The Annals of Statistics
2023, Vol. 51, No. 6, 2342–2365
https://doi.org/10.1214/23-AOS2328

OPTIMAL SUBGROUP SELECTION

BY HENRY W. J. REEVE[1,a], TIMOTHY I. CANNINGS[2,b] AND RICHARD J. SAMWORTH[3,c]

[1]*School of Mathematics, University of Bristol,* [a]*henry.reeve@bristol.ac.uk*
[2]*School of Mathematics and Maxwell Institute for Mathematical Sciences, The University of Edinburgh,* [b]*timothy.cannings@ed.ac.uk*
[3]*Statistical Laboratory, University of Cambridge,* [c]*r.samworth@statslab.cam.ac.uk*

In clinical trials and other applications, we often see regions of the feature space that appear to exhibit interesting behaviour, but it is unclear whether these observed phenomena are reflected at the population level. Focusing on a regression setting, we consider the subgroup selection challenge of identifying a region of the feature space on which the regression function exceeds a pre-determined threshold. We formulate the problem as one of constrained optimisation, where we seek a low-complexity, data-dependent selection set on which, with a guaranteed probability, the regression function is uniformly at least as large as the threshold; subject to this constraint, we would like the region to contain as much mass under the marginal feature distribution as possible. This leads to a natural notion of regret, and our main contribution is to determine the minimax optimal rate for this regret in both the sample size and the Type I error probability. The rate involves a delicate interplay between parameters that control the smoothness of the regression function, as well as exponents that quantify the extent to which the optimal selection set at the population level can be approximated by families of well-behaved subsets. Finally, we expand the scope of our previous results by illustrating how they may be generalised to a treatment and control setting, where interest lies in the heterogeneous treatment effect.

1. Introduction. Consider a clinical trial that assesses the effectiveness of a drug or vaccine. It will typically be the case that efficacy is heterogeneous across the population, in the sense that the probability of a successful outcome depends on several recorded covariates. As a consequence, we may be unable to recommend the treatment for all individuals; nevertheless, it may be too conservative to reject it entirely. It is very tempting to trawl through the data to identify a subset of the population for which the treatment appears to perform well, but statisticians are well versed in the dangers of this type of data snooping (Altman (2015), Feinstein (1998), Gabler et al. (2016), Kaufman and MacLehose (2013), Lipkovich, Dmitrienko and D'Agostino (2017), Rothwell (2005), Senn and Harrell (1997), Wang et al. (2007), Zhang et al. (2015)).

The aim of this paper is to study a *subgroup selection* problem, where we seek to identify a subset of the population for which a regression function exceeds a pre-determined threshold. In the clinical trial example above, this threshold would represent the level at which the treatment is deemed effective. Subgroup selection forms an important component of the more general field of *subgroup analysis* (Herrera et al. (2011), Ting et al. (2020), Wang et al. (2007)), which refers to the problem of understanding the association between a response and subgroups of subjects under study, as defined by one or more subgrouping variables. The main challenge is to provide valid inference, given that the subgroup will be chosen after seeing the data (Lagakos (2006)).

Received February 2023; revised September 2023.
MSC2020 subject classifications. 62G05.
Key words and phrases. Subgroup selection, nonparametric inference, selective inference, FWER.

Our first contribution is to formulate subgroup selection as a constrained optimisation problem. Given independent covariate-response pairs and a family $\mathcal{A}$ of subsets of our feature space, we seek a data-dependent selection set $\hat{A}$ taking values in $\mathcal{A}$ with the Type I error control property that, with probability at least $1-\alpha$, the regression function is uniformly no smaller than the level τ on $\hat{A}$; subject to this constraint, we would like the proportion of the population belonging to $\hat{A}$ to be as large as possible. In practice, $\mathcal{A}$ would typically be chosen to be of relatively low complexity, so as to lead to an interpretable decision rule.

After introducing this new framework, our first result (Proposition 1 in Section 2) reveals the extent of the challenge. We show that if our regression function belongs to a Hölder class, but the corresponding Hölder constant is unknown, then there is a sense in which no algorithm that respects the Type I error guarantee can do better in terms of power than one that ignores the data. We therefore work initially over Hölder classes of known smoothness β, and with a known upper bound λ on the Hölder constant; see Definition 1. This enables us to define a data-dependent selection set that satisfies our Type I error guarantee. The idea is to construct, for each hyper-cube B in a suitable collection within our feature space $\mathbb{R}^d$, a p-value for testing the null hypothesis that the regression function is not uniformly above the level τ on B. The p-values are then combined via Holm's procedure (Holm (1979)) to identify a finite union of hyper-cubes that satisfy our Type I error control property. Our final selection set $\hat{A}_{\mathrm{OSS}}$ maximises the empirical measure among all elements of $\mathcal{A}$ that lie within this finite union of hyper-cubes.

Next, we define a notion of regret $R_\tau(\hat{A})$ that quantifies the power discrepancy between a particular algorithm $\hat{A}$ and an oracle choice. Our aim is to study the optimal regret that can be attained while maintaining Type I error control. We find that the minimax optimal regret is determined by a combination of the smoothness β (initially assumed to lie in $(0,1]$) and two further exponents $\kappa, \gamma > 0$ that quantify the extent to which the oracle selection set can be approximated by families of well-behaved subsets in $\mathcal{A}$. In particular, κ and γ control respectively the degree of concentration of the marginal measure, and the separation between the regression function and the critical level τ on these well-behaved subsets. See Definition 2 for a formal description.

Our main contribution in Section 2 is to establish in Theorem 2, that with a sample size of n, the minimax optimal rate of convergence of the regret over these distributional classes and over all algorithms that respect the Type I error guarantee at significance level $\alpha \in (0, 1/2)$ is of order[1]

$$\min\left\{\left(\frac{\log_+(n/\alpha)}{n}\right)^{\frac{\beta\kappa\gamma}{\kappa(2\beta+d)+\beta\gamma}} + \frac{1}{n^{1/2}}, 1\right\}. \tag{1}$$

The second term in the sum reflects the parametric rate, which corresponds to the difficulty of uniformly estimating the population measure of sets in a Vapnik–Chervonenkis class $\mathcal{A}$. The primary interest, however, is in the first term in the sum, which reveals an intricate interplay between the distributional parameters, the sample size and the significance level.

As mentioned above, the algorithm that achieves the upper bound in Theorem 2 takes λ and β as inputs. In Section 3, therefore, we describe how these parameters can be chosen in a data-driven manner. Since Proposition 1 reveals the impossibility of adaptation in full generality, in Section 3.1, we impose mild additional regularity conditions on our classes, and show that under a sample size condition, we can with high probability estimate the Hölder constant of the regression function to within a factor of 2. Moreover, in Section 3.2, we show that under

[1]Here, $\log_+ x := \log x$ when $x \geq e$ and $\log_+ x := 1$ otherwise. To be fully precise, the upper bound holds when $\mathcal{A}$ is a Vapnik–Chervonenkis class; the lower bound holds when $\beta\gamma(\kappa-1) < d\kappa$ and the class $\mathcal{A}$ consists of convex sets and contains all axis-aligned hyper-rectangles.

a self-similarity condition, we can also estimate β accurately, so that under a sample size condition, our fully data-driven algorithm maintains Type I error control, and has the same regret as the original algorithm up to a sub-logarithmic factor.

A limitation of our constructions for the upper bounds in Sections 2 and 3 are that they are unable to take advantage of higher orders of smoothness beyond $\beta = 1$. To overcome this, in Section 4, we introduce a modified algorithm based on a local polynomial approximation of the regression function, and prove in Theorem 11 that this new construction both respects the Type I error at significance level α and has a regret of optimal order (1) for general smoothness $\beta \in (0, \infty)$. The price we pay for this is a stronger assumption on the marginal feature distribution: we now ask for it to have a well-behaved density with respect to Lebesgue measure (though we do not require this density to be bounded away from zero on its support).

The lower bound constructions for Theorems 2 and 11 are addressed in Section 5. They involve three different finite collections of distributions within our classes, each designed to highlight different aspects of the challenge. The first is a two-point construction, with both distributions having regression functions that are close to τ on disconnected regions, but with each such function only being uniformly above τ on one of these regions; this identifies the dependence of the lower bound on α. The second extends this construction to many distributions, each having its own region where the regression function is uniformly above τ, which underlines the necessity of the logarithmic factor in n in (1). Finally, the third family, which identifies the parametric rate, is another two-point construction with a shared regression function, but whose marginal feature distributions assign slightly different masses to the different connected components of the τ-super level set of this regression function.

Finally, in Section 6, we consider the more general setting where individuals may belong to either a treatment or control group, and where interest lies in the heterogeneous treatment effect. We show that this heterogeneous treatment effect plays a very similar role to that of the regression function in earlier sections, so that our results generalise almost immediately. Proofs of all of our results, as well as auxiliary results and their proofs, are deferred to the Supplementary Material (Reeve, Cannings and Samworth (2023)).

One of the interesting messages of our work from an applied perspective is that, when carefully formulated, it is possible to make formally-justified, post-hoc observations concerning subgroup analyses from clinical studies. When attempted without due care, such observations have been rightly criticised in the medical literature; for example:

> Analyses must be predefined, carefully justified, and limited to a few clinically important questions, and post-hoc observations should be treated with scepticism irrespective of their statistical significance. (Rothwell (2005))

> The statisticians are right in denouncing subgroups that are formed post hoc from exercises in pure data dredging. (Feinstein (1998))

A standard approach to handle subgroup analysis is via statistical tests of interaction (2001 (2001, 2004), Kehl and Ulm (2006)). Zhang et al. (2017) propose a procedure to select a subgroup defined by a half-space that seeks to maximise the expected difference in treatment effect in the context of an adaptive signature design trial. Several other methods have been proposed for studying subgroups defined through heterogeneous treatment effects. For instance, Foster, Taylor and Ruberg (2011) propose an approach to identify subgroups having enhanced treatment effect via the construction of 'virtual twins', while Ballarini et al. (2018) consider maximum likelihood and Lasso-type approaches for estimating a difference in treatment effect in a parametric linear model setting. Su et al. (2009), Dusseldorp, Conversano and Van Os (2010), Lipkovich et al. (2011) and Seibold, Zeileis and Hothorn (2016) propose tree-based procedures to explore the heterogeneity structure of a treatment effect across subgroups that are defined after seeing the data; Huber, Benda and Friede (2019) provide

a simulation comparison of the relative performance of these methods, as well as the algorithm for adaptive refinement by directed peeling proposed by Patel et al. (2016), Crump et al. (2008) and Watson and Holmes (2020) introduce tests of the global null hypothesis of no treatment effect heterogeneity (no subgroups).

One can think of subgroup selection in our context as a super-level set estimation problem, with a key feature being the asymmetry of the way in which we handle cases where $\hat{A}$ contains regions where the regression function is below τ, and where it misses regions where the regression function is at least at level τ. This is motivated by applications such as clinical trials, where the primary concern is the retention of Type I error control despite the post-selection inference. In this respect, our framework has some similarities with that of Neyman–Pearson classification (Cannon et al. (2002), Scott and Nowak (2005), Tong, Feng and Zhao (2016), Xia et al. (2021)). There, our covariate-response pairs (X, Y) take values in $\mathbb{R}^d \times \{0, 1\}$, and we seek a classifier $C : \mathbb{R}^d \to \{0, 1\}$ that minimises $\mathbb{P}(C(X) = 0|Y = 1)$ subject to an upper bound on $\mathbb{P}(C(X) = 1|Y = 0)$. Thus, as in our setting, the way in which the two types of error are handled is asymmetric. On the other hand, as well as allowing continuous responses, our notions of loss are very different. In particular, in our context, we incur a Type I error whenever our selected set $\hat{A}$ contains a single point that does not belong to the τ-super-level set of the regression function. In other words, our framework provides guarantees at an individual level, instead of on average over sub-populations. This may well be ethically and practically advantageous, for example, in medical contexts, as discussed above.

Related work on the estimation of super-level sets of a regression function includes Cavalier (1997), Scott and Davenport (2007), Willett and Nowak (2007), Gotovos et al. (2013), Laloë and Servien (2013), Zanette, Zhang and Kochenderfer (2018) and Dau, Laloë and Servien (2020); likewise, in a density estimation context, there is a large literature on highest density region estimation (Chen, Genovese and Wasserman (2017), Doss and Weng (2018), Hyndman (1996), Mason and Polonik (2009), Polonik (1995), Qiao (2020), Qiao and Polonik (2019), Rodríguez-Casal and Saavedra-Nieves (2019), Samworth and Wand (2010), Tsybakov (1997)). The formulations of the problems studied in these works are rather different from ours, tending to focus on measures of the set difference or Hausdorff distance between the estimated and true sets of interest. Mammen and Polonik (2013) study bootstrap confidence regions for level sets of nonparametric functions, with a particular emphasis on kernel estimation of density level sets.

We conclude this Introduction with some notation used throughout the paper. We adopt the convention that $\inf \varnothing := \infty$, and write $[n] := \{1, \dots, n\}$ for $n \in \mathbb{N} \cup \{0\}$, with $[0] := \varnothing$. Given a set S, we denote its power set by $\mathrm{Pow}(S)$ and its cardinality and complement by $|S|$ and S^{c}, respectively. If S has a strict total ordering, and $g : S \to \mathbb{R}$ is a function that attains its maximum, then we write $\mathrm{sargmax}\{g(s) : s \in S\}$ for the smallest element of $\mathrm{argmax}\{g(s) : s \in S\}$. The σ-algebra of Borel measurable subsets of $\mathbb{R}^d$ is denoted by $\mathcal{B}(\mathbb{R}^d)$. We write $\dim_{\mathrm{VC}}(\mathcal{A})$ for the Vapnik–Chervonenkis dimension of a class of sets $\mathcal{A}$ (e.g., Vershynin (2018), Chapter 2). We also let $\mathcal{A}_{\mathrm{hpr}}$ and $\mathcal{A}_{\mathrm{conv}}$ denote the class of compact axis-aligned hyper-rectangles in $\mathbb{R}^d$ (i.e., sets of the form $\prod_{j=1}^d [a_j, b_j]$ for some $a_j \leq b_j$ and $j \in [d]$), and the set of convex subsets of $\mathbb{R}^d$, respectively.

Given $x \in \mathbb{R}$, we write $x_+ := \max(x, 0)$ and $\log_+ x := \log x$ when $x \geq e$ and $\log_+ x := 1$ otherwise. Let $\|\cdot\|_\infty$ and $\|\cdot\|_2$ denote the supremum and Euclidean norms on $\mathbb{R}^d$, respectively. We denote the d-dimensional Lebesgue on $\mathbb{R}^d$ by $\mathcal{L}_d$, and let $V_d := \mathcal{L}_d(\{x \in \mathbb{R}^d : \|x\|_2 \leq 1\}) = \pi^{d/2}/\Gamma(1 + d/2)$. Given a set $S \subseteq \mathbb{R}^d$, we denote its ℓ_∞-norm *diameter* by $\mathrm{diam}_\infty(S) := \sup_{x,y \in S} \|x - y\|_\infty$ and write $\mathrm{dist}_\infty(x, S) := \inf_{y \in S} \|x - y\|_\infty$. For $r > 0$, let $B_r(x)$ and $\bar{B}_r(x)$ denote the open and closed ℓ_∞ balls of radius r about $x \in \mathbb{R}^d$, respectively. For a function $f : \mathbb{R}^d \to \mathbb{R}$, and for $\xi \in \mathbb{R}$, we also let $\mathcal{X}_\xi(f) := \{x \in \mathbb{R}^d : f(x) \geq \xi\}$ denote

its super-level set at level ξ. Given a symmetric matrix $A \in \mathbb{R}^{q\times q}$, we write A^+ for its Moore–Penrose pseudo-inverse and $\lambda_{\min}(A)$ and $\lambda_{\max}(A)$ for its minimal and maximal eigenvalues, respectively.

For $p \in [0,1]$, we let $\mathrm{Bern}(p)$ denote the Bernoulli distribution on $\{0,1\}$ with mean p. Given a Borel probability measure μ on $\mathbb{R}^d$, we write $\mathrm{supp}(\mu)$ for its *support*, that is, the intersection of all closed sets $C \subseteq \mathbb{R}^d$ with $\mu(C) = 1$. Given Borel subsets B_0, $B_1 \subseteq \mathbb{R}^d$ and a measure μ on $\mathbb{R}^d$, we write $B_0 \subseteq B_1$ if $\mu(B_0 \setminus B_1) = 0$ and $B_0 \nsubseteq B_1$ if $\mu(B_0 \setminus B_1) > 0$; the dependence on μ in our notation here is left implicit since it will be clear from context, and we are thus equating sets whose symmetric difference has μ-measure zero. For probability measures P, Q on a measurable space $(\Omega, \mathcal{F})$, we denote their total variation distance by $\mathrm{TV}(P,Q) := \sup_{B\in\mathcal{F}} |P(B) - Q(B)|$. If these measures are absolutely continuous with respect to a σ-finite measure μ, with Radon–Nikodym derivatives f and g, respectively, then we write $\mathrm{H}(P,Q) := \{\int_\Omega (f^{1/2} - g^{1/2})^2 \, d\mu\}^{1/2}$ for their Hellinger distance, and $\chi^2(P,Q) := \int_\Omega f^2/g \, d\mu - 1$ for their χ^2-divergence. For $a \in [0,1]$, $b \in (0,1)$, we define $\mathrm{kl}(a,b)$ to be the Kullback–Leibler divergence between the $\mathrm{Bern}(a)$ and $\mathrm{Bern}(b)$ distributions, that is, for $a \in (0,1)$,

$$\mathrm{kl}(a,b) := a\log\biggl(\frac{a}{b}\biggr) + (1-a)\log\biggl(\frac{1-a}{1-b}\biggr),$$

with $\mathrm{kl}(0,b) := -\log(1-b)$ and $\mathrm{kl}(1,b) := -\log b$.

2. Subset selection framework and minimax rates. Suppose that the covariate-response pair (X,Y) has joint Borel probability distribution P on $\mathbb{R}^d \times [0,1]$. Let $\mu \equiv \mu_P$ denote the marginal distribution of X. We say that $\eta \equiv \eta_P : \mathbb{R}^d \to [0,1]$ is a *regression function* for P if η is a version of the conditional expectation $\mathbb{E}(Y|X)$. In other words, $\eta : \mathbb{R}^d \to [0,1]$ is a Borel measurable function such that $\int_B \eta(x)\, d\mu(x) = \int_{B\times[0,1]} y \, dP(x,y)$ for all $B \in \mathcal{B}(\mathbb{R}^d)$. We let $\mathcal{A} \subseteq \mathcal{B}(\mathbb{R}^d)$ denote our class of candidate selection sets, and assume that $\varnothing \in \mathcal{A}$. Given a threshold $\tau \in (0,1)$, and recalling the notation $\mathcal{X}_\tau(\eta) := \{x \in \mathbb{R}^d : \eta(x) \geq \tau\}$ for the τ-super level set of η, an ideal output set in our class would have measure

$$M_\tau \equiv M_\tau(P,\mathcal{A}) := \sup\bigl\{\mu(A) : A \in \mathcal{A} \cap \mathrm{Pow}\bigl(\mathcal{X}_\tau(\eta)\bigr)\bigr\}.$$

Since P is unknown, it will typically not be possible to output such an ideal subset. Instead, we will assume that the practitioner has access to a sample $\mathcal{D} \equiv ((X_1,Y_1),\ldots,(X_n,Y_n))$ of independent copies of (X,Y). We define the class of *data-dependent selection sets*, denoted $\hat{\mathcal{A}}_n$, to be the set of functions $\hat{A} : (\mathbb{R}^d \times [0,1])^n \to \mathcal{A}$ such that $(x, D) \mapsto \mathbb{1}_{\hat{A}(D)}(x)$ is a Borel measurable function on $\mathbb{R}^d \times (\mathbb{R}^d \times [0,1])^n$. Given a family $\mathcal{P}$ of distributions on $\mathbb{R}^d \times [0,1]$ and a significance level $\alpha \in (0,1)$, we relax the hard requirement that our output set should be a subset of $\mathcal{X}_\tau(\eta)$ by seeking a data-dependent selection set $\hat{A} \in \hat{\mathcal{A}}_n$, with

$$\inf_{P\in\mathcal{P}} \mathbb{P}_P\bigl(\hat{A}(\mathcal{D}) \subseteq \mathcal{X}_\tau(\eta)\bigr) \geq 1-\alpha. \tag{2}$$

Note that the condition $A \subseteq \mathcal{X}_\tau(\eta)$ is independent of our choice of regression function (Lemma S35). When (2) holds, we will say that $\hat{A}$ controls the Type I error at level α over the class $\mathcal{P}$, and denote the set of data-dependent selection sets that satisfy this requirement as $\hat{\mathcal{A}}_n(\alpha,\mathcal{P})$. For $\hat{A} \in \hat{\mathcal{A}}_n(\alpha,\mathcal{P})$, we would also like that for each $P \in \mathcal{P}$, the random quantity $\mu(\hat{A}(\mathcal{D}))$ should be close to M_τ, that is, we will seek upper bounds for the regret

$$R_\tau(\hat{A}) \equiv R_\tau(\hat{A},P,\mathcal{A}) := M_\tau - \mathbb{E}_P\bigl\{\mu\bigl(\hat{A}(\mathcal{D})\bigr) | \hat{A}(\mathcal{D}) \subseteq \mathcal{X}_\tau(\eta)\bigr\}.$$

In several places below, we abbreviate $\hat{A}(\mathcal{D})$ as $\hat{A}$ where the argument is clear from context.

Our first result reveals that even Lipschitz restrictions on the regression function in our class $\mathcal{P}$ do not suffice to obtain a data-dependent selection set $\hat{A}$ that satisfies both (2) and $\mathbb{P}_P(\mu(\hat{A}) > 0) > \alpha$ for some $P \in \mathcal{P}$. The negative implication is that, regardless of smoothness properties of the true regression function, the regret of any $\hat{A}$ satisfying (2) can be no smaller than the infimum of the regrets of all selection sets that ignore the data while still controlling the Type I error over our Lipschitz class.

Given a probability measure μ on $\mathbb{R}^d$, we let $\mathcal{P}_{\mathrm{Lip}}(\mu)$ denote the set of all Borel probability distributions on $\mathbb{R}^d \times [0,1]$ with marginal μ on $\mathbb{R}^d$, and for which the corresponding regression function η is Lipschitz. We say that $\bar{A} \in \hat{\mathcal{A}}_n$ is *data independent* if $\mathbb{1}_{\{\bar{A}(\mathcal{D})=A\}}$ and $\mathcal{D}$ are independent for all $A \in \mathcal{A}$, and write $\bar{\mathcal{A}}$ for the set of data-independent selection sets.

PROPOSITION 1. *Let μ be a distribution on $\mathbb{R}^d$ without atoms and take $\mathcal{A} \subseteq \mathcal{B}(\mathbb{R}^d)$ with $\dim_{\mathrm{VC}}(\mathcal{A}) < \infty$. Further, let $\hat{A} \in \hat{\mathcal{A}}_n(\alpha, \mathcal{P}_{\mathrm{Lip}}(\mu))$. Then for all $P \in \mathcal{P}_{\mathrm{Lip}}(\mu)$, we have*

$$\mathbb{P}_P\big(\mu(\hat{A}) = 0 | \hat{A} \subseteq \mathcal{X}_\tau(\eta)\big) \geq \mathbb{P}_P\big(\{\mu(\hat{A}) = 0\} \cap \{\hat{A} \subseteq \mathcal{X}_\tau(\eta)\}\big) \geq 1 - \alpha. \tag{3}$$

Hence,

$$R_\tau(\hat{A}) \geq M_\tau \cdot (1 - \alpha) = \inf\big\{R_\tau(\bar{A}) : \bar{A} \in \bar{\mathcal{A}} \cap \hat{\mathcal{A}}_n\big(\alpha, \mathcal{P}_{\mathrm{Lip}}(\mu)\big)\big\}. \tag{4}$$

In the light of Proposition 1, we will assume initially that our regression function belongs to a Hölder class for which both the Hölder exponent and the associated constant are known. In this section, we will work with smoothness exponents that are at most 1.

DEFINITION 1 (Hölder class). Given $\beta \in (0,1]$, $\lambda \in (0,\infty)$ and $A \subseteq \mathbb{R}^d$, we let $\mathcal{F}_{\mathrm{Höl}}(\beta, \lambda, A)$ denote the set of all continuous functions $\eta : \mathbb{R}^d \to [0,1]$ such that

$$|\eta(x') - \eta(x)| \leq \lambda \cdot \|x' - x\|_\infty^\beta,$$

for all $x, x' \in A$. We then let $\mathcal{P}_{\mathrm{Höl}}(\beta, \lambda, \tau)$ denote the class of all distributions P on $\mathbb{R}^d \times [0,1]$ with a regression function $\eta \in \mathcal{F}_{\mathrm{Höl}}(\beta, \lambda, \mathcal{X}_\tau(\eta))$.

Observe that in this definition, the Hölder smoothness condition is only required to hold on the restriction of η to $\mathcal{X}_\tau(\eta)$. While our algorithms will control the Type I error over Hölder classes, we will see that the optimal regret for a data-dependent selection set depends on further aspects of the underlying data generating mechanism. To describe the relevant classes, we first define a function $\omega \equiv \omega_{\mu,d} : \mathbb{R}^d \to [0,1]$ by

$$\omega(x) := \inf_{r \in (0,1)} \frac{\mu(\bar{B}_r(x))}{r^d}.$$

Borrowing the terminology of Reeve, Cannings and Samworth (2021), we will refer to ω as a *lower density*, even though our definition is slightly different as we work with an ℓ_∞-ball instead of a Euclidean ball. A nice feature of this definition is that it allows us to avoid assuming that μ is absolutely continuous with respect to Lebesgue measure; see Reeve, Cannings and Samworth (2021) for several basic properties of lower density functions.

We are now in a position to define what we refer to as an *approximable* class of distributions; these are ones for which we can approximate M_τ well by $\mu(A)$, where $A \in \mathcal{A}$ is both such that the lower density on A is not too small and such that the regression function on A is bounded away from the critical threshold τ.

DEFINITION 2 (Approximable class). Given $\mathcal{A} \subseteq \mathcal{B}(\mathbb{R}^d)$, $\kappa, \gamma > 0$, $\tau \in (0,1)$ and $C_{\mathrm{App}} \geq 1$, let $\mathcal{P}_{\mathrm{App}}(\mathcal{A}, \kappa, \gamma, \tau, C_{\mathrm{App}})$ denote the class of all distributions P on $\mathbb{R}^d \times [0,1]$ with marginal μ on $\mathbb{R}^d$ and a regression function $\eta : \mathbb{R}^d \to [0,1]$ such that

$$\sup\{\mu(A) : A \in \mathcal{A} \cap \mathrm{Pow}(\mathcal{X}_\xi(\omega) \cap \mathcal{X}_{\tau+\Delta}(\eta))\} \geq M_\tau - C_{\mathrm{App}} \cdot (\xi^\kappa + \Delta^\gamma),$$

for all $\xi, \Delta > 0$.

We now provide several examples of distributions belonging to appropriate approximable classes. The proofs of the claims in these examples are given in Section S3.

EXAMPLE 1. *Let $\mathbb{P}(Y=0) = \mathbb{P}(Y=1) = 1/2$, and $X|Y=r \sim N((-1)^{r-1}\nu, 1)$ for $r \in \{0,1\}$ and some $\nu > 0$. Fix some $\tau \in (0,1)$, and let $\mathcal{A}_{\mathrm{int}}$ denote the set of all closed intervals in $\mathbb{R}$. Then the distribution P of (X,Y) belongs to $\mathcal{P}_{\mathrm{App}}(\mathcal{A}_{\mathrm{int}}, \kappa, \gamma, \tau, C_{\mathrm{App}})$ with $\kappa = \gamma = 1$, for a suitably large choice of C_{App}, depending only on ν and τ.*

EXAMPLE 2. *Let μ denote the uniform distribution on $[0,1]^d$ and fix $\tau \in (0,1)$. Suppose that $\eta : \mathbb{R}^d \to [0,1]$ is coordinate-wise increasing, that $\mathcal{S}_\tau := \{x \in [0,1]^d : \eta(x) = \tau\} \neq \varnothing$, and that there exist $\delta, \epsilon \in (0,1]$ and $\gamma > 0$ such that $\eta(x) - \tau \geq \epsilon \cdot \mathrm{dist}_\infty(x, \mathcal{S}_\tau)^{1/\gamma}$, for every $x \in \mathcal{X}_\tau(\eta) \cap [0,1]^d$ with $\mathrm{dist}_\infty(x, \mathcal{S}_\tau) \leq \delta$. If P denotes a distribution on $\mathbb{R}^d \times [0,1]$ with marginal μ on $\mathbb{R}^d$ and regression function η, then $P \in \mathcal{P}_{\mathrm{App}}(\mathcal{A}_{\mathrm{hpr}}, \kappa, \gamma, \tau, C_{\mathrm{App}})$ for arbitrarily large $\kappa > 0$, provided that $C_{\mathrm{App}} \geq 2d/(\epsilon^\gamma \delta)$.*

EXAMPLE 3. *Consider the family of distributions $\{\mu_\kappa : \kappa \in (0,\infty)\}$ on $\mathbb{R}^d$ with densities of the form $x \mapsto g_\kappa(\|x\|_\infty)$, where $g_\kappa : [0,\infty) \to [0,\infty)$ is given by*

$$g_\kappa(y) := \begin{cases} (\kappa/2^d) \cdot \{1 + (1-\kappa)y^d\}^{-1/(1-\kappa)} & \text{if } \kappa \in (0,1), \\ (1/2^d) \cdot e^{-y^d} & \text{if } \kappa = 1, \\ (\kappa/2^d) \cdot \{1 - (\kappa-1)y^d\}^{1/(\kappa-1)} \mathbb{1}_{\{y \leq 1/(\kappa-1)^{1/d}\}} & \text{if } \kappa \in (1,\infty). \end{cases}$$

Now, for $\gamma > 0$ and $\tau \in (0,1)$, define the regression function $\eta_\gamma : \mathbb{R}^d \to [0,1]$ by

$$\eta_\gamma(x) := 0 \vee \{\tau + \lambda \cdot \mathrm{sgn}(x_1)|x_1|^{1/\gamma}\} \wedge 1.$$

Writing $P_{\kappa,\gamma}$ for the distribution on $\mathbb{R}^d \times \{0,1\}$ with marginal μ_κ on $\mathbb{R}^d$ and regression function η_γ, we have that $P_{\kappa,\gamma} \in \mathcal{P}_{\mathrm{App}}(\mathcal{A}_{\mathrm{hpr}}, \kappa, \gamma, \tau, C_{\mathrm{App}})$ for $C_{\mathrm{App}} \equiv C_{\mathrm{App}}(d, \kappa, \gamma, \lambda) > 0$ sufficiently large.

Example 1 is designed to be a simple setting of our problem, where we can take $\gamma = \kappa = 1$. Example 2 illustrates the way that the growth of η as we move away from $\eta^{-1}(\tau)$ affects the parameter γ of our class, while Example 3 shows the effect of the tail behaviour of the marginal density on $\mathbb{R}^d$ on the parameter κ. Proposition S6 in the Supplementary Material provides general conditions under which our joint distribution belongs to $\mathcal{P}_{\mathrm{App}}(\mathcal{A}, \kappa, \gamma, \tau, C_{\mathrm{App}})$ with $\gamma = 1$. In essence, the $\gamma = 1$ setting occurs when the gradient of the regression function never vanishes on the boundary $\eta^{-1}(\tau)$ of $\mathcal{X}_\tau(\eta)$.

We can now state the main theorem of this section, which reveals the minimax optimal rate of convergence for the regret over $\mathcal{P}_{\mathrm{Höl}}(\beta, \lambda, \tau) \cap \mathcal{P}_{\mathrm{App}}(\mathcal{A}, \kappa, \gamma, \tau, C_{\mathrm{App}})$ for a data-dependent selection set in $\hat{\mathcal{A}}_n(\alpha, \mathcal{P}_{\mathrm{Höl}}(\beta, \lambda, \tau))$.

THEOREM 2. *Take $\beta \in (0,1]$, $\lambda \geq 1$, $\kappa, \gamma > 0$, $\tau \in (0,1)$ and $C_{\mathrm{App}} \geq 1$.*

(i) Upper bound: *Let* $\mathcal{A} \subseteq \mathcal{B}(\mathbb{R}^d)$ *satisfy* $\dim_{\mathrm{VC}}(\mathcal{A}) < \infty$ *and* $\varnothing \in \mathcal{A}$. *Then there exists* $C \geq 1$, *depending only on* d, κ, γ, τ, C_{App} *and* $\dim_{\mathrm{VC}}(\mathcal{A})$, *such that for all* $n \in \mathbb{N}$ *and* $\alpha \in (0, 1/2]$, *we have*

$$\inf_{\hat{A}} \sup_{P} R_\tau(\hat{A}) \leq C \cdot \min\left\{ \left(\frac{\lambda^{d/\beta} \cdot \log_+(n/\alpha)}{n} \right)^{\frac{\beta\kappa\gamma}{\kappa(2\beta+d)+\beta\gamma}} + \frac{1}{n^{1/2}}, 1 \right\}, \tag{5}$$

where the infimum in (5) *is taken over* $\hat{\mathcal{A}}_n(\tau, \alpha, \mathcal{P}_{\mathrm{Höl}}(\beta, \lambda, \tau))$ *and the supremum is taken over* $\mathcal{P}_{\mathrm{Höl}}(\beta, \lambda, \tau) \cap \mathcal{P}_{\mathrm{App}}(\mathcal{A}, \kappa, \gamma, \tau, C_{\mathrm{App}})$.

(ii) Lower bound: *Now suppose that* $\beta\gamma(\kappa - 1) < d\kappa$, $\epsilon_0 \in (0, 1/2)$, $\tau \in (\epsilon_0, 1 - \epsilon_0)$ *and* $\alpha \in (0, 1/2 - \epsilon_0]$. *Then there exists* $c > 0$, *depending only on* d, β, κ, γ, C_{App} *and* ϵ_0, *such that for any* $\mathcal{A} \subseteq \mathcal{B}(\mathbb{R}^d)$ *satisfying* $\mathcal{A}_{\mathrm{hpr}} \subseteq \mathcal{A} \subseteq \mathcal{A}_{\mathrm{conv}}$ *and any* $n \in \mathbb{N}$, *we have*

$$\inf_{\hat{A}} \sup_{P} R_\tau(\hat{A}) \geq c \cdot \min\left\{ \left(\frac{\lambda^{d/\beta} \cdot \log_+(n/(\lambda^{d/\beta}\alpha))}{n} \right)^{\frac{\beta\kappa\gamma}{\kappa(2\beta+d)+\beta\gamma}} + \frac{1}{n^{1/2}}, 1 \right\}, \tag{6}$$

where, again, the infimum in (6) *is taken over* $\hat{\mathcal{A}}_n(\tau, \alpha, \mathcal{P}_{\mathrm{Höl}}(\beta, \lambda, \tau))$ *and the supremum is taken over* $\mathcal{P}_{\mathrm{Höl}}(\beta, \lambda, \tau) \cap \mathcal{P}_{\mathrm{App}}(\mathcal{A}, \kappa, \gamma, \tau, C_{\mathrm{App}})$.

Since $\dim_{\mathrm{VC}}(\mathcal{A}_{\mathrm{hpr}}) < \infty$ (e.g., Shalev-Shwartz and Ben-David (2014), Exercise 5 in Section 6.8), the choice $\mathcal{A} = \mathcal{A}_{\mathrm{hpr}}$ provides a natural example satisfying both the lower and upper bounds in Theorem 2.

In order to introduce the algorithm that achieves the upper bound, we define the *empirical marginal distribution* $\hat{\mu}_n$ and *empirical regression function* $\hat{\eta}_n$, for $B \subseteq \mathbb{R}^d$, by

$$\hat{\mu}_n(B) := \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \in B\}}, \tag{7}$$

$$\hat{\eta}_n(B) := \frac{1}{n \cdot \hat{\mu}_n(B)} \sum_{i=1}^{n} Y_i \cdot \mathbb{1}_{\{X_i \in B\}} \tag{8}$$

whenever $\hat{\mu}_n(B) > 0$, and $\hat{\eta}_n(B) := 1/2$ otherwise. The main idea is to associate, to each $B \subseteq \mathbb{R}^d$, a p-value for a test of the hypothesis that the regression function is uniformly above the level τ on B. More precisely, for $B \subseteq \mathbb{R}^d$, we define

$$\hat{p}_n(B) \equiv \hat{p}_{n,\beta,\lambda}(B) := \exp\{-n \cdot \hat{\mu}_n(B) \cdot \mathrm{kl}(\hat{\eta}_n(B), \tau + \lambda \cdot \mathrm{diam}_\infty(B)^\beta)\}, \tag{9}$$

whenever $\hat{\eta}_n(B) > \tau + \lambda \cdot \mathrm{diam}_\infty(B)^\beta$, and $\hat{p}_n(B) := 1$ otherwise. Lemma 3 below confirms that $\hat{p}_n(B)$ is indeed a p-value (even conditionally on $\mathcal{D}_X \equiv (X_i)_{i \in [n]}$).

LEMMA 3. *Fix* $\beta \in (0, 1]$, $\lambda \in [1, \infty)$ *and* $P \in \mathcal{P}_{\mathrm{Höl}}(\beta, \lambda, \tau)$ *with a regression function* $\eta \in \mathcal{F}_{\mathrm{Höl}}(\beta, \lambda, \mathcal{X}_\tau(\eta))$. *Then given* $B \in \mathcal{B}(\mathbb{R}^d)$ *with* $B \nsubseteq \mathcal{X}_\tau(\eta)$, *and any* $\alpha \in (0, 1)$, *we have*

$$\mathbb{P}_P\big(\hat{p}_n(B) \leq \alpha | \mathcal{D}_X\big) \leq \alpha.$$

We now exploit these p-values to specify a data-dependent selection set $\hat{A}$ that controls the Type I error over $\mathcal{P}_{\mathrm{Höl}}(\beta, \lambda, \tau)$. First, define a set of hyper-cubes

$$\mathcal{H} := \left\{ 2^{-q} \prod_{j=1}^{d} [2a_j - 1, 2a_j + 3) : (a_1, \ldots, a_d) \in \mathbb{Z}^d, q \in \mathbb{N} \right\}.$$

Now, given $n \in \mathbb{N}$ and $\boldsymbol{x}_{1:n} = (x_i)_{i \in [n]} \in (\mathbb{R}^d)^n$, we define

$$\mathcal{H}(\boldsymbol{x}_{1:n}) := \big\{ B \in \mathcal{H} : \{x_1, \ldots, x_n\} \cap B \neq \varnothing \text{ and } \mathrm{diam}_\infty(B) \geq 1/n \big\},$$

Algorithm 1: The data-dependent selection set $\hat{A}_{\mathrm{OSS}}$

Input: Data $\mathcal{D} = ((X_1, Y_1), \ldots, (X_n, Y_n)) \in (\mathbb{R}^d \times [0,1])^n$, an ordered set $\mathcal{A}$ of subsets of $\mathbb{R}^d$ with $\varnothing \in \mathcal{A}$, $(\tau, \alpha) \in (0,1)^2$, Hölder parameters $(\beta, \lambda) \in (0,1] \times [1, \infty)$;
Compute $\hat{p}_n(B)$ for each $B \in \mathcal{H}(\mathcal{D}_X)$ using (9) and let $\hat{L} := |\mathcal{H}(\mathcal{D}_X)|$;
Enumerate $\mathcal{H}(\mathcal{D}_X)$ as $(B_{(\ell)})_{\ell \in [\hat{L}]}$, in such a way that $\hat{p}_n(B_{(\ell)}) \leq \hat{p}_n(B_{(\ell')})$ for $\ell \leq \ell'$;
if $\hat{L} \cdot \hat{p}_n(B_{(1)}) \leq \alpha$ **then**
  Compute $\ell_\alpha := \max\{\ell \in [\hat{L}] : (\hat{L} + 1 - \ell') \cdot \hat{p}_n(B_{(\ell')}) \leq \alpha \ \forall \ell' \leq \ell\}$;
  Choose $\hat{A}_{\mathrm{OSS}}(\mathcal{D}) := \mathrm{sargmax}\{\hat{\mu}_n(A) : A \in \mathcal{A} \cap \mathrm{Pow}(\bigcup_{\ell \in [\ell_\alpha]} B_{(\ell)})\}$;
else
  Set $\hat{A}_{\mathrm{OSS}}(\mathcal{D}) = \varnothing$;
end
Result: The selected set $\hat{A}_{\mathrm{OSS}}(\mathcal{D})$.

so that $|\mathcal{H}(\boldsymbol{x}_{1:n})| \leq 2^d n(2 + \log_2 n)$. The overall algorithm, which applies Holm's procedure (Holm (1979)) to the p-values in (9) and is denoted by $\hat{A}_{\mathrm{OSS}} \in \hat{\mathcal{A}}_n$, is given in Algorithm 1. Note that in this algorithm, there is no loss of generality in assuming that $\mathcal{A}$ is ordered, by the well-ordering theorem, which is equivalent to the axiom of choice. Of course, in most practical settings, there would be a natural ordering on $\mathcal{A}$ induced by an injective map from $\mathcal{A}$ to a Euclidean space with lexicographic ordering; for example, if $\mathcal{A}$ is the set of hyper-rectangles in $\mathbb{R}^d$, then this map could take a given hyper-rectangle $A = \prod_{j=1}^d [a_j, b_j]$ to $(a_1, b_1, \ldots, a_d, b_d) \in \mathbb{R}^{2d}$. We remark also that in Algorithm 1, it may be the case that there exist $B_1, B_2 \in \mathcal{H}(\mathcal{D}_X)$ with $B_1 \subseteq B_2$ and $B_2 \in \{B_{(1)}, \ldots, B_{(\ell_\alpha)}\}$, while $B_1 \notin \{B_{(1)}, \ldots, B_{(\ell_\alpha)}\}$. This causes no difficulties for our theory.

Proposition 4 below provides part of the proof of the upper bound in Theorem 2.

PROPOSITION 4. *Let $\alpha \in (0,1)$ and $(\beta, \lambda) \in (0,1] \times [1, \infty)$. Then the data-dependent selection set $\hat{A}_{\mathrm{OSS}}$ controls the Type I error at level α over $\mathcal{P}_{\mathrm{Höl}}(\beta, \lambda, \tau)$; in other words, $\hat{A}_{\mathrm{OSS}} \in \hat{\mathcal{A}}_n(\tau, \alpha, \mathcal{P}_{\mathrm{Höl}}(\beta, \lambda, \tau))$.*

Proposition 5 complements Proposition 4 by bounding the regret $R_\tau(\hat{A}_{\mathrm{OSS}})$, and together these results prove the upper bound in Theorem 2. In fact, we provide a high-probability bound as well as an expectation bound.

PROPOSITION 5. *Take $\tau, \alpha \in (0,1)$, $\beta \in (0,1]$, $\lambda \geq 1$, $\kappa, \gamma > 0$, $C_{\mathrm{App}} \geq 1$ and $\mathcal{A} \subseteq \mathcal{B}(\mathbb{R}^d)$ with $\dim_{\mathrm{VC}}(\mathcal{A}) < \infty$ and $\varnothing \in \mathcal{A}$. There exists $\tilde{C} \geq 1$, depending only on d, κ, γ, τ, C_{App} and $\dim_{\mathrm{VC}}(\mathcal{A})$, such that for all $P \in \mathcal{P}_{\mathrm{Höl}}(\beta, \lambda, \tau) \cap \mathcal{P}_{\mathrm{App}}(\mathcal{A}, \kappa, \gamma, \tau, C_{\mathrm{App}})$, $n \in \mathbb{N}$ and $\delta \in (0,1)$, we have*

$$\mathbb{P}_P\left[M_\tau - \mu(\hat{A}_{\mathrm{OSS}}) > \tilde{C}\left\{\left(\frac{\lambda^{d/\beta}}{n} \cdot \log_+\left(\frac{n}{\alpha \wedge \delta}\right)\right)^{\frac{\beta\kappa\gamma}{\kappa(2\beta+d)+\beta\gamma}} + \left(\frac{\log_+(1/\delta)}{n}\right)^{1/2}\right\}\right] \leq \delta.$$

As a consequence, for $\alpha \in (0, 1/2]$,

$$R_\tau(\hat{A}_{\mathrm{OSS}}) \leq C\left\{\left(\frac{\lambda^{d/\beta}}{n} \cdot \log_+\left(\frac{n}{\alpha}\right)\right)^{\frac{\beta\kappa\gamma}{\kappa(2\beta+d)+\beta\gamma}} + \frac{1}{n^{1/2}}\right\},$$

where $C > 0$ depends only on $\tilde{C}$.

In the lower bound part of Theorem 2, we have the condition $\beta\gamma(\kappa-1) < d\kappa$. A constraint of this form is natural in light of the tension between β, κ and γ. In particular, large values of κ and γ mean that little μ-mass in $\mathcal{X}_\tau(\eta)$ is lost by restricting to sets $A \in \mathcal{A}$ for which the lower density of A is not too small, and the regression function on A is uniformly well above τ; but the smoothness of the regression function constrains the rate of change of η and, therefore, the extent to which this is possible. This intuition is formalised in Lemma S33, where we prove that $\beta\gamma(\kappa-1) \leq d\kappa$ provided there exists a distribution in our class for which the pre-image of τ under η is nonempty, and μ is sufficiently well behaved. Since the lower bound construction for the proof of Theorem 2(ii) is common to both the setting of this section and that of the upcoming Section 4 on higher-order smoothness, we will defer discussion of this construction until Section 5.

3. Choice of λ and β. Our original algorithm for computing $\hat{A}_{\mathrm{OSS}}$ takes λ and β as inputs. In cases where a practitioner is unable to make informed default choices, it is natural to seek to understand the effect of overspecification and underspecification of these parameters, as well as to seek data-driven estimators. To study the first of these questions, fix $\beta \in (0, 1]$ and $\lambda \geq 1$, as well as $\beta' \in (0, \beta]$ and $\lambda' \geq \lambda$. Since $\mathcal{P}_{\mathrm{Höl}}(\beta', \lambda', \tau) \supseteq \mathcal{P}_{\mathrm{Höl}}(\beta, \lambda, \tau)$, if we apply Algorithm 1 with inputs β' and λ', then $\hat{A}_{\mathrm{OSS}} \in \hat{\mathcal{A}}_n(\tau, \alpha, \mathcal{P}_{\mathrm{Höl}}(\beta', \lambda', \tau)) \subseteq \hat{\mathcal{A}}_n(\tau, \alpha, \mathcal{P}_{\mathrm{Höl}}(\beta, \lambda, \tau))$; in other words, if we underspecify the smoothness, then we continue to control the Type I error. On the other hand, we may pay a price in terms of a suboptimal rate of convergence for the regret: in the bounds in Proposition 5, we would need to replace λ and β with λ' and β', respectively. Conversely, if we over-specify the smoothness, then we no longer have guaranteed Type I error control.

In the remainder of this section, we tackle in turn the problems of estimating λ and β from the data.

3.1. *Choice of* λ. In view of Proposition 1, we will need to impose some additional (mild) restrictions on our classes of distributions, as well as a sample size condition, in order to estimate λ effectively. First, however, we describe our algorithm to estimate λ for fixed β, and demonstrate a sense in which it provides a slightly conservative estimate (Corollary 7). It is convenient, for $u \in \mathbb{R} \setminus \{0\}$, to define $u/0 := \infty$ if $u > 0$ and $u/0 := -\infty$ if $u < 0$. For each $i, k \in [n]$ we let $r_{i,k} \equiv r_{i,k}(\mathcal{D}_X) := \inf\{r \in (0, \infty) : |\{X_j\}_{j\in[n]} \cap \bar{B}_r(X_i)| \geq k\}$ denote the kth nearest neighbour distance from X_i within $\mathcal{D}_X$. Recalling the definition of the empirical regression function $\hat{\eta}_n$ from (8), given $\beta \in (0, 1]$, $\delta \in (0, 1)$ and $i, j, k, \ell \in [n]$, we define

$$\hat{\phi}_{n,\beta,\delta}(i, j, k, \ell) := \frac{\hat{\eta}_n(\bar{B}_{r_{i,k}}(X_i)) - \hat{\eta}_n(\bar{B}_{r_{j,\ell}}(X_j)) - \sqrt{2\log(4n^2/\delta)/(k \wedge \ell)}}{\|X_i - X_j\|_\infty^\beta + r_{i,k}^\beta + r_{j,\ell}^\beta},$$

$$\hat{\psi}_{n,\beta,\delta}(i, k) := \frac{1}{(2r_{i,k})^\beta} \cdot \{\hat{\eta}_n(\bar{B}_{r_{i,k}}(X_i)) - \tau - \sqrt{\log(4n^2/\delta)/(2k)}\}.$$

We then set

$$\hat{\lambda}_{n,\beta,\delta} \equiv \hat{\lambda}_{n,\beta,\delta}(\mathcal{D}) := 1 \vee \max_{(i,j,k,\ell)\in[n]^4} \min\{\hat{\phi}_{n,\beta,\delta}(i, j, k, \ell), \hat{\psi}_{n,\beta,\delta}(i, k), \hat{\psi}_{n,\beta,\delta}(j, \ell)\}.$$

To explain the idea behind this construction, suppose that P has continuous regression function η with Hölder constant λ on $\mathcal{X}_\tau(\eta)$, and marginal distribution μ on $\mathbb{R}^d$. Note that $\hat{\eta}_n(\bar{B}_{r_{i,k}}(X_i))$ is the k-nearest neighbour regression estimate of $\eta(X_i)$ when the nearest neighbour distances of X_i are distinct. Thus, $\hat{\eta}_n(\bar{B}_{r_{i,k}}(X_i)) - \sqrt{\log(4n^2/\delta)/(2k)} - \lambda \cdot r_{i,k}^\beta$ is a lower confidence bound for $\eta(X_i)$ and $\hat{\eta}_n(\bar{B}_{r_{j,\ell}}(X_j)) + \sqrt{\log(4n^2/\delta)/(2\ell)} + \lambda \cdot r_{j,\ell}^\beta$ is an upper confidence bound for $\eta(X_j)$. It follows that when $X_i, X_j \in \mathcal{X}_\tau(\eta)$, we have with high probability

that $\hat{\phi}_{n,\beta,\delta}(i,j,k,\ell)$ does not exceed λ for any $k,\ell \in [n]$. On the other hand, if $X_i \notin \mathcal{X}_\tau(\eta)$, then with high probability, $\hat{\psi}_{n,\beta,\delta}(i,k) \leq \lambda$, and similarly with j,ℓ in place of i,k. Thus, with high probability $\hat{\lambda}_{n,\beta,\delta} \leq \lambda$.

Now suppose that there exist well-separated points $x_0, x_1 \in \mathcal{X}_\tau(\eta)$ such that μ is well behaved near both x_0 and x_1, and $\eta(x_0) - \eta(x_1) = \tilde{\lambda} \cdot \|x_0 - x_1\|_\infty^\beta$ for some $\tilde{\lambda} \leq \lambda$, then with high probability, for a sufficiently large sample size, there will also exist data points X_i near x_0 and X_j near x_1, along with $k,\ell \in [n]$ for which $\hat{\phi}_{n,\beta,\delta}(i,j,k,\ell)$ is not too much less than $\tilde{\lambda}$. If, in addition, $\eta(x_0)$ is not too close to τ then with high probability, $\hat{\psi}_{n,\beta,\delta}(i,k)$ will also not be much less than $\tilde{\lambda}$ for an appropriately chosen $k \in [n]$, and similarly for x_1, X_j and ℓ in place of x_0, X_i and k, respectively. Overall, this ensures that $\hat{\lambda}_{n,\beta,\delta}$ will be nearly as large as $\tilde{\lambda}$ with high probability, for a sufficiently large sample size.

THEOREM 6. *Let $(\beta,\lambda) \in (0,1] \times [1,\infty)$ and $P \in \mathcal{P}_{\mathrm{Höl}}(\beta,\lambda,\tau)$. Then, for $\delta \in (0,1)$,*

$$\mathbb{P}_P\left(\sup_{x_0,x_1 \in \mathcal{X}_\tau(\eta - \Delta_n)} \left\{ \frac{|\eta(x_0) - \eta(x_1)| - \Delta_n(x_0) \vee \Delta_n(x_1)}{\|x_0 - x_1\|_\infty^\beta} \right\} \leq \hat{\lambda}_{n,\beta,\delta} \leq \lambda \right) \geq 1 - \delta,$$

where $\Delta_n \equiv \Delta_{n,\beta,\lambda} : \mathbb{R}^d \to [0,\infty]$ is defined by

$$\Delta_n(x) := 192 \cdot \lambda^{d/(2\beta+d)} \cdot \left(\frac{\log(2n/\delta)}{n \cdot \omega(x)} \right)^{\beta/(2\beta+d)}, \tag{10}$$

with the convention that $\Delta_n(x) := \infty$ if $\omega(x) = 0$.

An attraction of Theorem 6 is that it makes no assumptions on P apart from the corresponding regression function satisfying the Hölder condition of Definition 1. We now show that, under further conditions on our class, it is possible to control the lower confidence bound for $\hat{\lambda}_{n,\beta,\delta}$ to guarantee that with high probability it is within a factor of 2 of the appropriate Hölder constant. More precisely, for a distribution P on $\mathbb{R}^d \times [0,1]$ having marginal distribution μ on $\mathbb{R}^d$ and continuous regression function η, and for $\beta \in (0,1]$, we define the β-Hölder constant of P to be

$$\underline{\lambda}_\beta(P) := \sup\left\{ \frac{|\eta(x_0) - \eta(x_1)|}{\|x_0 - x_1\|_\infty^\beta} : x_0, x_1 \in \mathrm{supp}(\mu) \cap \mathcal{X}_\tau(\eta), x_0 \neq x_1 \right\} \vee 1.$$

We write $P \in \mathcal{P}_{\mathrm{Reg}}(\tau)$ if η is continuous and satisfies $\eta^{-1}([\tau,\tau+\varepsilon)) \subseteq \mathrm{supp}(\mu)$ for some $\varepsilon > 0$. Further, for $\lambda \in [1,\infty)$ and $\boldsymbol{\epsilon} = (\epsilon_0,\epsilon_1,\epsilon_2) \in (0,1]^3$, we write $P \in \mathcal{P}^+_{\mathrm{Höl}}(\beta,\lambda,\tau,\boldsymbol{\epsilon})$ if $\underline{\lambda}_\beta(P) \leq \lambda$ and either $\underline{\lambda}_\beta(P) = 1$, or there exist $x_0, x_1 \in \mathcal{X}_{\tau+\epsilon_0}(\eta)$ with $\|x_0 - x_1\|_\infty \geq \epsilon_1$, as well as $\min\{\omega(x_0), \omega(x_1)\} \geq \epsilon_2$ and

$$|\eta(x_0) - \eta(x_1)| \geq \frac{3}{4} \cdot \underline{\lambda}_\beta(P) \cdot \|x_0 - x_1\|_\infty^\beta.$$

The idea here is that if $P \in \mathcal{P}^+_{\mathrm{Höl}}(\beta,\lambda,\tau,\boldsymbol{\epsilon})$ and $\underline{\lambda}_\beta(P) > 1$, then we can find a well-separated pair of points that nearly attains the supremum in the definition of $\underline{\lambda}_\beta(P)$, as well as belonging comfortably to the τ-super level set of η and having μ assign sufficient mass in small neighbourhoods of each of the points. In Lemma S11, we show that

$$\mathcal{P}_{\mathrm{Reg}}(\tau) \cap \mathcal{P}_{\mathrm{Höl}}(\beta,\lambda,\tau) \subseteq \bigcup_{\boldsymbol{\epsilon} \in (0,\infty)^3} \mathcal{P}^+_{\mathrm{Höl}}(\beta,\lambda,\tau,\boldsymbol{\epsilon}),$$

so that, under the mild condition that $P \in \mathcal{P}_{\mathrm{Reg}}(\tau)$, the additional restriction enforced by $P \in \mathcal{P}^+_{\mathrm{Höl}}(\beta,\lambda,\tau,\boldsymbol{\epsilon})$ for small $\epsilon_0,\epsilon_1,\epsilon_2 > 0$ amounts to very little more than asking for P to satisfy the Hölder condition of Definition 1 for some $\lambda \geq 1$.

COROLLARY 7. *Fix* $\beta \in (0,1]$, $\lambda \in [1,\infty)$, $\boldsymbol{\epsilon} = (\epsilon_0, \epsilon_1, \epsilon_2) \in (0,1]^3$ *and take* $P \in \mathcal{P}_{\mathrm{Reg}}(\tau) \cap \mathcal{P}^{+}_{\mathrm{Höl}}(\beta, \lambda, \tau, \boldsymbol{\epsilon})$. *Let* $n \in \mathbb{N}$ *and* $\delta \in (0,1)$ *be such that*

$$\frac{n}{\log(2n/\delta)} \geq \frac{1}{\epsilon_2} \cdot \max\left\{ \left(\frac{192}{\epsilon_0}\right)^{(2\beta+d)/\beta} \lambda^{d/\beta}, \left(\frac{768}{\epsilon_1}\right)^{(2\beta+d)/\beta} \right\}. \tag{11}$$

Then

$$\mathbb{P}_P\left(\frac{\underline{\lambda}_\beta(P)}{2} \leq \hat{\lambda}_{n,\beta,\delta} \leq \underline{\lambda}_\beta(P) \right) \geq 1 - \delta.$$

Corollary 7 reveals that when $P \in \mathcal{P}_{\mathrm{Reg}}(\tau) \cap \mathcal{P}^{+}_{\mathrm{Höl}}(\beta, \lambda, \tau, \boldsymbol{\epsilon})$, the estimator $\hat{\lambda}_{n,\beta,\delta}$ is reliable in the sense that with high probability, it is within a factor of 2 of the desired $\underline{\lambda}_\beta(P)$ for a sufficiently large sample size. However, in order to control Type I error we will in fact use the estimator $2\hat{\lambda}_{n,\beta,\delta}$, which has the benefit of being slightly conservative, that is, it tends to overestimate $\underline{\lambda}_\beta(P)$, while still being within a factor of 2 of the desired $\underline{\lambda}_\beta(P)$. Theorem 8 summarises our overall guarantees when applying Algorithm 1 in this context.

THEOREM 8. *Fix* $\alpha \in (0,1)$, $d \in \mathbb{N}$, $\beta \in (0,1]$, $\lambda \in [1,\infty)$ *and* $\boldsymbol{\epsilon} = (\epsilon_0, \epsilon_1, \epsilon_2) \in (0,1]^3$. *Suppose that* $n \in \mathbb{N}$ *satisfies* (11) *with* $\alpha_n := (\alpha/2) \wedge (1/n)$ *in place of* δ. *Let* $\hat{A}'_{\mathrm{OSS}}$ *denote the output of Algorithm* 1 *with inputs* $\mathcal{D}$, $\mathcal{A}$, τ, α_n, β *and* $2\hat{\lambda}_{n,\beta,\alpha_n}$. *Then*:

(i) Type I error: $\hat{A}'_{\mathrm{OSS}} \in \hat{\mathcal{A}}_n(\tau, \alpha, \mathcal{P}_{\mathrm{Reg}}(\tau) \cap \mathcal{P}^{+}_{\mathrm{Höl}}(\beta, \lambda, \tau, \boldsymbol{\epsilon}))$.

(ii) Regret: *Now suppose further that* $\mathcal{A}$ *satisfies* $\dim_{\mathrm{VC}}(\mathcal{A}) < \infty$ *and* $\varnothing \in \mathcal{A}$, *that* $\alpha \in (0, 1/2]$, *and fix* κ, γ *and* C_{App}. *There exists* $C \geq 1$, *depending only on* d, κ, γ, τ, C_{App} *and* $\dim_{\mathrm{VC}}(\mathcal{A})$, *such that for any* $P \in \mathcal{P}_{\mathrm{Reg}}(\tau) \cap \mathcal{P}^{+}_{\mathrm{Höl}}(\beta, \lambda, \tau, \boldsymbol{\epsilon}) \cap \mathcal{P}_{\mathrm{App}}(\mathcal{A}, \kappa, \gamma, \tau, C_{\mathrm{App}})$, *we have*

$$R_\tau(\hat{A}'_{\mathrm{OSS}}) \leq C\left\{ \left(\frac{\underline{\lambda}_\beta(P)^{d/\beta}}{n} \cdot \log_+\left(\frac{n}{\alpha}\right) \right)^{\frac{\beta\kappa\gamma}{\kappa(2\beta+d)+\beta\gamma}} + \frac{1}{n^{1/2}} \right\}. \tag{12}$$

The main message of Theorem 8 is that replacing λ with $2\hat{\lambda}_{n,\beta,\alpha/2}$ in Algorithm 1 retains both the Type I error validity and the regret guarantees when the sample size is sufficiently large and provided that we make the slight restrictions to our classes of distributions mentioned above. An immediate consequence of this result together with Lemma S11 is that for any $P \in \mathcal{P}_{\mathrm{Reg}}(\tau) \cap \bigcup_{\lambda \in [1,\infty)} \mathcal{P}_{\mathrm{Höl}}(\beta, \lambda, \tau)$, there exists $N_\alpha(P) \in \mathbb{N}$ such that for all $n \geq N_\alpha(P)$ we have both Type I error control, that is, $\mathbb{P}_P(\hat{A}(\mathcal{D}) \subseteq \mathcal{X}_\tau(\eta)) \geq 1 - \alpha$, and the regret guarantee (12). An attraction of this approach to choosing the input λ is that we avoid sample splitting.

3.2. *Choice of* β. Since the algorithm described in Section 3.1 takes $\beta \in (0,1]$ as an input, we now describe a data-driven algorithm for estimating this parameter. In contrast to identifying the Hölder constant, identifying the Hölder exponent requires an analysis of the behaviour of the regression function at multiple scales. This motivates the introduction of distributional classes for which the corresponding regression functions exhibit 'self-similar' behaviour; see Definition 3 below. Related ideas have appeared in the adaptive confidence band literature (Bull (2012), Giné and Nickl (2010), Gur, Momeni and Wager (2022), Picard and Tribouley (2000)). First, given a Borel measure μ on $\mathbb{R}^d$ and $c_0 \in (0,\infty)$, we let

$$\mathcal{R}_\circ(\mu, c_0) := \bigcap_{r \in (0,1)} \{x \in \mathbb{R}^d : \mu(\bar{B}_r(x)) \geq c_0 \cdot r^d\}.$$

Thus, $\mathcal{R}_\circ(\mu, c_0)$ denotes the set of points x for which the μ-measure of small balls centred at x can be bounded below by their volumes, up to constants.

DEFINITION 3 (Self-similar Hölder class). Given $\beta \in (0,1]$, $\lambda \in [1,\infty)$, $\lambda_0 \in (0,\infty)$, $c_0, r_0 \in (0,1]$, we let $\mathcal{P}^{\dagger}_{\mathrm{Höl}}(\beta,\lambda,\lambda_0,c_0,r_0)$ denote the class of all distributions P on $\mathbb{R}^d \times [0,1]$ with regression function $\eta \in \mathcal{F}_{\mathrm{Höl}}(\beta,\lambda,\mathbb{R}^d)$ such that for all $r \in (0,r_0]$ there exist $x_0, x_1 \in \mathcal{R}_\circ(\mu,c_0)$ such that $\|x_0 - x_1\|_\infty \le r$ and $|\eta(x_0) - \eta(x_1)| \ge \lambda_0 \cdot r^\beta$.

We view the regression functions η of distributions $P \in \mathcal{P}^{\dagger}_{\mathrm{Höl}}(\beta,\lambda,\lambda_0,c_0,r_0)$ as self-similar since the fluctuations permitted by the β-Hölder constraint are exhibited at every sufficiently small scale. For $\delta \in (0,1)$, we aim to define estimators $(\hat{\beta}_{n,\delta}, \hat{\lambda}_{n,\hat{\beta}_{n,\delta},\delta})$ of (β,λ). To this end, for $i,j,k,\ell \in [n]$, let

$$\hat{\varepsilon}^{\dagger}_{n,\beta,\delta}(i,j,k,\ell) := -\log\bigl(\|X_i - X_j\|_\infty + r_{i,k} + r_{j,\ell}\bigr),$$

$$\hat{\varphi}^{\dagger}_{n,\beta,\delta}(i,j,k,\ell) := -\log\Bigl(\bigl|\hat{\eta}_n\bigl(\bar{B}_{r_{i,k}}(X_i)\bigr) - \hat{\eta}_n\bigl(\bar{B}_{r_{j,\ell}}(X_j)\bigr)\bigr| - \sqrt{2\log(4n^2/\delta)/(k \wedge \ell)}\Bigr),$$

with the convention that $-\log z := \infty$ for $z \le 0$. We then define

$$\hat{\Gamma}^{\dagger}_{n,\delta} := \bigl\{\bigl(\hat{\varepsilon}^{\dagger}_{n,\beta,\delta}(i,j,k,\ell), \hat{\varphi}^{\dagger}_{n,\beta,\delta}(i,j,k,\ell)\bigr) : (i,j,k,\ell) \in [n]^4, \hat{\varphi}^{\dagger}_{n,\beta,\delta}(i,j,k,\ell) < \infty\bigr\}.$$

Finally, letting $f : \mathbb{N} \to [1,\infty)$ denote any increasing function satisfying $f(n) \to \infty$ as $n \to \infty$, we define

$$\hat{\beta}_{n,\delta} \equiv \hat{\beta}_{n,\delta}(\mathcal{D}) := 0 \vee \max_{(u_0,v_0) \in \hat{\Gamma}^{\dagger}_{n,\delta} : u_0 \le \frac{\log n}{6+2d}} \min_{(u_1,v_1) \in \hat{\Gamma}^{\dagger}_{n,\delta} : u_1 \ge 2u_0} \frac{v_1 - v_0 - \log f(n)}{u_1 - u_0},$$

with the conventions that $\max \varnothing := -\infty$ and $\min \varnothing := \infty$. Theorem 9 below provides a high-probability guarantee on the performance of $\hat{\beta}_{n,\delta}$.

THEOREM 9. *Fix $\beta \in (0,1]$, $d \in \mathbb{N}$, $\lambda \in [1,\infty)$ as well as $\lambda_0 \in (0,\lambda]$, $c_0, r_0 \in (0,1]$ and $P \in \mathcal{P}^{\dagger}_{\mathrm{Höl}}(\beta,\lambda,\lambda_0,c_0,r_0)$. Then for $n \in \mathbb{N}$ such that $f(n) \ge 14\lambda/\lambda_0$ and*

$$n \ge \max\biggl\{\frac{1}{r_0^{(7+2d)}}, \frac{8(16\lambda)^{d/\beta}\log(4n^2/\delta)}{2^{7d/2\beta}c_0}, \biggl(\frac{2^{10}7^{2\beta+d}(16\lambda/\lambda_0)^{d/\beta}\log(4n^2/\delta)}{c_0 \cdot \lambda_0^2}\biggr)^{3+d}\biggr\}, \tag{13}$$

we have

$$\mathbb{P}_P\biggl(\beta - \frac{2(7+2d)\log f(n)}{\log n} \le \hat{\beta}_{n,\delta} \le \beta\biggr) \ge 1 - \delta.$$

Theorem 9 ensures that $\hat{\beta}_{n,\delta}$ is a uniformly consistent estimator over each of our self-similar classes. Moreover, analogously to Corollary 7, it tends to slightly underestimate the true Hölder exponent, which is again advantageous for establishing our overall guarantees on the performance of Algorithm 1 with this data-driven choice of β.

THEOREM 10. *Fix $\alpha \in (0,1)$, $\beta \in (0,1]$, $d \in \mathbb{N}$, $\lambda \in [1,\infty)$, $\lambda_0 \in (0,\lambda]$, $c_0, r_0 \in (0,1]$ and $\boldsymbol{\epsilon} = (\epsilon_0,\epsilon_1,\epsilon_2) \in (0,1]^3$. Let $\tilde{\alpha}_n := (\alpha/3) \wedge (1/n)$. Suppose that $n \in \mathbb{N}$ satisfies $14\lambda/\lambda_0 \le f(n) \le n^{\frac{\log(9/7)}{2(7+2d)\log(1/\epsilon_2)}}$ and*

$$n \ge \max\biggl\{\frac{1}{r_0^{(7+2d)}}, \frac{8(16\lambda)^{d/\beta}\log(12n^3/\alpha)}{2^{7d/2\beta}c_0}, \biggl(\frac{2^{10}7^{2\beta+d}(16\lambda/\lambda_0)^{d/\beta}\log(12n^3/\alpha)}{c_0 \cdot \lambda_0^2}\biggr)^{3+d},$$
$$\frac{\log(6n^2/\alpha)}{\epsilon_2} \cdot \biggl\lceil f(n)^{2(7+2d)} \cdot \biggl(192 \cdot \max\biggl\{\frac{\lambda}{\epsilon_0}, \frac{12}{\epsilon_1}\biggr\}\biggr)^{2+d}\biggr\rceil^{1/\beta}\biggr\}.$$

Let $\hat{A}''_{\mathrm{OSS}}$ *denote the output of Algorithm 1 with input* $\tilde{\alpha}_n$ *in place of* α, $\hat{\beta}_{n,\tilde{\alpha}_n}$ *in place of* β *and* $2\hat{\lambda}_{n,\hat{\beta}_{n,\tilde{\alpha}_n},\tilde{\alpha}_n}$ *in place of* λ. *Then*:

(i) Type I error: $\hat{A}''_{\mathrm{OSS}} \in \hat{\mathcal{A}}_n(\tau, \alpha, \mathcal{P}_{\mathrm{Reg}}(\tau) \cap \mathcal{P}^{+}_{\mathrm{Höl}}(\beta, \lambda, \tau, \boldsymbol{\epsilon}) \cap \mathcal{P}^{\dagger}_{\mathrm{Höl}}(\beta, \lambda, \lambda_0, c_0, r_0))$.

(ii) Regret: *Now suppose further that* $\mathcal{A}$ *satisfies* $\dim_{\mathrm{VC}}(\mathcal{A}) < \infty$ *and* $\varnothing \in \mathcal{A}$, *that* $\alpha \in (0, 1/2]$, *and fix* κ, γ *and* C_{App}. *There exists* $C \geq 1$, *depending only on* d, κ, γ, τ, C_{App} *and* $\dim_{\mathrm{VC}}(\mathcal{A})$, *such that for* $P \in \mathcal{P}_{\mathrm{Reg}}(\tau) \cap \mathcal{P}^{+}_{\mathrm{Höl}}(\beta, \lambda, \tau, \boldsymbol{\epsilon}) \cap \mathcal{P}^{\dagger}_{\mathrm{Höl}}(\beta, \lambda, \lambda_0, c_0, r_0) \cap \mathcal{P}_{\mathrm{App}}(\mathcal{A}, \kappa, \gamma, \tau, C_{\mathrm{App}})$, *we have*

$$R_\tau(\hat{A}''_{\mathrm{OSS}}) \leq C\biggl\{ f(n)^{4\gamma(7+2d)/d}\biggl(\frac{\underline{\lambda}_\beta(P)^{d/\beta}}{n} \cdot \log_+\biggl(\frac{n}{\alpha}\biggr)\biggr)^{\frac{\beta\kappa\gamma}{\kappa(2\beta+d)+\beta\gamma}} + \frac{1}{n^{1/2}}\biggr\}. \tag{14}$$

Theorem 10 confirms that when applying Algorithm 1 with our data-driven choices of β and λ, we maintain large-sample Type I error control over appropriate classes, and only lose a sub-logarithmic factor in n in terms of regret.

4. Higher-order smoothness. In this section, we explain how the procedure and analysis of Section 2 can be modified and extended to cover a general smoothness level $\beta > 0$ for the regression function. Given $\nu = (\nu_1, \ldots, \nu_d)^\top \in \mathbb{N}_0^d$ and $x = (x_1, \ldots, x_d)^\top \in \mathbb{R}^d$, we define $\|\nu\|_1 := \sum_{j=1}^d \nu_j$, $\nu! := \prod_{j=1}^d \nu_j!$ and $x^\nu := \prod_{j=1}^d x_j^{\nu_j}$. For an $\|\nu\|_1$-times differentiable function $g : \mathbb{R}^d \to \mathbb{R}$, define $\partial_x^\nu(g) := \frac{\partial^{\|\nu\|_1} g}{\partial x_1^{\nu_1} \ldots \partial x_d^{\nu_d}}(x)$. Given $\beta \in (0, \infty)$, we let $\mathcal{V}(\beta) := \{\nu \in \mathbb{N}_0^d : \|\nu\|_1 \leq \lceil\beta\rceil - 1\}$, so that $|\mathcal{V}(\beta)| = \binom{\lceil\beta\rceil + d - 1}{d}$, and for a $(\lceil\beta\rceil - 1)$-times differentiable function $g : \mathbb{R}^d \to \mathbb{R}$, let $\mathcal{T}_x^\beta[g] : \mathbb{R}^d \to \mathbb{R}$ denote the associated Taylor polynomial at $x \in \mathbb{R}^d$, defined by

$$\mathcal{T}_x^\beta[g](x') := \sum_{\nu \in \mathcal{V}(\beta)} \frac{(x' - x)^\nu}{\nu!} \cdot \partial_x^\nu(g),$$

for $x' \in \mathbb{R}^d$.

DEFINITION 4 (General Hölder class). Given $(\beta, \lambda) \in (0, \infty) \times [1, \infty)$, let $\tilde{\mathcal{P}}_{\mathrm{Höl}}(\beta, \lambda)$ denote the class of all distributions P on $\mathbb{R}^d \times [0, 1]$ such that the associated regression function $\eta : \mathbb{R}^d \to [0, 1]$ is $(\lceil\beta\rceil - 1)$-differentiable and satisfies

$$|\eta(x') - \mathcal{T}_x^\beta[\eta](x')| \leq \lambda \cdot \|x' - x\|_\infty^\beta,$$

for all $x, x' \in \mathbb{R}^d$. Moreover, we let $\mathcal{P}_{\mathrm{Höl}}(\beta, \lambda) := \bigcap_{\beta' \in (0,\beta]} \tilde{\mathcal{P}}_{\mathrm{Höl}}(\beta', \lambda)$.

Throughout this section, and in contrast to Section 2, we will require that the marginal distribution μ is absolutely continuous with respect to Lebesgue measure, and write f_μ for its density. Given a probability measure μ on $\mathbb{R}^d$ and some $\upsilon \in (0, 1)$, we define

$$\mathcal{R}_\upsilon(\mu) := \bigcap_{r \in (0,1)} \Bigl\{x \in \mathbb{R}^d : \mu\bigl(B_r(x)\bigr) \geq \upsilon \cdot r^d \cdot \sup_{x' \in B_{(1+\upsilon)r}(x)} f_\mu(x')\Bigr\}. \tag{15}$$

To provide some intuition about $\mathcal{R}_\upsilon(\mu)$, consider first a simple example where μ denotes the $N(0, \sigma^2)$ distribution. In that case, the point $x = 0$ belongs to $\mathcal{R}_\upsilon(\mu)$ if and only if $\upsilon \leq \sqrt{2\pi}\sigma\{2\Phi(1/\sigma) - 1\}$. In particular, we must have $\upsilon \leq \sqrt{2\pi}\sigma$, and (since we only consider $\upsilon < 1$), it suffices that $\upsilon \leq 2\sigma$. More generally, if $S \subseteq \mathrm{supp}(\mu)$ is a (c_0, r_0)-regular set ((Audibert and Tsybakov (2007)), equation (2.1)) and if μ is absolutely continuous with respect to $\mathcal{L}_d$ with corresponding density f_μ satisfying the condition that $K_\mu :=$

$\sup_{x\in\mathrm{supp}(\mu)} f_\mu(x)/\inf_{x\in S} f_\mu(x) < \infty$, then $S \subseteq \mathcal{R}_\upsilon(\mu)$ for $\upsilon \le c_0 \cdot V_d \cdot (r_0 \wedge 1)^d \cdot K_\mu^{-1}$. Moreover, we can still have $\mathcal{R}_\upsilon(\mu) = \mathrm{supp}(\mu)$ even when μ is not compactly supported and there is no uniform positive lower bound for f_μ on its support. For instance, the family of probability measures $\{\mu_\kappa : \kappa \in (0,1)\}$ on $\mathbb{R}^d$ considered in Example 3 satisfies $\mathcal{R}_\upsilon(\mu_\kappa) = \mathbb{R}^d$ for $\upsilon \le 2^d \cdot \{1+3^d(1-\kappa)\}^{-1/(1-\kappa)}$. Finally, if $\log f_\mu$ if L-Lipschitz with respect to the supremum norm, then $\mathcal{R}_\upsilon(\mu) = \mathbb{R}^d$ provided that $\upsilon \le 2^d e^{-3L}$.

We are now in a position to define an appropriate definition of approximable classes for regression functions with higher-order smoothness.

DEFINITION 5 (Approximable classes for higher-order smoothness). Given $\mathcal{A} \subseteq \mathcal{B}(\mathbb{R}^d)$ and $(\kappa, \gamma, \upsilon, C_{\mathrm{App}}) \in (0,\infty)^2 \times (0,1) \times [1,\infty)$, we let $\mathcal{P}^+_{\mathrm{App}}(\mathcal{A}, \kappa, \gamma, \upsilon, \tau, C_{\mathrm{App}})$ denote the class of all distributions P on $\mathbb{R}^d \times [0,1]$ with marginal μ on $\mathbb{R}^d$ and a continuous regression function $\eta : \mathbb{R}^d \to [0,1]$ such that

$$\sup\{\mu(A) : A \in \mathcal{A} \cap \mathrm{Pow}(\mathcal{R}_\upsilon(\mu) \cap \mathcal{X}_\xi(f_\mu) \cap \mathcal{X}_{\tau+\Delta}(\eta))\} \ge M_\tau - C_{\mathrm{App}} \cdot (\xi^\kappa + \Delta^\gamma), \tag{16}$$

for all $(\xi, \Delta) \in (0,\infty)^2$.

Finally, then we can state the main theorem of this section.

THEOREM 11. *Take* $(\tau,\alpha) \in (0,1)^2$, $(\beta,\lambda) \in (0,\infty) \times [1,\infty)$ *and* $(\kappa,\gamma,\upsilon,C_{\mathrm{App}}) \in (0,\infty)^2 \times (0,1) \times [1,\infty)$.

(i) Upper bound: *Let* $\mathcal{A} \subseteq \mathcal{B}(\mathbb{R}^d)$ *satisfy* $\dim_{\mathrm{VC}}(\mathcal{A}) < \infty$ *and* $\varnothing \in \mathcal{A}$. *Then there exists* $C \ge 1$, *depending only on* d, β, κ, γ, υ, C_{App} *and* $\dim_{\mathrm{VC}}(\mathcal{A})$, *such that for all* $n \in \mathbb{N}$ *and* $\alpha \in (0, 1/2]$, *we have*

$$\inf_{\hat{A}} \sup_P R_\tau(\hat{A}) \le C \cdot \min\left\{\left(\frac{\lambda^{d/\beta} \cdot \log_+(n/\alpha)}{n}\right)^{\frac{\beta\kappa\gamma}{\kappa(2\beta+d)+\beta\gamma}} + \frac{1}{n^{1/2}}, 1\right\}, \tag{17}$$

where the infimum in (17) *is taken over* $\hat{\mathcal{A}}_n(\tau,\alpha,\mathcal{P}_{\mathrm{Höl}}(\beta,\lambda))$ *and the supremum is taken over* $\mathcal{P}_{\mathrm{Höl}}(\beta,\lambda) \cap \mathcal{P}^+_{\mathrm{App}}(\mathcal{A},\kappa,\gamma,\upsilon,\tau,C_{\mathrm{App}})$.

(ii) Lower bound: *Now suppose that* $\beta\gamma(\kappa-1) < d\kappa$, $\epsilon_0 \in (0,1/2)$, $\tau \in (\epsilon_0, 1-\epsilon_0)$, $\alpha \in (0, 1/2-\epsilon_0]$ *and* $\upsilon \in (0, (4d^{1/2})^{-d}]$. *Then there exists* $c > 0$, *depending only on* d, β, κ, γ, C_{App} *and* ϵ_0, *such that for any* $\mathcal{A} \subseteq \mathcal{B}(\mathbb{R}^d)$ *satisfying* $\mathcal{A}_{\mathrm{hpr}} \subseteq \mathcal{A} \subseteq \mathcal{A}_{\mathrm{conv}}$ *and any* $n \in \mathbb{N}$, *we have*

$$\inf_{\hat{A}} \sup_P R_\tau(\hat{A}) \ge c \cdot \min\left\{\left(\frac{\lambda^{d/\beta} \cdot \log_+\{n/(\lambda^{d/\beta}\alpha)\}}{n}\right)^{\frac{\beta\kappa\gamma}{\kappa(2\beta+d)+\beta\gamma}} + \frac{1}{n^{1/2}}, 1\right\}, \tag{18}$$

where, again, the infimum in (18) *is taken over* $\hat{\mathcal{A}}_n(\tau,\alpha,\mathcal{P}_{\mathrm{Höl}}(\beta,\lambda))$ *and the supremum is taken over* $\mathcal{P}_{\mathrm{Höl}}(\beta,\lambda) \cap \mathcal{P}^+_{\mathrm{App}}(\mathcal{A},\kappa,\gamma,\upsilon,\tau,C_{\mathrm{App}})$.

In order to prove the upper bound in Theorem 11, we will introduce a modified algorithm. The key alteration is a different choice of p-values that now makes use of data points outside (as well as within) our hyper-cube of interest to test whether or not the regression function is uniformly above τ on the hyper-cube. Given $\beta \in (0,\infty)$, $x, x' \in \mathbb{R}^d$, $h \in (0,1]$ and $P \in \mathcal{P}_{\mathrm{Höl}}(\beta,\lambda)$ with regression function η, we let

$$\Phi^\beta_{x,h}(x') := \left(\left(\frac{x'-x}{h}\right)^\nu\right)_{\nu\in\mathcal{V}(\beta)} \in \mathbb{R}^{\mathcal{V}(\beta)} \quad \text{and} \quad w^\beta_{x,h} := \left(\frac{h^{\|\nu\|_1}}{\nu!} \cdot \partial^\nu_x(\eta)\right)_{\nu\in\mathcal{V}(\beta)} \in \mathbb{R}^{\mathcal{V}(\beta)},$$

so that $\mathcal{T}_x^\beta[g](x') = \langle w_{x,h}^\beta, \Phi_{x,h}^\beta(x')\rangle$ for all $x, x' \in \mathbb{R}^d$ and $h \in (0,1]$, where $\langle\cdot,\cdot\rangle$ denotes the Euclidean inner product. Moreover, if we let $e_0 := (\mathbb{1}_{\{\nu=(0,\dots,0)^\top\}})_{\nu\in\mathcal{V}(\beta)} \in \mathbb{R}^{\mathcal{V}(\beta)}$, then $\eta(x) = \langle e_0, w_{x,h}^\beta\rangle$. A natural estimator of $w_{x,h}^\beta$ is the local polynomial estimator obtained by taking $\mathcal{N}_{x,h} := \{i \in [n] : X_i \in \bar{B}_h(x)\}$ and letting

$$\hat{w}_{x,h}^\beta \in \operatorname*{argmin}_{w\in\mathbb{R}^{\mathcal{V}(\beta)}} \sum_{i\in\mathcal{N}_{x,h}} \big(Y_i - \langle w, \Phi_{x,h}^\beta(X_i)\rangle\big)^2.$$

In fact, it will be convenient to choose a particular element of this argmin: if we define

$$V_{x,h}^\beta := \sum_{i\in\mathcal{N}_{x,h}} Y_i \cdot \Phi_{x,h}^\beta(X_i) \in \mathbb{R}^{\mathcal{V}(\beta)},$$

$$Q_{x,h}^\beta := \sum_{i\in\mathcal{N}_{x,h}} \Phi_{x,h}^\beta(X_i)\Phi_{x,h}^\beta(X_i)^\top \in \mathbb{R}^{\mathcal{V}(\beta)\times\mathcal{V}(\beta)},$$

then we will take $\hat{w}_{x,h}^\beta := (Q_{x,h}^\beta)^+ V_{x,h}^\beta$. Thus, $\hat{\eta}(x) := 0 \vee (1 \wedge \langle e_0, \hat{w}_{x,h}^\beta\rangle)$ is an estimator of $\eta(x)$. Next, we associate a p-value to closed hyper-cubes $B \subseteq \mathbb{R}^d$ with $\operatorname{diam}_\infty(B) \le 1$ as follows. Let $x \in \mathbb{R}^d$ and $r \in [0, 1/2]$ denote the centre and ℓ_∞-radius of B, so that $B = \bar{B}_r(x)$. Let $h := (2r)^{1\wedge\frac{1}{\beta}} \in [0,1]$, and define

$$\begin{aligned}\hat{p}_n^+(B) &\equiv \hat{p}_{n,\beta,\lambda}^+(B)\\ &:= \exp\Bigg\{-\frac{2}{e_0^\top (Q_{x,h}^\beta)^{-1} e_0}\\ &\qquad\times \Big(\hat{\eta}(x) - \tau - \lambda\Big(1 + 2\sqrt{e_0^\top (Q_{x,h}^\beta)^{-1} e_0 \cdot |\mathcal{N}_{x,h}|}\Big) r^{\beta\wedge 1}\Big)^2\Bigg\},\end{aligned} \tag{19}$$

whenever $Q_{x,h}^\beta$ is invertible and $\hat{\eta}(x) \ge \tau + \lambda(1 + 2\sqrt{e_0^\top (Q_{x,h}^\beta)^{-1} e_0 \cdot |\mathcal{N}_{x,h}|}) r^{\beta\wedge 1}$, and $\hat{p}_n^+(B) := 1$ otherwise. Lemma S17 in Section S5 shows that these are indeed p-values.

We will also make use of an alternative set of hyper-cubes

$$\mathcal{H}^+ := \Bigg\{2^{-q}\prod_{j=1}^d [a_j, a_j+1] : (a_1,\dots,a_d) \in \mathbb{Z}^d, q \in \mathbb{N}\Bigg\}.$$

Now, given $n \in \mathbb{N}$, $\boldsymbol{x}_{1:n} = (x_i)_{i\in[n]} \in (\mathbb{R}^d)^n$, we define

$$\mathcal{H}^+(\boldsymbol{x}_{1:n}) := \big\{B \in \mathcal{H}^+ : \{x_1,\dots,x_n\} \cap B \ne \varnothing \text{ and } \operatorname{diam}_\infty(B) \ge 1/n\big\}, \tag{20}$$

so that $|\mathcal{H}^+(\boldsymbol{x}_{1:n})| \le 2^d n \log_2 n$, and $|\mathcal{H}^+(\mathcal{D}_X)| \le n\log_2 n$ with probability 1 when μ is absolutely continuous with respect to Lebesgue measure. We denote our modified procedure for general smoothness, obtained by applying Algorithm 1 with the p-values (19) and the hyper-cubes given by (20), as $\hat{A}_{\mathrm{OSS}}^+$.

Propositions 12 and 13 are the analogues of Propositions 4 and 5 for $\hat{A}_{\mathrm{OSS}}^+$ and, in combination, prove the upper bound in Theorem 11.

PROPOSITION 12. *Let $\tau \in (0,1)$, $\alpha \in (0,1)$ and $(\beta,\lambda) \in (0,\infty)\times[1,\infty)$. Then $\hat{A}_{\mathrm{OSS}}^+ \in \hat{\mathcal{A}}_n(\tau,\alpha,\mathcal{P}_{\mathrm{Höl}}(\beta,\lambda))$.*

PROPOSITION 13. *Take $\alpha \in (0,1)$, $(\beta,\lambda) \in (0,\infty)\times[1,\infty)$, $(\kappa,\gamma,\upsilon,C_{\mathrm{App}}) \in (0,\infty)^2\times(0,1)\times[1,\infty)$ and $\mathcal{A} \subseteq \mathcal{B}(\mathbb{R}^d)$ with $\dim_{\mathrm{VC}}(\mathcal{A}) < \infty$ and $\varnothing \in \mathcal{A}$. There exists*

$\tilde{C} \geq 1$, *depending only on* d, β, κ, γ, υ, C_{App} *and* $\dim_{\mathrm{VC}}(\mathcal{A})$, *such that for all* $P \in \mathcal{P}_{\mathrm{Höl}}(\beta, \lambda) \cap \mathcal{P}^{+}_{\mathrm{App}}(\mathcal{A}, \kappa, \gamma, \upsilon, \tau, C_{\mathrm{App}})$, $n \in \mathbb{N}$ *and* $\delta \in (0,1)$, *we have*

$$\mathbb{P}_P\biggl[M_\tau - \mu(\hat{A}^{+}_{\mathrm{OSS}}) > \tilde{C}\biggl\{\biggl(\frac{\lambda^{d/\beta}}{n} \cdot \log_+\biggl(\frac{n}{\alpha \wedge \delta}\biggr)\biggr)^{\frac{\beta\kappa\gamma}{\kappa(2\beta+d)+\beta\gamma}} + \biggl(\frac{\log_+(1/\delta)}{n}\biggr)^{1/2}\biggr\}\biggr] \leq \delta.$$

As a consequence, for $\alpha \in (0, 1/2]$,

$$R_\tau(\hat{A}^{+}_{\mathrm{OSS}}) \leq C\biggl\{\biggl(\frac{\lambda^{d/\beta}}{n} \cdot \log_+\biggl(\frac{n}{\alpha}\biggr)\biggr)^{\frac{\beta\kappa\gamma}{\kappa(2\beta+d)+\beta\gamma}} + \frac{1}{n^{1/2}}\biggr\},$$

where $C > 0$ *depends only on* $\tilde{C}$.

5. Lower bound constructions. As mentioned at the end of Section 2, our lower bound constructions are common to both Theorem 2 and Theorem 11. In fact, both lower bounds will follow from Propositions 14 and 15 below. As shorthand, given $\mathcal{A} \subseteq \mathcal{B}(\mathbb{R}^d)$, $\tau \in (0,1)$, $(\beta, \kappa, \gamma) \in (0, \infty)^3$, $\upsilon \in (0,1)$ and $(\lambda, C_{\mathrm{App}}) \in [1, \infty)^2$, we write

$$\begin{aligned}&\mathcal{P}^{\dagger}(\mathcal{A}, \beta, \kappa, \gamma, \upsilon, \lambda, \tau, C_{\mathrm{App}})\\ &\quad := \mathcal{P}_{\mathrm{Höl}}(\beta, \lambda) \cap \mathcal{P}_{\mathrm{App}}(\mathcal{A}, \kappa, \gamma, \tau, C_{\mathrm{App}}) \cap \mathcal{P}^{+}_{\mathrm{App}}(\mathcal{A}, \kappa, \gamma, \upsilon, \tau, C_{\mathrm{App}}).\end{aligned}$$

PROPOSITION 14. *Take* $\epsilon_0 \in (0, 1/2)$, $\tau \in (\epsilon_0, 1-\epsilon_0)$, $\beta > 0$, $\lambda \geq 1$, $\kappa, \gamma > 0$, $C_{\mathrm{App}} \geq 1$ *with* $\beta\gamma(\kappa - 1) < d\kappa$, *and* $\upsilon \in (0, (4d^{1/2})^{-d}]$. *Suppose that* $\mathcal{A} \subseteq \mathcal{B}(\mathbb{R}^d)$ *satisfies* $\mathcal{A}_{\mathrm{hpr}} \subseteq \mathcal{A} \subseteq \mathcal{A}_{\mathrm{conv}}$.

(i) *There exists* $c_0 > 0$, *depending only on* $d, \beta, \gamma, \kappa, C_{\mathrm{App}}$ *and* ϵ_0, *such that, for every* $\alpha \in (0, 1/8]$, $n \in \mathbb{N}$ *and* $\hat{A} \in \hat{\mathcal{A}}_n(\tau, \alpha, \mathcal{P}^{\dagger}(\mathcal{A}, \beta, \kappa, \gamma, \upsilon, \lambda, \tau, C_{\mathrm{App}}))$, *we can find* $P \in \mathcal{P}^{\dagger}(\mathcal{A}, \beta, \kappa, \gamma, \upsilon, \lambda, \tau, C_{\mathrm{App}})$ *with regression function* $\eta : \mathbb{R}^d \to [\tau - \epsilon_0/2, \tau + \epsilon_0/2]$ *and marginal distribution* μ *on* $\mathbb{R}^d$, *satisfying*

$$\mathbb{E}_P\bigl[\{M_\tau(P, \mathcal{A}) - \mu(\hat{A})\} \cdot \mathbb{1}_{\{\hat{A} \subseteq \mathcal{X}_\tau(\eta)\}}\bigr] \geq c_0 \cdot \biggl(\frac{\lambda^{d/\beta}\log(1/(4\alpha))}{n} \wedge 1\biggr)^{\frac{\beta\kappa\gamma}{\kappa(2\beta+d)+\beta\gamma}}.$$

(ii) *There exists* $c_1 > 0$, *depending only on* $d, \beta, \kappa, \gamma, C_{\mathrm{App}}$ *and* ϵ_0, *such that, given* $\alpha \in (0, \frac{1}{2} - \epsilon_0]$, $n \in \mathbb{N}$ *and* $\hat{A} \in \hat{\mathcal{A}}_n(\tau, \alpha, \mathcal{P}^{\dagger}(\mathcal{A}, \beta, \kappa, \gamma, \upsilon, \lambda, \tau, C_{\mathrm{App}}))$, *we can find* $P \in \mathcal{P}^{\dagger}(\mathcal{A}, \beta, \kappa, \gamma, \upsilon, \lambda, \tau, C_{\mathrm{App}})$ *with regression function* $\eta : \mathbb{R}^d \to [\tau - \epsilon_0/2, \tau + \epsilon_0/2]$ *and marginal distribution* μ *on* $\mathbb{R}^d$, *satisfying*

$$\mathbb{E}_P\bigl[\{M_\tau(P, \mathcal{A}) - \mu(\hat{A})\} \cdot \mathbb{1}_{\{\hat{A} \subseteq \mathcal{X}_\tau(\eta)\}}\bigr] \geq c_1 \cdot \biggl(\frac{\lambda^{d/\beta}\log_+(n/\lambda^{d/\beta})}{n} \wedge 1\biggr)^{\frac{\beta\kappa\gamma}{\kappa(2\beta+d)+\beta\gamma}}.$$

Two remarks are in order. First, note that for any $\hat{A} \in \hat{\mathcal{A}}_n$, we have

$$\begin{aligned}R_\tau(\hat{A}) &= \frac{\mathbb{E}_P[\{M_\tau(P, \mathcal{A}) - \mu(\hat{A})\} \cdot \mathbb{1}_{\{\hat{A} \subseteq \mathcal{X}_\tau(\eta)\}}]}{\mathbb{P}_P(\hat{A} \subseteq \mathcal{X}_\tau(\eta))}\\ &\geq \mathbb{E}_P\bigl[\{M_\tau(P, \mathcal{A}) - \mu(\hat{A})\} \cdot \mathbb{1}_{\{\hat{A} \subseteq \mathcal{X}_\tau(\eta)\}}\bigr],\end{aligned}$$

so Proposition 14 does indeed yield lower bounds on the worst-case regret. Second,

$$\hat{\mathcal{A}}_n\bigl(\tau, \alpha, \mathcal{P}^{\dagger}(\mathcal{A}, \beta, \kappa, \gamma, \upsilon, \lambda, \tau, C_{\mathrm{App}})\bigr) \supseteq \hat{\mathcal{A}}_n\bigl(\tau, \alpha, \mathcal{P}_{\mathrm{Höl}}(\beta, \lambda)\bigr),$$

so it suffices to provide a lower bound for the regret when $\hat{A}$ belongs to the larger set; note also that $\mathcal{P}_{\mathrm{Höl}}(\beta, \lambda) \subseteq \mathcal{P}_{\mathrm{Höl}}(\beta, \lambda, \tau)$ for $\beta \in (0, 1]$.

To lay the groundwork for the proof of Proposition 14, let $L \in \mathbb{N}$, $r \in (0, \infty)$, $w \in (0, (2r)^{-d} \wedge 1)$, $s \in (0, 1 \wedge (r/2)]$ and $\theta \in (0, \epsilon_0/2]$. Our goal is to define a collection of probability distributions $\{P^\ell \equiv P^\ell_{L,r,w,s,\theta} : \ell \in [L]\}$ on $\mathbb{R}^d \times [0,1]$ as illustrated in Figure 1; we will show that these distributions belong to $\mathcal{P}^\dagger(\mathcal{A}, \beta, \kappa, \gamma, \upsilon, \lambda, \tau, C_{\mathrm{App}})$ for appropriate choices of L, r, w, s and θ, and are such that any data-dependent selection set $\hat{A} \in \hat{\mathcal{A}}_n(\tau, \alpha, \mathcal{P}^\dagger(\mathcal{A}, \beta, \kappa, \gamma, \upsilon, \lambda, \tau, C_{\mathrm{App}}))$ must satisfy the lower bound in Proposition 14. To this end, let $r_\sharp(w) := \frac{1}{2}(\{(4\sqrt{d})^d - 2^d\}r^d + w^{-1})^{1/d}$ and choose $\{z_1, \ldots, z_L\} \subseteq \mathbb{R}^d$ such that $\|z_\ell - z_{\ell'}\|_\infty > 2(r_\sharp(w) + 1)$ for all distinct $\ell, \ell' \in [L]$. We introduce sets $K^0_r(1), \ldots, K^0_r(L) \subseteq \mathbb{R}^d$, $K^1_r(1), \ldots, K^1_r(L) \subseteq \mathbb{R}^d$ defined by $K^0_r(\ell) := \bar{B}_r(z_\ell)$ and $K^1_r(\ell) := \bar{B}_{r_\sharp(w)}(z_\ell) \setminus B_{2d^{1/2}r}(z_\ell)$ for $\ell \in [L]$. We also define the probability measure $\mu_{L,r,w}$ on $\mathbb{R}^d$ to be the uniform distribution on $J_{L,r,w} := \bigcup_{(\ell,j)\in[L]\times\{0,1\}} K^j_r(\ell)$; since $\mathcal{L}_d(K^0_r(\ell) \cup K^1_r(\ell)) = (2r)^d + (2r_\sharp(w))^d - (4d^{1/2}r)^d = w^{-1}$ for all $\ell \in [L]$, it follows that the density of $\mu_{L,r,w}$ with respect to $\mathcal{L}_d$ takes the constant value w/L on $J_{L,r,w}$.

Now define a function $h : [0,1] \to [0,1]$ by $h(z) := e^{-z^2/(1-z^2)}$ for $z \in [0,1)$ and $h(1) := 0$, so that $h(0) = 1$, $\max_{k\in\mathbb{N}} \max_{z\in\{0,1\}} |h^{(k)}(z)| = 0$ and

$$A_m := \max_{k\in[m]} \sup_{z\in[0,1]} |h^{(k)}(z)| \in (0, \infty) \tag{21}$$

for each $m \in \mathbb{N}$. This allows us to define regression functions $\eta^\ell_{L,r,w,s,\theta} : \mathbb{R}^d \to [0,1]$ for $\ell \in [L]$ by

$$\eta^\ell_{L,r,w,s,\theta}(x) := \begin{cases} \tau + \theta & \text{if } \|x - z_\ell\|_2 \le d^{1/2}r, \\ \tau - \theta & \text{if } \|x - z_{\ell'}\|_2 \le s \text{ with } \ell' \in [L]\setminus\{\ell\}, \\ \tau + \theta - 2\theta h\Big(\dfrac{\|x - z_{\ell'}\|_2}{s} - 1\Big) & \text{if } s < \|x - z_{\ell'}\|_2 \le 2s \text{ with } \ell' \in [L]\setminus\{\ell\}, \\ \tau + \theta & \text{if } 2s < \|x - z_{\ell'}\|_2 \le d^{1/2}r \text{ with } \ell' \in [L]\setminus\{\ell\}, \\ \tau - \theta + 2\theta h\Big(\dfrac{\|x - z_{\ell'}\|_2}{d^{1/2}r} - 1\Big) & \text{if } d^{1/2}r < \|x - z_{\ell'}\|_2 < 2d^{1/2}r \text{ with } \ell' \in [L], \\ \tau - \theta & \text{otherwise.} \end{cases} \tag{22}$$

Thus, $\eta^\ell_{L,r,w,s,\theta}$ is infinitely differentiable and uniformly above the level τ on $K^0_r(\ell)$, but both takes a value below τ at the centre $z_{\ell'}$ of each $K^0_r(\ell')$ with $\ell' \in [L]\setminus\{\ell\}$, and is uniformly below the level τ on each $K^1_r(\ell')$ with $\ell' \in [L]$. Finally, then for $\ell \in [L]$, we can let $P^\ell_{L,r,w,s,\theta}$ denote the unique Borel probability distribution on $\mathbb{R}^d \times \{0,1\}$ with marginal $\mu_{L,r,w}$ on $\mathbb{R}^d$ and regression function $\eta^\ell_{L,r,w,s,\theta}$. Figure 1 provides an illustration of the regression functions used in this construction.

Lemmas S28 and S30 verify that $\{P^\ell_{L,r,w,s,\theta} : \ell \in [L]\} \subseteq \mathcal{P}^\dagger(\mathcal{A}, \beta, \kappa, \gamma, \upsilon, \lambda, \tau, C_{\mathrm{App}})$ for appropriate choices of r, w, s and θ. Moreover, Proposition S32 reveals both that the chi-squared divergences between pairs of distributions in our construction are small, and yet that the distributions are sufficiently different that any set $A \in \mathcal{A} \cap \mathrm{Pow}(\mathcal{X}_\tau(\eta^\ell_{L,r,w,s,\theta}) \cap \mathcal{X}_\tau(\eta^{\ell'}_{L,r,w,s,\theta}))$ for distinct $\ell, \ell' \in [L]$ must have much smaller μ-measure than M_τ. To conclude, we apply a constrained risk inequality due to Brown and Low (1996) in the proof of Proposition 14(i) and a version of Fano's lemma in the proof of Proposition 14(ii).

Proposition 15 provides the final (parametric) part of the lower bounds in Theorems 2 and 11.

PROPOSITION 15. *Take* $\epsilon_0 \in (0, 1/2]$, $\tau \in (\epsilon_0, 1 - \epsilon_0)$ $(\beta, \lambda) \in (0, \infty) \times [1, \infty)$, $(\kappa, \gamma, \upsilon, C_{\mathrm{App}}) \in (0, \infty)^2 \times (0,1) \times [1, \infty)$ *with* $\beta\gamma(\kappa - 1) < d\kappa$ *and* $\upsilon \in (0, 4^{-d}]$. *Suppose that* $\mathcal{A} \subseteq \mathcal{B}(\mathbb{R}^d)$ *satisfies* $\mathcal{A}_{\mathrm{hpr}} \subseteq \mathcal{A} \subseteq \mathcal{A}_{\mathrm{conv}}$. *Then there exists* $c_2 > 0$, *depending*

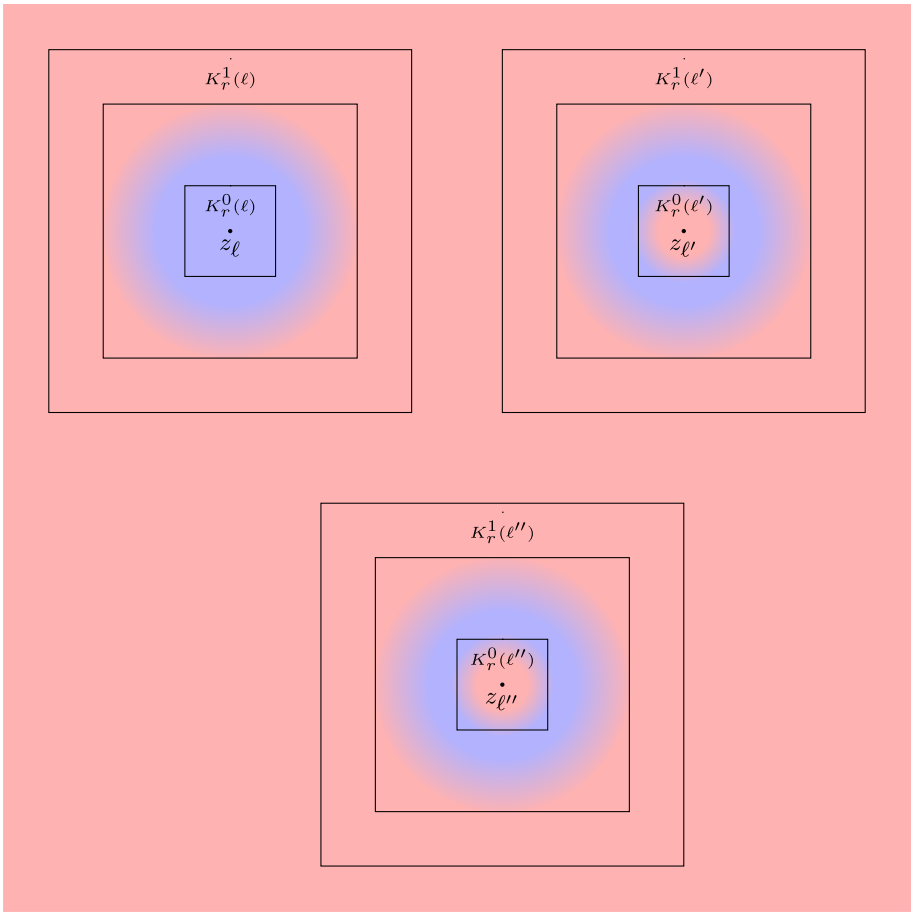

FIG. 1. *Illustration of the lower bound construction of* $P^{\ell}_{L,r,w,s,\theta}$ *in the proof of Proposition 14. Blue and red regions correspond to the regression function* $\eta^{\ell}_{L,r,w,s,\theta}$ *being above and below* τ*, respectively. Note the different behaviour in the* ℓ*th region* $K_r^0(\ell)$ *from the others. The marginal measure* $\mu^{\ell}_{L,r,w,s,\theta}$ *on* $\mathbb{R}^d$ *is uniformly distributed on* $\bigcup_{\ell\in[L]}(K_r^0(\ell)\cup K_r^1(\ell))$*; the boundaries of these regions are denoted with black lines.*

only on ϵ_0, d, β, κ, γ, λ *and* C_{App}, *such that for any* $n\in\mathbb{N}$, $\alpha\in(0,1/2-\epsilon_0]$ *and* $\hat{A}\in\hat{\mathcal{A}}_n(\tau,\alpha,\mathcal{P}^{\dagger}(\mathcal{A},\beta,\kappa,\gamma,\upsilon,\lambda,\tau,C_{\mathrm{App}}))$, *we can find* $P\in\mathcal{P}^{\dagger}(\mathcal{A},\beta,\kappa,\gamma,\upsilon,\lambda,\tau,C_{\mathrm{App}})$ *with regression function* $\eta:\mathbb{R}^d\to[\tau-\epsilon_0/2,\tau+\epsilon_0/2]$ *and marginal distribution* μ *on* $\mathbb{R}^d$ *that satisfies*

$$\mathbb{E}_P\big[\{M_\tau(P,\mathcal{A})-\mu(\hat{A})\}\cdot\mathbb{1}_{\{\hat{A}\subseteq\mathcal{X}_\tau(\eta)\}}\big]\geq\frac{c_2}{\sqrt{n}}.$$

The construction for the proof of Proposition 15 is somewhat different from those in the proof of Proposition 14 and is illustrated in Figure 2: it hinges on the difficulty of estimating $\mu(A)$ for $A\in\mathcal{A}$. To formalise this idea, given $t\in[1,\infty)$, $\theta\in(0,\epsilon_0/2]$, $s\in(0,1]$ and $\zeta\in[0,\frac{s^d}{2\{(2t)^d+2s^d\}}]$, we first define a pair of distributions $\{P^{\ell}_{t,\theta,s,\zeta}\}_{\ell\in\{-1,1\}}$ on $\mathbb{R}^d\times[0,1]$. Define

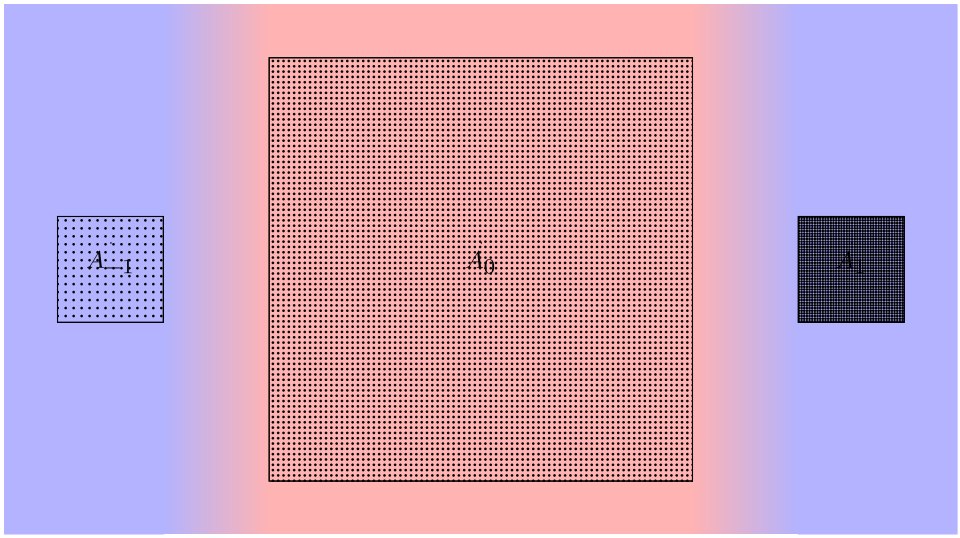

FIG. 2. *Illustration of the lower bound construction of* $P_\zeta^1 \equiv P_{t,\theta,s,\zeta}^1$ *in the proof of Proposition* 15. *Blue and red regions correspond to the regression function* $\eta_{t,\theta,s}$ *being above and below* τ, *respectively. The density of dots is greatest on* A_1 *and smallest on* A_{-1}, *reflecting the greater marginal density of* μ_ζ^1 *on* A_1; *for* P_ζ^{-1}, *the density of dots would be reversed.*

$\eta \equiv \eta_{t,\theta,s} : \mathbb{R}^d \to [0,1]$ by

$$\eta(x) \equiv \eta_{t,\theta,s}(x_1,\dots,x_d) := \begin{cases} \tau+\theta & \text{for } x_1 \le -t-s, \\ \tau+\theta\left\{1-2h\left(\dfrac{-x_1-t}{s}\right)\right\} & \text{for } -t-s < x_1 \le -t, \\ \tau-\theta & \text{for } -t < x_1 \le t, \\ \tau+\theta\left\{1-2h\left(\dfrac{x_1-t}{s}\right)\right\} & \text{for } t < x_1 \le t+s, \\ \tau+\theta & \text{for } x_1 \ge t+s. \end{cases}$$

Define $A_0 := [-t,t]^d$, $A_{-1} := [-t-2s, -t-s] \times [-\frac{s}{2}, \frac{s}{2}]^{d-1}$ and $A_1 := [t+s, t+2s] \times [-\frac{s}{2}, \frac{s}{2}]^{d-1}$. For $\ell \in \{-1,1\}$, let $\mu_\zeta^\ell \equiv \mu_{t,s,\zeta}^\ell$ be the Lebesgue absolutely continuous measure supported on $A_{-1} \cup A_0 \cup A_1 \subseteq \mathbb{R}^d$ with piecewise constant density $f_{\mu_\zeta^\ell} : \mathbb{R}^d \to [0,\infty)$ given by

$$f_{\mu_\zeta^\ell}(x) := \begin{cases} \dfrac{1}{(2t)^d + 2s^d} + \dfrac{\zeta \cdot j \cdot \ell}{s^d} & \text{for } x \in A_j \text{ with } j \in \{-1,0,1\}, \\ 0 & \text{for } x \notin A_{-1} \cup A_0 \cup A_1. \end{cases}$$

Now for $\ell \in \{-1,1\}$, let $P_\zeta^\ell \equiv P_{t,\theta,s,\zeta}^\ell$ denote the unique distribution on $\mathbb{R}^d \times \{0,1\}$ with marginal μ_ζ^ℓ on $\mathbb{R}^d$ and regression function η. Figure 2 illustrates this construction.

In the proof of Proposition 15, we will show that $\{P_{t,\theta,s,\zeta}^\ell : \ell \in \{-1,1\}\} \subseteq \mathcal{P}^\dagger(\mathcal{A}, \beta, \kappa, \gamma, \upsilon, \lambda, \tau, C_{\mathrm{App}})$ for suitable t, θ, s and ζ. Moreover, $P_{t,\theta,s,\zeta}^{-1}$ and $P_{t,\theta,s,\zeta}^1$ are close in chi-squared divergence, but nevertheless we cannot have both $\mu_\zeta^{-1}(A)$ and $\mu_\zeta^1(A)$ close to $M_\tau(P_\zeta^{-1}, \mathcal{A}) = M_\tau(P_\zeta^1, \mathcal{A})$ for $A \in \mathcal{A} \cap \mathrm{Pow}(\mathcal{X}_\tau(\eta))$. Hence, any data-dependent selection set $\hat{A}$ that satisfies our Type I error guarantee must incur large regret for at least one of these distributions.

6. Application to study of heterogeneous treatment effects. The aim of this section is to show how our previous results may be applied to the two-arm setting with a treatment

and control, where we are interested in regions of substantial treatment effect. To this end, let $\tilde{P}$ denote the distribution of a random triple $(X, T, \tilde{Y})$ taking values in $\mathbb{R}^d \times \{0,1\} \times [0,1]$, where X represents covariates, T is a treatment indicator and $\tilde{Y}$ denotes the corresponding response. Assume that $X \sim \mu$, and that the function $\pi : \mathbb{R}^d \to [0,1]$ given by $\pi(x) := \mathbb{P}(T = 1 | X = x)$ is known. For $\ell \in \{0,1\}$, let $\tilde{\eta}^\ell(x) := \mathbb{E}(\tilde{Y} | X = x, T = \ell)$. The *heterogeneous treatment effect* is the function $\varphi : \mathbb{R}^d \to [-1,1]$ defined by $\varphi(x) := \tilde{\eta}^1(x) - \tilde{\eta}^0(x)$ for $x \in \mathbb{R}^d$. Given $t \in [-1,1]$ and a class of sets $\mathcal{A} \subseteq \mathcal{B}(\mathbb{R}^d)$, our primary interest is in identifying subsets $A \in \mathcal{A}$ that are contained in $\mathcal{X}_t(\varphi) := \{x \in \mathbb{R}^d : \varphi(x) \geq t\}$ based on data $\tilde{\mathcal{D}} := ((X_1, T_1, \tilde{Y}_1), \ldots, (X_n, T_n, \tilde{Y}_n)) \sim \tilde{P}^{\otimes n}$.

Given a family $\mathcal{P}$ of distributions on $\mathbb{R}^d \times \{0,1\} \times [0,1]$ and a significance level $\alpha \in (0,1)$, we let $\hat{\mathcal{A}}_n^{\mathrm{HTE}}(t, \alpha, \mathcal{P})$ denote the set of functions $\hat{A} : (\mathbb{R}^d \times \{0,1\} \times [0,1])^n \to \mathcal{A}$ such that $(x, \tilde{D}) \mapsto \mathbb{1}_{\hat{A}(\tilde{D})}(x)$ is a Borel measurable function on $\mathbb{R}^d \times (\mathbb{R}^d \times \{0,1\} \times [0,1])^n$ and we have the Type I error guarantee that

$$\inf_{\tilde{P} \in \mathcal{P}} \mathbb{P}_{\tilde{P}}\big(\hat{A}(\tilde{\mathcal{D}}) \subseteq \mathcal{X}_t(\varphi)\big) \geq 1 - \alpha.$$

Similar to our formulation in Section 2, we seek $\hat{A} \in \hat{\mathcal{A}}_n^{\mathrm{HTE}}(t, \alpha, \mathcal{P})$ with low regret

$$R_t^\varphi(\hat{A}) := \sup\{\mu(A) : A \in \mathcal{A} \cap \mathrm{Pow}\big(\mathcal{X}_t(\varphi)\big)\} - \mathbb{E}_{\tilde{P}}\{\mu\big(\hat{A}(\tilde{\mathcal{D}})\big) | \hat{A}(\tilde{\mathcal{D}}) \subseteq \mathcal{X}_t(\varphi)\}.$$

Given $(\beta, \lambda) \in (0,\infty) \times [1,\infty)$ and a Borel measurable function $\pi : \mathbb{R}^d \to [0,1]$, we let $\mathcal{P}_{\mathrm{Höl}}^{\mathrm{HTE}}(\beta, \lambda, \pi)$ denote the class of distributions $\tilde{P}$ on $\mathbb{R}^d \times \{0,1\} \times [0,1]$ such that φ is (β, λ)-Hölder (see Definition 4), and such that $\pi(x) = \mathbb{P}_{\tilde{P}}(T = 1 | X = x)$ for all $x \in \mathbb{R}^d$. Similarly, given $(\kappa, \gamma, \upsilon, C_{\mathrm{App}}) \in (0,\infty)^2 \times (0,1) \times [1,\infty)$, we let $\mathcal{P}_{\mathrm{App}}^{\mathrm{HTE}}(\mathcal{A}, t, \kappa, \gamma, \upsilon, C_{\mathrm{App}})$ denote the class of all distributions $\tilde{P}$ such that μ is absolutely continuous, with Lebesgue density f_μ, and such that (16) holds with φ in place of η, with t in place of τ, and with $\sup\{\mu(A) : A \in \mathcal{A} \cap \mathrm{Pow}(\mathcal{X}_t(\varphi))\}$ in place of M_τ.

The following result on the minimax rate of regret in this heterogeneous treatment effect context is an almost immediate corollary of Theorem 11.

COROLLARY 16. *Take $\zeta_0 \in (0, 1/2)$, $t \in [-(1-\zeta_0), 1-\zeta_0]$, $(\beta, \lambda) \in (0,\infty) \times [1,\infty)$, $(\kappa, \gamma, \upsilon, C_{\mathrm{App}}) \in (0,\infty)^2 \times (0,1) \times [1,\infty)$ with $\beta\gamma(\kappa - 1) < d\kappa$ and $\upsilon \in (0, (4d^{1/2})^{-d}]$, and let $\pi : \mathbb{R}^d \to [\zeta_0, 1-\zeta_0]$ be Borel measurable. Let $\mathcal{A} \subseteq \mathcal{B}(\mathbb{R}^d)$ satisfy $\mathcal{A}_{\mathrm{hpr}} \subseteq \mathcal{A} \subseteq \mathcal{A}_{\mathrm{conv}}$, $\dim_{\mathrm{VC}}(\mathcal{A}) < \infty$ and $\varnothing \in \mathcal{A}$. Given $n \in \mathbb{N}$ and $\alpha \in (0, 1/2 - \zeta_0]$, we have*

$$\inf_{\hat{A}} \sup_{\tilde{P}} R_t^\varphi(\hat{A}) \asymp \min\left\{\left(\frac{\log_+(n/\alpha)}{n}\right)^{\frac{\beta\kappa\gamma}{\kappa(2\beta+d)+\beta\gamma}} + \frac{1}{n^{1/2}}, 1\right\}, \tag{23}$$

where the infimum in (23) is taken over $\hat{\mathcal{A}}_n^{\mathrm{HTE}}(t, \alpha, \mathcal{P}_{\mathrm{Höl}}^{\mathrm{HTE}}(\beta, \lambda, \pi))$, the supremum is taken over $\mathcal{P}_{\mathrm{Höl}}^{\mathrm{HTE}}(\beta, \lambda, \pi) \cap \mathcal{P}_{\mathrm{App}}^{\mathrm{HTE}}(\mathcal{A}, t, \kappa, \gamma, \upsilon, C_{\mathrm{App}})$. In (23), $\asymp$ indicates that the ratio of the left- and right-hand sides is bounded above and below by positive quantities depending only on d, β, λ, κ, γ, υ, C_{App}, ζ_0 and $\dim_{\mathrm{VC}}(\mathcal{A})$.

To establish the upper bound in Corollary 16, we reduce the problem to the setting of Section 4 by letting $\rho_{\min} := \min\{\inf_{x \in \mathbb{R}^d} \pi(x), 1 - \sup_{x \in \mathbb{R}^d} \pi(x)\}$ and introducing proxy labels

$$Y := \frac{1}{2}\left\{1 + \frac{\rho_{\min}}{\pi(X)(1-\pi(X))} \cdot \big(T - \pi(X)\big)\tilde{Y}\right\},$$

so that Y takes values in $[0,1]$ and satisfies both $\eta(x) := \mathbb{E}(Y | X = x) = \frac{1}{2}(1 + \rho_{\min} \cdot \varphi(x))$ and $\mathcal{X}_t(\varphi) = \mathcal{X}_\tau(\eta)$ with $\tau := \frac{1}{2}(1 + \rho_{\min} \cdot t)$. The upper bound then follows from Theorem 11(i).

To deduce the lower bound, we convert distributions P of random pairs (X, Y) into distributions $\tilde{P}$ of random triples $(X, T, \tilde{Y})$ for which $\mathbb{P}_{\tilde{P}}(T = 1|X = x, Y = y) = \pi(x)$ and $\tilde{Y} := T \cdot Y + (1 - T) \cdot (1 - Y)$, so that the corresponding heterogeneous treatment effect satisfies $\varphi(x) = 2\eta(x) - 1$. We may therefore deduce the lower bound in Corollary 16 from Theorem 11(ii), applied with $\tau := (1 + t)/2$.

Acknowledgments. We thank the anonymous reviewers for constructive feedback that helped to improve the paper.

Funding. The second author was supported by Engineering and Physical Sciences Research Council (EPSRC) New Investigator Award EP/V002694/1.

The third author was supported by Engineering and Physical Sciences Research Council (EPSRC) Programme Grant EP/N031938/1, EPSRC Fellowship EP/P031447/1 and European Research Council Advanced Grant 101019498.

SUPPLEMENTARY MATERIAL

Supplement to "Optimal subgroup selection" (DOI: 10.1214/23-AOS2328SUPP; .pdf). Supplementary information.

REFERENCES

ALTMAN, D. G. (2015). Clinical trials: Subgroup analyses in randomized trials—more rigour needed. *Nat. Rev. Clin. Oncol.* **12** 506–507. https://doi.org/10.1038/nrclinonc.2015.133

AUDIBERT, J.-Y. and TSYBAKOV, A. B. (2007). Fast learning rates for plug-in classifiers. *Ann. Statist.* **35** 608–633. MR2336861 https://doi.org/10.1214/009053606000001217

BALLARINI, N. M., ROSENKRANZ, G. K., JAKI, T., KÖNIG, F. and POSCH, M. (2018). Subgroup identification in clinical trials via the predicted individual treatment effect. *PLoS ONE* **13** e0205971. https://doi.org/10.1371/journal.pone.0205971

BROOKES, S. T., WHITELY, E., EGGER, M., SMITH, G. D., MULHERAN, P. A. and PETERS, T. J. (2004). Subgroup analyses in randomized trials: Risks of subgroup-specific analyses; power and sample size for the interaction test. *J. Clin. Epidemiol.* **57** 229–236. https://doi.org/10.1016/j.jclinepi.2003.08.009

BROOKES, S. T., WHITLEY, E., PETERS, T. J., MULHERAN, P. A., EGGER, M. and SMITH, G. D. (2001). Subgroup analyses in randomised controlled trials: Quantifying the risks of false-positives and false-negatives. *Health Technol. Assess.* **5** 1–56. https://doi.org/10.3310/hta5330

BROWN, L. D. and LOW, M. G. (1996). A constrained risk inequality with applications to nonparametric functional estimation. *Ann. Statist.* **24** 2524–2535. MR1425965 https://doi.org/10.1214/aos/1032181166

BULL, A. D. (2012). Honest adaptive confidence bands and self-similar functions. *Electron. J. Stat.* **6** 1490–1516. MR2988456 https://doi.org/10.1214/12-EJS720

CANNON, A., HOWSE, J., HUSH, D. and SCOVEL, C. (2002). Learning with the Neyman–Pearson and min-max criteria. Los Alamos National Laboratory, Tech. Rep. LA-UR 02–2951.

CAVALIER, L. (1997). Nonparametric estimation of regression level sets. *Statistics* **29** 131–160. MR1484386 https://doi.org/10.1080/02331889708802579

CHEN, Y.-C., GENOVESE, C. R. and WASSERMAN, L. (2017). Density level sets: Asymptotics, inference, and visualization. *J. Amer. Statist. Assoc.* **112** 1684–1696. MR3750891 https://doi.org/10.1080/01621459.2016.1228536

CRUMP, R. K., HOTZ, V. J., IMBENS, G. W. and MITNIK, O. A. (2008). Nonparametric tests for treatment effect heterogeneity. *Rev. Econ. Stat.* **90** 389–405.

DAU, H. D., LALOË, T. and SERVIEN, R. (2020). Exact asymptotic limit for kernel estimation of regression level sets. *Statist. Probab. Lett.* **161** 108721. MR4065485 https://doi.org/10.1016/j.spl.2020.108721

DOSS, C. R. and WENG, G. (2018). Bandwidth selection for kernel density estimators of multivariate level sets and highest density regions. *Electron. J. Stat.* **12** 4313–4376. MR3892342 https://doi.org/10.1214/18-ejs1501

DUSSELDORP, E., CONVERSANO, C. and VAN OS, B. J. (2010). Combining an additive and tree-based regression model simultaneously: STIMA. *J. Comput. Graph. Statist.* **19** 514–530. MR2759902 https://doi.org/10.1198/jcgs.2010.06089

FEINSTEIN, A. R. (1998). The problem of cogent subgroups: A clinicostatistical tragedy. *J. Clin. Epidemiol.* **51** 297–299. https://doi.org/10.1016/s0895-4356(98)00004-3

FOSTER, J. C., TAYLOR, J. M. G. and RUBERG, S. J. (2011). Subgroup identification from randomized clinical trial data. *Stat. Med.* **30** 2867–2880. MR2844689 https://doi.org/10.1002/sim.4322

GABLER, N. B., DUAN, N., RANESES, E., SUTTNER, L., CIARAMETARO, M., COONEY, E., DUBOIS, R. W., HALPERN, S. D. and KRAVITZ, R. L. (2016). No improvement in the reporting of clinical trial subgroup effects in high-impact general medical journals. *Trials* **17** 1–12.

GINÉ, E. and NICKL, R. (2010). Confidence bands in density estimation. *Ann. Statist.* **38** 1122–1170. MR2604707 https://doi.org/10.1214/09-AOS738

GOTOVOS, A., CASATI, N., HITZ, G. and KRAUSE, A. (2013). Active learning for level set estimation. In *Twenty-Third International Joint Conference on Artificial Intelligence* 1344–1350.

GUR, Y., MOMENI, A. and WAGER, S. (2022). Smoothness-adaptive contextual bandits. *Oper. Res.* **70** 3198–3216. MR4538513 https://doi.org/10.1287/opre.2021.2215

HERRERA, F., CARMONA, C. J., GONZÁLEZ, P. and DEL JESUS, M. J. (2011). An overview on subgroup discovery: Foundations and applications. *Knowl. Inf. Syst.* **29** 495–525.

HOLM, S. (1979). A simple sequentially rejective multiple test procedure. *Scand. J. Stat.* **6** 65–70. MR0538597

HUBER, C., BENDA, N. and FRIEDE, T. (2019). A comparison of subgroup identification methods in clinical drug development: Simulation study and regulatory considerations. *Pharm. Stat.* **18** 600–626. https://doi.org/10.1002/pst.1951

HYNDMAN, R. J. (1996). Computing and graphing highest density regions. *Amer. Statist.* **50** 120–126.

KAUFMAN, J. S. and MACLEHOSE, R. F. (2013). Which of these things is not like the others? *Cancer* **119** 4216–4222.

KEHL, V. and ULM, K. (2006). Responder identification in clinical trials with censored data. *Comput. Statist. Data Anal.* **50** 1338–1355. MR2224375 https://doi.org/10.1016/j.csda.2004.11.015

LAGAKOS, S. W. (2006). The challenge of subgroup analyses–reporting without distorting. *N. Engl. J. Med.* **354** 1667–1669. https://doi.org/10.1056/NEJMp068070

LALOË, T. and SERVIEN, R. (2013). Nonparametric estimation of regression level sets using kernel plug-in estimator. *J. Korean Statist. Soc.* **42** 301–311. MR3255389 https://doi.org/10.1016/j.jkss.2012.10.001

LIPKOVICH, I., DMITRIENKO, A. and D'AGOSTINO, R. B. SR. (2017). Tutorial in biostatistics: Data-driven subgroup identification and analysis in clinical trials. *Stat. Med.* **36** 136–196. MR3580950 https://doi.org/10.1002/sim.7064

LIPKOVICH, I., DMITRIENKO, A., DENNE, J. and ENAS, G. (2011). Subgroup identification based on differential effect search—a recursive partitioning method for establishing response to treatment in patient subpopulations. *Stat. Med.* **30** 2601–2621. MR2815438 https://doi.org/10.1002/sim.4289

MAMMEN, E. and POLONIK, W. (2013). Confidence regions for level sets. *J. Multivariate Anal.* **122** 202–214. MR3189318 https://doi.org/10.1016/j.jmva.2013.07.017

MASON, D. M. and POLONIK, W. (2009). Asymptotic normality of plug-in level set estimates. *Ann. Appl. Probab.* **19** 1108–1142. MR2537201 https://doi.org/10.1214/08-AAP569

PATEL, S., HEE, S. W., MISTRY, D., JORDAN, J., BROWN, S., DRITSAKI, M., ELLARD, D. R., FRIEDE, T., LAMB, S. E. et al. (2016). Identifying back pain subgroups: Developing and applying approaches using individual patient data collected within clinical trials. *Programme Grants for Applied Research* **4**.

PICARD, D. and TRIBOULEY, K. (2000). Adaptive confidence interval for pointwise curve estimation. *Ann. Statist.* **28** 298–335. MR1762913 https://doi.org/10.1214/aos/1016120374

POLONIK, W. (1995). Measuring mass concentrations and estimating density contour clusters—an excess mass approach. *Ann. Statist.* **23** 855–881. MR1345204 https://doi.org/10.1214/aos/1176324626

QIAO, W. (2020). Asymptotics and optimal bandwidth for nonparametric estimation of density level sets. *Electron. J. Stat.* **14** 302–344. MR4048601 https://doi.org/10.1214/19-EJS1668

QIAO, W. and POLONIK, W. (2019). Nonparametric confidence regions for level sets: Statistical properties and geometry. *Electron. J. Stat.* **13** 985–1030. MR3934621 https://doi.org/10.1214/19-EJS1543

REEVE, H. W. J., CANNINGS, T. I. and SAMWORTH, R. J. (2021). Adaptive transfer learning. *Ann. Statist.* **49** 3618–3649. MR4352543 https://doi.org/10.1214/21-aos2102

REEVE, H. W. J., CANNINGS, T. I. and SAMWORTH, R. J. (2023). Supplement to "Optimal subgroup selection." https://doi.org/10.1214/23-AOS2328SUPP

RODRÍGUEZ-CASAL, A. and SAAVEDRA-NIEVES, P. (2019). Minimax Hausdorff estimation of density level sets. ArXiv preprint. Available at arXiv:1905.02897.

ROTHWELL, P. M. (2005). Subgroup analysis in randomised controlled trials: Importance, indications, and interpretation. *Lancet* **365** 176–186.

SAMWORTH, R. J. and WAND, M. P. (2010). Asymptotics and optimal bandwidth selection for highest density region estimation. *Ann. Statist.* **38** 1767–1792. MR2662359 https://doi.org/10.1214/09-AOS766

SCOTT, C. and DAVENPORT, M. (2007). Regression level set estimation via cost-sensitive classification. *IEEE Trans. Signal Process.* **55** 2752–2757. MR1500201 https://doi.org/10.1109/TSP.2007.893758

SCOTT, C. and NOWAK, R. (2005). A Neyman–Pearson approach to statistical learning. *IEEE Trans. Inf. Theory* **51** 3806–3819. MR2239000 https://doi.org/10.1109/TIT.2005.856955

SEIBOLD, H., ZEILEIS, A. and HOTHORN, T. (2016). Model-based recursive partitioning for subgroup analyses. *Int. J. Biostat.* **12** 45–63. MR3505686 https://doi.org/10.1515/ijb-2015-0032

SENN, S. and HARRELL, F. (1997). On wisdom after the event. *J. Clin. Epidemiol.* **50** 749–751. https://doi.org/10.1016/s0895-4356(97)00023-1

SHALEV-SHWARTZ, S. and BEN-DAVID, S. (2014). *Understanding Machine Learning*: *From Theory to Algorithms*. Cambridge Univ. Press, Cambridge.

SU, X., TSAI, C.-L., WANG, H., NICKERSON, D. M. and LI, B. (2009). Subgroup analysis via recursive partitioning. *J. Mach. Learn. Res.* **10** 141–158.

TING, N., CAPPELLERI, J. C., HO, S. and CHEN, D.-G. (2020). *Design and Analysis of Subgroups with Biopharmaceutical Applications*. Springer, Berlin.

TONG, X., FENG, Y. and ZHAO, A. (2016). A survey on Neyman–Pearson classification and suggestions for future research. *Wiley Interdiscip. Rev.*: *Comput. Stat.* **8** 64–81. MR3465999 https://doi.org/10.1002/wics.1376

TSYBAKOV, A. B. (1997). On nonparametric estimation of density level sets. *Ann. Statist.* **25** 948–969. MR1447735 https://doi.org/10.1214/aos/1069362732

VERSHYNIN, R. (2018). *High-Dimensional Probability*: *An Introduction with Applications in Data Science*. *Cambridge Series in Statistical and Probabilistic Mathematics* **47**. Cambridge Univ. Press, Cambridge. MR3837109 https://doi.org/10.1017/9781108231596

WANG, R., LAGAKOS, S. W., WARE, J. H., HUNTER, D. J. and DRAZEN, J. M. (2007). Statistics in medicine–reporting of subgroup analyses in clinical trials. *N. Engl. J. Med.* **357** 2189–2194. https://doi.org/10.1056/NEJMsr077003

WATSON, J. A. and HOLMES, C. C. (2020). Machine learning analysis plans for randomised controlled trials: Detecting treatment effect heterogeneity with strict control of type I error. *Trials* **21** 1–10.

WILLETT, R. M. and NOWAK, R. D. (2007). Minimax optimal level-set estimation. *IEEE Trans. Image Process.* **16** 2965–2979. MR2472804 https://doi.org/10.1109/TIP.2007.910175

XIA, L., ZHAO, R., WU, Y. and TONG, X. (2021). Intentional control of type I error over unconscious data distortion: A Neyman–Pearson approach to text classification. *J. Amer. Statist. Assoc.* **116** 68–81. MR4227675 https://doi.org/10.1080/01621459.2020.1740711

ZANETTE, A., ZHANG, J. and KOCHENDERFER, M. J. (2018). Robust super-level set estimation using Gaussian processes. In *Joint European Conference on Machine Learning and Knowledge Discovery in Databases* 276–291. Springer, Berlin.

ZHANG, S., LIANG, F., LI, W. and HU, X. (2015). Subgroup analyses in reporting of phase III clinical trials in solid tumors. *J. Clin. Oncol.* **33** 1697–1702.

ZHANG, Z., LI, M., LIN, M., SOON, G., GREENE, T. and SHEN, C. (2017). Subgroup selection in adaptive signature designs of confirmatory clinical trials. *J. R. Stat. Soc. Ser. C. Appl. Stat.* **66** 345–361. MR3611691 https://doi.org/10.1111/rssc.12175