OPTIMAL ADAPTIVE INFERENCE IN RANDOM DESIGN BINARY REGRESSION

BY RAJARSHI MUKHERJEE AND SUBHABRATA SEN

Department of Statistics, Stanford University

We construct confidence sets for the regression function in non-parametric 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.

In many epidemiological studies, a binary response variable $Y$ is independently observed on a population of individuals along with multiple covariates $X$ 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 (Rosenbaum and Rubin, 1983), inverse probability weighted inference (Robins, Rotnitzky and Zhao, 1994) 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 Tsiatis (2007).

Historically, regression models with binary outcomes have been approached through both parametric (McCullagh and Nelder, 1989) and nonparametric lenses (Antoniadis and Leblanc, 2000; Signorini and Jones, 2004). 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 $(x_i, y_i)_{i=1}^n$, where $x_i \in [0, 1]^d$ and $y_i \in \{0, 1\}$. Consider the binary regression model

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

\[
l(y, x | f, g) = f(x)^y(1 - f(x))^{1-y}g(x).
\] (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 (Burnashev (1979), Ingster (1994), Spokoiny (1996), Lepskii and Spokoiny (1999), Ingster and Sapatinas (2009), Ingster and Suslina (2012), for goodness of fit testing, Lepski (1992); Lepskii (1992, 1993), Donoho and Johnstone (1994, 1995), Lepskii and Spokoiny (1997), Johnstone (2002), Cai (2012) for adaptive estimation, and Li (1989), Low (1997), Baraud (2004), Cai and Low (2004), Robins and Van Der Vaart (2006), Giné and Nickl (2010), Hoffmann and Nickl (2011), Bull and Nickl (2013), Szabó et al. (2015) 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 Carpentier (2013) 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 \( \frac{d}{2} \) derivatives was developed by Ingster and Sapatinas (2009).

(c) We provide theory for adaptive confidence sets which complements those obtained in density (Bull and Nickl, 2013) and sequence models (Carpentier, 2013; Robins and Van Der Vaart, 2006) (see Theorem 1.3). A part of the adaptation theory for Hölder balls was sketched briefly in Robins et al. (2008) 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 Bull and Nickl (2013), 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 Bull and Nickl (2013) 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 Carpentier (2015).

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 \( \|h\|_q := (\int_{[0,1]^d} |h(x)|^q \, dx)^{1/q} \) and \( \|h\|_\infty := \sup_{x \in [0,1]^d} |h(x)| \) denote the usual \( L_q \) and \( L_\infty \) seminorm 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,

\[ \mathcal{P}(\beta, \gamma, M, M', B_L, B_U) = \left\{ (f, g) : f \in B^{\beta}_{2, \infty}(M), g \in B^{\gamma}_{2, \infty}(M'), 0 < f < 1, 0 < B_L \leq g \leq B_U, \int g(x) \, dx = 1 \right\}. \]  

(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 Bull and Nickl (2013). 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 \(\hat{g} = \hat{g}(x_1, \ldots, x_n, B_U, \gamma_{\min}, \gamma_{\max})\) of the design density \(g\) and constant \(C = C(M', \gamma_{\min}, \gamma_{\max}, B_U)\) such that for each \(\gamma \in [\gamma_{\min}, \gamma_{\max}]\) and \(\beta > 0\),

\[ \sup_{P \in \mathcal{P}(\beta, \gamma)} \mathbb{E}_P \left[ \|\hat{g} - g\|_2^2 \right] \leq C n^{-\frac{2\gamma}{2\gamma + d}}, \]

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

and there exists constants \(0 < B'_{L} \leq B'_{U}\) (depending on \(B_L, B_U\)) such that \(B'_{L} \leq \hat{g} \leq B'_{U}\) almost surely. Further, there exists a universal constant \(c > 0\) such that

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

2. Let \(0 < \beta_{\min} \leq \beta_{\max}, \gamma_{\min} < \gamma_{\max}\) be given. If \(\gamma_{\min} > \beta_{\max}\), there exists a sequence of estimators \(\hat{f} = \hat{f}(x_1, y_1, \ldots, x_n, y_n, M, M', B_U, B_L, \beta_{\min}, \beta_{\max}, \gamma_{\max})\) and constant \(C = C(M, M', B_U, B_L, \beta_{\min}, \beta_{\max}, \gamma_{\max})\) such that for every \(\beta \in [\beta_{\min}, \beta_{\max}]\) and \(\gamma \in [\gamma_{\min}, \gamma_{\max}]\)

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

\[ \liminf_{n \to \infty} \inf_{P \in \mathcal{P}(\beta, \gamma)} \mathbb{P}[\hat{f} \in B^\beta_{2, \infty}(C)] = 1, \]
and there exist constants $C_L \leq C_U$ such that $C_L \leq \hat{f} \leq C_U$ almost surely. Further, there exists a constant $c > 0$, independent of $n$, such that

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

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 Bull and Nickl (2013). It is worth noting that results of the kind stating that $\hat{g} \in B^2_{2,\infty}(M^*)$ 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 $\hat{g} \in B^2_{2,\infty}(M^*)$ 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 Baraud (2002) 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 $\gamma_{\text{max}}$ stems from deciding the regularity of the wavelet basis used. Once we fix a wavelet basis with regularity $S > \gamma_{\text{max}}$, the dependence of our constants on $\gamma_{\text{max}}$ 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

$$\mathcal{P}_0(\beta, \gamma) = \{(f, g): f \equiv 1/2, g \in B^\gamma_{2,\infty}(M'), B_L < g < B_U, \int g(x)dx = 1\},$$

$$\mathcal{P}(\beta, \gamma, \rho_n^2) = \{(f, g): f \in B^\gamma_{2,\infty}(M), \|f - 1/2\|_2^2 > \rho_n^2, g \in B^\gamma_{2,\infty}(M'), B_L < g < B_U, \int g(x)dx = 1\}.$$

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

$$\mathcal{P}(\beta_1, \beta_2, \gamma, \rho_n^2) = \left\{(f, g): f \in B^{\beta_2}_{2,\infty}(M), \|f - B^{\beta_1}_{2,\infty}(M)\|_2^2 > \rho_n^2, g \in B^\gamma_{2,\infty}(M'), B_L < g < B_U, \int g(x)dx = 1\right\}.$$

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

**Theorem 1.2.**

1. Consider the testing problem

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

for $\gamma > \beta$. We have,
• For any $0 < \alpha < 1$, there exists $D > 0$ sufficiently large (depending on $\alpha, M, M'$) and a test $\phi$ such that for $\rho_n^2 = Dn^{-\frac{4\beta_1}{4\beta_1 + \delta}}$

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

• For any test $\phi$ which satisfies (1.2) introduced above, the corresponding sequence $\rho_n^2$ satisfies

$$\liminf_{n \to \infty} \rho_n^2 \gtrsim n^{-\frac{4\beta_1}{4\beta_1 + \delta}}.$$  

2. Consider the testing problem

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

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

• For any $0 < \alpha < 1$, there exists $D > 0$ sufficiently large (depending on $\alpha, M, M'$) and a test $\phi$ such that for $\rho_n^2 = Dn^{-\frac{4\beta_1}{4\beta_1 + \delta}}$

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

• For any test $\phi$ which satisfies (1.4) introduced above, the corresponding sequence $\rho_n^2$ satisfies

$$\liminf_{n \to \infty} \rho_n^2 \gtrsim n^{-\frac{4\beta_1}{4\beta_1 + \delta}}.$$  

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 Bull and Nickl (2013); Carpentier (2015), 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 Carpentier (2015) 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 $\left\| f - B_{2, \infty}(M) \right\|_2^2$. Instead, our proof shows that under the alternative, the quantity $\|\Pi(fg_\beta|L)\|_2^2$ 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 Carpentier (2015) 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 $C_n = C(x_1, y_1, \cdots, x_n, y_n)$ is a random measurable subset of $L^2$. We define the $L^2$ radius of a set $C$ as

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

We seek to determine the maximal parameter spaces $P_n$ so that adaptive confidence sets exist. We define a confidence set $C_n = C_n(x_1, y_1, \cdots, x_n, y_n)$ to be honest over a sequence of models $P_n$ if

$$\inf_{P \in P_n} \mathbb{E}_P[|f \in C_n| \geq 1 - \alpha - r_n],$$

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 $P_n$ if there exists a constant $C$ depending on the known parameters of the model space $P_n$ such that

$$\sup_{P \in P_n \cap P(\beta, \gamma)} \mathbb{P}_P[|C_n|^2 \geq C n^{-\frac{2j}{2j + d}}] \leq \alpha',$$

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 Bull and Nickl (2013). With $N = \lceil \log_2 \left( \frac{\beta_{\max}}{\beta_{\min}} \right) \rceil$ define $\beta_j = 2^{j-1}\beta_{\min}$, $j = 1, \ldots, N$. With this notation, we define

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

$$\mathcal{P}_n(M^*, M', \gamma) = \left\{ (f, g) : f \in \mathcal{F}_n(M^*), 0 < f < 1, g \in B_{2, \infty}^{\gamma}(M'), B_L \leq g \leq B_U, \int g(x) dx = 1 \right\},$$

where $B_{2, \infty}^{\beta_j}(M, M^* \rho_n(\beta_j)) = \left\{ h \in B_{2, \infty}^{\beta_j}(M) : \| h - B_{2, \infty}^{\beta_j+1}(M) \|_2 \geq M^* \rho_n(\beta_j) \right\}$, $\rho_n(\beta) = n^{-\frac{2j/d}{2j/d + 1}}$, 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 $\bigcup_{\gamma \in [\gamma_{\min}, \gamma_{\max}]} \mathcal{P}_n(M^*, M', \gamma)$, whenever $M^*$ is large enough.

Theorem 1.3 is proved in Section 4.3. It is of interest to determine whether the models $\cup_\gamma \mathcal{P}_n$, 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 $\cup_\gamma \mathcal{P}_n$ 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 (Bull and Nickl, 2013, 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

$$\mathbb{E}(y|x) = \theta(h(x)), \quad y \in \{0, 1\}, \ x \sim g.$$  

for \(\theta\) a distribution function of a symmetric random variable (probit, logistic etc.), our results still go through provided \(\theta\) satisfies some regularity conditions. In general for a smooth function \(\theta\), 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 Robins and Van Der Vaart (2006), Bull and Nickl (2013), and Carpentier (2015) 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 (Efroymovich et al. (1996), Efroymovich and Pinsker (1996), Antoniadis, Besbeas and Sapatinas (2001), Antoniadis and Sapatinas (2001), Galtchouk and Pergamenschikov (2008), Galtchouk and Pergamenschikov (2009)). 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 \(x\). Although our assumption is not sharp, we believe that such an assumption on the marginal density of \(x\) 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
\( \gamma_{\text{min}} > 2\beta_{\text{max}} \) in Theorem 1.3 can be relaxed. Using results from Robins et al. (2008) 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 \( \beta_{\text{min}} < \beta_{\text{max}} < 2\beta_{\text{min}} \). 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 i.e. \( 2\beta_{\text{min}} < \beta_{\text{max}} \). In a future project we plan to investigate this issue with special focus on using higher order influence functions (Robins et al., 2008, 2015). Indeed, even in the case of non-adaptive inference, Robins et al. (2008, 2015) 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 (Bull et al., 2012; Giné and Nickl, 2010; Kerkyacharian, Nickl and Picard, 2012; Nickl and Szabó, 2014; Picard and Tribouley, 2000; Ray, 2014; Szabó et al., 2015). 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

\[
  h(x) = \sum_{k \in \mathbb{Z}^d} \langle h, \psi_{0,k}^0 \rangle \psi_{0,k}^0(x) + \sum_{l=0}^{\infty} \sum_{k \in \mathbb{Z}^d} \sum_{v \in \{0,1\}^d - \{0\}^d} \langle h, \psi_{1,k}^v \rangle \psi_{1,k}^v(x),
\]

where the base functions \( \psi_{l,k}^v \) are orthogonal for different indices \( (l, k, v) \) and are scaled and translated versions of the \( 2^d \)-\( S \)-regular base functions \( \psi_{0,0}^v \) with \( S > \beta \), i.e., \( \psi_{l,k}^v(x) = 2^{ld/2} \psi_{0,0}^{2l} (2^l x - k) = \prod_{j=1}^d 2^{j/2} \psi_{0,0}^{2^l v_j} (2^l x_j - k_j) \) 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 (Cohen, Daubechies and Vial, 1993). In view of the compact support of the wavelets, for each resolution level \( l \) and index \( v \), only \( O(2^{ld}) \) base elements \( \psi_{l,k}^v \) are non-zero on \([0,1]\); let us denote the corresponding set of indices \( k \) by \( Z_l \) obtaining the representation,

\[
  h(x) = \sum_{k \in Z_{l_0}} \langle h, \psi_{J_0,k}^0 \rangle \psi_{J_0,k}^0(x) + \sum_{l=J_0}^{\infty} \sum_{k \in Z_l} \sum_{v \in \{0,1\}^d - \{0\}^d} \langle h, \psi_{l,k}^v \rangle \psi_{l,k}^v(x),
\]

where \( J_0 = J_0(S) \geq 1 \) is such that \( 2^{J_0} > S \) (Cohen, Daubechies and Vial, 1993; Giné and Nickl, 2015). Thereafter, letting for any \( h \in L_2[0,1]^d \), \( \| \langle h, \psi_{l,.}^v \rangle \|_2 \) be the vector \( L_2 \) norm of the vector
to 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 Bull and Nickl (2013). In particular, we are interested in tail bounds of U-statistics based on kernel \( R(O_1, O_2) = L (O_1) K_{V_j} (X_1, X_2) L (O_2) \) and \( R(O_1, O_2) = L (O_1) K_{W_j} (X_1, X_2) L (O_2) \) where \( O = (Y, X) \) and \( Y \in \mathbb{R}, X \in [0, 1]^d \). Assume that \( |L(O)| \leq B \) (which corresponds to our situation).

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

\[
\mathbb{P} \left( \left| \frac{1}{n(n-1)} \sum_{i_1 \neq i_2} R(O_{i_1}, O_{i_2}) - \mathbb{E}(R(O_1, O_2)) \right| \geq t \right) \leq e^{-C n t^2} + \frac{C t^2}{\sqrt{n}} + \frac{C t}{\sqrt{n}} + \frac{C \sqrt{t}}{\sqrt{n}}
\]
where \( a_1 = \frac{1}{n-1} 2^{\frac{4d}{n}} \), \( a_2 = \frac{1}{n-1} \left( \sqrt{\frac{2d}{n}} + 1 \right) \), \( a_3 = \frac{1}{n-1} \left( \sqrt{\frac{2d}{n}} + \frac{2d}{n} \right) \), \( R(O_1, O_2) = L(O_1)K_{V_j}(X_1, X_2)L(O_2) \) or \( R(O_1, O_2) = L(O_1)K_{W_j}(X_1, X_2)L(O_2) \) with \( K_{V_j} \) and \( K_{W_j} \) constructed using compactly supported wavelet bases of regularity \( S \), \( O = (Y, X) \), \( |L(O)| \leq B \) almost surely \( O \), and \( X \in [0, 1]^d \) has density \( g \) such that \( g(x) \leq B_U \) for all \( x \in [0, 1]^d \).

4.2.1. Proof of Part 1. We will first introduce a test with the desired properties. We use the statistic

\[
T = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} (y_i - 1/2)K_{V_{j_0}}(x_i, x_j)(y_j - 1/2).
\]

Here, we choose \( j_0 = \left\lfloor \frac{2}{4 \beta + d} \log_2 n \right\rfloor \). 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 \( \alpha \) 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,

\[
\mathbb{E}_P[T] = \| \Pi_{V_{j_0}}((f - 1/2)g) \|_2^2 = \| (f - 1/2)g \|_2^2 - \| \Pi_{V_{j_0}}((f - 1/2)g) \|_2^2,
\]

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

Thus we have,

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

\[
\mathbb{P}_P[|T| \leq C \frac{2^{j_0 d/2}}{n}] \leq \mathbb{P}_P \left[ |T - E_P[T]| \geq E_P[T] - C \frac{2^{j_0 d/2}}{n} \right],
\]

\[
\leq \mathbb{P}_P \left[ |T - E_P[T]| \geq B_L^2 \rho_n^2 - C' \frac{n^{4d/4 + d}}{n^{4d/4}} \right],
\]

where \( C' \) depends on \( B_L, M, M' \). The proof is completed by an application of Lemma 4.1, upon setting \( \rho_n^2 = Dn^{-\frac{4d}{4 + d}} \) 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 \( \rho_n^2 \ll n^{-\frac{4d}{4 + d}} \). The proof of the lower bound is then based on Theorem 2.1 of Robins et al. (2009). In particular, let \( H: [0, 1]^d \to \mathbb{R} \) be a \( C^\infty \) function supported on \([0, \frac{1}{2}]^d \) such that \( \int H(x) dx = 0 \) and \( \int H^2(x) dx = 1 \) and let \( k = \lfloor c_0 n^{\frac{d}{4 + d}} \rfloor \) for some \( c_0 > 0 \). Now suppose that \( \Omega_1, \ldots, \Omega_k \) be the translates of the cube \( k^{-\frac{d}{2}} \left[ 0, \frac{1}{2} \right]^d \) that are disjoint and contained in \([0, 1]^d \). Letting \( x_1, \ldots, x_k \) denote the bottom left corners of these cubes, we set for \( \lambda = (\lambda_1, \ldots, \lambda_k) \in \{-1, +1\}^k \),

\[
f_\lambda = \frac{1}{2} + (\frac{1}{k})^{\frac{d}{2}} \sum_{j=1}^k \lambda_j H((x - x_j)k^{\frac{1}{d}}).
\]
The construction ensures that \( f_\lambda \in B_{2,\infty}^B(M) \) (\( H \) can be chosen to guarantee desired \( M \)) for every \( \lambda = (\lambda_1, \ldots, \lambda_k) \in \{-1, +1\}^k \) and \( \| f_\lambda - \frac{1}{2} \|_2^2 = \left( \frac{1}{k} \right)^{2\frac{d}{2}} \). Therefore, by the choice of \( k \), each \( f_\lambda \) corresponds to a measure in the alternative hypothesis. Choose \( \pi \) to be the uniform prior on \( \{-1, +1\}^k \). We use the notation of Theorem 2.1 of Robins et al. (2009), 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 \( P_\lambda \) and \( Q_\lambda \) be the probability measure on \( \{0, 1\} \times [0, 1]^d \) corresponding to likelihood (0.2) for \( f = f_\lambda \) and \( f = \sqrt{\frac{1}{2}} \) respectively, its obvious that \( P_\lambda(\chi_j) = Q_\lambda(\chi_j) = p_j \) (say), since \( \int H(x)dx = 0 \). Also, \( p_j \in \frac{1}{2} [B, \overline{B}] \) for fixed constants \( B, \overline{B} \). Moreover, \( \delta = \max_j \int_{\chi_j} \frac{(q - p)^2}{p \lambda \overline{p}} \) since \( p = \int p_\lambda d\pi(\lambda) = \int f_\lambda^j (1 - f_\lambda)^{1-y} d\pi(\lambda) = \frac{1}{2} = q \). Finally, \( p_\lambda - q = p_\lambda - p = (f_\lambda - \frac{1}{2})^y (\frac{1}{2} - f_\lambda)^{1-y} \) implies that \( a = b = \max_j \sup_\lambda \int_{\Omega_j} \frac{(f_\lambda - \frac{1}{2})^2}{p_j} \in k^{-\frac{2d}{2}} [B, \overline{B}] \). Therefore, by Theorem 2.1 of Robins et al. (2009) if \( \rho(P_1, P_2) \) denotes the Hellinger affinity between two probability measures \( P_1, P_2 \) defined on the same probability space

\[
\rho \left( \int P_\lambda d(\pi(\lambda)), \int Q_\lambda d(\pi(\lambda)) \right) \geq 1 - C \frac{n^2}{k} k^{-\frac{2d}{2}} ,
\]

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 \( f_\lambda \)’s constructed above (Tsybakov, 2008).

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 Carpentier (2015), though the details are considerably different.

Similar to the argument for the previous part, we set \( j_0 = \left\lceil \frac{2}{4^{\frac{d+2}{d}} \log_2 n} \right\rceil \). We assume that we have data \( \{x_i, y_i\}_{i=1}^{2n} \). 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, \( \mathbb{E}_{i,j}[:j] \) will denote the expectation with respect to the \( i \)th half of the sample, with the other half held fixed, under the distribution \( P \). For \( J_0 \leq l \leq j_0 \), we construct the test statistics

\[
T_n(l) = \frac{1}{n(n-1)} \sum_{1 \leq i \neq j \leq n} \frac{y_i}{g(x_i)} K_{W_1}(x_i, x_j) \frac{g_j}{\hat{g}(x_j)}.
\]

By Markov inequality, there exists a constant \( C^* \) such that

\[
\mathbb{P}_P[\|\hat{g} - g\|_2^2 > C^*2 n^{-\frac{2r+2}{2r+\sigma}}] < \frac{\alpha}{4} .
\]

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 $\zeta$ sufficiently large such that

$$
\mathbb{P}_P\left[ \forall 0 \leq l \leq j_0, |T_n(l) - \|\Pi_{W_1}(f_{g/\tilde{g}})\|_2| \leq \zeta \sqrt{\frac{2(l+j_0)d/2}{n^2} + \frac{2l^d/4 \|\Pi_{W_1}(f_{g/\tilde{g}})\|_2^2}{n}} \right] \geq 1 - 3\alpha/4.
$$

Lemma 4.3. 

- Under $H_0$, sup$_{0 \leq l \leq j_0} \left( \|\Pi_{W_1}(f_{g/\tilde{g}})\|_2 - \left( \frac{M}{2^{l\beta_1}} + \frac{C^*}{B_L} n^{-\frac{\gamma}{2l+\alpha}} \right) \right) \leq 0$ with probability at least $(1 - \alpha/4)$.

- Let $\{T_l : 0 \leq l \leq j_0\}$ be a sequence of numbers satisfying $\sum_{l=0}^{j_0} T_l \leq \frac{2}{3} \sqrt{Dn^{-\frac{2\beta_2}{2l+\alpha}}}$. Then under $(f, g) \in H_1$, with probability at least $1 - \alpha/4$, there exists $0 \leq l \leq j_0$ such that $\|\Pi_{W_1}(f_{g/\tilde{g}})\|_2 \geq \frac{M}{2^{l\beta_1}} + T_l$.

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

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

$$
\tilde{C}_l = \left( \frac{M}{2^{l\beta_1}} + \frac{C^*}{B_L} n^{-\frac{\gamma}{2l+\alpha}} + \zeta \frac{2(l+j_0)d/8}{\sqrt{n}} \right)^2
$$

for $\zeta$ 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 $0 \leq l \leq j_0$,

$$
T_n(l) \leq \|\Pi_{W_1}(f_{g/\tilde{g}})\|_2^2 + \zeta \frac{2(l+j_0)d/4}{n} \|\Pi_{W_1}(f_{g/\tilde{g}})\|_2 + \zeta 2^{l/4} \|\Pi_{W_1}(f_{g/\tilde{g}})\|_2 \frac{\|\Pi_{W_1}(f_{g/\tilde{g}})\|_2}{\sqrt{n}}
$$

$$
\leq \left( \frac{M}{2^{l\beta_1}} + \frac{C^*}{B_L} n^{-\frac{\gamma}{2l+\alpha}} \right)^2 + \zeta^2 \left( \frac{2(l+j_0)d/4}{n} + 2 \zeta \frac{2(l+j_0)d/8}{\sqrt{n}} \right) \left( \frac{M}{2^{l\beta_1}} + \frac{C^*}{B_L} n^{-\frac{\gamma}{2l+\alpha}} \right)
$$

$$
\leq \left( \frac{M}{2^{l\beta_1}} + \frac{C^*}{B_L} n^{-\frac{\gamma}{2l+\alpha}} + \zeta \frac{2(l+j_0)d/8}{\sqrt{n}} \right)^2,
$$

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

To control the Type II error, we fix $(f, g) \in P(\beta_1, \beta_2, \gamma, \rho_n^2)$. Using Lemma 4.3, there exists $0 \leq l \leq j_0$ such that

$$
\|\Pi_{W_1}(f_{g/\tilde{g}})\|_2 \geq \frac{M}{2^{l\beta_1}} + T_l,
$$
where we choose \( \tau_l = C_1(n^{-2/(\gamma + \alpha)} + \frac{2(l+j_0)d/8}{\sqrt{n}}) \). Thus with probability at least \((1 - \alpha)\), we have,

\[
T_n(l) \geq \|\Pi W_l(f \frac{g}{g})\|^2 - \frac{2(l+j_0)d/2}{n^2} + 2l/d/4 \|\Pi W_l(f \frac{g}{g})\|^2/n, 
\]

\[
\geq \|\Pi W_l(f \frac{g}{g})\|^2 \left(\|\Pi W_l(f \frac{g}{g})\|^2_2 - \frac{2(d/8)}{\sqrt{n}}(l+j_0)d/4 \right) - \frac{2(l+j_0)d/4}{n}, 
\]

\[
\geq \left(\frac{M}{2l^{2l}} + \eta_1\right) \left(\frac{M}{2l^{2l}} + \eta_1 - \frac{2(d/8)}{\sqrt{n}}(l+j_0)d/4 \right) - \frac{2(l+j_0)d/4}{n}, 
\]

\[
\geq \left(\frac{M}{2l^{2l}} + \eta_1\right) \left(\frac{M}{2l^{2l}} + \eta_1/2 - \frac{2(d/8)}{\sqrt{n}}(l+j_0)d/4 \right), 
\]

where we choose \( C_1 > 2\zeta \). Now, choosing \( C_1 \) even larger, specifically \( C_1^2 > 4\zeta \), it follows that

\[
\zeta \frac{2(l+j_0)d/4}{n} \leq \tau_l^2/4. 
\]

Thus for some \( J_0 \leq l \leq j_0 \), with probability at least \( 1 - \alpha \),

\[
T_n(l) \geq \left(\frac{M}{2l^{2l}} + \frac{\eta_1}{2}\right)^2 
\geq \left(\frac{M}{2l^{2l}} + \frac{C^*}{B_L^2}n^{-2/(\gamma + \alpha)} + \frac{2(d/8)}{\sqrt{n}}(l+j_0)d/4 \right)^2, 
\]

provided we choose \( C_1 > 2C^*/B_L^2 \). 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.

### 4.2.2.1. 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 Chebychev’s inequality. The proof follows by an argument similar to (Carpentier, 2015, Lemma 4.2). We have, for \( J_0 \leq l \leq j_0 \),

\[
\mathbb{E}_{1,P}[T_n(l)] = \|\Pi W_l(f \frac{g}{g})\|^2 \quad \text{var}_{1,P}[T_n(l)] \leq C(S, B_L, B_U) \left(\frac{\|\Pi W_l(f \frac{g}{g})\|^2/n}{n^2} + \frac{2l/d}{n^2}\right), 
\]

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 Chebychev’s inequality,

\[
\mathbb{P}_P \left[ \exists l, J_0 \leq l \leq j_0, |T_n(l) - \|\Pi W_l(f \frac{g}{g})\|^2| > \frac{\|\Pi W_l(f \frac{g}{g})\|^2}{n}\sqrt{\frac{2l/(l+j_0)d/2}{n^2} + 2l/d/4 \|\Pi W_l(f \frac{g}{g})\|^2/n} \right] 
\]

\[
\leq \sum_{l=J_0}^{J_0} \mathbb{E}_P \left[ \mathbb{P}_{1,P} \left[ |T_n(l) - \|\Pi W_l(f \frac{g}{g})\|^2| > \frac{\|\Pi W_l(f \frac{g}{g})\|^2}{n}\sqrt{\frac{2l/(l+j_0)d/2}{n^2} + 2l/d/4 \|\Pi W_l(f \frac{g}{g})\|^2/n} \right] \right] 
\]

\[
\leq \sum_{l=J_0}^{J_0} \mathbb{E}_P \left[ \frac{\text{var}_{1,P}[T_n(l)]}{\zeta^2 \frac{2l/(l+j_0)d/2}{n^2} + 2l/d/4 \|\Pi W_l(f \frac{g}{g})\|^2/n} \right] \leq \frac{C(S, B_L, B_U)}{\zeta^2} \sum_{l=J_0}^{J_0} [2^{-(l-j_0)/2} + 2^{-ld/4}]. 
\]
The proof follows upon noting that $\sum_{l=j_0}^{j_0} 2^{-(j_0-l)d/2} \leq \frac{2^{d/2}}{2^{j_0/2} - 1}$ and $\sum_{l=j_0}^{j_0} 2^{-ld/4} \leq \frac{2^{j_0d/4}}{1 - 2^{-d/4}}$.

4.2.2.2. Proof of Lemma 4.3. Let us consider $f \in B^{\beta_1}_{2,\infty}(M)$. Setting $\tilde{\Delta} = \frac{2^2}{g}$, we have, for $J_0 \leq l \leq j_0$, 

$$\|\Pi_{W_1}(f \frac{g}{h})\|_2 \leq \|\Pi_{W_1}(f)\|_2 + \|\Pi_{W_1}(f \tilde{\Delta})\|_2.$$  

For $f \in B^{\beta_1}_{2,\infty}(M)$, it follows from definition that $\|\Pi_{W_1}(f)\|_2 \leq \frac{M}{2\epsilon_1}$. 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 $(1 - \alpha/4)$, 

$$\|\Pi_{W_1}(f \tilde{\Delta})\|_2^2 \leq \|f \tilde{\Delta}\|_2^2 \leq \left(\frac{C^*}{B^*_L}\right)^2 n^{-\frac{2}{\tau_1 + \alpha}}.$$  

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

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

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

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

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

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 (Carpentier, 2015, Lemma 4.1), it is easy to see that there exists $J_0 \leq l \leq j_0$ such that $\|\Pi_{W_1}(f \frac{g}{h})\|_2 \geq \frac{M}{2\epsilon_1} + \tau_l$. We choose 

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

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

4.3. Proof of Theorem 1.3. This proof idea is motivated by Bull and Nickl (2013). For $\beta \in [\beta_{\min}, \beta_{\max}]$ and $2\beta_{\max} < \gamma < \gamma_{\max}$ consider $P = (f, g) \in P_n(M^*, M', \gamma) \cap P(\beta, \gamma)$. Recall the finite grid $\{\beta_1, \cdots, \beta_N\}$ used for the construction of the parameter spaces $P_n(M^*, M', \gamma)$. We define $F_n(M^*, j) = B^{\beta_j}_{2,\infty}(M, M^* \rho_n(\beta_j)), j = 1, \cdots, N - 1, F_n(M^*, N) = B^{\beta_N}_{2,\infty}(M)$. In addition, we set, for $j \in \{1, \cdots, N\}$, 

$$P_n(j, M^*, M', \gamma) = \{(h, g): h \in F_n(M^*, j), 0 < h < 1, g \in B^{3}_{2,\infty}(M'), B_L < g < B_U, \int g(x) dx = 1\}.$$
Recall further the test $\Psi$ 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 $\Psi$, maintaining the validity of the results. The test $\Psi$ with $\beta = \beta_j$ will be referred to as $\Psi(j)$. We first test the hypothesis $H_0: h \in \mathcal{F}_n(M^*, 1)$ at level $\alpha/4N$. If we reject $H_0$, we set $\hat{\beta} = \beta_1$ and stop. Otherwise we continue. At the $j$th 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 Robins and Van Der Vaart (2006).

Without loss of generality, we assume we have data $\{x_i, y_i: 1 \leq i \leq 3n\}$. 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 $\{\hat{f} \in B_{2, \infty}(C')\}$ and $\{\hat{g} \in B_{2, \infty}(C')\}$ which happen with probability at least $1 - r_n$ (for $C'$ large enough depending on $M, M', B_U, B_U, \gamma_{\max}$ uniformly over $\mathcal{P}_n(M^*, M', \gamma) \cap \mathcal{P}(\beta, \gamma)$, for some vanishing sequence $r_n$). Finally, we set $j_1 = \left\lceil \frac{2\log_2(n)}{4\beta + d} \right\rceil$. Using the data $\{(x_i, y_i): 1 \leq i \leq n\}$, we construct the following U-statistic.

$$U_n = \frac{1}{n(n-1)} \sum_{2n+1 \leq i_1 \neq i_2 \leq 3n} (y_{i_1} - \hat{f}(x_{i_1})) \hat{g}(x_{i_1}) (y_{i_2} - \hat{f}(x_{i_2})) \hat{g}(x_{i_2}).$$

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

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

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

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

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

$$\sup_{P \in \mathcal{P}_n(i_0, M^*, M', \gamma)} \mathbb{P}_P[\hat{\beta} > \beta_{i_0}] \leq \sum_{i > i_0} \sup_{P \in \mathcal{P}_n(i_0, M^*, M', \gamma)} \mathbb{E}_P[1 - \Psi(i)] \leq \frac{\alpha}{4}.$$
Combining, we have, \( \sup_{P \in \mathcal{P}(i_0, M^*, M', \gamma)} \mathbb{P}_P[\hat{\beta} \neq \beta_{i_0}] \leq \frac{\alpha}{2} \). Now, we have,

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

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

\[
\mathbb{E}_{P,\{2,3\}}[\hat{U}_n] = \|\Pi_{V_{j_1}}(f - \hat{f}_n)\cos g\|^2
\]

\[
= \|\Pi_{V_{j_1}}(\hat{f}_n - f)\|^2 + \|\Pi_{V_{j_1}}((\hat{f}_n - f)\Delta)\|^2
\]

\[
+ 2\left\langle \Pi_{V_{j_1}}(\hat{f}_n - f), \Pi_{V_{j_1}}((\hat{f}_n - f)\Delta)\right\rangle
\]

\[
= \|\hat{f}_n - f\|^2 - \Pi_{V_{j_1}}((\hat{f}_n - f)\|^2 + \|\Pi_{V_{j_1}}((\hat{f}_n - f)\Delta)\|^2
\]

\[
+ 2\left\langle \Pi_{V_{j_1}}(\hat{f}_n - f), \Pi_{V_{j_1}}((\hat{f}_n - f)\Delta)\right\rangle
\]

\[
\geq \|\hat{f}_n - f\|^2 - C(M, M', \beta_{\max})n^{-\frac{4\beta_{i_0}}{\beta_{i_0} + d}}
\]

\[
- 2\left\|\Pi_{V_{j_1}}(\hat{f}_n - f)\right\|_2 \left\|\Pi_{V_{j_1}}((\hat{f}_n - f)\Delta)\right\|_2
\]

\[
\geq \|\hat{f}_n - f\|^2 - C(M, M', \beta_{\max}, B_L, B_U)\left(n^{-\frac{4\beta_{i_0}}{\beta_{i_0} + d}} + n^{-\frac{\beta_{i_0}}{\beta_{i_0} + d} - \frac{\gamma}{\beta_{i_0} + d}}\right).
\]

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

\[
\hat{U}_n - \mathbb{E}_{P,\{2,3\}}[\hat{U}_n] = L + R,
\]

\[
L = \frac{2}{n} \sum_{i=2n+1}^{3n} \sum_{k \in Z_{j_1}} \sum_{v \in \{0,1\}^d} \left[ \frac{(y_i - \hat{f}_n(x_i))}{g(x_i)} \psi_{j_1,k}(x_i) - \langle (\hat{f}_n - f)\cos g, \psi_{j_1,k} \rangle \right] \langle (\hat{f}_n - f)\cos g, \psi_{j_1,k} \rangle,
\]

\[
R = \frac{1}{n(n-1)} \sum_{2n+1 \leq i_1 \neq i_2 \leq 3n} \sum_{k \in Z_{j_1}} \sum_{v \in \{0,1\}^d} \left[ \frac{(y_{i_1} - \hat{f}_n(x_{i_1}))}{g(x_{i_1})} \psi_{j_1,k}(x_{i_1}) - \langle (\hat{f}_n - f)\cos g, \psi_{j_1,k} \rangle \right] \times \left[ \frac{(y_{i_2} - \hat{f}_n(x_{i_2}))}{g(x_{i_2})} \psi_{j_1,k}(x_{i_2}) - \langle (\hat{f}_n - f)\cos g, \psi_{j_1,k} \rangle \right].
\]

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 \( f, \hat{f}_n, g_n \), we have that

\[
\text{var}_{P,\{2,3\}}[L] \leq \frac{C(S, B_L, B_U)}{n} \left\|\Pi_{V_{j_1}}(f - \hat{f}_n)\cos g\right\|^2.
\]
By a proof similar to that of controlling $\Lambda_1$ in Lemma B.2, we have that for a deterministic constant $C(\beta_{\max}, \gamma_{\max}, M, B_L, B_U)$

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

Finally, we set

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

By an application of Chebychev inequality, we have,

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

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

Based on our construction, we have,

$$\mathbb{P}_{P}(f \in C_n(\beta_{d_0})) = \mathbb{P}_{P}\left(\|f - \hat{f}\|^2_2 \leq \hat{U}_n + C(M, M', \beta_{\max}, B_L, B_U)\left(n^{-\frac{4\beta_{d_0}}{2} - \frac{\gamma_{\min}}{2\min + d}} + n^{-\frac{\beta_{d_0}}{2} - \frac{\gamma_{\min}}{2\min + d}}\right) + z(\alpha)\tau_n(f)\right) \geq \mathbb{P}_{P}[\|\hat{U}_n - \mathbb{E}_{P,\{2,3\}}[\hat{U}_n]\| \leq z(\alpha)\tau_n(f)] \geq (1 - \frac{\alpha}{2}).$$

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 $n^{-\frac{2\beta_{d_0}}{2}} = o(n^{-\frac{2\beta}{2\min + d}})$ (as $\beta < \beta_{d_0 + 1} < 2\beta_{d_0}$) and $n^{-\frac{\beta_{d_0}}{2} - \frac{\gamma_{\min}}{2\min + d}}$, which, by some tedious algebra, is also $o(n^{-\frac{2\beta}{2\min + d}})$ since $\beta < \beta_{d_0 + 1} < 2\beta_{d_0}, \gamma_{\min} > 2\beta_{d_0 + 1}$. The random part of $\tau_n(f^2)$ is also $o_P(n^{-\frac{2\beta}{2\min + d}})$ as $\hat{f}_n$ is an adaptive estimator and $\|\frac{Q}{g}\|_{\infty} \leq \frac{B_U}{B_L(B_L)} = C(B_U, B_L).$ Finally, the leading term for the diameter is contributed by

$$\mathbb{E}_{P}[\hat{U}_n] = \mathbb{E}_{P}\left[\|\Pi_{\mathcal{V}_1}(f - \hat{f}_n)\|_2^2\right] \leq C(B_U, B_L)\|\hat{f}_n - f\|_2^2,$$

which is $O_P(n^{-\frac{2\beta}{2\min + d}})$ 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 \(\{x_i, y_i\}_{i=1}^{2n}\). 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, \(E_{i,P}[\cdot]\) will denote the expectation with respect to the \(i^{th}\) 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 \(\gamma_{\text{max}}\) for the desired approximation and moment properties to hold. As a result our constants depend on \(\gamma_{\text{max}}\).

Let \(2^{j_{\text{min}}d} = \lceil n^{2^{-j_{\text{max}}}/d+1} \rceil\), \(2^{j_{\text{max}}d} = \lceil n^{2^{-j_{\text{max}}}/d+1} \rceil\), \(2^{l_{\text{min}}d} = \lceil n^{2^{-1}/d+1} \rceil\), and \(2^{l_{\text{max}}d} = \lceil n^{2^{-1}/d+1} \rceil\) and define \(T_1 = [j_{\text{min}}, j_{\text{max}}]\cap \mathbb{N}\) and \(T_2 = [l_{\text{min}}, l_{\text{max}}]\cap \mathbb{N}\). Let \(\hat{g}_l = \frac{1}{n} \sum_{i=n+1}^{2n} K_{V_l}(x_i, x)\).

Now, let

\[
\hat{l} = \min \left\{ j \in T_2: \|\hat{g}_j - \hat{g}_l\|_2 \leq C^* \sqrt{\frac{2^{jd}}{n}}, \forall l \in T_2 \text{ s.t. } l \geq j \right\}.
\]

where \(C^*\) is a constant (depending on \(\gamma_{\text{max}}, B_U\)) that can be determined from the proof hereafter. Thereafter, consider the Lepski-type estimator \(\tilde{g} = \hat{g}_{\hat{l}}\) (Lepski, 1991, 1992). The following lemma states the mean squared properties of \(\tilde{g}\).

**Lemma 4.4.** (Theorem 2 of Bull and Nickl (2013)) For any \(\gamma_{\text{min}} \leq \gamma \leq \gamma_{\text{max}},\)

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

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

Although the proof of Lemma 4.4 can be found in Bull and Nickl (2013), 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 \(\lim \inf_{n \to \infty} \inf_{P \in P(\beta,\gamma)} \mathbb{P}_P[\tilde{g} \in B_{2,\infty}(C)] = 1\) for large enough constant \(C\). Indeed, for any \(C > 0\) and \(l' \geq J_0\), (letting for any \(h \in L_2[0,1]^d, \|\langle h, \psi_{l',v}^\nu \rangle\|_2\) be the vector \(L_2\) norm of the vector \((\langle h, \psi_{l',v}^\nu \rangle)_{k' \in Z_{l'}, v \in \{0,1\}^d - \{0\}^d}\). We have,

\[
\mathbb{P}_P \left( 2^{l'\gamma} \|\langle \tilde{g}, \psi_{l',v}^\nu \rangle\|_2 > C \right) = \sum_{l=l_{\text{min}}}^{l_{\text{max}}} \mathbb{P}_P \left( 2^{l'\gamma} \|\langle \hat{g}_l, \psi_{l',v}^\nu \rangle\|_2 > C, \hat{l} = l \right) I \left( l' \leq l \right)
\]

\[
= \sum_{l=l_{\text{min}}}^{l_{\text{max}}} \mathbb{P}_P \left( 2^{l'\gamma} \|\langle \hat{g}_l, \psi_{l',v}^\nu \rangle\|_2 > C, \hat{l} = l \right) I \left( l' \leq l \right) + \sum_{l=l_{\text{max}}+1}^{l_{\text{max}}} \mathbb{P}_P \left( 2^{l'\gamma} \|\langle \hat{g}_l, \psi_{l',v}^\nu \rangle\|_2 > C, \hat{l} = l \right) I \left( l' \leq l \right)
\]

\[
\leq \sum_{l=l_{\text{min}}}^{l_{\text{max}}} \mathbb{P}_P \left( 2^{l'\gamma} \|\langle \hat{g}_l, \psi_{l',v}^\nu \rangle\|_2 > C \right) I \left( l' \leq l \right) + \sum_{l=l_{\text{max}}+1}^{l_{\text{max}}} \mathbb{P}_P \left( \hat{l} = l \right) I \left( l' \leq l \right)
\]

\[
\leq \sum_{l=l_{\text{min}}}^{l_{\text{max}}} \mathbb{P}_P \left( 2^{l'\gamma} \|\langle \hat{g}_l, \psi_{l',v}^\nu \rangle\|_2 > C \right) I \left( l' \leq l \right) + \sum_{l>l_{\text{max}}} e^{-C^*2^{l'/d}} I \left( l' \leq l \right)
\]

(4.10)
where the last inequality follows from (C.7) for a suitable $C'$ (depending on $B_U$ and the wavelet basis choice). Now,

$$
\mathbb{P}_P \left( 2^l \gamma \| \langle \hat{g}_l, \psi_{l, \cdot} \rangle \|_2 > C \right)
\leq \mathbb{P}_P \left( 2^l \gamma \| \langle \hat{g}_l, \psi_{l, \cdot} \rangle \|_2 > C/2 \right) + \mathbb{P}_P \left( 2^l \gamma \| \mathbb{E}_P \left( \langle \hat{g}_l, \psi_{l, \cdot} \rangle \right) \|_2 > C/2 \right)
= \mathbb{P}_P \left( 2^l \gamma \| \langle \hat{g}_l, \psi_{l, \cdot} \rangle \|_2 > C/2 \right)
$$

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

$$
\mathbb{P}_P \left( 2^l \gamma \| \langle \hat{g}_l, \psi_{l, \cdot} \rangle \|_2 > C \right) \leq \sum_{l=l_{\min}}^{l^*} \mathbb{P}_P \left( 2^l \gamma \| \langle \hat{g}_l, \psi_{l, \cdot} \rangle \|_2 > C/2 \right) I \left( l' \leq l \right) + \sum_{l^*}^{l^*} 2 e^{-C^2 d / 2} I \left( l' \leq l \right). \tag{4.11}
$$

It remains to control $\| \langle \hat{g}_l, \psi_{l, \cdot} \rangle \|_2 > \mathbb{E}_P \left( \langle \hat{g}_l, \psi_{l, \cdot} \rangle \right)$ appropriately. To this end, note that when $l' \leq l$,

$$
\| \langle \hat{g}_l, \psi_{l, \cdot} \rangle \|_2^2 = \frac{1}{n^2} \sum_{i=n+1}^{2n} \sum_{k', v} \left( \psi_{i, k', v}^v(x_i) - \mathbb{E}_P \left( \psi_{i, k', v}^v(x_i) \right) \right)^2 + \frac{1}{n^2} \sum_{n+1 \leq i_1, i_2 \leq 2n, k', v} \left( \psi_{i_1, k', v}^v(x_{i_1}) - \mathbb{E}_P \left( \psi_{i_1, k', v}^v(x_{i_1}) \right) \right) \left( \psi_{i_2, k', v}^v(x_{i_2}) - \mathbb{E}_P \left( \psi_{i_2, k', v}^v(x_{i_2}) \right) \right).
$$

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.

$$
\mathbb{P}_P \left( 2^l \gamma \| \langle \hat{g}_l, \psi_{l, \cdot} \rangle \|_2 > C^2 / 4 \right)
\leq \mathbb{P}_P \left( \frac{1}{n^2} \sum_{i=n+1}^{2n} \left( \psi_{i, k', v}^v(x_i) - \mathbb{E}_P \left( \psi_{i, k', v}^v(x_i) \right) \right)^2 \right) + \mathbb{P}_P \left( \sum_{n+1 \leq i_1, i_2 \leq 2n, k', v} \left( \psi_{i_1, k', v}^v(x_{i_1}) - \mathbb{E}_P \left( \psi_{i_1, k', v}^v(x_{i_1}) \right) \right) \left( \psi_{i_2, k', v}^v(x_{i_2}) - \mathbb{E}_P \left( \psi_{i_2, k', v}^v(x_{i_2}) \right) \right) \right) > \frac{C^2 / 8}{2^l \gamma}
= I + II.
$$

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

$$
\sum_{k', v} \left( \psi_{i, k', v}^v(x) - \mathbb{E}_P \left( \psi_{i, k', v}^v(x) \right) \right)^2 \leq C \left( \psi_{0, 0}^0, \psi_{0, 0}^1, \gamma_{\max} \right) 2^l d,
$$

where the last inequality follows from the properties of the wavelet basis. Therefore, we have

$$
\mathbb{P}_P \left( 2^l \gamma \| \langle \hat{g}_l, \psi_{l, \cdot} \rangle \|_2 > C \right) \leq \sum_{l=l_{\min}}^{l^*} \mathbb{P}_P \left( 2^l \gamma \| \langle \hat{g}_l, \psi_{l, \cdot} \rangle \|_2 > C/2 \right) I \left( l' \leq l \right) + \sum_{l^*}^{l^*} 2 e^{-C^2 d / 2} I \left( l' \leq l \right). \tag{4.11}
$$
and therefore
\[ E_P \sum_{k',v} \left( \psi_{l,k'}^v(x) - E_P \left( \psi_{l,k'}^v(x) \right) \right)^2 \leq C(\psi_{0,0}, \psi_{0,0}, \gamma_{\text{max}}) 2^{l'd}. \]

Therefore by Hoeffding's Inequality,
\[ I \leq \mathbb{P}_P \left( \frac{1}{n} \sum_{i=n+1}^{2n} \left( \psi_{l,k'}^v(x_i) - E_P \left( \psi_{l,k'}^v(x_i) \right) \right)^2 > \frac{nC^2/8}{2^{l'd}} \right) \]
\[ \leq 2e^{-C(\psi_{0,0}, \psi_{0,0}, \gamma_{\text{max}}) 2^{l'd} \left( \frac{ac^2/8}{2^{l'd}} \right)^2}. \]

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

\[ II \leq e^{\frac{c\alpha(l')^2}{a_1(l')^2}} + e^{\frac{c\alpha(l')}{a_2(l')}} + e^{\frac{C\sqrt{a(l')}}{\sqrt{a_3(l')}}} \]

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

Therefore, for $C > 2M'$

\[ \sum_{l' \geq J_0} \mathbb{P}_P \left( 2^{l'/2} \| \langle \tilde{g}, \psi_{l',.} \rangle \|_2 > C \right) \]
\[ \leq \sum_{l' \geq J_0} \sum_{l'=l_{\min}}^{l*} \mathbb{P}_P \left( 2^{l'/2} \| \langle \tilde{g}_l, \psi_{l',.} \rangle - E_P \left( \langle \tilde{g}_l, \psi_{l',.} \rangle \right) \|_2 > C/2 \right) \mathcal{I} \left( l' \leq l \right) + \sum_{l' \geq J_0} \sum_{l' > l*} 2e^{-C'2^{l'/2}} \mathcal{I} \left( l' \leq l \right) \]
\[ \leq \sum_{l=\min}^{l*} \sum_{l' = J_0}^{l} 2e^{\frac{-c\alpha(l')}{a_1(l')^2}} + e^{\frac{-c\alpha(l')}{a_2(l')}} + e^{\frac{C\sqrt{a(l')}}{\sqrt{a_3(l')}}} + \sum_{l=\min}^{l*} \sum_{l' = J_0}^{l} 2e^{-C'2^{l'/2}}. \] (4.12)

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 $B_2^{\gamma}(C)$, completes the proof of $\liminf_{n \to \infty} \inf_{P \in \mathcal{P}([\beta, \gamma])} \mathbb{P}_P [ \langle \tilde{g}, \psi_{\gamma}(\cdot, \cdot) \rangle \in B_2^{\gamma}(C) ] = 1$ for sufficiently large constant $C$ depending on $(M', B_U, \gamma_{\text{max}})$.

However this $\tilde{g}$ does not satisfy the desired point-wise bounds. To achieve this let $\psi$ be a $C^\infty$ function such that $\psi(x) \equiv x$ while $\frac{B_U}{2} \leq \psi(x) \leq 2B_U$ for all $x$. Finally, consider the estimator $\hat{g}(x) = \psi(\tilde{g}(x))$. We note that $(g(x) - \hat{g}(x))^2 \leq (g(x) - \tilde{g}(x))^2$— 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 \( h \in B_{2,\infty}^\beta(M), \psi(h) \in B_{2,\infty}^\beta(C(M, \beta)), \) where \( C(M, \beta) \) is a universal constant dependent only on \( M, \beta \) and independent of \( h \in B_{2,\infty}^\beta(M) \).

The lower bound of the minimax estimation error follows in our case by the results of Bull and Nickl (2013), 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 \( \hat{f}_j(x) = \frac{1}{n} \sum_{i=1}^n \frac{y_i}{g(x_i)} K_{\gamma_j}(x_i, x) \). Now, let

\[
\hat{j} = \min \left\{ j \in \mathcal{T}_1 : \| \hat{f}_j - \hat{f}_l \|_2 \leq C^{**} \sqrt{\frac{2d}{n}}, \forall l \in \mathcal{T}_1 \text{ s.t. } l \geq j \right\}.
\]

where \( C^{**} \) is a suitable constant (depending on \( M, \beta \)). Set \( \hat{j} = \hat{j} \).

Let \( j^* = \min \left\{ j : C_{1+2^{-j\beta}} \leq C_{2+} \sqrt{\frac{2d}{n}} \right\} \), and note that for any \( x \in [0, 1]^d \),

\[
\int \left| \mathbb{E}_{P, 1} \left( \hat{f}_j(x) - f(x) \right)^2 dx \right| = \int \left| \mathbb{E}_{P, 1} \left( \left( \frac{g}{\hat{g}} \right) V_j \right)(x) - f(x) \right|^2 dx
\]

\[
= \int \left| \mathbb{E}_{P, 1} \left( \left( \frac{g}{\hat{g}} - 1 \right) V_j \right)(x) - \mathbb{E}_{P, 1} \left( \left( \frac{g}{\hat{g}} - 1 \right) V_j \right)(x) \right|^2 dx
\]

\[
= \int \left| \mathbb{E}_{P, 1} \left( \left( \frac{g}{\hat{g}} - 1 \right) V_j \right)(x) \right|^2 dx + \int \left| \mathbb{E}_{P, 1} \left( \left( \frac{g}{\hat{g}} - 1 \right) V_j \right)(x) \right|^2 dx
\]

\[
\leq \left\| \mathbb{E}_{P, 1} \left( \left( \frac{g}{\hat{g}} - 1 \right) V_j \right)(x) \right\|^2 + \left\| \mathbb{E}_{P, 1} \left( \left( \frac{g}{\hat{g}} - 1 \right) V_j \right)(x) \right\|^2 + C_1^2 M_2^{-j\beta}.
\]

Therefore,

\[
\mathbb{E}_{P, 2} \int \left| \mathbb{E}_{P, 1} \left( \hat{f}_j(x) - f(x) \right)^2 dx \right| \leq \mathbb{E}_{P, 2} \left\| \left( \frac{g}{\hat{g}} - 1 \right) \right\|_2^2 + C_1^2 M_2^{-j\beta}
\]

\[
\leq \left( \frac{B_U}{B_L} \right)^2 (C(M, \beta))^{2} n^{-\frac{2\gamma + d}{2\gamma + d}} + C_1^2 M_2^{-j\beta}. \tag{4.13}
\]

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

\[
\mathbb{E}_{P, 2} \int \left| \mathbb{E}_{P, 1} \left( \hat{f}_{j^*}(x) \right) - f(x) \right|^2 dx \leq C_1^2 2^{-2j^* \gamma} \beta. \tag{4.14}
\]

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

\[
\mathbb{E}_{P, 1} \left( \left| \hat{f}_j(x) - \mathbb{E}_{P, 1} \left( \hat{f}_j(x) \right) \right|^q \right) \]
where the last inequality in the above display follows by using standard facts about compactly supported wavelet basis having regularity larger than $\gamma_{\text{max}}$ (Härdle et al., 1998) and the fact that the constructed $\hat{g}$ from the second half of the sample lies point-wise in $[B_L, 2B_U]$. The constant $C_2$ therefore depends on $q$, the wavelet basis used, $B_U$ and $B_L$. Therefore, by the choice of $j \in T_1$, we have that for all $x \in [0,1]^d$,

$$\mathbb{E}_{P,1} \left( |\hat{f}_j(x) - \hat{f}_j(x) |^q \right) \leq C_2^q \left( \frac{2^{jd} n}{q} \right)^{q/2}.$$  \hspace{1cm} (4.16)

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

$$\mathbb{E}_P \left( \|\hat{f}_j - f \|_2^2 \right) = \mathbb{E}_{P,2} \int \mathbb{E}_{P,1} \left( |\hat{f}_j - f |^2 \right) dx$$

$$= \mathbb{E}_{P,2} \left[ \int \mathbb{E}_{P,1} \left( |\hat{f}_j(x) - \hat{f}_j(x) |^2 \right) dx + \int \mathbb{E}_{P,1} \left( |\mathbb{E}_{P,1} (\hat{f}_j(x)) - f(x) |^2 \right) dx \right]$$

$$\leq C_1^2 2^{-2j^* \beta} + C_2^2 \left( \frac{2^{jd} n}{q} \right)^{q/2} \leq 2^{d+1}(C_1^2 + C_2^2) n^{-\frac{2q}{2d+2}}.$$  \hspace{1cm} (4.17)

By Cauchy-Schwarz inequality,

$$\mathbb{E}_P \left( \|\tilde{f} - f \|_2^2 \mathbb{I} \left( \tilde{j} > j^* \right) \right) \leq 2 \mathbb{E}_P \left( \|\tilde{f} - \tilde{f}_{j^*} \|_2^2 \mathbb{I} \left( \tilde{j} > j^* \right) \right) + 2 \mathbb{E}_P \left( \|\tilde{f}_{j^*} - f \|_2^2 \right)$$

$$\leq 2((C^{**})^2 + 2^{d+1}(C_1^2 + C_2^2)) n^{-\frac{2q}{2d+2}}.$$  \hspace{1cm} (4.18)

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

$$\mathbb{E}_{P,1} \left( \left| \hat{f}_j(x) - \mathbb{E}_{P,1} (\hat{f}_j(x)) \right| \right) \leq C(B_U, B_L, \gamma_{\text{max}}) \left[ \left( \frac{2^{jd} n}{q} \right)^{3/2} + \left( \frac{2^{jd} n}{q} \right)^{2} \right]$$

$$\leq C(B_U, B_L, \gamma_{\text{max}}) \hspace{1cm} (4.19)$$

by our choice of $2^{j_{\text{max}} d}$. Also, by standard arguments Härdle et al. (1998), $|\mathbb{E}_{P,1} (\hat{f}_j(x)) | = |\mathbb{E} (\hat{f}_{\frac{2}{\gamma} V_j}(x)) | \leq C(B_U, B_L, \gamma_{\text{max}})$ for all $x \in [0,1]^d$. Therefore by (4.19),

$$\mathbb{E}_P \left( \|\tilde{f}_j - f \|_2^2 \right) \leq 8 \mathbb{E}_{P,2} \left[ \int \mathbb{E}_{P,1} \left( \left| \hat{f}_j(x) - \mathbb{E}_{P,1} (\hat{f}_j(x)) \right| \right)^4 dx + \int \left( \mathbb{E}_{P,1} (\hat{f}_j(x)) - f(x) \right)^4 dx \right]$$
\[ C(B_U, B_L, \gamma_{\text{max}}). \]

Also, for any constant \( C'' \)

\[
\mathbb{P}_P \left( \hat{j} = l \right) \leq \sum_{j > j^*} \mathbb{P}_P \left( \| \hat{f}_j - \hat{f}_{j^*} \|_2 > C'' \sqrt{\frac{2j^d}{n}} \right)
\]

\[
\leq \sum_{j > j^*} \mathbb{E}_{P,2} \left\{ \left| \mathbb{P}_{P,1} \left( \| \hat{f}_j - \hat{f}_{j^*} \|_2 > C'' \sqrt{\frac{2j^d}{n}} - \| \mathbb{E}_{P,1} \left( \hat{f}_{j^*} \right) - \mathbb{E}_{P,1} \left( \hat{f}_j \right) \|_2 \right) \right| \right\}
\]

where the inequality in the last display holds by Lemma A.2 since \( \max \left\{ y, \frac{1}{y(x)} \right\} \leq C(B_L) \), for a \( C > 0 \) (depending on \( M, M', B_U, B_L, \gamma_{\text{max}}, C'', C''\) if \( C'' \) is chosen large enough (depending on \( M, M', B_U, B_L, \gamma_{\text{max}} \)) 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,

\[
\sum_{j > j^*} \mathbb{P}_{P,2} \left( \| \mathbb{E} \left( f \frac{g}{g} | V_{j^*} \right) - \mathbb{E} \left( f \frac{g}{g} | V_j \right) \|_2 > C'' \sqrt{\frac{2j^d}{n}} \right)
\]

\[
\leq \sum_{j > j^*} \mathbb{P}_{P,2} \left( \| \mathbb{E} \left( f | V_{j^*} \right) - \mathbb{E} \left( f | V_j \right) \|_2 > C'' \sqrt{\frac{2j^d}{n}} \right)
\]

\[
+ \sum_{j > j^*} \mathbb{P}_{P,2} \left( \| \mathbb{E} \left( f \left( \frac{g}{g} - 1 \right) | V_{j^*} \right) - \mathbb{E} \left( f \left( \frac{g}{g} - 1 \right) | V_j \right) \|_2 > C'' \sqrt{\frac{2j^d}{n}} \right)
\]

\[
= I + II
\]

Since \( f \in B^g_{2,\infty} (M) \) and choice of \( j^* \), we have from C.2 that for \( C'' \) chosen sufficiently large (depending on \( M, M' \) and \( \gamma_{\text{max}} \)), 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

\[ II \leq \sum_{j > j^*} \mathbb{P}_{P_2} \left( \| \Pi \left( f \left( \frac{g}{g} - 1 \right) | V_j \right) \|_2 > \frac{C''}{4} \sqrt{\frac{2j^* d}{n}} \right) + \sum_{j > j^*} \mathbb{P}_{P_2} \left( \| \Pi \left( f \left( \frac{g}{g} - 1 \right) | V_j \right) \|_2 > \frac{C''}{4} \sqrt{\frac{2j^* d}{n}} \right) \]

\[ \leq 2 \sum_{j > j^*} \mathbb{P}_{P_2} \left( \| \hat{g} - g \|_2 > \frac{B_L C''}{4B_U} \sqrt{\frac{2j^* d}{n}} \right). \]

(4.22)

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

\[ \sup_{P \in \mathcal{P}(\beta, \gamma)} \mathbb{P}_P \left( \| \hat{g} - g \|_2 > \frac{B_L C''}{4B_U} \sqrt{\frac{2j^* d}{n}} \right) \leq C_1 (l_{\max} - l_{\min})^2 \left( e^{-C_2 2^{d_{\gamma_{\min}} - l_{\max}} d_{\beta_{\max}}} + e^{-C_3 2^{d_{\min}} d_{\beta_{\max}}} \right). \]

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 \( \hat{f} \) depending on \( M, M', B_U, B_L, \beta_{\min}, \beta_{\max}, \gamma_{\max} \), such that for every \( (\beta, \gamma) \in [\beta_{\min}, \beta_{\max}] \times [\gamma_{\min}, \gamma_{\max}] \),

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

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

The proof of \( \liminf_{n \to \infty} \inf_{P \in \mathcal{P}(\beta, \gamma)} \mathbb{P}_P[\hat{f} \in B_{2, \infty}^d(C)] = 1 \), can be done along the lines of the proof of \( \liminf_{n \to \infty} \inf_{P \in \mathcal{P}(\beta, \gamma)} \mathbb{P}_P[\hat{g} \in B_{2, \infty}^d(C)] = 1 \), 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 \( C \), \( \hat{f}_g \in B_{2, \infty}^d(C) \), with suitably high probability uniformly over \( P \in \mathcal{P}(\beta, \gamma) \).

The construction of a \( \hat{f} \) from this \( f \) and demonstrating its desired properties is very similar to the derivation of \( \hat{g} \) from \( g \), 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

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

This completes the proof of Theorem 1.1.

**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.

Antoniadis, A., Besbeas, P. and Sapatinas, T. (2001). Wavelet shrinkage for natural exponential families with cubic variance functions. *Sankhyā: The Indian Journal of Statistics, Series A* 309–327.

Antoniadis, A. and Leblanc, F. (2000). Nonparametric wavelet regression for binary response. *Statistics: A Journal of Theoretical and Applied Statistics* 34 183–213.

Antoniadis, A. and Sapatinas, T. (2001). Wavelet shrinkage for natural exponential families with quadratic variance functions. *Biometrika* 88 805–820.

Baraud, Y. (2002). Model selection for regression on a random design. *ESAIM: Probability and Statistics* 6 127–146.

Baraud, Y. (2004). Confidence balls in Gaussian regression. *Annals of statistics* 32 528–551.

Bull, A. D. et al. (2012). Honest adaptive confidence bands and self-similar functions. *Electronic Journal of Statistics* 6 1490–1516.

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

Burnashev, M. (1979). On the minimax detection of an inaccurately known signal in a white Gaussian noise background. *Theory of Probability & Its Applications* 24 107–119.

Cai, T. T. (2012). Minimax and adaptive inference in nonparametric function estimation. *Statistical Science* 27 31–50.

Cai, T. T. and Low, M. G. (2004). An adaptation theory for nonparametric confidence intervals. *Annals of statistics* 1805–1840.

Carpentier, A. (2013). Honest and adaptive confidence sets in $L_p$. *Electronic Journal of Statistics* 7 2875–2923.

Carpentier, A. (2015). Testing the regularity of a smooth signal. *Bernoulli* 21 465–488.

Cohen, A., Daubechies, I. and Vial, P. (1993). Wavelets on the interval and fast wavelet transforms. *Applied and computational harmonic analysis* 1 54–81.

Donoho, D. L. and Johnstone, J. M. (1994). Ideal spatial adaptation by wavelet shrinkage. *Biometrika* 81 425–455.

Donoho, D. L. and Johnstone, I. M. (1995). Adapting to unknown smoothness via wavelet shrinkage. *Journal of the american statistical association* 90 1200–1224.

Efroymovich, S. et al. (1996). On nonparametric regression for IID observations in a general setting. *The Annals of Statistics* 24 1126–1144.

Efroymovich, S. and Pinsker, M. (1996). Sharp-optimal and adaptive estimation for heteroscedastic nonparametric regression. *Statistics Sinica* 6 925–942.

Galtchouk, L. and Pergamenshchikov, S. (2008). Adaptive asymptotically efficient estimation in heteroscedastic nonparametric regression via model selection. *arXiv preprint arXiv:0810.1173*.

Galtchouk, L. and Pergamenshchikov, S. (2009). Sharp non-asymptotic oracle inequalities for non-parametric heteroscedastic regression models. *Journal of Nonparametric Statistics* 21 1–18.

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

Giné, E. and Nickl, R. (2010). Confidence bands in density estimation. *The Annals of Statistics* 38 1122–1170.

Giné, E. and Nickl, R. (2011). Rates of contraction for posterior distributions in L r-metrics, 1 r. *The Annals of Statistics* 2883–2911.

Giné, E. and Nickl, R. (2015). Mathematical Foundations of Infinite-Dimensional Statistical Models. Härdele, W., Kerkyacharian, G., Tsybakov, A. and Picard, D. (1998). *Wavelets, approximation, and statistical applications*. Springer.

Hoffmann, M. and Nickl, R. (2011). On adaptive inference and confidence bands. *The Annals of Statistics* 2383–2409.

Houdré, C. and Reynaud-Bouret, P. (2003). Exponential inequalities, with constants, for U-statistics of order two. 55–69.

Ingster, Y. I. (1994). Minimax detection of a signal in p metrics. *Journal of Mathematical Sciences* 68 503–515.

Ingster, Y. I. and Sapatinas, T. (2009). Minimax goodness-of-fit testing in multivariate nonparametric regression. *Mathematical Methods of Statistics* 18 241–269.

Ingster, Y. and Suslina, I. A. (2012). *Nonparametric goodness-of-fit testing under Gaussian models* 169. Springer Science & Business Media.

Johnstone, I. (2002). Function estimation and gaussian sequence models. *Unpublished manuscript* 2 2.
BINARY REGRESSION

Kerkyacharian, G., Nickl, R. and Picard, D. (2012). Concentration inequalities and confidence bands for needlet density estimators on compact homogeneous manifolds. Probability Theory and Related Fields 153 363–404.

Lepski, O. V. (1992). On problems of adaptive estimation in white Gaussian noise. Topics in nonparametric estimation 12 87–106.

Lepski, O. (1991). On a problem of adaptive estimation in Gaussian white noise. Theory of Probability & Its Applications 36 682–697.

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

Lepski, O. V. (1993). On asymptotically exact testing of nonparametric hypotheses Technical Report, Université catholique de Louvain, Center for Operations Research and Econometrics (CORE).

Lepski, O. V. and Spokoiny, V. (1997). Optimal pointwise adaptive methods in nonparametric estimation. The Annals of Statistics 2512–2546.

Lepski, O. V. and Spokoiny, V. G. (1999). Minimax nonparametric hypothesis testing: the case of an inhomogeneous alternative. Bernoulli 5 333–358.

Li, K.-C. (1989). Honest confidence regions for nonparametric regression. The Annals of Statistics 1001–1008.

Low, M. G. (1997). On nonparametric confidence intervals. The Annals of Statistics 25 2547–2554.

McCullagh, P. and Nelder, J. A. (1989). Generalized linear models 37. CRC press.

Ni, K.-C. and Szabó, B. (2014). A sharp adaptive confidence ball for self-similar functions. arXiv preprint arXiv:1406.3994.

Petrov, V. V. (1995). Limit Theorems of Probability Theory. Sequences of Independent Random Variables, vol. 4 of. Oxford Studies in Probability.

Picard, D. and Tribouley, K. (2000). Adaptive confidence interval for pointwise curve estimation. Annals of statistics 298–335.

Ray, K. (2014). Bernstein-von Mises theorems for adaptive Bayesian nonparametric procedures. arXiv preprint arXiv:1407.3397.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coefficients when some regressors are not always observed. Journal of the American Statistical Association 89 846–866.

Robins, J. and Van Der Vaart, A. (2006). Adaptive nonparametric confidence sets. The Annals of Statistics 34 229–253.

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

Robins, J., Tchetgen, E. T., Li, L., Van der Vaart, A. et al. (2009). Semiparametric minimax rates. Electronic Journal of Statistics 3 1305–1321.

Robins, J., Li, L., Mukherjee, R., Tchetgen, E. T. and van der Vaart, A. (2015). Higher Order Estimating Equations for High-dimensional Models. arXiv preprint arXiv:1512.02174.

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.

Signorini, D. and Jones, M. (2004). Kernel estimators for univariate binary regression. Journal of the American Statistical Association 99 119–126.

Simon, J. (1990). Sobolev, Besov and Nikolskii fractional spaces: imbeddings and comparisons for vector valued spaces on an interval. Annali di Matematica Pura ed Applicata 157 117–148.

Spokoiny, V. G. (1996). Adaptive hypothesis testing using wavelets. The Annals of Statistics 24 2477–2498.

Szabó, B., Van der Vaart, A., Van Zanten, J. et al. (2015). Frequentist coverage of adaptive nonparametric Bayesian credible sets. The Annals of Statistics 43 1391–1428.

Triebel, H. (2006). Theory of function spaces. III, vol. 100 of Monographs in Mathematics.

Tsiatis, A. (2007). Semiparametric theory and missing data. Springer Science & Business Media.

Tsybakov, A. B. (2008). Introduction to nonparametric estimation. Springer Science & Business Media.
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 (Petrov, 1995).

**Lemma A.1.** If $O_1, \ldots, O_n \sim \mathbb{P}$ are iid random vectors such that $|L(O)| \leq B$ almost surely $\mathbb{P}$, then for $q \geq 2$ one has for large enough constants $C(B)$ and $C(B, q)$

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

and

$$\mathbb{E}\left(\left|\frac{1}{n} \sum_{i=1}^{n} (L(O_i) - \mathbb{E}(L(O_i)))\right|^q\right) \leq C(B, q)n^{q/2}.$$ 

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 Giné and Nickl (2011) or Theorem 5.1.13 of Giné and Nickl (2015).

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

$$\mathbb{E}(\|\hat{m} - \mathbb{E}(\hat{m})\|_2) \leq C \sqrt{\frac{d}{n}},$$

and for any $x > 0$

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

APPENDIX B: PROOFS OF U-STATISTICS DEVIATION RESULTS

The following tail bound for second order degenerate U-statistics (Giné and Nickl, 2015) is due to Giné, Latala and Zinn (2000) with constants by Houdré and Reynaud-Bouret (2003) 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

$$P[\left|\sum_{i \neq j} R(W_1, W_2)\right| \geq C(A_1 \sqrt{u} + A_2 u + A_3 u^{3/2} + A_4 u^2)] \leq 6 \exp(-u),$$
where, we have,
\[
\Lambda_1^2 = \frac{n(n-1)}{2} E[R^2(W_1, W_2)],
\]
\[
\Lambda_2 = n \sup \{ E[R(W_1, W_2)\xi(W_1)\xi(W_2)]; E[\xi^2(W_1)] \leq 1, E[\xi^2(W_1)] \leq 1 \},
\]
\[
\Lambda_3 = \|nE[R^2(W_1, \cdot)]\|_{\mathcal{B}^1},
\]
\[
\Lambda_4 = \|R\|_{\infty}.
\]

We use this lemma to establish Lemma 4.1.

**Proof.** Let us analyze \( R(O_1, O_2) = L(O_1) K_{V_j}(X_1, X_2) L(O_2) \) first. The proof for \( R(O_1, O_2) = L(O_1) K_W(X_1, X_2) L(O_2) \) is analogous. By Hoeffding’s decomposition one has
\[
\frac{1}{n(n-1)} \sum_{i_1 \neq i_2} R(O_{i_1}, O_{i_2}) - \mathbb{E}(R(O_1, O_2))
\]
\[
= \frac{2}{n} \sum_{i_1=1}^{n} \left[ \mathbb{E}_{O_{i_1}} R(O_{i_1}, O_{i_2}) - \mathbb{E}R(O_{i_1}, O_{i_2}) \right]
\]
\[
+ \frac{1}{n(n-1)} \sum_{i_1 \neq i_2} \left[ R(O_{i_1}, O_{i_2}) - \mathbb{E}_{O_{i_1}} R(O_{i_1}, O_{i_2}) - \mathbb{E}_{O_{i_2}} R(O_{i_1}, O_{i_2}) + \mathbb{E}R(O_{i_1}, O_{i_2}) \right]
\]
\[
: = T_1 + T_2
\]

B.0.3. Analysis of \( T_1 \). Noting that \( T_1 = \frac{2}{n} \sum_{i_1=1}^{n} H(O_{i_1}) \) where \( H(O_{i_1}) = \mathbb{E}(R(O_{i_1}, O_{i_2} | O_{i_1}) - \mathbb{E}R(O_{i_1}, O_{i_2}) \) we control \( T_1 \) by standard Hoeffding’s Inequality. First note that,
\[
|H(O_{i_1})| = \left| \sum_{k \in \mathcal{Z}_j \text{, } v \in \{0,1\}^d} \left[ L(O_{i_1}) \psi_{jk}^v (X_{i_1}) \mathbb{E} \left( \psi_{jk}^v (X_{i_2}) L(O_{i_2}) \right) - \mathbb{E} \left( \psi_{jk}^v (X_{i_2}) L(O_{i_2}) \right) \right]^2 \right|
\]
\[
\leq \sum_{k \in \mathcal{Z}_j} \sum_{v \in \{0,1\}^d} \left| L(O_{i_1}) \psi_{jk}^v (X_{i_1}) \mathbb{E} \left( \psi_{jk}^v (X_{i_2}) L(O_{i_2}) \right) \right| + \sum_{k \in \mathcal{Z}_j} \sum_{v \in \{0,1\}^d} \left( \mathbb{E} \left( \psi_{jk}^v (X_{i_2}) L(O_{i_2}) \right) \right)^2
\]

First, by standard compactness argument for the wavelet bases,
\[
|\mathbb{E} \left( \psi_{jk}^v (X) L(O) \right)| \leq \int |\mathbb{E} (L(O)|X = x) (2^{-2d} \prod_{l=1}^{d} q_{0l}^{v_l} (2^j, x_l - k_l)) | |g(x)| dx
\]
\[
\leq C(B, B_U, S) 2^{-\frac{4d}{2}}.
\]
(B.1)

Therefore,
\[
\sum_{k \in \mathcal{Z}_j} \sum_{v \in \{0,1\}^d} \left( \mathbb{E} \left( \psi_{jk}^v (X_{i_2}) L(O_{i_2}) \right) \right)^2 \leq C(B, B_U, S)
\]
(B.2)
Also, using the fact that for each fixed $x \in [0, 1]^d$, the number indices $k \in \mathcal{Z}_j$ such that $x$ belongs to support of at least one of $\psi_{jk}$ is bounded by a constant depending only on $\psi_{00}$ and $\psi_{00}$. Therefore combining (B.1) and (B.2),

$$
\sum_{k \in \mathcal{Z}_j} \sum_{v \in \{0, 1\}^d} |L(O_{i_1}) \psi_{jk}^v (X_{i_1}) | \leq C(B, B_U, S) 2^{-\frac{id}{2}} 2^d = C(B, B_U, S).
$$

(B.3)

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

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

(B.4)

B.0.4. Analysis of $T_2$. Since $T_2$ is a degenerate U-statistics, it’s analysis is based on Lemma B.1. In particular,

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

where

$$
R^* (O_{i_1}, O_{i_2}) = \sum_{k \in \mathcal{Z}_j} \sum_{v \in \{0, 1\}^d} \left\{ \left( L(O_{i_1}) \psi_{jk}^v (X_{i_1}) - E (\psi_{jk}^v (X_{i_1}) | L(O_{i_1}) \right) \right\} \times \left( L(O_{i_2}) \psi_{jk}^v (X_{i_2}) - E (\psi_{jk}^v (X_{i_2}) | L(O_{i_2}) \right)
$$

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

$$
\Lambda_1^2 \leq C \frac{n(n-1)}{2} 2^{hd}, \Lambda_2 \leq Cn, \Lambda_3^2 \leq Cn 2^{hd}, \Lambda_4 \leq C 2^{hd}.
$$

**Proof.** First we control $\Lambda_1$. To this end, note that by simple calculations, using bounds on $L, g$, and orthonormality of $\psi_{jk}$’s we have,

$$
\Lambda_1^2 = \frac{n(n-1)}{2} E \left( \{R^* (O_{i_1}, O_{i_2})\}^2 \right) \leq 3n(n-1) E \left( R^2 (O_{i_1}, O_{i_2}) \right)
$$

$$
= 3n(n-1) E \left( L^2 (O_{i_1}) K_{ij}^v (X_{1}, X_{2}) L^2 (O_{i_2}) \right)
$$

$$
\leq 3n(n-1) B^4 \int \int \left[ \sum_{k \in \mathcal{Z}_j} \sum_{v \in \{0, 1\}^d} \psi_{jk}^v (x_1) \psi_{jk}^v (x_2) \right]^2 g(x_1) g(x_2) dx_1 dx_2
$$

$$
\leq 3n(n-1) B^4 B_U^2 \int \int \left[ \sum_{k \in \mathcal{Z}_j} \sum_{v \in \{0, 1\}^d} \psi_{jk}^v (x_1) \psi_{jk}^v (x_2) \right]^2 dx_1 dx_2
$$

$$
= 3n(n-1) B^4 B_U^2 \sum_{k \in \mathcal{Z}_j} \sum_{v \in \{0, 1\}^d} \int (\psi_{jk}^v (x_1))^2 dx_2 \int (\psi_{jk}^v (x_2))^2 dx_2
$$
Next we control \( \Lambda_2 = n \sup \{ \mathbb{E} (R^* (O_1, O_2) \cdot \xi (O_1) \cdot \xi (O_2)) : \mathbb{E} (\xi^2 (O_1)) \leq 1, \mathbb{E} (\xi^2 (O_2)) \leq 1 \} \).

To this end, we first control
\[
\mathbb{E} (L(O_1) K_{V_j} (X_1, X_2) L(O_2) \cdot \xi (O_1) \cdot \xi (O_2))
\]
\[
= \left| \int \mathbb{E} (L(O_1) \cdot \xi (O_1) | X_1 = x_1) K_{V_j} (x_1, x_2) \mathbb{E} (L(O_2) \cdot \xi (O_2) | X_2 = x_2) g(x_1) g(x_2) dx_1 dx_2 \right|
\]
\[
= \left| \int \mathbb{E} (L(O) \cdot \xi (O) | X = x) \Pi (\mathbb{E} (L(O) \cdot \xi (O) | X = x) g(x) | V_j) g(x) dx \right|
\]
\[
\leq \left( \int \mathbb{E}^2 (L(O) \cdot \xi^2 (O) | X = x) g^2(x) dx \right)^{\frac{1}{2}} \left( \int \Pi^2 (\mathbb{E} (L(O) \cdot \xi (O) | X = x) g(x) | V_j) dx \right)^{\frac{1}{2}}
\]
\[
\leq \left( \int \mathbb{E}^2 (L(O) \cdot \xi^2 (O) | X = x) g^2(x) dx \right)^{\frac{1}{2}} \left( \int \mathbb{E} (L^2 (O) \cdot \xi^2 (O) | X = x) g^2(x) dx \right)^{\frac{1}{2}}
\]
\[
\leq B^2 B_U \mathbb{E} (\xi^2 (O_1)) \mathbb{E} (\xi^2 (O_2)) \leq B^2 B_U
\]

Above we have used Cauchy-Schwarz Inequality, Jensen’s Inequality, and the fact that projections contract norm. Also,
\[
\mathbb{E} (L(O_1) K_{V_j} (X_1, X_2) L(O_2)) \cdot \xi (O_1) \cdot \xi (O_2))
\]
\[
= \mathbb{E} (L(O_1) \Pi (\mathbb{E} (L(O_1) g(X_1) | X_1) V_j) \cdot \xi (O_1) \cdot \xi (O_2))
\]
\[
= \mathbb{E} [L(O_1) \Pi (\mathbb{E} (L(O_1) g(X_1) | X_1) V_j) \cdot \xi (O_1)] || \mathbb{E} (\xi (O_2))]
\]
\[
\leq \left| \int \Pi (\mathbb{E} (L(O) \cdot \xi (O) | X = x) g(x) | V_j) \Pi (\mathbb{E} (L(O) | X = x) g(x) | V_j) dx \right| \leq B^2 B_U,
\]

where the last step once again uses contraction property of projection, Jensen’s Inequality, and bounds on \( L \) and \( g \). Finally, by Cauchy-Schwarz Inequality and (B.2),
\[
\mathbb{E} \left[ \mathbb{E} (L(O_1) K_{V_j} (X_1, X_2) L(O_2)) \cdot \xi (O_1) \cdot \xi (O_2) \right] \leq \sum_{k \in Z_j} \sum_{v \in \{0, 1\}^d} \mathbb{E}^2 (L(O) \psi^v_{jk} (X)) \leq C(B, B_U, S).
\]

This completes the proof of \( \Lambda_2 \leq C(B, B_U, S)n \). Turning to \( \Lambda_3 = n \| \mathbb{E} \left[ (R^* (O_1, \cdot))^2 \right] \|_\infty \) we have that
\[
(R^* (O_1, o_2))^2 \leq 2 [R(O_1, o_2) - \mathbb{E} (R(O_1, O_2) | O_1)]^2 + 2 [\mathbb{E} (R(O_1, O_2) | O_2 = o_2) - \mathbb{E} (R(O_1, O_2))]^2
\]

Now,
\[
\mathbb{E} [R(O_1, o_2) - \mathbb{E} (R(O_1, O_2) | O_1)]^2
\]
\[
\leq 2 \mathbb{E} \left( L^2 (O_1) K_{V_j}^2 (X_1, x_2) L^2 (o_2) \right) + 2 \mathbb{E} \left( \sum_{k \in Z_j} \sum_{v \in \{0, 1\}^d} L(O_1) \psi_{jk}^v (X_1) \mathbb{E} (\psi_{jk}^v (X_2) L(O_2)) \right)^2
\]
we have

\[ \psi_{jk}(x_3) = \psi_{jk}(x_2) + \psi_{jk}(x_1) \]

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

\[ \left[ \mathbb{E}(R(O_1, O_2)|O_2 = o_2) - \mathbb{E}(R(O_1, O_2)) \right]^2 = \left[ \sum_{k \in Z_j \atop v \in \{0,1\}^d} \mathbb{E}(L(O_1)\psi_{jk}(X_1)) \right)^2 \leq C(B, B_U, S). \]

This completes the proof of controlling \( \Lambda_3 \). Finally, using compactness of the wavelet basis,

\[ ||R(\cdot, \cdot)||_\infty \leq B^2 \sup_{x_1, x_2} \sum_{k \in Z_j \atop v \in \{0,1\}^d} |\psi_{jk}(x_1)||\psi_{jk}(x_2)| \leq C(B, B_U, S)2^{jd} \]

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

Therefore, using Lemma \( B.1 \) and Lemma \( B.2 \) we have

\[ \mathbb{P}\left( |T_2| \geq \frac{C(B, B_U, S)}{n-1} \left( \sqrt{2^{jd}t + t} + \sqrt{\frac{2^{jd}}{n}t^2} + \frac{2^{jd}}{n}t^2 \right) \right) \leq 6e^{-t}. \]

Finally using \( 2t^2 \leq t + t^2 \) we have,

\[ \mathbb{P}_f \left[ |T_2| > a_1\sqrt{t} + a_2t + a_3t^2 \right] \leq 6e^{-t} \quad (B.5) \]

where \( a_1 = \frac{C(B, B_U, S)}{n-1}2^{jd} \), \( a_2 = \frac{C(B, B_U, S)}{n-1} \left( \sqrt{\frac{2^{jd}}{n}t^2} + 1 \right) \), \( a_3 = \frac{C(B, B_U, S)}{n-1} \left( \sqrt{\frac{2^{jd}}{n}t^2} + \frac{2^{jd}}{n}t^2 \right) \). Now if \( h(t) \) is such that \( a_1\sqrt{h(t)} + a_2h(t) + a_3h^2(t) \leq t \), then one has by \( (B.5) \),

\[ \mathbb{P}(|T_2| \geq t) \leq \mathbb{P}_f \left[ |T_2| \geq a_1\sqrt{h(t)} + a_2h(t) + a_3h^2(t) \right] \leq 6e^{-6h(t)}. \]

Indeed, there exists such an \( h(t) \) such that \( h(t) = b_1t^2 \wedge b_2t \wedge b_3\sqrt{t} \) 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

\[ \mathbb{P}(|T_2| \geq t) \leq e^{-\frac{Ct^2}{a_1^2}} + e^{-\frac{Ct}{a_2^2}} + e^{-\frac{Ct}{\sqrt{a_3}}}. \quad (B.6) \]

B.0.5. 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 $\hat{g}$, we begin with standard bias variance analysis for the candidate estimators $\hat{g}$. Therefore, we have the following bias-variance decomposition.

$$\|E_P(\hat{g}) - g\|_2^2 = \|\Pi (g|V) - g\|_2^2 \leq C_1^2 M^2 2^{-2l^*d/2}.$$  \hspace{1cm} (C.1)  

Above we have used the fact that

$$\sup_{h \in B^2_2(M)} \|h - \Pi(h|V)\|_2 \leq C_1 M 2^{-l^*}.$$  \hspace{1cm} (C.2)  

Also by Rosenthal’s Inequality (Petrov, 1995), there exists a constant $C(q)$ for $q \geq 2$ such that

$$E_P(|\hat{g}(x) - E_P(\hat{g}(x))|^q) \leq C(q) \left[ \frac{2n}{q^q} \left\{ \sum_{i=n+1}^{2n} \sup_{B^2_2(M)} E_P(|K_{V_i}(x)|^q) \right\} \right]^{q/2},$$

where the last inequality follows using standard facts about compactly supported wavelet basis having regularity larger than $\gamma_{\text{max}}$ (Härdle et al., 1998) with a constant $C_2$ that depends only on $q$ and the wavelet basis used. Therefore, for $q \geq 2$, by the choice of $l \in T_2$, we have that for all $x \in [0, 1]^d$,

$$E_P(|\hat{g}(x)| - E_P(\hat{g}(x))|^q) \leq C_2 (\frac{2^{ld}}{n})^{q/2}.$$  \hspace{1cm} (C.3)  

Therefore, we have the following bias-variance decomposition.

$$E_P(\|\hat{g} - g\|_2^2) = \int E_P(|\hat{g}(x) - g(x)|^2) \, dx \leq C_1^2 M^2 2^{-2l^*} + C_2 (\frac{2^{ld}}{n})^{q/2}.$$

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

$$\|E_P(\hat{g}^{l^*}) - g\|_2^2 \leq C_1^2 M^2 2^{-2l^*} \leq C_2^2 \left( \frac{2^{ld}}{n} \right)^2 \leq 2 C_2^2 \left( \frac{C_1 M^2}{C_2} \right)^{\frac{2d}{d+4}} (n^{-\frac{2d}{d+4}})$$
Therefore, by definition of \( \bar{l} \) and \( l^* \),
\[
\mathbb{E}_P \left( \| \bar{g} - g \|_2^2 \mathbb{I} \left( \bar{l} \leq l^* \right) \right) = \mathbb{E}_{P,2} \left( \| \bar{g} - g \|_2^2 \mathbb{I} \left( \bar{l} \leq l^* \right) \right) \\
\leq 2 \mathbb{E}_{P,2} \left( \| \bar{g} - \hat{g}_r \|_2^2 \mathbb{I} \left( \bar{l} \leq l^* \right) \right) + 2 \mathbb{E}_{P,2} (\| \hat{g}_r - g \|_2^2) \\
\leq 2^{d+1} ((C^*)^2 + 2) C_2^2 \left( \frac{C_1}{C_2} M' \right)^{\frac{2d}{2d+4}} n^{-\frac{2d}{2d+4}}. \tag{C.5}
\]

Using Cauchy-Schwarz inequality, we have,
\[
\mathbb{E}_P \left( \| \bar{g} - g \|_2^2 \mathbb{I} \left( \bar{l} > l^* \right) \right) \leq \sum_{l=l^*}^{j_{\text{max}}} \sqrt{\mathbb{E}_{P,2} (\| \hat{g}_l - g \|_2^2)} \sqrt{\mathbb{P}_{P,2} (\bar{l} = l)}. \tag{C.6}
\]

Now, by (C.1), (C.3), choice of \( l \in T_2 \), and Jensen’s Inequality
\[
\mathbb{E}_{P,2} (\| \hat{g}_l - g \|_2^2) = \mathbb{E}_{P,2} \left( \int |\hat{g}_l(x) - g(x)|^2 dx \right) \leq \mathbb{E}_{P,2} \int |\hat{g}_l(x) - g(x)|^4 dx \\
\leq C_1^4 M'^4 2^{-4\gamma} + C_2^4 \left( \frac{2ld}{n} \right)^2 \leq C_1^4 M'^4 + C_2^4.
\]

Next, note that for \( l > l^* \),
\[
\mathbb{P}_{P,2} (\bar{l} = l) \leq \sum_{l>l^*} \mathbb{P}_{P,2} (\| \hat{g}_l - \hat{g}_r \|_2 > C^* \sqrt{\frac{2ld}{n}}) \\
\leq \sum_{l>l^*} \mathbb{P}_{P,2} \left( \| \hat{g}_r - \mathbb{E}_{P,2} (\hat{g}_r) \|_2 > \frac{C^*}{2} \sqrt{\frac{2ld}{n}} - \| \mathbb{E}_{P,2} (\hat{g}_r) - \mathbb{E}_{P,2} (\hat{g}_l) \|_2 \right) \\
\leq \sum_{l>l^*} \mathbb{P}_{P,2} \left( \| \hat{g}_r - \mathbb{E}_{P,2} (\hat{g}_r) \|_2 > \frac{C^*}{2} \sqrt{\frac{2ld}{n}} - \| \mathbb{E}_{P,2} (\hat{g}_r) - \mathbb{E}_{P,2} (\hat{g}_l) \|_2 \right) \\
\leq \sum_{l>l^*} \mathbb{P}_{P,2} \left( \| \hat{g}_r - \mathbb{E}_{P,2} (\hat{g}_r) \|_2 > \frac{C^*}{2} \sqrt{\frac{2ld}{n}} - 2C_2 \sqrt{\frac{2ld}{n}} \right) \\
\leq \sum_{l>l^*} \mathbb{P}_{P,2} \left( \| \hat{g}_r - \mathbb{E}_{P,2} (\hat{g}_r) \|_2 > \left( \frac{C^*}{2} - 2C_2 \right) \sqrt{\frac{2ld}{n}} \right) \leq \sum_{l>l^*} 2e^{-C^* 2ld/2}, \tag{C.7}
\]
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 Giné and Nickl (2011). Finally combining equations (C.5), (C.6) and (C.7), we have the existence of an estimator \( \hat{g} \) depending on \( B_U, \gamma_{\min}, \) and \( \gamma_{\max}, \) such that for every \( (\beta, \gamma) \in [\beta_{\min}, \beta_{\max}] \times [\gamma_{\min}, \gamma_{\max}], \)

\[
\sup_{P \in \mathcal{P}(\beta, \gamma)} \mathbb{E}_P \left\| g - \hat{g} \right\|_2^2 \leq C n^{-\frac{2}{2+\alpha}},
\]

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 \( \Delta^r_h = \Delta_h(\Delta_h^{r-1}) \) for \( r \geq 2, \) where \( \Delta_h^1 = \Delta. \) Next, for \( t > 0 \) and \( r \) a natural number greater than \( \beta, \) we define the modulus of smoothness \( \omega_r(f, t) = \sup_{|h| \leq t} \| \Delta^r_h(f) \|_2. \) Finally, we define the Besov semi-norm \( |f|_{B^\beta_{2,\infty}} = \sup_{t > 0} \omega_r(f, t)/t^\beta. \) Finally, we define

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

It is a standard fact (Härdle et al., 1998) 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 \( \phi \) and independent of \( f. \) Next, we control the term \( |\phi(f)|_{B^\beta_{2,\infty}}. \) Using Mean Value Theorem, we have,

\[
\Delta_h(\phi(f))(x) = \phi(f(x + h)) - \phi(f(x)) = \phi'(\xi)\Delta_h(f)(x),
\]

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, \( \Delta^r_h(f)(x) = \sum_{k=0}^{r} \binom{r}{k} (-1)^{r-k} f(x + kh). \) Setting \( r = \lceil \beta \rceil, \) we have, by Taylor expansion for \( \phi, \)

\[
\Delta^r_h(\phi(f))(x) = \sum_{k=0}^{r} \binom{r}{k} (-1)^{r-k} \phi(f(x + kh))
\]

\[
= \sum_{k=0}^{r} \binom{r}{k} (-1)^{r-k} \left[ \phi'(f(x)) \Delta_k h(f)(x) + \frac{\phi''(\xi(x))}{2} (\Delta_k h(f)(x))^2 \right]
\]

\[
= \phi'(f(x)) \Delta^r_h(\phi(f))(x) + \Sigma(x, h).
\]

Thus we have, \( \| \Delta^r_h(\phi(f)) \|_2 \leq \| \phi \|_\infty \| \Delta^r_h(\phi(f)) \|_2 + \| \Sigma(\cdot, h) \|_2. \) To control \( \| \Sigma \|_2, \) we use the fact that \( B^\beta_{2,\infty}(M) \subset B^{\beta-1/2}_{\infty,\infty}(C(M, \beta)), \) where \( \subset \) stands for the usual embedding operation (results of similar
flavor can be found in Simon (1990); Triebel (2006). This naturally implies that \( (\Delta_{kh}(f)(x))^2 \lesssim (kh)^{2\beta-1} \). Thus, we have,

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

This completes the proof.

C.3. Proof of Lemma 4.6. Indeed, \( \hat{g} = \psi(\tilde{g}) \), where \( \psi(x) \) is \( C^\infty \) 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 \( l \in T_2 \) suitably concentrates around \( g \) in a radius of the order of \( \sqrt{\frac{d^d}{n}} \) 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 \( 2^{l_d} \ll 2^{l_2} \) for any \((j, l) \in T_1 \times T_2 \).

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 (Tsypkov, 2008). 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

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

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

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

where \( \alpha_{i,v}^v = (\alpha_{i,k}^v) \) 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 Tsypkov (2008)) to select \( (\alpha_{i,v}^v) \) with mutual separation at least \( \frac{1}{n^{2^d - 1}} \). The Varshamov-Gilbert Lemma guarantees the existence of such a subset with size at least \( 2^{\frac{d}{2^{\beta+d}}} \). 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},

\[
d((f, g), (f', g')) \geq \frac{(2^d - 2)\varepsilon^2}{8} \frac{2^\beta}{n^{\frac{2^d}{2^{\beta+d}}}}.
\]
We denote the joint distribution of \( \{x_l, y_l: 1 \leq l \leq n\} \) 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 
\[
1 + \chi^2((f_i, g_i), (f_0, g_0)) = E_0\left[\left(\frac{f_i(x_1) y_1 (1-f_i(x_1)) (1-y_1)}{1/2}\right)^2\right] = 4E_0[f_i^2(x_1) + (1-f(x_i))^2],
\]
where \( E_0[\cdot] \) represents the expectation with respect to \((f_0, g_0)\). Setting \( f_i = 1/2 + \psi_i \), we have, 
\[
1 + \chi^2((f_i, g_i), (f_0, g_0)) = 1 + 4E_0[\psi_i(x_1)^2] \leq 1 + 4(2^d - 2)\varepsilon^2 n^{-\frac{2d}{2d+3}}.
\]

Thus \( \chi^2(\mathcal{P}_i, \mathcal{P}_0) \leq \exp(4(2^d - 2)\varepsilon^2 n^{-\frac{2d}{2d+3}}) \leq \delta M \), 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 Tsybakov (2008).