the estimator $\hat{f}$ outlined in Theorem 1.1 and $\hat{\beta}$ described above are constructed from the first, while the adaptive estimator of the design density $\hat{g}$ introduced in Theorem 1.1 is constructed from the second part. We condition on the events $\{\hat{f} \in B^{\beta}_{2,\infty}(C')\}$ and $\{\hat{g} \in B^{\gamma}_{2,\infty}(C')\}$ which happen with probability at least $1 - r_n$ (for C' large enough depending on M, M', B_U, B_L, $\gamma_{\max}$) uniformly over $\mathcal{P}_n(M^*, M', \gamma) \cap \mathcal{P}(\beta, \gamma)$, for some vanishing sequence r_n. Finally, we set $j_1 = \lceil \frac{2}{4\hat{\beta}+d} \log_2(n) \rceil$. Using the data $\{(\mathbf{x}_i, y_i) : 1 \le i \le n\}$, we construct the following U-statistic.

$$\hat{U}_n = \frac{1}{n(n-1)} \sum_{2n+1 \le i_1 \ne i_2 \le 3n} \frac{(y_{i_1} - \hat{f}(\mathbf{x}_{i_1}))}{\hat{g}(\mathbf{x}_{i_1})} K_{V_{j_1}}(\mathbf{x}_{i_1}, \mathbf{x}_{i_2}) \frac{(y_{i_2} - \hat{f}(\mathbf{x}_{i_2}))}{\hat{g}(\mathbf{x}_{i_2})}.$$

For any $h \in L^2$, we define $\tau_n^2(h) = \frac{C_1}{n}\|h - \hat{f}_n\|_2^2 + \frac{C_2 2^{j_1}}{n(n-1)}$, for constants C_1, C_2 to be chosen later. Finally, we define the set

$$C_n(\beta) = \left\{ h : \|h - \hat{f}\|_2^2 \le \hat{U}_n + C(M, B_L, B_U)\big(n^{-\frac{4\beta}{4\beta+d}} + n^{-\frac{\beta}{2\beta+d}} n^{-\frac{\gamma_{\min}}{2\gamma_{\min}+d}}\big) + z(\alpha)\tau_n(h) \right\},$$

with $z(\alpha) \ge 1/\alpha$. We will show that the set $C_n(\hat{\beta})$ is a confidence set with the desired properties. Throughout the rest of the proof $\mathbb{E}_{P,S}[\cdot]$ for $P \in \mathcal{P}_n(M^*, M', \gamma)$ and $S \subset \{1, 2, 3\}$ will denote expectation under the distribution P conditional on the subset of the data corresponding to the subset S.

Let $i_0 = i_0(f) \in \{1, \ldots, N\}$ denote the unique index such that $f \in \mathcal{F}_n(M^*, i_0)$. We prove that uniformly over $P \in \mathcal{P}_n(M^*, M', \gamma) \cap \mathcal{P}(\beta, \gamma)$, $\mathbb{P}_P(\hat{\beta} \ne \beta_{i_0}) \le \alpha/2$. Indeed, $\hat{\beta} < \beta_{i_0}$ implies that one of the test $\Psi(j)$, $j = 1, \ldots, i_0 - 1$ has rejected the true null hypothesis. Thus,

$$\sup_{P \in \mathcal{P}_n(i_0, M^*, M', \gamma)} \mathbb{P}_P[\hat{\beta} < \beta_{i_0}] \le \sum_{i < i_0} \sup_{P \in \mathcal{P}_n(i_0, M^*, M', \gamma)} \mathbb{E}_P\big[\Psi(i)\big] < \frac{\alpha}{4}.$$

Similarly, $\hat{\beta} > \beta_{i_0}$ essentially implies that one of the tests $\Psi(i)$, $i > i_0$ fails to reject the null hypothesis. Therefore

$$\sup_{P \in \mathcal{P}_n(i_0, M^*, M', \gamma)} \mathbb{P}_P[\hat{\beta} > \beta_{i_0}] \le \sum_{i > i_0} \sup_{P \in \mathcal{P}_n(i_0, M^*, M', \gamma)} \mathbb{E}_P\big[1 - \Psi(i)\big] \le \frac{\alpha}{4}.$$

Combining, we have, $\sup_{P \in \mathcal{P}_n(i_0, M^*, M', \gamma)} \mathbb{P}_P[\hat{\beta} \ne \beta_{i_0}] \le \frac{\alpha}{2}$. Now, we have,

$$\mathbb{P}_P\big[f \in C_n(\hat{\beta})\big] \ge \mathbb{P}_P\big[f \in C_n(\beta_{i_0})\big] - \frac{\alpha}{2}.$$

Thus, honesty of the confidence set follows provided we establish that $\mathbb{P}_P[f \in \hat{C}_n(\beta_{i_0})] \ge 1 - \alpha/2$ uniformly on $\mathcal{P}_n(i_0, M^*, M', \gamma)$. To this end, we note that setting $\hat{\Delta} = \frac{\hat{g} - g}{g}$, we have that

for a deterministic constant $C(M, B_L, B_U)$,

$$\begin{aligned}
\mathbb{E}_{P,\{2,3\}}[\hat{U}_n] &= \left\| \Pi_{V_{j_1}}(f - \hat{f}_n)\frac{g}{\hat{g}} \right\|_2^2 \\
&= \left\| \Pi_{V_{j_1}}(\hat{f}_n - f) \right\|_2^2 + \left\| \Pi_{V_{j_1}}\big((\hat{f}_n - f)\hat{\Delta}\big) \right\|_2^2 \\
&\quad + 2\big\langle \Pi_{V_{j_1}}(\hat{f}_n - f), \Pi_{V_{j_1}}\big((\hat{f}_n - f)\hat{\Delta}\big) \big\rangle \\
&= \|\hat{f}_n - f\|_2^2 - \left\| \Pi_{V_{j_1}^{\perp}}(\hat{f}_n - f) \right\|_2^2 + \left\| \Pi_{V_{j_1}}\big((\hat{f}_n - f)\hat{\Delta}\big) \right\|_2^2 \\
&\quad + 2\big\langle \Pi_{V_{j_1}}(\hat{f}_n - f), \Pi_{V_{j_1}}\big((\hat{f}_n - f)\hat{\Delta}\big) \big\rangle \\
&\geq \|\hat{f}_n - f\|_2^2 - C\big(M, M', \beta_{\max}\big) n^{-\frac{4\beta_{i_0}}{4\beta_{i_0}+d}} \\
&\quad - 2\left\| \Pi_{V_{j_1}}(\hat{f}_n - f) \right\|_2 \left\| \Pi_{V_{j_1}}\big((\hat{f}_n - f)\hat{\Delta}\big) \right\|_2 \\
&\geq \|\hat{f}_n - f\|_2^2 - C\big(M, M', \beta_{\max}, B_L, B_U\big)\big(n^{-\frac{4\beta_{i_0}}{4\beta_{i_0}+d}} + n^{-\frac{\beta_{i_0}}{2\beta_{i_0}+d}} n^{-\frac{\gamma}{2\gamma+d}}\big).
\end{aligned}$$

Further, we have, using Hoeffding decomposition conditional on samples $\{2, 3\}$,

$$\hat{U}_n - E_{P,\{2,3\}}[\hat{U}_n] = L + R,$$

$$L = \frac{2}{n} \sum_{i=2n+1}^{3n} \sum_{\substack{k \in \mathcal{Z}_{j_1}, \\ v \in \{0,1\}^d}} \left[\frac{(y_i - \hat{f}_n(\mathbf{x}_i))}{\hat{g}(\mathbf{x}_i)} \psi_{j_1,k}^v(\mathbf{x}_i) - \left\langle (\hat{f}_n - f)\frac{g}{\hat{g}}, \psi_{j_1,k}^v \right\rangle \right] \left\langle (\hat{f}_n - f)\frac{g}{\hat{g}}, \psi_{j_1,k}^v \right\rangle,$$

$$\begin{aligned}
R = \frac{1}{n(n-1)} \sum_{2n+1 \leq i_1 \neq i_2 \leq 3n} \sum_{\substack{k \in \mathcal{Z}_{j_1}, \\ v \in \{0,1\}^d}} &\left[\frac{(y_{i_1} - \hat{f}_n(\mathbf{x}_{i_1}))}{\hat{g}(\mathbf{x}_{i_1})} \psi_{j_1,k}^v(\mathbf{x}_{i_1}) - \left\langle (\hat{f}_n - f)\frac{g}{\hat{g}}, \psi_{j_1,k}^v \right\rangle \right] \\
&\times \left[\frac{(y_{i_2} - \hat{f}_n(\mathbf{x}_{i_2}))}{\hat{g}(\mathbf{x}_{i_2})} \psi_{j_1,k}^v(\mathbf{x}_{i_2}) - \left\langle (\hat{f}_n - f)\frac{g}{\hat{g}}, \psi_{j_1,k}^v \right\rangle \right].
\end{aligned}$$

Using the orthogonality of the linear and non-linear term in Hoeffding's decomposition, we can bound the variance of $\hat{U}_n$ by the sum of the variances of L and R. The variance of the linear term may be bounded by the second moment and using the boundedness of f, $\hat{f}_n$, $\hat{g}_n$, we have that

$$\operatorname{var}_{P,\{2,3\}}[L] \leq \frac{C(S, B_L, B_U)}{n} \left\| \Pi_{V_{j_1}}(f - \hat{f}_n)\frac{g}{\hat{g}} \right\|_2^2.$$

By a proof similar to that of controlling Λ_1 in Lemma B.2, we have that for a deterministic constant $C(\beta_{\max}, \gamma_{\max}, M, B_L, B_U)$

$$\operatorname{var}_{P,\{2,3\}}[R] \leq \frac{C(S, B_L, B_U) 2^{j_1}}{n(n-1)}.$$

Finally, we set

$$\tau_n(f)^2 = \frac{C(S, B_L, B_U)}{n} \big\|(f - \hat{f}_n)\big\|_2^2 + \frac{C(S, B_L, B_U) 2^{j_1}}{n(n-1)}.$$

By an application of Chebyshev inequality, we have,

$$\mathbb{P}_{P,\{2,3\}}\big[\big|\hat{U}_n - \mathbb{E}_{P,\{2,3\}}[\hat{U}_n]\big| > C\tau_n(f)\big] \le \frac{\mathrm{var}_{P,\{2,3\}}[\hat{U}_n]}{C^2\tau_n(f)^2} \le \frac{1}{C^2}.$$

Thus for C chosen appropriately, the above probability may be controlled at any pre-specified level $\alpha/2$.

Based on our construction, we have for a $C' = C(M, M', \beta_{\max}, B_L, B_U)$,

$$\begin{aligned}\mathbb{P}_P\big(f \in C_n(\beta_{i_0})\big) &= \mathbb{P}_P\big(\|f - \hat{f}\|_2^2 \le \hat{U}_n + C'\big(n^{-\frac{4\beta_{i_0}}{4\beta_{i_0}+d}} + n^{-\frac{\beta_{i_0}}{2\beta_{i_0}+d}} n^{-\frac{\gamma_{\min}}{2\gamma_{\min}+d}}\big) + z(\alpha)\tau_n(f)\big) \\ &\ge \mathbb{P}_P\big[\big|\hat{U}_n - \mathbb{E}_{P,\{2,3\}}[\hat{U}_n]\big| \le z(\alpha)\tau_n(f)\big] \ge \left(1 - \frac{\alpha}{2}\right).\end{aligned}$$

Finally, we establish that the L^2 diameter of this set adapts to the underlying smoothness. Assume $P \in \mathcal{P}_n(M^*, M', \gamma) \cap \mathcal{P}(\beta, \gamma)$ and the following calculations are uniform over this parameter space. The deterministic terms in the diameter term are respectively, of the order $n^{-\frac{2\beta_{i_0}}{4\beta_{i_0}+d}} = o(n^{-\frac{\beta}{2\beta+d}})$ (as $\beta < \beta_{i_0+1} < 2\beta_{i_0}$) and $n^{-\frac{\beta_{i_0}}{2\beta_{i_0}+d}} n^{-\frac{\gamma_{\min}}{2\gamma_{\min}+d}}$ which, by some tedious algebra, is also $o(n^{-\frac{2\beta}{2\beta+d}})$ since $\beta < \beta_{i_0+1} < 2\beta_{i_0}$, $\gamma_{\min} > 2\beta_{i_0+1}$. The random part of $\tau_n(f)^2$ is also $o_P(n^{-\frac{2\beta}{2\beta+d}})$ as $\hat{f}_n$ is an adaptive estimator and $\|\frac{g}{\hat{g}}\|_\infty \le \frac{B_U}{B_L'(B_L)} = C(B_U, B_L)$. Finally, the leading term for the diameter is contributed by

$$\mathbb{E}_P[\hat{U}_n] = \mathbb{E}_P\left[\left\|\Pi_{V_{j_1}}(f - \hat{f}_n)\frac{g}{\hat{g}}\right\|_2^2\right] \le C(B_U, B_L)\|\hat{f}_n - f\|_2^2,$$

which is $O_P(n^{-\frac{2\beta}{2\beta+d}})$ as $\hat{f}_n$ is adaptive. This completes the proof.

4.4. Proof of Theorem 1.1

4.4.1. *Proof of part* 1

Without loss of generality assume that we have data $\{\mathbf{x}_i, y_i\}_{i=1}^{2n}$. We split it into two equal parts and use the second part to construct the estimator $\hat{g}$ of the design density g. Throughout the proof, $\mathbb{E}_{i,P}[\cdot]$ will denote the expectation with respect to the ith half of the sample, with the other half held fixed, under the distribution P. Throughout, we choose the regularity of our wavelet bases to be larger than $\gamma_{\max}$ for the desired approximation and moment properties to hold. As a result our constants depend on $\gamma_{\max}$.

Let $2^{j_{\min} d} = \lfloor n^{\frac{1}{2\beta_{\max}/d+1}} \rfloor$, $2^{j_{\max} d} = \lfloor n^{\frac{1}{2\beta_{\min}/d+1}} \rfloor$, $2^{l_{\min} d} = \lfloor n^{\frac{1}{2\gamma_{\max}/d+1}} \rfloor$, and $2^{l_{\max} d} = \lfloor n^{\frac{1}{2\gamma_{\min}/d+1}} \rfloor$ and define $\mathcal{T}_1 = [j_{\min}, j_{\max}] \cap \mathbb{N}$ and $\mathcal{T}_2 = [l_{\min}, l_{\max}] \cap \mathbb{N}$. Let $\hat{g}_l = \frac{1}{n} \sum_{i=n+1}^{2n} K_{V_l}(\mathbf{x}_i, x)$.

Now, let

$$\hat{l} = \min\left\{ j \in \mathcal{T}_2 : \|\hat{g}_j - \hat{g}_l\|_2 \le C^* \sqrt{\frac{2^{ld}}{n}}, \forall l \in \mathcal{T}_2 \text{ s.t. } l \ge j \right\},$$

where C^* is a constant (depending on $\gamma_{\max}$, B_U) that can be determined from the proof hereafter. Thereafter, consider the Lepski-type estimator $\tilde{g} := \hat{g}_{\hat{l}}$ [34,35]. The following lemma states the mean squared properties of $\tilde{g}$.

Lemma 4.4 (Theorem 2 of [7]). *For any $\gamma_{\min} \le \gamma \le \gamma_{\max}$,*

$$\sup_{P \in \mathcal{P}(\beta,\gamma)} \mathbb{E}_P \|\tilde{g} - g\|_2^2 \le (C)^{\frac{2d}{2\gamma+d}} n^{-\frac{2\gamma}{2\gamma+d}},$$

with a large enough positive constant C depending on M and B_U.

Although the proof of Lemma 4.4 can be found in [7], since we need certain steps of the proof in our subsequent analysis, we provide the proof again in the Appendix C.1.

Now we prove that $\liminf_{n\to\infty} \inf_{P \in \mathcal{P}(\beta,\gamma)} \mathbb{P}_P[\tilde{g} \in B^{\gamma}_{2,\infty}(C)] = 1$ for large enough constant C. Indeed, for any $C > 0$ and $l' \ge J_0$, (letting for any $h \in L_2[0,1]^d$, $\|\langle h, \boldsymbol{\psi}_{l',.}\rangle\|_2$ be the vector L_2 norm of the vector

$$\left(\left\langle h, \boldsymbol{\psi}^v_{l',k'}\right\rangle : k' \in \mathcal{Z}_{l'}, v \in \{0,1\}^d - \{0\}^d\right).$$

Then,

$$\begin{aligned}
&\mathbb{P}_P\left(2^{l'\gamma} \left\|\langle \tilde{g}, \boldsymbol{\psi}_{l',.}\rangle\right\|_2 > C\right) \\
&= \sum_{l=l_{\min}}^{l_{\max}} \mathbb{P}_P\left(2^{l'\gamma} \left\|\langle \hat{g}_l, \boldsymbol{\psi}_{l',.}\rangle\right\|_2 > C, \hat{l} = l\right) \mathcal{I}\left(l' \le l\right) \\
&= \sum_{l=l_{\min}}^{l^*} \mathbb{P}_P\left(2^{l'\gamma} \left\|\langle \hat{g}_l, \boldsymbol{\psi}_{l',.}\rangle\right\|_2 > C, \hat{l} = l\right) \mathcal{I}\left(l' \le l\right) \\
&\quad + \sum_{l=l^*+1}^{l_{\max}} \mathbb{P}_P\left(2^{l'\gamma} \left\|\langle \hat{g}_l, \boldsymbol{\psi}_{l',.}\rangle\right\|_2 > C, \hat{l} = l\right) \mathcal{I}\left(l' \le l\right) \\
&\le \sum_{l=l_{\min}}^{l^*} \mathbb{P}_P\left(2^{l'\gamma} \left\|\langle \hat{g}_l, \boldsymbol{\psi}_{l',.}\rangle\right\|_2 > C\right) \mathcal{I}\left(l' \le l\right) + \sum_{l=l^*+1}^{l_{\max}} \mathbb{P}_P(\hat{l} = l) \mathcal{I}\left(l' \le l\right) \\
&\le \sum_{l=l_{\min}}^{l^*} \mathbb{P}_P\left(2^{l'\gamma} \left\|\langle \hat{g}_l, \boldsymbol{\psi}_{l',.}\rangle\right\|_2 > C\right) \mathcal{I}\left(l' \le l\right) + \sum_{l>l^*} 2e^{-C' 2^{ld/2}} \mathcal{I}\left(l' \le l\right),
\end{aligned} \tag{4.10}$$

where the last inequality follows from (C.7) for a suitable C' (depending on B_U and the wavelet basis choice). Now,

$$\begin{aligned}
&\mathbb{P}_P\big(2^{l'\gamma}\big\|\langle\hat{g}_l,\boldsymbol{\psi}_{l',.}\rangle\big\|_2>C\big)\\
&\quad\le\mathbb{P}_P\big(2^{l'\gamma}\big\|\langle\hat{g}_l,\boldsymbol{\psi}_{l',.}\rangle-\mathbb{E}_P\big(\langle\hat{g}_l,\boldsymbol{\psi}_{l',.}\rangle\big)\big\|_2>C/2\big)\\
&\qquad+\mathbb{P}_P\big(2^{l'\gamma}\big\|\mathbb{E}_P\big(\langle\hat{g}_l,\boldsymbol{\psi}_{l',.}\rangle\big)\big\|_2>C/2\big)\\
&\quad=\mathbb{P}_P\big(2^{l'\gamma}\big\|\langle\hat{g}_l,\boldsymbol{\psi}_{l',.}\rangle-\mathbb{E}_P\big(\langle\hat{g}_l,\boldsymbol{\psi}_{l',.}\rangle\big)\big\|_2>C/2\big)
\end{aligned}$$

if $C>2M'$. Therefore, from (4.10), one has for any $C>2M'$,

$$\begin{aligned}
&\mathbb{P}_P\big(2^{l'\gamma}\big\|\langle\hat{g},\boldsymbol{\psi}_{l',.}\rangle\big\|_2>C\big)\\
&\quad\le\sum_{l=l_{\min}}^{l^*}\mathbb{P}_P\big(2^{l'\gamma}\big\|\langle\hat{g}_l,\boldsymbol{\psi}_{l',.}\rangle-\mathbb{E}_P\big(\langle\hat{g}_l,\boldsymbol{\psi}_{l',.}\rangle\big)\big\|_2>C/2\big)\mathcal{I}\big(l'\le l\big)\\
&\qquad+\sum_{l>l^*}2e^{-C'2^{ld/2}}\mathcal{I}\big(l'\le l\big).
\end{aligned}\tag{4.11}$$

It remains to control $\|\langle\hat{g}_l,\boldsymbol{\psi}_{l',.}\rangle-\mathbb{E}_P(\langle\hat{g}_l,\boldsymbol{\psi}_{l',.}\rangle)\|_2$ appropriately. To this end, note that when $l'\le l$,

$$\begin{aligned}
&\big\|\langle\hat{g}_l,\boldsymbol{\psi}_{l',.}\rangle-\mathbb{E}_P\big(\langle\hat{g}_l,\boldsymbol{\psi}_{l',.}\rangle\big)\big\|_2^2\\
&\quad=\frac{1}{n^2}\sum_{i=n+1}^{2n}\sum_{k',v}\big(\boldsymbol{\psi}^v_{l',k'}(\mathbf{x}_i)-\mathbb{E}_P\big(\boldsymbol{\psi}^v_{l',k'}(\mathbf{x}_i)\big)\big)^2\\
&\qquad+\frac{1}{n^2}\sum_{n+1\le i_1\ne i_2\le 2n}\sum_{k',v}\big(\boldsymbol{\psi}^v_{l',k'}(\mathbf{x}_{i_1})-\mathbb{E}_P\big(\boldsymbol{\psi}^v_{l',k'}(\mathbf{x}_{i_1})\big)\big)\big(\boldsymbol{\psi}^v_{l',k'}(\mathbf{x}_{i_2})-\mathbb{E}_P\big(\boldsymbol{\psi}^v_{l',k'}(\mathbf{x}_{i_2})\big)\big).
\end{aligned}$$

Note that the second term of the above summand is a type U-statistics of order 2 analyzed in Lemma 4.1. We make use of this fact below.

$$\begin{aligned}
&\mathbb{P}_P\big(2^{2l'\gamma}\big\|\langle\hat{g}_l,\boldsymbol{\psi}_{l',.}\rangle-\mathbb{E}_P\big(\langle\hat{g}_l,\boldsymbol{\psi}_{l',.}\rangle\big)\big\|_2^2>C^2/4\big)\\
&\quad\le\mathbb{P}_P\left(\frac{1}{n^2}\sum_{i=n+1}^{2n}\sum_{k',v}\big(\boldsymbol{\psi}^v_{l',k'}(\mathbf{x}_i)-\mathbb{E}_P\big(\boldsymbol{\psi}^v_{l',k'}(\mathbf{x}_i)\big)\big)^2>\frac{C^2/8}{2^{2l'\gamma}}\right)\\
&\qquad+\mathbb{P}_P\left(\left|\frac{1}{n^2}\sum_{\substack{n+1\le i_1\ne i_2\le 2n,\\ k',v}}\begin{array}{c}\big(\boldsymbol{\psi}^v_{l',k'}(\mathbf{x}_{i_1})-\mathbb{E}_P\big(\boldsymbol{\psi}^v_{l',k'}(\mathbf{x}_{i_1})\big)\big)\\ \times\big(\boldsymbol{\psi}^v_{l',k'}(\mathbf{x}_{i_2})-\mathbb{E}_P\big(\boldsymbol{\psi}^v_{l',k'}(\mathbf{x}_{i_2})\big)\big)\end{array}\right|>\frac{C^2/8}{2^{2l'\gamma}}\right)\\
&\quad=I+II.
\end{aligned}$$

To control I note that for any fixed $\mathbf{x} \in [0,1]^d$

$$\sum_{k',v} \big(\boldsymbol{\psi}^v_{l',k'}(\mathbf{x}) - \mathbb{E}_P\big(\boldsymbol{\psi}^v_{l',k'}(\mathbf{x})\big)\big)^2 \le C\big(\boldsymbol{\psi}^0_{0,0}, \boldsymbol{\psi}^1_{0,0}, \gamma_{\max}\big) 2^{l'd},$$

and therefore

$$\mathbb{E}_P \sum_{k',v} \big(\boldsymbol{\psi}^v_{l',k'}(\mathbf{x}) - \mathbb{E}_P\big(\boldsymbol{\psi}^v_{l',k'}(\mathbf{x})\big)\big)^2 \le C\big(\boldsymbol{\psi}^0_{0,0}, \boldsymbol{\psi}^1_{0,0}, \gamma_{\max}\big) 2^{l'd}.$$

Therefore by Hoeffding's Inequality,

$$\begin{aligned} I &\le \mathbb{P}_P\left(\frac{1}{n}\sum_{i=n+1}^{2n} \sum_{k',v} \big(\boldsymbol{\psi}^v_{l',k'}(\mathbf{x}_i) - \mathbb{E}_P\big(\boldsymbol{\psi}^v_{l',k'}(\mathbf{x}_i)\big)\big)^2 > \frac{nC^2/8}{2^{2l'\gamma}}\right) \\ &\le 2e^{-C(\boldsymbol{\psi}^0_{0,0}, \boldsymbol{\psi}^1_{0,0}, \gamma_{\max}) \frac{n}{2^{2l'd}} (\frac{nC^2/8}{2^{2l'\gamma}})^2}. \end{aligned}$$

Finally, arguing similar to Lemma 4.1 we also have that for a constant $C(B_U, \gamma_{\max})$

$$II \le e^{-\frac{Ct(l')^2}{a_1(l')^2}} + e^{-\frac{Ct(l')}{a_2(l')}} + e^{-\frac{C\sqrt{t(l')}}{\sqrt{a_3(l')}}},$$

where $t(l') = \frac{C^2/8}{2^{2l'\gamma}}$, $a_1(l') = \frac{1}{n-1} 2^{\frac{l'd}{2}}$, $a_2(l') = \frac{1}{n-1}(\sqrt{\frac{2^{l'd}}{n}} + 1)$, and $a_3(l') = \frac{1}{n-1}(\sqrt{\frac{2^{l'd}}{n}} + \frac{2^{l'd}}{n})$.

Therefore, for $C > 2M'$

$$\begin{aligned} &\sum_{l' \ge J_0} \mathbb{P}_P\big(2^{l'\gamma} \big\| \langle \tilde{g}, \boldsymbol{\psi}_{l',\cdot} \rangle \big\|_2 > C\big) \\ &\le \sum_{l' \ge J_0} \sum_{l=l_{\min}}^{l^*} \mathbb{P}_P\big(2^{l'\gamma} \big\| \langle \hat{g}_l, \boldsymbol{\psi}_{l',\cdot} \rangle - \mathbb{E}_P\big(\langle \hat{g}_l, \boldsymbol{\psi}_{l',\cdot} \rangle\big) \big\|_2 > C/2\big) \mathcal{I}\big(l' \le l\big) \\ &\quad + \sum_{l' \ge J_0} \sum_{l > l^*} 2e^{-C' 2^{ld/2}} \mathcal{I}\big(l' \le l\big) \qquad (4.12) \\ &\le \sum_{l=l_{\min}}^{l^*} \sum_{l'=J_0}^{l} 2e^{-C(\boldsymbol{\psi}^0_{0,0}, \boldsymbol{\psi}^1_{0,0}, \gamma_{\max}) \frac{n}{2^{2l'd}} (\frac{nC^2/8}{2^{2l'\gamma}})^2} \\ &\quad + \sum_{l=l_{\min}}^{l^*} \sum_{l'=J_0}^{l} \big(e^{-\frac{Ct(l')^2}{a_1(l')^2}} + e^{-\frac{Ct(l')}{a_2(l')}} + e^{-\frac{C\sqrt{t(l')}}{\sqrt{a_3(l')}}}\big) + \sum_{l=l_{\min}}^{l^*} \sum_{l'=J_0}^{l} 2e^{-C' 2^{ld/2}}. \end{aligned}$$

Some tedious calculations now show that the last term in the display above converges to 0 uniformly in $P \in \mathcal{P}(\beta, \gamma)$ as $n \to \infty$. This, along with the definition of $B^{\gamma}_{2,\infty}(C)$, completes the proof of $\liminf_{n\to\infty} \inf_{P \in \mathcal{P}(\beta,\gamma)} \mathbb{P}_P[\tilde{g} \in B^{\gamma}_{2,\infty}(C)] = 1$ for sufficiently large constant C depending on $(M', B_U, \gamma_{\max})$.

However this $\tilde{g}$ does not satisfy the desired point-wise bounds. To achieve this let ψ be a C^∞ function such that $\psi(x)|_{[B_L,B_U]} \equiv x$ while $\frac{B_L}{2} \le \psi(\mathbf{x}) \le 2B_U$ for all x. Finally, consider the estimator $\hat{g}(\mathbf{x}) = \psi(\widetilde{g}(\mathbf{x}))$. We note that $(g(\mathbf{x}) - \hat{g}(\mathbf{x}))^2 \le (g(\mathbf{x}) - \widetilde{g}(\mathbf{x}))^2$ – thus $\hat{g}$ is adaptive to the smoothness of the design density. The boundedness of $\hat{g}$ follows immediately from the construction. Finally, we wish to show that almost surely, the constructed estimator belongs to the Besov space with the same smoothness, possibly of a different radius. This is captured by the next lemma. The proof is deferred to Section C.2.

Lemma 4.5. *For all $h \in B^\beta_{2,\infty}(M)$, $\psi(h) \in B^\beta_{2,\infty}(C(M,\beta))$, where $C(M,\beta)$ is a universal constant dependent only on M, β and independent of $h \in B^\beta_{2,\infty}(M)$.*

The lower bound of the minimax estimation error follows in our case by the results of [7], by setting $f \equiv 0$ in the prior used for the construction of the lower bound.

4.4.2. *Proof of part* 2

For the construction of $\hat{f}$, construct the estimator $\hat{g}$ of the design density g as above from second part of the sample and let $\hat{f}_j(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^n \frac{y_i}{\hat{g}(\mathbf{x}_i)} K_{V_j}(\mathbf{x}_i, \mathbf{x})$. Now, let

$$\hat{j} = \min\left\{ j \in \mathcal{T}_1 : \|\hat{f}_j - \hat{f}_l\|_2 \le C^{**}\sqrt{\frac{2^{ld}}{n}}, \forall l \in \mathcal{T}_1 \text{ s.t. } l \ge j \right\},$$

where C^{**} is a suitable constant (depending on B_U, B_L, $\gamma_{\max}$) to be decided later. Thereafter, consider the estimator $\widetilde{f} := \hat{f}_{\hat{j}}$.

Let $j^* = \min\{j : C_{1*}2^{-j\beta} \le C_{2*}\sqrt{\frac{2^{jd}}{n}}\}$, and note that for any $\mathbf{x} \in [0,1]^d$,

$$\begin{aligned}
&\int \left|\mathbb{E}_{P,1}\left(\hat{f}_j(\mathbf{x})\right) - f(\mathbf{x})\right|^2 d\mathbf{x} \\
&= \int \left| \Pi\left(f\frac{g}{\hat{g}} | V_j \right)(\mathbf{x}) - f(\mathbf{x}) \right|^2 d\mathbf{x} \\
&= \int \left| \Pi\left(f\left(\frac{g}{\hat{g}} - 1\right) \middle| V_j \right)(\mathbf{x}) - \Pi\left(f|V_j^\perp\right)(\mathbf{x}) \right|_2^2 d\mathbf{x} \\
&= \int \left| \Pi\left(f\left(\frac{g}{\hat{g}} - 1\right) \middle| V_j \right)(\mathbf{x}) \right|^2 d\mathbf{x} + \int \left|\Pi\left(f|V_j^\perp\right)(\mathbf{x})\right|_2^2 d\mathbf{x} \\
&= \left\| \Pi\left(f\left(\frac{g}{\hat{g}} - 1\right) \middle| V_j \right) \right\|_2^2 + \left\|\Pi\left(f|V_j^\perp\right)\right\|_2^2 \\
&\le \left\| f\left(\frac{g}{\hat{g}} - 1\right) \right\|_2^2 + C_1^2 M^2 2^{-j\beta}.
\end{aligned}$$

Therefore,

$$\begin{aligned}
&\mathbb{E}_{P,2}\int \left|\mathbb{E}_{P,1}\big(\hat{f}_j(\mathbf{x})\big) - f(\mathbf{x})\right|^2 d\mathbf{x} \\
&\leq \mathbb{E}_{P,2}\left\| f\left(\frac{g}{\hat{g}} - 1\right)\right\|_2^2 + C_1^2 M^2 2^{-j\beta} \\
&\leq \left(\frac{B_U}{B'_L}\right)^2 \big(C\big(M', B_U\big)\big)^{\frac{2}{2\gamma+d}} n^{-\frac{2\gamma}{2\gamma+d}} + C_1^2 M^2 2^{-j\beta}.
\end{aligned} \tag{4.13}$$

Since $\gamma_{\min} > \beta_{\max}$, we have from the definition of j^* (4.13) that there exists a constant C_{1*} depending on M, M', B_U, B_L, $\gamma_{\max}$ such that

$$\mathbb{E}_{P,2}\int \left|\mathbb{E}_{P,1}\big(\hat{f}_{j^*}(\mathbf{x})\big) - f(\mathbf{x})\right|^2 d\mathbf{x} \leq C_{1*}^2 2^{-2j^*\beta}. \tag{4.14}$$

Also by Rosenthal's (Lemma A.1) and Jensen's Inequality, there exists a constant $C(q)$ for $q \geq 2$ such that

$$\begin{aligned}
&\mathbb{E}_{P,1}\big(\big|\hat{f}_j(\mathbf{x}) - \mathbb{E}_{P,1}\big(\hat{f}_j(\mathbf{x})\big)\big|^q\big) \\
&\leq \frac{C(q)}{n^q}\Bigg[\sum_{i=n+1}^{2n} \mathbb{E}_{P,1}\left(\left|\frac{y_i}{\hat{g}(\mathbf{x}_i)} K_{V_j}(\mathbf{x}_i, \mathbf{x})\right|^q\right) \\
&\quad + \left(\sum_{i=n+1}^{2n} \mathbb{E}_{P,1}\left(\left|\frac{y_i}{\hat{g}(\mathbf{x}_i)} K_{V_j}(\mathbf{x}_i, \mathbf{x})\right|^2\right)\right)^{q/2}\Bigg] \\
&\leq \frac{C_{2*}^q/2}{n^q} \times \big[n\big(2^{jd}\big)^{q-1} + n^{q/2}\big(2^{jd}\big)^{q/2}\big],
\end{aligned} \tag{4.15}$$

where the last inequality in the above display follows by using standard facts about compactly supported wavelet basis having regularity larger than $\gamma_{\max}$ [24] and the fact that the constructed $\hat{g}$ from the second half of the sample lies point-wise in $[\frac{B_L}{2}, 2B_U]$. The constant C_{2*} therefore depends on q, the wavelet basis used, B_U and B_L. Therefore, by the choice of $j \in \mathcal{T}_1$, we have that for all $\mathbf{x} \in [0,1]^d$,

$$\mathbb{E}_{P,1}\big(\big|\hat{f}_j(\mathbf{x}) - \mathbb{E}_{P,1}\big(\hat{f}_j(\mathbf{x})\big)\big|^q\big) \leq C_{2*}^q \left(\frac{2^{jd}}{n}\right)^{q/2}. \tag{4.16}$$

Therefore, using (4.14) and (4.16), we have the following bias-variance decomposition bound.

$$\begin{aligned}
\mathbb{E}_P\big(\|\hat{f}_{j^*} - f\|_2^2\big) &= \mathbb{E}_{P,2}\int \mathbb{E}_{P,1}\big(\big|\hat{f}_{j^*}(\mathbf{x}) - f(\mathbf{x})\big|^2\big)\, d\mathbf{x} \\
&= \mathbb{E}_{P,2}\Bigg[\int \mathbb{E}_{P,1}\big(\big|\hat{f}_{j^*}(\mathbf{x}) - \mathbb{E}_{P,1}\big(\hat{f}_{j^*}(\mathbf{x})\big)\big|^2\big)\, d\mathbf{x}
\end{aligned}$$

$$+\int \mathbb{E}_{P,1}\big(\big|\mathbb{E}_{P,1}\big(\hat{f}_{j^*}(\mathbf{x})\big)-f(\mathbf{x})\big|^2\big)\,d\mathbf{x}\Bigg]$$
$$\le C_{1*}^2 2^{-2j^*\beta}+C_{2*}^2\left(\frac{2^{j^*d}}{n}\right)\le 2^{d+1}\big(C_{1*}^2+C_{2*}^2\big)n^{-\frac{2\beta}{2\beta+d}}.$$

Therefore, by definition of $\hat{j}$ and j^*,

$$\begin{aligned}\mathbb{E}_P\big(\|\tilde{f}-f\|_2^2\mathcal{I}\big(\hat{j}\le j^*\big)\big)&\le 2\mathbb{E}_P\big(\|\tilde{f}-\hat{f}_{j^*}\|_2^2\mathcal{I}\big(\hat{j}\le j^*\big)\big)+2\mathbb{E}_P\big(\|\hat{f}_{j^*}-f\|_2^2\big)\\&\le 2\big(\big(C^{**}\big)^2+2^{d+1}\big(C_{1*}^2+C_{2*}^2\big)\big)n^{-\frac{2\beta}{2\beta+d}}.\end{aligned}\tag{4.17}$$

By Cauchy–Schwarz inequality,

$$\mathbb{E}_P\big(\|\tilde{f}-f\|_2^2\mathcal{I}\big(\hat{j}>j^*\big)\big)\le\sum_{j=j^*+1}^{j_{\max}}\sqrt{\mathbb{E}_P\big(\|\hat{f}_j-f\|_2^4\big)}\sqrt{\mathbb{P}_P(\hat{j}=j)}.\tag{4.18}$$

Now, by (4.15) with $q=2$

$$\begin{aligned}\mathbb{E}_{P,1}\big(\hat{f}_j(\mathbf{x})-\mathbb{E}_{P,1}\big(\hat{f}_j(\mathbf{x})\big)\big)^4&\le C(B_U,B_L,\gamma_{\max})\left[\left(\frac{2^{jd}}{n}\right)^3+\left(\frac{2^{jd}}{n}\right)^2\right]\\&\le C(B_U,B_L,\gamma_{\max})\end{aligned}\tag{4.19}$$

by our choice of $2^{j_{\max}d}$. Also, by standard arguments [24], $|\mathbb{E}_{P,1}(\hat{f}_j(\mathbf{x}))|=|\Pi(f\frac{g}{\hat{g}}|V_j)(\mathbf{x})|\le C(B_U,B_L,\gamma_{\max})$ for all $\mathbf{x}\in[0,1]^d$. Therefore by (4.19),

$$\begin{aligned}&\mathbb{E}_P\big(\|\hat{f}_j-f\|_2^4\big)\\&\le 8\mathbb{E}_{P,2}\Bigg[\int\mathbb{E}_{P,1}\big(\hat{f}_j(\mathbf{x})-\mathbb{E}_{P,1}\big(\hat{f}_j(\mathbf{x})\big)\big)^4\,d\mathbf{x}+\int\big(\mathbb{E}_{P,1}\big(\hat{f}_j(\mathbf{x})\big)-f(\mathbf{x})\big)^4\,d\mathbf{x}\Bigg]\\&\le C(B_U,B_L,\gamma_{\max}).\end{aligned}$$

Also, for any constant C''

$$\begin{aligned}\mathbb{P}_P(\hat{j}=l)&\le\sum_{j>j^*}\mathbb{P}_P\left(\|\hat{f}_j-\hat{f}_{j^*}\|_2>C^{**}\sqrt{\frac{2^{jd}}{n}}\right)\\&\le\sum_{j>j^*}\mathbb{E}_{P,2}\left\{\begin{array}{l}\mathbb{P}_{P,1}\left(\begin{array}{c}\big\|\hat{f}_{j^*}-\mathbb{E}_{P,1}(\hat{f}_{j^*})\big\|_2>\dfrac{C^{**}}{2}\sqrt{\dfrac{2^{jd}}{n}}\\-\big\|\mathbb{E}_{P,1}(\hat{f}_{j^*})-\mathbb{E}_{P,1}(\hat{f}_j)\big\|_2\end{array}\right)\\+\mathbb{P}_{P,1}\left(\big\|\hat{f}_j-\mathbb{E}_{P,1}(\hat{f}_j)\big\|_2>\dfrac{C^{**}}{2}\sqrt{\dfrac{2^{jd}}{n}}\right)\end{array}\right\}\end{aligned}$$

$$\leq \sum_{j>j^*} \mathbb{E}_{P,2} \left\{ \mathbb{P}_{P,1} \left(\begin{array}{c} \|\hat{f}_{j^*} - \mathbb{E}_{P,1}(\hat{f}_{j^*})\|_2 > \dfrac{C^{**}}{2}\sqrt{\dfrac{2^{jd}}{n}} \\ - \left\| \Pi\left(f\dfrac{g}{\hat{g}} \middle| V_{j^*} \right) - \Pi\left(f\dfrac{g}{\hat{g}} \middle| V_j \right) \right\|_2 \end{array} \right) \right.$$
$$\left. + \mathbb{P}_{P,1}\left(\|\hat{f}_j - \mathbb{E}_{P,1}(\hat{f}_j)\|_2 > \frac{C^{**}}{2}\sqrt{\frac{2^{jd}}{n}} \right) \right\} \tag{4.20}$$

$$\leq \sum_{j>j^*} \mathbb{E}_{P,2} \left\{ \begin{array}{c} \mathbb{P}_{P,1}\left(\|\hat{f}_{j^*} - \mathbb{E}_{P,1}(\hat{f}_{j^*})\|_2 > \left(\dfrac{C^{**}}{2}\sqrt{\dfrac{2^{jd}}{n}} - C''\sqrt{\dfrac{2^{j^*d}}{n}} \right) \right) \\ + \mathbb{P}_{P,1}\left(\|\hat{f}_j - \mathbb{E}_{P,1}(\hat{f}_j)\|_2 > \dfrac{C^{**}}{2}\sqrt{\dfrac{2^{jd}}{n}} \right) \\ + \mathcal{I}\left(\left\| \Pi\left(f\dfrac{g}{\hat{g}} \middle| V_{j^*} \right) - \Pi\left(f\dfrac{g}{\hat{g}} \middle| V_j \right) \right\|_2 > C''\sqrt{\dfrac{2^{j^*d}}{n}} \right) \end{array} \right\}$$
$$\leq \sum_{j>j^*} 2e^{-C2^{jd/2}} + \sum_{j>j^*} \mathbb{P}_{P,2}\left(\left\| \Pi\left(f\frac{g}{\hat{g}} \middle| V_{j^*} \right) - \Pi\left(f\frac{g}{\hat{g}} \middle| V_j \right) \right\|_2 > C''\sqrt{\frac{2^{j^*d}}{n}} \right),$$

where the inequality in the last display holds by Lemma A.2 since $\max\{y, \frac{1}{\hat{g}(\mathbf{x})}\} \leq C(B_L)$, for a $C > 0$ (depending on M, M', B_U, B_L, $\gamma_{\max}$, C'', C^{**}) if C^{**} is chosen large enough (depending on M, M', B_U, B_L, $\gamma_{\max}$) such that $C^{**} > 2C''$. C'' will be chosen later in the proof to be large enough depending on the known parameters of the problem, which in turn will imply that C^{**} can be chosen large enough depending on the known parameters of the problem as well. Finally,

$$\begin{aligned} &\sum_{j>j^*} \mathbb{P}_{P,2}\left(\left\| \Pi\left(f\frac{g}{\hat{g}} \middle| V_{j^*} \right) - \Pi\left(f\frac{g}{\hat{g}} \middle| V_j \right) \right\|_2 > C''\sqrt{\frac{2^{j^*d}}{n}} \right) \\ &\leq \sum_{j>j^*} \mathbb{P}_{P,2}\left(\|\Pi(f|V_{j^*}) - \Pi(f|V_j)\|_2 > \frac{C''}{2}\sqrt{\frac{2^{j^*d}}{n}} \right) \\ &\quad + \sum_{j>j^*} \mathbb{P}_{P,2}\left(\left\| \Pi\left(f\left(\frac{g}{\hat{g}} - 1\right) \middle| V_{j^*} \right) - \Pi\left(f\left(\frac{g}{\hat{g}} - 1\right) \middle| V_j \right) \right\|_2 > \frac{C''}{2}\sqrt{\frac{2^{j^*d}}{n}} \right) \\ &= I + II. \end{aligned} \tag{4.21}$$

Since $f \in B^{\beta}_{2,\infty}(M)$ and choice of j^*, we have from (C.2) that for C'' chosen sufficiently large (depending on M, M' and $\gamma_{\max}$), one has that $I = 0$. Control of II is more delicate, but can be

handled as below. Using the fact that projection contracts norm, we have

$$
\begin{aligned}
II &\le \sum_{j>j^*} \mathbb{P}_{P,2}\left(\left\|\Pi\left(f\left(\frac{g}{\hat{g}}-1\right)\Big|V_{j^*}\right)\right\|_2 > \frac{C''}{4}\sqrt{\frac{2^{j^*d}}{n}}\right) \\
&\quad + \sum_{j>j^*} \mathbb{P}_{P,2}\left(\left\|\Pi\left(f\left(\frac{g}{\hat{g}}-1\right)\Big|V_{j}\right)\right\|_2 > \frac{C''}{4}\sqrt{\frac{2^{j^*d}}{n}}\right) \qquad (4.22) \\
&\le 2\sum_{j>j^*} \mathbb{P}_{P,2}\left(\|\hat{g}-g\|_2 > \frac{B'_L C''}{4B_U}\sqrt{\frac{2^{j^*d}}{n}}\right).
\end{aligned}
$$

The last term in the above display can be bounded using the following lemma.

Lemma 4.6. *Assume* $\gamma_{\min} > \beta_{\max}$. *Then for constants* $C_1, C_2, C_3 > 0$ (*depending on* M, M', B_U, B_L, $\gamma_{\max}$) *one has*

$$
\begin{aligned}
&\sup_{P\in\mathcal{P}(\beta,\gamma)} \mathbb{P}_{P,2}\left(\|\hat{g}-g\|_2 > \frac{B'_L C''}{4B_U}\sqrt{\frac{2^{j^*d}}{n}}\right) \\
&\le C_1(l_{\max}-l_{\min})^2\left(e^{-C_2 2^{(j^*-l_{\max})d/2}} + e^{-C_3 2^{l_{\min}d/2}}\right).
\end{aligned}
$$

The proof of the lemma involves arguments exactly similar to those involved in the proof of Theorem 1.1 Part 1 and a sketch of the arguments is deferred to Appendix C.

Plugging in the result of Lemma 4.6 into (4.22), and thereafter using the facts that $\gamma_{\min} > \beta_{\max}$, $l_{\max}$, $j_{\max}$ are both poly logarithmic in nature, along with equations (4.21), (4.20), (4.19), and (4.18), followed by some straightforward but tedious algebra, we have the existence of an estimator $\tilde{f}$ depending on M, M', B_U, B_L, $\beta_{\min}$, $\beta_{\max}$, $\gamma_{\max}$, such that for every $(\beta,\gamma) \in [\beta_{\min},\beta_{\max}]\times[\gamma_{\min},\gamma_{\max}]$,

$$
\sup_{P\in\mathcal{P}(\beta,\gamma)} \mathbb{E}_P\|\tilde{f}-f\|_2^2 \le Cn^{-\frac{2\beta}{2\beta+d}},
$$

with a large enough positive constant C depending on M, M', B_U, B_L, $\beta_{\min}$, $\gamma_{\max}$.

The proof of $\liminf_{n\to\infty}\inf_{P\in\mathcal{P}(\beta,\gamma)}\mathbb{P}_P[\tilde{f}\in B^{\beta}_{2,\infty}(C)]=1$, can be done along the lines of the proof of $\liminf_{n\to\infty}\inf_{P\in\mathcal{P}(\beta,\gamma)}\mathbb{P}_P[\tilde{g}\in B^{\gamma}_{2,\infty}(C)]=1$, since using (4.12) and the fact that $\gamma_{\min}>\beta_{\max}$ one can show using arguments similar to proof of Lemma 4.5 that for sufficiently large C, $f\frac{g}{\hat{g}}\in B^{\beta}_{2,\infty}(C)$, with suitably high probability uniformly over $P\in\mathcal{P}(\beta,\gamma)$.

The construction of a $\hat{f}$ from this $\tilde{f}$ and demonstrating its desired properties is very similar to the derivation of $\hat{g}$ from $\tilde{g}$, and hence is omitted.

Next, we derive the lower bound on the estimation error. The proof will be deferred to the Appendix C.4.

Lemma 4.7. *There exists a constant $c > 0$, independent of n, such that*

$$\inf_{\hat{f}_n} \sup_{P \in \mathcal{P}(\beta,\gamma)} \mathbb{E}_P\big[\|\hat{f}_n - f\|_2^2\big] \geq cn^{-\frac{2\beta}{2\beta+d}}. \tag{4.23}$$

This completes the proof of Theorem 1.1.

Appendix A: Technical lemmas

Since the estimators arising in this paper also have a linear term, we will need the following standard Bernstein and Rosenthal type tail and moment bounds [42].

Lemma A.1. *If $\mathbf{O}_1, \ldots, \mathbf{O}_n \sim \mathbb{P}$ are i.i.d. random vectors such that $|L(\mathbf{O})| \leq B$ almost surely $\mathbb{P}$, then for $q \geq 2$ one has for large enough constants $C(B)$ and $C(B,q)$*

$$\mathbb{P}\left(\left|\frac{1}{n}\sum_{i=1}^{n}\big(L(\mathbf{O}_i) - \mathbb{E}\big(L(\mathbf{O}_i)\big)\big)\right| \geq t\right) \leq 2e^{-nt^2/C(B)},$$

and

$$\begin{aligned}
&\mathbb{E}\left(\left|\sum_{i=1}^{n}\big(L(\mathbf{O}_i) - \mathbb{E}\big(L(\mathbf{O}_i)\big)\big)\right|^q\right) \\
&\leq \left[\sum_{i=1}^{n}\mathbb{E}\big(\big|L(\mathbf{O}_i) - \mathbb{E}\big(L(\mathbf{O}_i)\big)\big|^q\big) + \left[\sum_{i=1}^{n}\mathbb{E}\big(\big|L(\mathbf{O}_i) - \mathbb{E}\big(L(\mathbf{O}_i)\big)\big|^2\big)\right]^{q/2}\right] \\
&\leq C(B,q)n^{\frac{q}{2}}.
\end{aligned}$$

We will also need the following concentration inequality for linear estimators based on wavelet projection kernels, proof of which can be done along the lines of proof of Equation (27) of [22] or Theorem 5.1.13 of [23].

Lemma A.2. *Consider i.i.d. observations $\mathbf{O}_i = (Y, \mathbf{X})_i$, $i = 1, \ldots, n$ where $\mathbf{X}_i \in [0,1]^d$ with marginal density g. Let $\hat{m}(\mathbf{x}) = \frac{1}{n}\sum_{i=1}^{n} L(\mathbf{O}_i)K_{V_l}(\mathbf{X}_i, \mathbf{x})$, such that $\max\{\|g\|_\infty, \|L\|_\infty\} \leq B_U$. If $\frac{2^{ld}}{n} \leq 1$, there exists $C, C_1, C_2 > 0$, depending on B_U and scaling functions $\psi^0_{0,0}$, $\psi^1_{0,0}$ respectively, such that*

$$\mathbb{E}\big(\big\|\hat{m} - \mathbb{E}(\hat{m})\big\|_2\big) \leq C\sqrt{\frac{2^{ld}}{n}},$$

and for any $x > 0$

$$\mathbb{P}\left(n\big\|\hat{m} - \mathbb{E}(\hat{m})\big\|_2 > \frac{3}{2}n\mathbb{E}\big(\big\|\hat{m} - \mathbb{E}(\hat{m})\big\|_2\big) + \sqrt{C_1 n 2^{ld/2}x} + C_2 2^{ld/2}x\right) \leq e^{-x}.$$

Appendix B: Proofs of U-statistics deviation results

The following tail bound for second order degenerate U-statistics [23] is due to [20] with constants by [26] and is crucial for our calculations.

Lemma B.1. *Let U_n be a degenerate U-statistic of order 2 with kernel R based on an i.i.d. sample $W_1, \ldots, W_n$. Then there exists a constant C independent of n, such that*

$$P\left[\left|\sum_{i\neq j} R(W_1, W_2)\right| \geq C\left(\Lambda_1\sqrt{u} + \Lambda_2 u + \Lambda_3 u^{3/2} + \Lambda_4 u^2\right)\right] \leq 6\exp(-u),$$

where, we have,

$$\begin{aligned}
\Lambda_1^2 &= \frac{n(n-1)}{2} E\big[R^2(W_1, W_2)\big],\\
\Lambda_2 &= n\sup\big\{E\big[R(W_1, W_2)\zeta(W_1)\xi(W_2)\big] : E\big[\zeta^2(W_1)\big] \leq 1, E\big[\xi^2(W_1)\big] \leq 1\big\},\\
\Lambda_3 &= \big\|nE[R^2(W_1, \cdot)\big\|_\infty^{\frac{1}{2}},\\
\Lambda_4 &= \|R\|_\infty.
\end{aligned}$$

We use this lemma to establish Lemma 4.1.

Proof. Let us analyze $R(\mathbf{O}_1, \mathbf{O}_2) = L(\mathbf{O}_1)K_{V_j}(\mathbf{X}_1, \mathbf{X}_2)L(\mathbf{O}_2)$ first. The proof for $R(\mathbf{O}_1, \mathbf{O}_2) = L(\mathbf{O}_1)K_{W_j}(\mathbf{X}_1, \mathbf{X}_2)L(\mathbf{O}_2)$ is analogous. By Hoeffding's decomposition one has

$$\begin{aligned}
&\frac{1}{n(n-1)}\sum_{i_1\neq i_2} R(\mathbf{O}_{i_1}, \mathbf{O}_{i_2}) - \mathbb{E}\big(R(\mathbf{O}_1, \mathbf{O}_2)\big)\\
&= \frac{2}{n}\sum_{i_1=1}^{n}\big[\mathbb{E}_{\mathbf{O}_{i_1}} R(\mathbf{O}_{i_1}, \mathbf{O}_{i_2}) - \mathbb{E}R(\mathbf{O}_{i_1}, \mathbf{O}_{i_2})\big]\\
&\quad + \frac{1}{n(n-1)}\sum_{i_1\neq i_2}\begin{bmatrix} R(\mathbf{O}_{i_1}, \mathbf{O}_{i_2}) - \mathbb{E}_{\mathbf{O}_{i_1}} R(\mathbf{O}_{i_1}, \mathbf{O}_{i_2})\\ -\mathbb{E}_{\mathbf{O}_{i_2}} R(\mathbf{O}_{i_1}, \mathbf{O}_{i_2}) + \mathbb{E}R(\mathbf{O}_{i_1}, \mathbf{O}_{i_2})\end{bmatrix}\\
&:= \mathrm{T}_1 + T_2.
\end{aligned}$$

B.1. Analysis of T_1

Noting that $T_1 = \frac{2}{n}\sum_{i_1=1}^{n} H(\mathbf{O}_{i_1})$ where $H(\mathbf{O}_{i_1}) = \mathbb{E}(R(\mathbf{O}_{i_1}, \mathbf{O}_{i_2}|\mathbf{O}_{i_1})) - \mathbb{E}R(\mathbf{O}_{i_1}, \mathbf{O}_{i_2})$ we control T_1 by standard Hoeffding's Inequality. First note that,

$$\big|H(\mathbf{O}_{i_1})\big| = \left|\sum_{k\in\mathcal{Z}_j}\sum_{v\in\{0,1\}^d}\big[L(\mathbf{O}_{i_1})\psi_{jk}^v(\mathbf{X}_{i_1})\mathbb{E}\big(\psi_{jk}^v(\mathbf{X}_{i_2})L(\mathbf{O}_{i_2})\big) - \big(\mathbb{E}\big(\psi_{jk}^v(\mathbf{X}_{i_2})L(\mathbf{O}_{i_2})\big)\big)^2\big]\right|$$

$$\leq \sum_{k\in\mathcal{Z}_j}\sum_{v\in\{0,1\}^d}\left|L(\mathbf{O}_{i_1})\psi^v_{jk}(\mathbf{X}_{i_1})\mathbb{E}\big(\psi^v_{jk}(\mathbf{X}_{i_2})L(\mathbf{O}_{i_2})\big)\right| + \sum_{k\in\mathcal{Z}_j}\sum_{v\in\{0,1\}^d}\big(\mathbb{E}\big(\psi^v_{jk}(\mathbf{X}_{i_2})L(\mathbf{O}_{i_2})\big)\big)^2.$$

First, by standard compactness argument for the wavelet bases,

$$\begin{aligned}\left|\mathbb{E}\big(\psi^v_{jk}(\mathbf{X})L(\mathbf{O})\big)\right| &\leq \int\left|\mathbb{E}\big(L(\mathbf{O})|\mathbf{X}=\mathbf{x}\big)\left(2^{\frac{jd}{2}}\prod_{l=1}^{d}\psi^{v_l}_{00}\big(2^j x_l - k_l\big)\right)\right|\big|g(\mathbf{x})\big|\,d\mathbf{x}\\ &\leq C(B, B_U, S)2^{-\frac{jd}{2}}.\end{aligned} \tag{B.1}$$

Therefore,

$$\sum_{k\in\mathcal{Z}_j}\sum_{v\in\{0,1\}^d}\big(\mathbb{E}\big(\psi^v_{jk}(\mathbf{X}_{i_2})L(\mathbf{O}_{i_2})\big)\big)^2 \leq C(B, B_U, S). \tag{B.2}$$

Also, using the fact that for each fixed $\mathbf{x}\in[0,1]^d$, the number indices $k\in\mathcal{Z}_j$ such that $\mathbf{x}$ belongs to support of at least one of ψ^v_{jk} is bounded by a constant depending only on ψ^0_{00} and ψ^1_{00}. Therefore combining (B.1) and (B.2),

$$\begin{aligned}&\sum_{k\in\mathcal{Z}_j}\sum_{v\in\{0,1\}^d}\left|L(\mathbf{O}_{i_1})\psi^v_{jk}(\mathbf{X}_{i_1})\mathbb{E}\big(\psi^v_{jk}(\mathbf{X}_{i_2})L(\mathbf{O}_{i_2})\big)\right|\\ &\leq C(B, B_U, S)2^{-\frac{jd}{2}}2^{\frac{jd}{2}} = C(B, B_U, S).\end{aligned} \tag{B.3}$$

Therefore, by (B.3) and Hoeffding's Inequality,

$$\mathbb{P}\big(|T_1|\geq t\big)\leq 2e^{-C(B,B_U,S)nt^2}. \tag{B.4}$$

B.2. Analysis of T_2

Since T_2 is a degenerate U-statistics, it's analysis is based on Lemma B.1. In particular,

$$T_2 = \frac{1}{n(n-1)}\sum_{i_1\neq i_2} R^*(\mathbf{O}_{i_1}, \mathbf{O}_{i_2}),$$

where

$$R^*(\mathbf{O}_{i_1}, \mathbf{O}_{i_2}) = \sum_{k\in\mathcal{Z}_j}\sum_{v\in\{0,1\}^d}\left\{\begin{array}{l}\big(L(\mathbf{O}_{i_1})\psi^v_{jk}(\mathbf{X}_{i_1}) - \mathbb{E}\big(\psi^v_{jk}(\mathbf{X}_{i_1})\mathbb{E}\big(L(\mathbf{O}_{i_1})|\mathbf{X}_{i_1}\big)\big)\big)\\ \times\big(L(\mathbf{O}_{i_2})\psi^v_{jk}(\mathbf{X}_{i_2}) - \mathbb{E}\big(\psi^v_{jk}(\mathbf{X}_{i_2})\mathbb{E}\big(L(\mathbf{O}_{i_2})|\mathbf{X}_{i_2}\big)\big)\big)\end{array}\right\}.$$

Letting $\Lambda_i, i=1,\ldots,4$ being the relevant quantities as in Lemma B.1, we have the following lemma.

Lemma B.2. *There exists a constant $C = C(B, B_U, S)$ such that*

$$\Lambda_1^2 \le C\frac{n(n-1)}{2}2^{jd}, \qquad \Lambda_2 \le Cn, \qquad \Lambda_3^2 \le Cn2^{jd}, \qquad \Lambda_4 \le C2^{\frac{jd}{2}}.$$

Proof. First we control Λ_1. To this end, note that by simple calculations, using bounds on L, g, and orthonormality of ψ_{jk}^v's we have,

$$\begin{aligned}
\Lambda_1^2 &= \frac{n(n-1)}{2}\mathbb{E}\big(\big\{R^*(\mathbf{O}_1,\mathbf{O}_2)\big\}^2\big) \le 3n(n-1)\mathbb{E}\big(R^2(\mathbf{O}_1,\mathbf{O}_2)\big)\\
&= 3n(n-1)\mathbb{E}\big(L^2(\mathbf{O}_1)K_{V_j}^2(\mathbf{X}_1,\mathbf{X}_2)L^2(\mathbf{O}_2)\big)\\
&\le 3n(n-1)B^4\int\int\Big[\sum_{k\in\mathcal{Z}_j}\sum_{v\in\{0,1\}^d}\psi_{jk}^v(\mathbf{x}_1)\psi_{jk}^v(\mathbf{x}_2)\Big]^2 g(\mathbf{x}_1)g(\mathbf{x}_2)\,d\mathbf{x}_1\,d\mathbf{x}_2\\
&\le 3n(n-1)B^4B_U^2\int\int\Big[\sum_{k\in\mathcal{Z}_j}\sum_{v\in\{0,1\}^d}\psi_{jk}^v(\mathbf{x}_1)\psi_{jk}^v(\mathbf{x}_2)\Big]^2 d\mathbf{x}_1\,d\mathbf{x}_2\\
&= 3n(n-1)B^4B_U^2\sum_{k\in\mathcal{Z}_j}\sum_{v\in\{0,1\}^d}\int\big(\psi_{jk}^v(\mathbf{x}_1)\big)^2 d\mathbf{x}_2\int\big(\psi_{jk}^v(\mathbf{x}_2)\big)^2 d\mathbf{x}_2\\
&\le C(B,B_U,S)n(n-1)2^{jd}.
\end{aligned}$$

Next we control

$$\Lambda_2 = n\sup\big\{\mathbb{E}\big(R^*(\mathbf{O}_1,\mathbf{O}_2)\zeta(\mathbf{O}_1)\xi(\mathbf{O}_2)\big) : \mathbb{E}\big(\zeta^2(\mathbf{O}_1)\big)\le 1, \mathbb{E}\big(\xi^2(\mathbf{O}_2)\big)\le 1\big\}.$$

To this end, we first control

$$\begin{aligned}
&\big|\mathbb{E}\big(L(\mathbf{O}_1)K_{V_j}(\mathbf{X}_1,\mathbf{X}_2)L(\mathbf{O}_2)\zeta(\mathbf{O}_1)\xi(\mathbf{O}_2)\big)\big|\\
&\quad= \Big|\int\mathbb{E}\big(L(\mathbf{O}_1)\zeta(\mathbf{O}_1)|\mathbf{X}_1=\mathbf{x}_1\big)K_{V_j}(\mathbf{x}_1,\mathbf{x}_2)\mathbb{E}\big(L(\mathbf{O}_2)\xi(\mathbf{O}_2)|\mathbf{X}_2=\mathbf{x}_2\big)g(\mathbf{x}_2)g(\mathbf{x}_2)\,d\mathbf{x}_1\,d\mathbf{x}_2\Big|\\
&\quad= \Big|\int\mathbb{E}\big(L(\mathbf{O})\zeta(\mathbf{O})|\mathbf{X}=\mathbf{x}\big)\Pi\big(\mathbb{E}\big(L(\mathbf{O})\xi(\mathbf{O})|\mathbf{X}=\mathbf{x}\big)g(\mathbf{x})|V_j\big)g(\mathbf{x})\,d\mathbf{x}\Big|\\
&\quad\le \Big(\int\mathbb{E}^2\big(L(\mathbf{O})\zeta(\mathbf{O})|\mathbf{X}=\mathbf{x}\big)g^2(\mathbf{x})\,d\mathbf{x}\Big)^{\frac12}\Big(\int\Pi^2\big(\mathbb{E}\big(L(\mathbf{O})\xi(\mathbf{O})|\mathbf{X}=\mathbf{x}\big)g(\mathbf{x})|V_j\big)\,d\mathbf{x}\Big)^{\frac12}\\
&\quad\le \Big(\int\mathbb{E}\big(L^2(\mathbf{O})\zeta^2(\mathbf{O})|\mathbf{X}=\mathbf{x}\big)g^2(\mathbf{x})\,d\mathbf{x}\Big)^{\frac12}\Big(\int\mathbb{E}\big(L^2(\mathbf{O})\xi^2(\mathbf{O})|\mathbf{X}=\mathbf{x}\big)g^2(\mathbf{x})\,d\mathbf{x}\Big)^{\frac12}\\
&\quad\le B^2B_U\sqrt{\mathbb{E}\big(\zeta^2(\mathbf{O}_1)\big)\mathbb{E}\big(\xi^2(\mathbf{O}_2)\big)} \le B^2B_U.
\end{aligned}$$

Above we have used Cauchy-Schwartz Inequality, Jensen's Inequality, and the fact that projections contract norm. Also,

$$\begin{aligned}
&\big|\mathbb{E}\big(\mathbb{E}\big(L(\mathbf{O}_1)K_{V_j}(\mathbf{X}_1,\mathbf{X}_2)L(\mathbf{O}_2)|\mathbf{O}_1\big)\zeta(\mathbf{O}_1)\xi(\mathbf{O}_2)\big)\big| \\
&= \big|\mathbb{E}\big[L(\mathbf{O}_1)\Pi\big(\mathbb{E}\big(L(\mathbf{O}_1)g(\mathbf{X}_1)|\mathbf{X}_1\big)|V_j\big)\zeta(\mathbf{O}_1)\xi(\mathbf{O}_2)\big]\big| \\
&= \big|\mathbb{E}\big[L(\mathbf{O}_1)\Pi\big(\mathbb{E}\big(L(\mathbf{O}_1)g(\mathbf{X}_1)|\mathbf{X}_1\big)|V_j\big)\zeta(\mathbf{O}_1)\big]\big|\big|\mathbb{E}\big(\xi(\mathbf{O}_2)\big)\big| \\
&\le \left|\int \Pi\big(\mathbb{E}\big(L(\mathbf{O})\zeta(\mathbf{O})|\mathbf{X}=\mathbf{x}\big)g(\mathbf{x})|V_j\big)\Pi\big(\mathbb{E}\big(L(\mathbf{O})|\mathbf{X}=\mathbf{x}\big)g(\mathbf{x})|V_j\big)\,d\mathbf{x}\right| \le B^2 B_U,
\end{aligned}$$

where the last step once again uses contraction property of projection, Jensen's Inequality, and bounds on L and g. Finally, by Cauchy-Schwartz Inequality and (B.2),

$$\begin{aligned}
&\mathbb{E}\big[\mathbb{E}\big(L(\mathbf{O}_1)K_{V_j}(\mathbf{X}_1,\mathbf{X}_2)L(\mathbf{O}_2)\big)\zeta(\mathbf{O}_1)\xi(\mathbf{O}_2)\big] \\
&\le \sum_{k\in\mathcal{Z}_j}\sum_{v\in\{0,1\}^d}\mathbb{E}^2\big(L(\mathbf{O})\psi_{jk}^v(\mathbf{X})\big) \le C(B,B_U,S).
\end{aligned}$$

This completes the proof of $\Lambda_2 \le C(B,B_U,S)n$. Turning to $\Lambda_3 = n\|\mathbb{E}[(R^*(\mathbf{O}_1,\cdot))^2]\|_\infty^{1/2}$ we have that

$$\begin{aligned}
\big(R^*(\mathbf{O}_1,\mathbf{o}_2)\big)^2 &\le 2\big[R(\mathbf{O}_1,\mathbf{o}_2)-\mathbb{E}\big(R(\mathbf{O}_1,\mathbf{O}_2)|\mathbf{O}_1\big)\big]^2 \\
&\quad + 2\big[\mathbb{E}\big(R(\mathbf{O}_1,\mathbf{O}_2)|\mathbf{O}_2=\mathbf{o}_2\big)-\mathbb{E}\big(R(\mathbf{O}_1,\mathbf{O}_2)\big)\big]^2.
\end{aligned}$$

Now,

$$\begin{aligned}
&\mathbb{E}\big[R(\mathbf{O}_1,\mathbf{o}_2)-\mathbb{E}\big(R(\mathbf{O}_1,\mathbf{O}_2)|\mathbf{O}_1\big)\big]^2 \\
&\le 2\mathbb{E}\big(L^2(\mathbf{O}_1)K_{V_j}^2(\mathbf{X}_1,\mathbf{x}_2)L^2(\mathbf{o}_2)\big)+2\mathbb{E}\bigg(\sum_{k\in\mathcal{Z}_j}\sum_{v\in\{0,1\}^d}L(\mathbf{O}_1)\psi_{jk}^v(\mathbf{X}_1)\mathbb{E}\big(\psi_{jk}^v(\mathbf{X}_2)L(\mathbf{O}_2)\big)\bigg)^2 \\
&\le 2B^4B_U^2\sum_{k\in\mathcal{Z}_j}\sum_{v\in\{0,1\}^d}\big(\psi_{jk}^v(\mathbf{x}_2)\big)^2+2\mathbb{E}\big(H^2(\mathbf{O}_2)\big)\le C(B,B_U,S)2^{jd},
\end{aligned}$$

where the last inequality follows from arguments along the line of (B.3). Also, using inequalities (B.2) and (B.3)

$$\begin{aligned}
&\big[\mathbb{E}\big(R(\mathbf{O}_1,\mathbf{O}_2)|\mathbf{O}_2=\mathbf{o}_2\big)-\mathbb{E}\big(R(\mathbf{O}_1,\mathbf{O}_2)\big)\big]^2 \\
&=\bigg[\sum_{k\in\mathcal{Z}_j}\sum_{v\in\{0,1\}^d}\mathbb{E}\big(L(\mathbf{O}_1)\psi_{jk}^v(\mathbf{X}_1)\big)\big(\mathbb{E}\big(L(\mathbf{O}_1)\psi_{jk}^v(\mathbf{X}_1)\big)-\psi_{jk}^v(\mathbf{x}_2)L(\mathbf{o}_2)\big)\bigg]^2 \\
&\le C(B,B_U,S).
\end{aligned}$$

This completes the proof of controlling Λ_3. Finally, using compactness of the wavelet basis,

$$\left\| R(\cdot,\cdot) \right\|_\infty \le B^2 \sup_{\mathbf{x}_1,\mathbf{x}_2} \sum_{k\in\mathcal{Z}_j} \sum_{v\in\{0,1\}^d} \left|\psi_{jk}^v(\mathbf{x}_1)\right| \left|\psi_{jk}^v(\mathbf{x}_2)\right| \le C(B, B_U, S) 2^{jd}.$$

Combining this with arguments similar to those leading to (B.3), we have $\Lambda_4 \le C(B, B_U, S)2^{jd}$. □

Therefore, using Lemma B.1 and Lemma B.2 we have

$$\mathbb{P}\left(|T_2| \ge \frac{C(B, B_U, S)}{n-1} \left(\sqrt{2^{jd} t} + t + \sqrt{\frac{2^{jd}}{n}} t^{\frac{3}{2}} + \frac{2^{jd}}{n} t^2 \right) \right) \le 6e^{-t}.$$

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

$$\mathbb{P}_f\left[|T_2| > a_1\sqrt{t} + a_2 t + a_3 t^2 \right] \le 6e^{-t}, \tag{B.5}$$

where $a_1 = \frac{C(B,B_U,S)}{n-1} 2^{\frac{jd}{2}}$, $a_2 = \frac{C(B,B_U,S)}{n-1}(\sqrt{\frac{2^{jd}}{n}} + 1)$, and $a_3 = \frac{C(B,B_U,S)}{n-1}(\sqrt{\frac{2^{jd}}{n}} + \frac{2^{jd}}{n})$. Now if $h(t)$ is such that $a_1\sqrt{h(t)} + a_2 h(t) + a_3 h^2(t) \le t$, then one has by (B.5),

$$\mathbb{P}\left[|T_2| \ge t \right] \le \mathbb{P}\left[|T_2| \ge a_1\sqrt{h(t)} + a_2 h(t) + a_3 h^2(t) \right] \le 6e^{-6h(t)}.$$

Indeed, there exists such an $h(t)$ such that $h(t) = b_1 t^2 \wedge b_2 t \wedge b_3\sqrt{t}$ where $b_1 = \frac{C(B,B_U,S)}{a_1^2}$, $b_2 = \frac{C(B,B_U,S)}{a_2}$, and $b_3 = \frac{C(B,B_U,S)}{\sqrt{a_3}}$. Therefore, there exists $C = C(B, B_U, S)$ such that

$$\mathbb{P}\left[|T_2| \ge t \right] \le e^{-\frac{Ct^2}{a_1^2}} + e^{-\frac{Ct}{a_2}} + e^{-\frac{C\sqrt{t}}{\sqrt{a_3}}}. \tag{B.6}$$

B.3. Combining bounds on T_1 and T_2

Applying union bound along with (B.4) and (B.6) completes the proof of Lemma 4.1. □

Appendix C: Remaining technical details for adaptive estimation

C.1. Proof of Lemma 4.4

To analyze the estimator $\tilde{g}$, we begin with standard bias variance analysis for the candidate estimators $\hat{g}_l$.

Note that for any $\mathbf{x} \in [0,1]^d$, using standard facts about compactly supported wavelet basis having regularity larger than $\gamma_{\max}$ [24], one has for a constant C_1 depending only on the wavelet basis used,

$$\left\|\mathbb{E}_P(\hat{g}_l) - g\right\|_2^2 = \left\|\Pi(g|V_l) - g\right\|_2^2 \le C_1^2 M'^2 2^{-2ld\frac{\gamma}{d}}. \tag{C.1}$$

Above we have used the fact that

$$\sup_{h \in B^{\gamma}_{2,\infty}(M)} \left\|h - \Pi(h|V_l)\right\|_2 \le C_1 M' 2^{-l\gamma}. \tag{C.2}$$

Also by Rosenthal's Inequality [42], there exists a constant $C(q)$ for $q \ge 2$ such that

$$\begin{aligned}
&\mathbb{E}_P\left(\left|\hat{g}_l(\mathbf{x}) - \mathbb{E}_P\left(\hat{g}_l(\mathbf{x})\right)\right|^q\right) \\
&\le \frac{C(q)}{n^q}\left[\sum_{i=n+1}^{2n} \mathbb{E}_P\left(\left|K_{V_l}(\mathbf{x}_i, \mathbf{x})\right|^q\right) + \left(\sum_{i=n+1}^{2n} \mathbb{E}_P\left(\left|K_{V_l}(\mathbf{x}_i, \mathbf{x})\right|^2\right)\right)^{q/2}\right] \\
&\le \frac{C_2^q/2}{n^q} \times \left[n\left(2^{ld}\right)^{q-1} + n^{q/2}\left(2^{ld}\right)^{q/2}\right],
\end{aligned}$$

where the last inequality follows using standard facts about compactly supported wavelet basis having regularity larger than $\gamma_{\max}$ [24] with a constant C_2 that depends only on q and the wavelet basis used. Therefore, for $q \ge 2$, by the choice of $l \in \mathcal{T}_2$, we have that for all $\mathbf{x} \in [0,1]^d$,

$$\mathbb{E}_P\left(\left|\hat{g}_l(\mathbf{x}) - \mathbb{E}_P\left(\hat{g}_l(\mathbf{x})\right)\right|^q\right) \le C_2^q \left(\frac{2^{ld}}{n}\right)^{q/2}. \tag{C.3}$$

Therefore, we have the following bias-variance decomposition.

$$\begin{aligned}
&\mathbb{E}_P\left(\|\hat{g}_l - g\|_2^2\right) \\
&= \int \mathbb{E}_P\left(\left|\hat{g}_l(\mathbf{x}) - g(\mathbf{x})\right|^2\right) d\mathbf{x} \\
&= \left[\int \mathbb{E}_P\left(\left|\hat{g}_l(\mathbf{x}) - \mathbb{E}_P\left(\hat{g}_l(\mathbf{x})\right)\right|^2\right) d\mathbf{x} + \int \mathbb{E}_P\left(\left|\mathbb{E}_P\left(\hat{g}_l(\mathbf{x})\right) - g(\mathbf{x})\right|^2\right) d\mathbf{x}\right] \\
&\le C_1^2 M'^2 2^{-2l\gamma} + C_2^2\left(\frac{2^{ld}}{n}\right).
\end{aligned} \tag{C.4}$$

Let $l^* = \min\{l : C_1 M' 2^{-l\gamma} \le C_2\sqrt{\frac{2^{ld}}{n}}\}$. This implies that

$$\left\|\mathbb{E}_P(\hat{g}_{l^*}) - g\right\|_2^2 \le C_1^2 M'^2 2^{-2l^*\gamma} \le C_2^2\left(\sqrt{\frac{2^{l^*d}}{n}}\right)^2 \le 2^d C_2^2\left(\frac{C_1}{C_2}M'\right)^{\frac{2d}{2\gamma+d}} n^{-\frac{2\gamma}{2\gamma+d}}.$$

Therefore, by definition of $\hat{l}$ and l^*,

$$\begin{aligned}\mathbb{E}_P\big(\|\tilde{g}-g\|_2^2\mathcal{I}\big(\hat{l}\le l^*\big)\big) &= \mathbb{E}_{P,2}\big(\|\tilde{g}-g\|_2^2\mathcal{I}\big(\hat{l}\le l^*\big)\big)\\ &\le 2\mathbb{E}_{P,2}\big(\|\tilde{g}-\hat{g}_{l^*}\|_2^2\mathcal{I}\big(\hat{l}\le l^*\big)\big)+2\mathbb{E}_{P,2}\big(\|\hat{g}_{l^*}-g\|_2^2\big) \qquad \text{(C.5)}\\ &\le 2^{d+1}\big((C^*)^2+2\big)C_2^2\bigg(\frac{C_1}{C_2}M'\bigg)^{\frac{2d}{2\gamma+d}}n^{-\frac{2\gamma}{2\gamma+d}}.\end{aligned}$$

Using Cauchy–Schwarz inequality, we have,

$$\mathbb{E}_P\big(\|\tilde{g}-g\|_2^2\mathcal{I}\big(\hat{l}>l^*\big)\big)\le\sum_{l=l^*}^{j_{\max}}\sqrt{\mathbb{E}_{P,2}\big(\|\hat{g}_l-g\|_2^4\big)}\sqrt{\mathbb{P}_{P,2}(\hat{l}=l)}. \qquad \text{(C.6)}$$

Now, by (C.1), (C.3), choice of $l\in\mathcal{T}_2$, and Jensen's Inequality

$$\begin{aligned}\mathbb{E}_{P,2}\big(\|\hat{g}_l-g\|_2^4\big)&=\mathbb{E}_{P,2}\bigg(\int\big|\hat{g}_l(\mathbf{x})-g(\mathbf{x})\big|^2\,d\mathbf{x}\bigg)^2\le\mathbb{E}_{P,2}\int\big|\hat{g}_l(\mathbf{x})-g(\mathbf{x})\big|^4\,d\mathbf{x}\\ &\le C_1^4M'^4 2^{-4l\gamma}+C_2^4\bigg(\frac{2^{ld}}{n}\bigg)^2\le C_1^4M'^4+C_2^4.\end{aligned}$$

Next, note that for $l>l^*$,

$$\begin{aligned}&\mathbb{P}_{P,2}(\hat{l}=l)\\ &\le\sum_{l>l^*}\mathbb{P}_{P,2}\bigg(\|\hat{g}_l-\hat{g}_{l^*}\|_2>C^*\sqrt{\frac{2^{ld}}{n}}\bigg)\\ &\le\sum_{l>l^*}\left\{\begin{array}{c}\mathbb{P}_{P,2}\bigg(\big\|\hat{g}_{l^*}-\mathbb{E}_{P,2}(\hat{g}_{l^*})\big\|_2>\dfrac{C^*}{2}\sqrt{\dfrac{2^{ld}}{n}}-\big\|\mathbb{E}_{P,2}(\hat{g}_{l^*})-\mathbb{E}_{P,2}(\hat{g}_l)\big\|_2\bigg)\\ +\mathbb{P}_{P,2}\bigg(\big\|\hat{g}_l-\mathbb{E}_{P,2}(\hat{g}_l)\big\|_2>\dfrac{C^*}{2}\sqrt{\dfrac{2^{ld}}{n}}\bigg)\end{array}\right\}\\ &\le\sum_{l>l^*}\left\{\begin{array}{c}\mathbb{P}_{P,2}\bigg(\big\|\hat{g}_{l^*}-\mathbb{E}_{P,2}(\hat{g}_{l^*})\big\|_2>\dfrac{C^*}{2}\sqrt{\dfrac{2^{ld}}{n}}-\big\|\Pi(g|V_{l^*})-\Pi(g|V_l)\big\|_2\bigg)\\ +\mathbb{P}\bigg(\big\|\hat{g}_l-\mathbb{E}_{P,2}(\hat{g}_l)\big\|_2>\dfrac{C^*}{2}\sqrt{\dfrac{2^{ld}}{n}}\bigg)\end{array}\right\} \qquad \text{(C.7)}\\ &\le\sum_{l>l^*}\left\{\begin{array}{c}\mathbb{P}_{P,2}\bigg(\big\|\hat{g}_{l^*}-\mathbb{E}_{P,2}(\hat{g}_{l^*})\big\|_2>\dfrac{C^*}{2}\sqrt{\dfrac{2^{ld}}{n}}-2C_2\sqrt{\dfrac{2^{l^*d}}{n}}\bigg)\\ +\mathbb{P}\bigg(\big\|\hat{g}_l-\mathbb{E}_{P,2}(\hat{g}_l)\big\|_2>\dfrac{C^*}{2}\sqrt{\dfrac{2^{ld}}{n}}\bigg)\end{array}\right\}\end{aligned}$$

$$\leq \sum_{l>l^*} \left\{ \begin{array}{l} \mathbb{P}_{P,2}\Big(\big\| \hat{g}_{l^*} - \mathbb{E}_{P,2}(\hat{g}_{l^*}) \big\|_2 > \Big(\frac{C^*}{2} - 2C_2 \Big) \sqrt{\frac{2^{ld}}{n}} \Big) \\ \qquad + \mathbb{P}\Big(\big\| \hat{g}_l - \mathbb{E}_{P,2}(\hat{g}_l) \big\|_2 > \frac{C^*}{2} \sqrt{\frac{2^{ld}}{n}} \Big) \end{array} \right\}$$

$$\leq \sum_{l>l^*} 2e^{-C2^{ld/2}},$$

for a $C > 0$ (depending on B_U and the wavelet basis choice) if C^* is chosen large enough (depending on M' and B_U) such that $C^* > 2C_2$. In the fourth and fifth of the above series of inequalities, we have used (C.2) and the definition of l^* respectively. The last line follows by an argument similar to results in Section 3.1 of [22]. Finally combining equations (C.5), (C.6) and (C.7), we have the existence of an estimator $\tilde{g}$ depending on B_U, $\gamma_{\min}$, and $\gamma_{\max}$, such that for every $(\beta, \gamma) \in [\beta_{\min}, \beta_{\max}] \times [\gamma_{\min}, \gamma_{\max}]$,

$$\sup_{P \in \mathcal{P}(\beta,\gamma)} \mathbb{E}_P \|\tilde{g} - g\|_2^2 \leq C n^{-\frac{2\gamma}{2\gamma+d}},$$

with a large enough positive constant C depending on M, B_U, $\gamma_{\min}$.

C.2. Proof of Lemma 4.5

We will utilize the equivalent definition of Besov space in terms of moduli of smoothness. We define the forward difference operator $\Delta_h(f)(x) = f(x+h) - f(x)$ and the operator $\Delta_h^r = \Delta_h(\Delta_h^{r-1})$ for $r \geq 2$, where $\Delta_h^1 = \Delta$. Next, for $t > 0$ and r a natural number greater than β, we define the modulus of smoothness $\omega_r(f,t) = \sup_{|h| \leq t} \|\Delta_h^r(f)\|_2$. Finally, we define the Besov semi-norm $|f|_{B^\beta_{2,\infty}} = \sup_{t>0} \omega_r(f,t)/t^\beta$. Finally, we define

$$B^\beta_{2,\infty}(M) = \big\{ f \in L^2 : \|f\|_{B^\beta_{2,\infty}} = \|f\|_2 + |f|_{B^\beta_{2,\infty}} \leq M \big\}. \tag{C.8}$$

It is a standard fact [24] that (C.8) is an equivalent definition of a Besov space. Further, the supremum in the definition of $|f|_{B^\beta_{2,\infty}}$ may be restricted to $0 < t < 1$. Throughout this proof, we work with $B^\beta_{2,\infty}(M)$ defined by (C.8) without loss of generality. We first consider the case when $0 < \beta < 1$. In this case, it is easy to see that $\|\phi(f)\|_2 < C(\phi)$, for some universal constant $C(\phi)$ depending on ϕ and independent of f. Next, we control the term $|\phi(f)|_{B^\beta_{2,\infty}}$. Using Mean Value Theorem, we have,

$$\Delta_h\big(\phi(f)\big)(x) = \phi\big(f(x+h)\big) - \phi\big(f(x)\big) = \phi'(\xi)\Delta_h(f)(x),$$

for some $\xi \in [\min\{f(x), f(x+h)\}, \max\{f(x), f(x+h)\}]$. This naturally implies $\omega_1(\phi(f), t) \leq \|\phi\|_\infty \omega_1(f,t)$, which gives us the desired claim in this case.

Next, we consider the case when $\beta > 1$. We note that for any $r \geq 1$, we have, $\Delta_h^r(f)(x) = \sum_{k=0}^r \binom{r}{k}(-1)^{r-k} f(x+kh)$. Setting $r = \lceil \beta \rceil$, we have, by Taylor expansion for ϕ,

$$\begin{aligned}\Delta_h^r\big(\phi(f)\big)(x) &= \sum_{k=0}^r \binom{r}{k}(-1)^{r-k}\phi\big(f(x+kh)\big) \\ &= \sum_{k=0}^r \binom{r}{k}(-1)^{r-k}\left[\phi'\big(f(x)\big)\Delta_{kh}(f)(x) + \frac{\phi''(\xi(x))}{2}\big(\Delta_{kh}(f)(x)\big)^2\right] \\ &= \phi'\big(f(x)\big)\Delta_h^r\big(\phi(f)\big)(x) + \Sigma(x,h).\end{aligned}$$

Thus we have, $\|\Delta_h^r(\phi(f))\|_2 \leq \|\phi'\|_\infty \|\Delta_h^r(\phi(f))\|_2 + \|\Sigma(\cdot,h)\|_2$. To control $\|\Sigma\|_2$, we use the fact that $B_{2,\infty}^\beta(M) \subset B_{\infty,\infty}^{\beta-1/2}(C(M,\beta))$, where $\subset$ stands for the usual embedding operation (results of similar flavor can be found in [52,55]). This naturally implies that $(\Delta_{kh}(f)(x))^2 \lesssim (kh)^{2\beta-1}$. Thus,we have,

$$\sup_{0<t<1} \frac{\sup_{|h|\leq t}\|\Sigma(\cdot,h)\|_2}{t^\beta} \leq C(M,\beta)t^{\beta-1}.$$

This completes the proof.

C.3. Proof of Lemma 4.6

Indeed, $\hat{g} = \psi(\tilde{g})$, where $\psi(x)$ is C^∞ function which is identically equal to x on $[B_L, B_U]$ and has universally bounded first derivative. Therefore, it is enough to prove Lemma 4.6 for $\tilde{g}$ instead of $\hat{g}$ and thereby invoking a simple first order Taylor series argument along with the fact that $\psi(g) \equiv g$ owing to the bounds on g. The proof of the lemma is therefore very similar to the proof of adaptivity of $\hat{g}$ (by dividing into cases where the chosen $\hat{l}$ is larger and smaller than l^* respectively and thereafter invoking Lemma A.2) and therefore we simply state the main idea and omit the details. The crux of the argument for proving Lemma 4.6 relies on the fact that by Lemma A.2, any $\hat{g}_l$ for $l \in \mathcal{T}_2$ suitably concentrates around g in a radius of the order of $\sqrt{\frac{2^{ld}}{n}}$, and Lepski's method chooses an index $\hat{l} \leq l^*$ with high probability. Thereafter one uses the fact that $\gamma_{\min} > \beta_{\max}$, and consequently $2^{ld} \ll 2^{jd}$ for any $(j,l) \in \mathcal{T}_1 \times \mathcal{T}_2$.

C.4. Proof of Lemma 4.7

The proof will follow the usual approach of lower bounding the estimation error by a related "testing" problem [57]. We will equip our parameter space with the distance function $d((f,g),(f',g')) = \sqrt{\|f-f'\|_2^2 + \|g-g'\|_2^2}$.

We will use M distributions in our derivation of the lower bound – M will be chosen appropriately later. The distributions $\mathcal{C} = \{(f_i, g_i) : 1 \leq i \leq M\}$ are chosen as follows: we set $g_i = 1$

for all i, that is, we set the design density to be uniform. Next, we set $j_0 = \lceil \frac{d}{2\beta+d} \log_2 n \rceil$. Let

$$f_i(x) = \frac{1}{2} + \varepsilon 2^{-j_0(1/2+\beta/d)} \sum_{k \in \mathcal{Z}_{j_0}} \sum_{v \in \{0,1\}^d - \{0\}} \alpha_{i,k}^v \psi_{j_0,k}^v(x),$$

where each $\alpha_{i,k}^v \in \{0, 1\}$. The constant $\varepsilon > 0$ is chosen sufficiently small such that $0 \le f_i \le 1$ for all $x \in [0, 1]^d$. Thus we have, for $(f, g), (f', g') \in \mathcal{C}$,

$$d\big((f, g), \big(f', g'\big)\big)^2 = \big\|f - f'\big\|_2^2 = \varepsilon^2 \frac{1}{n} \sum_{v \in \{0,1\}^d - \{0\}} \rho\big(\alpha_{i\cdot}^v, \alpha_{i'\cdot}^v\big),$$

where $\alpha_{i\cdot}^v = (\alpha_{i,k}^v)$ and $\rho(\cdot, \cdot)$ is the Hamming distance between two vectors on the hypercube. For each $v \in \{0, 1\}^d - \{0\}$, we apply the Varshamov–Gilbert Lemma (Lemma 2.9 of [57]) to select $(\alpha_{i,\cdot}^v)$ with mutual separation at least $\frac{1}{8} n^{\frac{d}{2\beta+d}}$. The Varshamov–Gilbert Lemma guarantees the existence of such a subset with size at least $2^{\frac{1}{8} n^{\frac{d}{2\beta+d}}}$. Thus we have, with $M = 2^{\frac{2^d-2}{8} n^{\frac{d}{2\beta+d}}}$, for any $(f, g), (f', g') \in \mathcal{C}$,

$$d\big((f, g), \big(f', g'\big)\big)^2 \ge \frac{(2^d - 2)\varepsilon^2}{8} n^{-\frac{2\beta}{2\beta+d}}.$$

We denote the joint distribution of $\{\mathbf{x}_l, y_l : 1 \le l \le n\}$ under the parameters (f_i, g_i) by $\mathcal{P}_i$. Thus we have, $\chi^2(\mathcal{P}_i, \mathcal{P}_0) = [1 + \chi^2((f_i, g_i), (f_0, g_0))]^n - 1$.

Finally, we note that $1 + \chi^2((f_i, g_i), (f_0, g_0)) = \mathbb{E}_0[(\frac{f_i(\mathbf{x}_1)^{y_1}(1-f_i(\mathbf{x}_1))^{(1-y_1)}}{1/2})^2] = 4\mathbb{E}_0[f_i^2(\mathbf{x}_1) + (1 - f(\mathbf{x}_i))^2]$, where $\mathbb{E}_0[\cdot]$ represents the expectation with respect to (f_0, g_0). Setting $f_i = 1/2 + \psi_i$, we have,

$$1 + \chi^2\big((f_i, g_i), (f_0, g_0)\big) = 1 + 4\mathbb{E}_0\big[\psi_i(\mathbf{x}_1)^2\big] \le 1 + 4\big(2^d - 2\big)\varepsilon^2 n^{-\frac{2\beta}{2\beta+d}}.$$

Thus $\chi^2(\mathcal{P}_i, \mathcal{P}_0) \le \exp(4(2^d - 2)\varepsilon^2 n^{\frac{d}{2\beta+d}}) \le \delta M$, for some $0 < \delta < 1/8$ if $\varepsilon > 0$ is chosen sufficiently small. This allows us to complete the proof by an application of Theorem 2.7 in [57].

Acknowledgments

The authors thank the Associate Editor and two anonymous referees for numerous helpful comments which substantially improved the content and presentation of the paper.

References

[1] Antoniadis, A., Besbeas, P. and Sapatinas, T. (2001). Wavelet shrinkage for natural exponential families with cubic variance functions. *Sankhyā Ser. A* **63** 309–327. MR1897045

[2] Antoniadis, A. and Leblanc, F. (2000). Nonparametric wavelet regression for binary response. *Statistics* **34** 183–213. MR1802727
[3] Antoniadis, A. and Sapatinas, T. (2001). Wavelet shrinkage for natural exponential families with quadratic variance functions. *Biometrika* **88** 805–820. MR1859411
[4] Baraud, Y. (2002). Model selection for regression on a random design. *ESAIM Probab. Stat.* **6** 127–146 (electronic). MR1918295
[5] Baraud, Y. (2004). Confidence balls in Gaussian regression. *Ann. Statist.* **32** 528–551. MR2060168
[6] Bull, A.D. (2012). Honest adaptive confidence bands and self-similar functions. *Electron. J. Stat.* **6** 1490–1516. MR2988456
[7] Bull, A.D. and Nickl, R. (2013). Adaptive confidence sets in L^2. *Probab. Theory Related Fields* **156** 889–919. MR3078289
[8] Burnashev, M.V. (1979). On the minimax detection of an inaccurately known signal in a white Gaussian noise background. *Theory Probab. Appl.* **24** 107–119.
[9] Cai, T.T. (2012). Minimax and adaptive inference in nonparametric function estimation. *Statist. Sci.* **27** 31–50. MR2953494
[10] Cai, T.T. and Low, M.G. (2004). An adaptation theory for nonparametric confidence intervals. *Ann. Statist.* **32** 1805–1840. MR2102494
[11] Carpentier, A. (2013). Honest and adaptive confidence sets in L_p. *Electron. J. Stat.* **7** 2875–2923. MR3148371
[12] Carpentier, A. (2015). Testing the regularity of a smooth signal. *Bernoulli* **21** 465–488. MR3322327
[13] Cohen, A., Daubechies, I. and Vial, P. (1993). Wavelets on the interval and fast wavelet transforms. *Appl. Comput. Harmon. Anal.* **1** 54–81. MR1256527
[14] Donoho, D.L. and Johnstone, I.M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* **81** 425–455. MR1311089
[15] Donoho, D.L. and Johnstone, I.M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *J. Amer. Statist. Assoc.* **90** 1200–1224. MR1379464
[16] Efromovich, S. (1996). On nonparametric regression for IID observations in a general setting. *Ann. Statist.* **24** 1125–1144. MR1401841
[17] Efromovich, S. and Pinsker, M. (1996). Sharp-optimal and adaptive estimation for heteroscedastic nonparametric regression. *Statist. Sinica* **6** 925–942. MR1422411
[18] Galtchouk, L. and Pergamenshchikov, S. (2008). Adaptive asymptotically efficient estimation in heteroscedastic nonparametric regression via model selection. Preprint. Available at arXiv:0810.1173.
[19] Galtchouk, L. and Pergamenshchikov, S. (2009). Sharp non-asymptotic oracle inequalities for nonparametric heteroscedastic regression models. *J. Nonparametr. Stat.* **21** 1–18. MR2483856
[20] Giné, E., Latala, R. and Zinn, J. (2000). Exponential and moment inequalities for U-statistics. In *High Dimensional Probability, II* (*Seattle, WA*, 1999). *Progress in Probability* **47** 13–38. Boston, MA: Birkhäuser. MR1857312
[21] Giné, E. and Nickl, R. (2010). Confidence bands in density estimation. *Ann. Statist.* **38** 1122–1170. MR2604707
[22] Giné, E. and Nickl, R. (2011). Rates on contraction for posterior distributions in L^r-metrics, $1 \leq r \leq \infty$. *Ann. Statist.* **39** 2883–2911. MR3012395
[23] Giné, E. and Nickl, R. (2015). Mathematical Foundations of Infinite-Dimensional Statistical Models.
[24] Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov, A. (1998). *Wavelets, Approximation, and Statistical Applications. Lecture Notes in Statistics* **129**. New York: Springer. MR1618204
[25] Hoffmann, M. and Nickl, R. (2011). On adaptive inference and confidence bands. *Ann. Statist.* **39** 2383–2409. MR2906872
[26] Houdré, C. and Reynaud-Bouret, P. (2003). Exponential inequalities, with constants, for U-statistics of order two. In *Stochastic Inequalities and Applications. Progress in Probability* **56** 55–69. Basel: Birkhäuser. MR2073426

[27] Ingster, Y. and Suslina, I.A. (2012). *Nonparametric Goodness-of-Fit Testing Under Gaussian Models*. Springer: Springer Science & Business Media.
[28] Ingster, Yu.I. (1990). Minimax detection of a signal in l_p-metrics. *J. Math. Sci.* **68** 503–515.
[29] Ingster, Yu.I. and Sapatinas, T. (2009). Minimax goodness-of-fit testing in multivariate nonparametric regression. *Math. Methods Statist.* **18** 241–269. MR2560455
[30] Johnstone, I.M. (2002). Function estimation and Gaussian sequence models. Unpublished Manuscript.
[31] Kerkyacharian, G., Nickl, R. and Picard, D. (2012). Concentration inequalities and confidence bands for needlet density estimators on compact homogeneous manifolds. *Probab. Theory Related Fields* **153** 363–404. MR2925578
[32] Lepski, O.V. and Spokoiny, V.G. (1997). Optimal pointwise adaptive methods in nonparametric estimation. *Ann. Statist.* **25** 2512–2546. MR1604408
[33] Lepski, O.V. and Spokoiny, V.G. (1999). Minimax nonparametric hypothesis testing: The case of an inhomogeneous alternative. *Bernoulli* **5** 333–358. MR1681702
[34] Lepskiĭ, O.V. (1990). A problem of adaptive estimation in Gaussian white noise. *Theory Probab. Appl.* **35** 454–466.
[35] Lepskiĭ, O.V. (1991). Asymptotically minimax adaptive estimation. I. Upper bounds. Optimally adaptive estimates. *Theory Probab. Appl.* **36** 682–697.
[36] Lepskiĭ, O.V. (1992). On problems of adaptive estimation in white Gaussian noise. In *Topics in Nonparametric Estimation*. *Adv. Soviet Math.* **12** 87–106. Providence, RI: Amer. Math. Soc. MR1191692
[37] Lepskii, O.V. (1993). On asymptotically exact testing of nonparametric hypotheses. Technical report, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE).
[38] Li, K.-C. (1989). Honest confidence regions for nonparametric regression. *Ann. Statist.* **17** 1001–1008. MR1015135
[39] Low, M.G. (1997). On nonparametric confidence intervals. *Ann. Statist.* **25** 2547–2554. MR1604412
[40] McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models* **37**. Boca Raton: CRC press.
[41] Nickl, R. and Szabó, B. (2014). A sharp adaptive confidence ball for self-similar functions. Preprint. Available at arXiv:1406.3994.
[42] Petrov, V.V. (1995). *Limit Theorems of Probability Theory. Sequences of Independent Random Variables*. *Oxford Studies in Probability* **4**. New York: The Clarendon Press. MR1353441
[43] Picard, D. and Tribouley, K. (2000). Adaptive confidence interval for pointwise curve estimation. *Ann. Statist.* **28** 298–335. MR1762913
[44] Ray, K. (2014). Bernstein–von Mises theorems for adaptive Bayesian nonparametric procedures. Preprint. Available at arXiv:1407.3397.
[45] Robins, J., Li, L., Mukherjee, R., Tchetgen, E.T. and van der Vaart, A. (2015). Higher order estimating equations for high-dimensional models. Preprint. Available at arXiv:1512.02174.
[46] Robins, J., Li, L., Tchetgen, E. and van der Vaart, A. (2008). Higher order influence functions and minimax estimation of nonlinear functionals. In *Probability and Statistics*: *Essays in Honor of David a. Freedman*. *Inst. Math. Stat. Collect.* **2** 335–421. Beachwood, OH: IMS. MR2459958
[47] Robins, J., Tchetgen, E., Li, L. and van der Vaart, A. (2009). Semiparametric minimax rates. *Electron. J. Stat.* **3** 1305–1321. MR2566189
[48] Robins, J. and van der Vaart, A. (2006). Adaptive nonparametric confidence sets. *Ann. Statist.* **34** 229–253. MR2275241
[49] Robins, J.M., Rotnitzky, A. and Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *J. Amer. Statist. Assoc.* **89** 846–866. MR1294730
[50] Rosenbaum, P.R. and Rubin, D.B. (1983). The central role of the propensity score in observational studies for causal effects. *Biometrika* **70** 41–55. MR0742974
[51] Signorini, D.F. and Jones, M.C. (2004). Kernel estimators for univariate binary regression. *J. Amer. Statist. Assoc.* **99** 119–126. MR2054291

[52] Simon, J. (1990). Sobolev, Besov and Nikolsskiĭ fractional spaces: Imbeddings and comparisons for vector valued spaces on an interval. *Ann. Mat. Pura Appl.* (4) **157** 117–148. MR1108473
[53] Spokoiny, V.G. (1996). Adaptive hypothesis testing using wavelets. *Ann. Statist.* **24** 2477–2498. MR1425962
[54] Szabó, B., van der Vaart, A.W. and van Zanten, J.H. (2015). Frequentist coverage of adaptive nonparametric Bayesian credible sets. *Ann. Statist.* **43** 1391–1428. MR3357861
[55] Triebel, H. (2006). Theory of function spaces. III, vol. 100 of Monographs in Mathematics.
[56] Tsiatis, A. (2007). *Semiparametric Theory and Missing Data*. Springer: Springer Science & Business Media.
[57] Tsybakov, A.B. (2008). *Introduction to Nonparametric Estimation*. Springer: Springer Science & Business Media.

Received January 2016 and revised July 2016