Near-optimal learning with average Hölder smoothness

Steve Hanneke
Department of Computer Science
Purdue University
steve.hanneke@gmail.com

Aryeh Kontorovich
Department of Computer Science
Ben-Gurion University of the Negev
karyeh@cs.bgu.ac.il

Guy Kornowski
Department of Computer Science and Applied Mathematics
Weizmann Institute of Science
guy.kornowski@weizmann.ac.il

Abstract

We generalize the notion of average Lipschitz smoothness proposed by Ashlagi et al. [2021] by extending it to Hölder smoothness. This measure of the "effective smoothness" of a function is sensitive to the underlying distribution and can be dramatically smaller than its classic "worst-case" Hölder constant. We consider both the realizable and the agnostic (noisy) regression settings, proving upper and lower risk bounds in terms of the average Hölder smoothness; these rates improve upon both previously known rates even in the special case of average Lipschitz smoothness. Moreover, our lower bound is tight in the realizable setting up to log factors, thus we establish the minimax rate. From an algorithmic perspective, since our notion of average smoothness is defined with respect to the unknown underlying distribution, the learner does not have an explicit representation of the function class, hence is unable to execute ERM. Nevertheless, we provide distinct learning algorithms that achieve both (nearly) optimal learning rates. Our results hold in any totally bounded metric space, and are stated in terms of its intrinsic geometry. Overall, our results show that the classic worst-case notion of Hölder smoothness can be essentially replaced by its average, yielding considerably sharper guarantees.

1 Introduction

A fundamental theme throughout learning theory and statistics is that "smooth" functions ought to be easier to learn than "rough" ones — an intuition that has been formalized and rigorously established in various frameworks [Györfi et al., 2002, Tsybakov, 2008, Giné and Nickl, 2021]. Hölder continuity is a natural and well-studied notion of smoothness that measures the extent to which nearby points can differ in function value and includes Lipschitz continuity as an important special case.

These global moduli of smoothness, while convenient for theoretical analysis, suffer from the shortcoming of being overly pessimistic. Indeed, being distribution-independent, they fail to distinguish a function that is highly oscillatory everywhere from one that is smooth over most of the probability mass; see Figure 1 for a simple illustration. Moreover, classically studied classes of *average smoothness* (e.g. Besov space) typically fix some distribution in advance (predominantly uniform), and then turn to consider smooth functions with respect to that single distribution. Thus, from a distribution-free statistical learning perspective — where the underlying distribution is assumed to be unknown — such classes fall short.

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

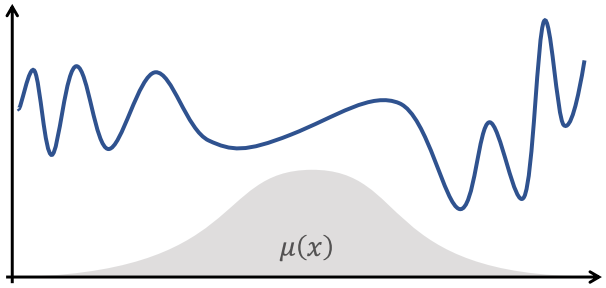

Figure 1: Illustration of a function and a measure μ exhibiting a large gap between "worst-case" smoothness (occurring in low density regions) and average-smoothness with respect to μ.

Seeking to address these drawbacks, Ashlagi et al. [2021] proposed a natural notion of average Lipschitz smoothness with respect to a distribution. Their average Lipschitz modulus can be considerably (even infinitely) smaller than the standard Lipschitz constant, while still being able to control the excess risk. However, the risk bounds obtained by Ashlagi et al. are far from optimal, while the optimal rates for distribution-free learning of average smoothness classes remained unknown. In particular, the cost of adapting to the smoothness with respect to the underlying distribution (in contrast to using classic worst-case smoothness) remained unclear so far.

Our contributions. In this work, we generalize the aforementioned notion of average Lipschitz smoothness by extending it to Hölder smoothness of any exponent $\beta \in (0, 1]$. After formally defining the average Hölder smoothness of a function with respect to a distribution in Section 2.1, our contributions can be summarized as follows:

- **Bracketing numbers upper bound (Theorem 3.1).** We establish a nearly-optimal distribution-free bound on the bracketing entropy of our proposed average-smooth function class, serving as the main crux on which we base our analyses throughout the paper. In particular, although it is known that asymptotically empirical covering numbers yield sharper bounds than bracketing numbers,[1] in the case of average smoothness we reveal that the latter are tight up to a logarithmic factor.
- **Realizable sample complexity (Theorem 3.4).** We derive a nearly-optimal sample complexity required for uniform convergence of average-Hölder functions in the realizable case, which was not previously known even in the special case of average Lipschitz functions.
- **Optimal realizable learning algorithm (Theorem 4.1).** Since our notion of average smoothness is defined with respect to the unknown sampling distribution, the learner does not have an explicit representation of the function class, and hence is unable to execute ERM.[2] We are able to overcome this obstacle by constructing a realizable nonparametric regression algorithm with a nearly-optimal learning rate. Such a rate was not previously known even in the special case of average Lipschitz smoothness.
- **Agnostic learning algorithm (Theorem 5.1).** We provide yet another learning algorithm for the fully agnostic (i.e. noisy) regression setting. Once again, our derived rate was not previously known even in the special case of average Lipschitz smoothness.
- **Matching lower bound (Theorem 6.1).** We prove a lower bound, showing that all the results mentioned above are tight up to logarithmic factors in the realizable case, establishing the (nearly) minimax risk rate for average-smooth classes.
- **Illustrative comparisons (Section 7).** Finally, we illustrate the extent to which the proposed smoothness notion is sharper than previously studied notions. We provide examples in which

[1] Uniform convergence of the $L_1(\mathcal{D})$ distance between the upper and lower bracket functions implies that, in the limit of sample size, the ε-bracket functions are almost surely an empirical $(1 + o(1))\varepsilon$-cover; see Section 2 for a reminder of relevant definitions.

[2] Indeed, the learner cannot know for sure whether any given non-classically Hölder function belongs to the average-Hölder class.

the "optimistic" average-Hölder constant is infinitely apart from both its "pessimistic" worst-case counterpart, or even the average-Lipschitz ($\beta = 1$) constant, exemplifying the substantial (possibly infinite) speed-ups in terms of learning rates.

1.1 Related work.

The sample complexities associated to distribution-free learning of (classic) Hölder classes is well covered in the literature, see for example the books by Györfi et al. [2002], Tsybakov [2008].

Previous notions of average smoothness include Bounded Variation (BV) [Appell et al., 2014] in dimensions one and higher [Kuipers and Niederreiter, 1974, Niederreiter and Talay, 2006]. One-dimensional BV has found learning-theoretic applications [Bartlett et al., 1997, Long, 2004, Anthony and Bartlett, 1999], but to our knowledge the higher-dimensional variants have not. Moreover, the positive results require μ to be uniformly distributed on a segment, and the aforementioned results break down for more general measures — especially if μ is not known to the learner.

Sobolev spaces, and the Sobolev-Slobodetskii norm in particular [Agranovich, 2015], bear some resemblance to our average Hölder smoothness. However, Ashlagi et al. [2021, Appendix I] demonstrate that from a learning-theoretic perspective this notion is inadequate for general (i.e., non-uniform or Lebesgue) measures, as it cannot be used to control sample complexity. Results for controlling bracketing in terms of various measures of average smoothness include Nickl and Pötscher [2007], who bound the bracketing numbers of Besov- and Sobolev-type classes and Malykhin [2010], who used the *averaged modulus of continuity* developed by Sendov and Popov [1988]; again, these are all defined under the Lebesgue measure. While it is easy to define these smoothness notions with respect to arbitrary distributions, we are not aware of any existing work to bound their corresponding sample complexity (or even their covering or bracketing numbers) in a distribution-independent manner. Moreover, the smoothness notion studied in this paper is defined over arbitrary metric spaces, whereas previous notions are typically restricted to Euclidean structures (or variants thereof). Despite of the considerable generality of our setting, we are able to provide tight bounds for all metric spaces alike, without requiring specialized analyses.

A seminal work on recovering functions with spatially inhomogeneous smoothness from noisy samples is Donoho and Johnstone [1998]. Arguably in the spirit of μ-dependent Hölder smoothness, some of the classic results on k-NN risk decay rates were refined by Chaudhuri and Dasgupta [2014] via an analysis that captures the interplay between the metric and the sampling distribution. Another related notion is that of *Probabilistic Lipschitzness* (PL) [Urner and Ben-David, 2013, Urner et al., 2013, Kpotufe et al., 2015], which seeks to relax a hard Lipschitz condition on the regression function. While PL is in the same spirit as our notion, one critical distinction from our work is that, while existing analyses of learning under PL have focused specifically on binary classification, our interest in the present work is learning real-valued functions.

As previously mentioned, the main feature setting this work apart from others studying regression under average smoothness is that our notion is defined with respect to a *general, unknown* measure μ. The notable exception is, of course, Ashlagi et al. [2021] — who introduced the framework of efficiently learning smooth-on-average functions with respect to an unknown distribution. Although extending their definition from Lipschitz to Hölder average smoothness was straightforward, optimal minimax rates are likely inaccessible via their techniques, which relied on empirical covering numbers. Estimating the magnitude of these random objects was a formidable challenge, and Ashlagi et al. were only able to do so to within an additive error decaying with sample size; this sampling noise appears to present an inherent obstruction to optimal rates. Thus, our results required a novel technique to overcome this obstruction, which we did by tightly controlling the bracketing entropy. Our Hölder-type extension is a direct adaptation of the Pointwise Minimum Slope Extension (PMSE) developed for the Lipschitz special case by Ashlagi et al., which in turn is closely related to the one introduced by Oberman [2008].

2 Preliminaries.

Setting. Throughout the paper we consider functions $f : \Omega \to [0, 1]$ where (Ω, ρ) is a metric space. We will consider a distribution $\mathcal{D}$ over $\Omega \times [0, 1]$ with marginal μ over Ω, such that (Ω, ρ, μ) forms a metric probability space (namely, μ is supported on the Borel σ-algebra induced by ρ). We associate

to any measurable function $f : \Omega \to [0,1]$ its L_1 risk $L_{\mathcal{D}}(f) := \mathbb{E}_{(X,Y)\sim\mathcal{D}} |f(X) - Y|$, and its empirical risk $L_S(f) := \frac{1}{n}\sum_{i=1}^n |f(X_i) - Y_i|$ with respect to a sample $S = (X_i, Y_i)_{i=1}^n \sim \mathcal{D}^n$. More generally, we associate to any measurable function its L_1 norm $\|f\|_{L_1(\mu)} := \mathbb{E}_{X\sim\mu} |f(X)|$, and given a sample $(X_1, \ldots, X_n) \sim \mu^n$ we denote its L_1 norm with respect to the empirical measure $\|f\|_{L_1(\mu_n)} := \frac{1}{n}\sum_{i=1}^n |f(X_i)|$.

We say that a distribution $\mathcal{D}$ over $\Omega \times [0,1]$ is *realizable* by a function class $\mathcal{F} \subset [0,1]^\Omega$ if there exists an $f^* \in \mathcal{F}$ such that $L_{\mathcal{D}}(f^*) = 0$. Thus, $f^*(X) = Y$ almost surely, where $(X, Y) \sim \mathcal{D}$.

Metric notions. The diameter of $A \subset \Omega$ is $\mathrm{diam}(A) := \sup_{x,x'\in A} \rho(x, x')$, and we denote by $B(x, r) := \{x' \in \Omega : \rho(x, x') \le r\}$ the closed ball around $x \in \Omega$ of radius $r > 0$. For $t > 0$, $A, B \subset \Omega$, we say that A is a *t-cover* of B if $B \subset \bigcup_{a\in A} B(a, t)$, and define the *$t$-covering number* $\mathcal{N}_B(t)$ to be the minimal cardinality of any t-cover of B. We say that $A \subset B \subset \Omega$ is a *t-packing* of B if $\rho(a, a') \ge t$ for all $a \ne a' \in A$. We call V a *t-net* of B if it is a t-cover and a t-packing. The induced *Voronoi partition* of B with respect to a net V is its partitioning into subsets sharing the same nearest neighbor in V (with ties broken in some consistent arbitrary manner). A metric space (Ω, ρ) is said to be *doubling* if there exists $d \in \mathbb{N}$ such that every r-ball in Ω is contained in the union of some d $r/2$-balls. The *doubling dimension* is defined as $\min_{d\ge 1} \log_2 d$ where the minimum is taken over d satisfying the doubling property.

Bracketing. Given any two functions $l, u : \Omega \to [0,1]$, we say that $f : \Omega \to [0,1]$ belongs to the *bracket* $[l, u]$ if $l \le f \le u$. A set of brackets $\mathcal{B}$ is said to cover a function class $\mathcal{F}$ if any function in $\mathcal{F}$ belongs to some bracket in $\mathcal{B}$. We say that $[l, u]$ is a t-bracket with respect to a norm $\|\cdot\|$ if $\|u - l\| \le t$. The *t-bracketing number* $\mathcal{N}_{[]}(\mathcal{F}, \|\cdot\|, t)$ is defined as the minimal cardinality of any set of t-brackets that covers $\mathcal{F}$. The logarithm of this quantity is called the *bracketing entropy*.

Remark 2.1 (Covering vs. bracketing)**.** *Having recalled two notions that quantify the "size" of a normed function space $(\mathcal{F}, \|\cdot\|)$ — namely, its covering and bracketing numbers — it is useful to note they are related through*

$$\mathcal{N}_{\mathcal{F}}(\varepsilon) \le \mathcal{N}_{[]}(\mathcal{F}, \|\cdot\|, 2\varepsilon) , \tag{1}$$

though no converse inequality of this sort holds in general. On the other hand, the main advantage of using bracketing numbers for generalization bounds is that it suffices to bound the ambient *bracketing numbers with respect to the distribution-specific metric, as opposed to the* empirical covering *numbers which are necessary to guarantee generalization [van der Vaart and Wellner, 1996, Section 2.1.1].*

Strong and weak mean. For any non-negative random variable Z we define its *weak mean* by $\mathbb{W}[Z] := \sup_{t>0} t \Pr[Z \ge t]$, and note that $\mathbb{W}[Z] \le \mathbb{E}[Z]$ by Markov's inequality. In the special case where Z has finite support of size $N \ge 3$ where each atom has mass $1/N$ we have the reverse inequality $\mathbb{E}[Z] \le 2\log(N)\mathbb{W}[Z]$ [Ashlagi et al., 2021, Lemma 22].

2.1 Average smoothness.

For $\beta \in (0, 1]$ and $f : \Omega \to \mathbb{R}$, we define its β-slope at $x \in \Omega$ to be $\Lambda_f^\beta(x) := \sup_{y\in\Omega\setminus\{x\}} \frac{|f(x)-f(y)|}{\rho(x,y)^\beta}$. Recall that f is called β-Hölder continuous if $\|f\|_{\mathrm{Höl}^\beta} := \sup_{x\in\Omega} \Lambda_f^\beta(x) < \infty$, with this quantity serving as its Hölder seminorm. In particular, when $\beta = 1$ these are exactly the Lipschitz functions equipped with the Lipschitz seminorm. For a metric probability space (Ω, ρ, μ), we consider the *average* β-slope to be the mean of $\Lambda_f^\beta(X)$ where $X \sim \mu$. Namely, we define

$$\begin{aligned}
\overline{\Lambda}_f^\beta(\mu) &:= \mathbb{E}_{X\sim\mu}[\Lambda_f^\beta(X)] , \\
\widetilde{\Lambda}_f^\beta(\mu) &:= \mathbb{W}_{X\sim\mu}[\Lambda_f^\beta(X)] = \sup_{t>0}\ t \cdot \mu(x : \Lambda_f^\beta(x) \ge t) .
\end{aligned}$$

Notably,

$$\widetilde{\Lambda}_f^\beta(\mu) \le \overline{\Lambda}_f^\beta(\mu) \le \|f\|_{\mathrm{Höl}^\beta} , \tag{2}$$

where each subsequent pair can be infinitely apart — as we demonstrate in Section 7. Having defined notions of averaged smoothness, we can further define their corresponding function spaces

$$\begin{aligned}
\mathrm{H\ddot{o}l}_L^\beta(\Omega) &:= \{f : \Omega \to [0,1] \,:\, \|f\|_{\mathrm{H\ddot{o}l}^\beta} \le L\}\,, \\
\overline{\mathrm{H\ddot{o}l}}_L^\beta(\Omega,\mu) &:= \left\{f : \Omega \to [0,1] \,:\, \overline{\Lambda}_f^\beta(\mu) \le L\right\}, \\
\widetilde{\mathrm{H\ddot{o}l}}_L^\beta(\Omega,\mu) &:= \left\{f : \Omega \to [0,1] \,:\, \widetilde{\Lambda}_f^\beta(\mu) \le L\right\}.
\end{aligned}$$

We occasionally omit μ when it is clear from context. Note that $\mathrm{H\ddot{o}l}_L^\beta(\Omega) \subset \overline{\mathrm{H\ddot{o}l}}_L^\beta(\Omega,\mu) \subset \widetilde{\mathrm{H\ddot{o}l}}_L^\beta(\Omega,\mu)$ due to Eq. (2), where both containments are strict in general. The special case of $\beta = 1$ recovers the average-Lipschitz spaces $\mathrm{Lip}_L(\Omega) \subset \overline{\mathrm{Lip}}_L(\Omega,\mu) \subset \widetilde{\mathrm{Lip}}_L(\Omega,\mu)$ studied by Ashlagi et al. [2021].

3 Generalization bounds

Our first goal is to bound the bracketing entropy (namely, the logarithm of the bracketing number) of average-Hölder classes. We present this bound in full generality in terms of the underlying metric space, as captured by its covering number (see Corollary 3.5 for the typical scaling of covering numbers). As we will soon establish, this bound implies nearly-tight generalization guarantees in terms of the average smoothness constant.

Theorem 3.1. *For any metric probability space (Ω,ρ,μ), any $\beta \in (0,1]$ and any $0 < \epsilon < L$, it holds*

$$\begin{aligned}
\log \mathcal{N}_{[]}(\overline{\mathrm{H\ddot{o}l}}_L^\beta(\Omega,\mu), L_1(\mu), \varepsilon) &\le \log \mathcal{N}_{[]}(\widetilde{\mathrm{H\ddot{o}l}}_L^\beta(\Omega,\mu), L_1(\mu), \varepsilon) \\
&\le \mathcal{N}_\Omega\left(\left(\frac{\varepsilon}{128L\log(1/\varepsilon)}\right)^{1/\beta}\right) \cdot \log\left(\frac{16\log_2(1/\varepsilon)}{\varepsilon}\right).
\end{aligned}$$

Crucially, the bound above does not depend on μ, allowing us to obtain distribution-free generalization guarantees. We defer the proof of Theorem 3.1 to Appendix B.1. We start by showing that bounding the bracketing entropy implies a generalization bound in the realizable case:

Proposition 3.2. *Suppose (Ω,ρ) is a metric space, $\mathcal{F} \subseteq [0,1]^\Omega$ is a function class, and let $\mathcal{D}$ be a distribution over $\Omega \times [0,1]$ which is realizable by $\mathcal{F}$, with marginal μ over Ω. Then with probability at least $1-\delta$ over drawing a sample $S \sim \mathcal{D}^n$ it holds that for all $f \in \mathcal{F}$:*

$$L_{\mathcal{D}}(f) \le 1.01 L_S(f) + \inf_{\alpha \ge 0}\left(\alpha + \frac{205\log \mathcal{N}_{[]}(\mathcal{F}, L_1(\mu), \alpha)}{n}\right) + \frac{205\log(1/\delta)}{n}.$$

Remark 3.3 (Constant is arbitrary)**.** *In Proposition 3.2 and in what follows, the constant multiplying $L_S(f)$ is arbitrary, and can be replaced by $(1+\gamma)$ for any $\gamma > 0$ at the expense of multiplying the remaining summands by γ^{-1}. In the next section we will provide a realizable regression algorithm that returns an approximate empirical risk minimizer f for which $L_S(f) \approx 0$, thus this constant will not matter for our purposes.*

We prove Proposition 3.2 in Appendix B.2. By combining Theorem 3.1 with Proposition 3.2 and setting $\alpha = \varepsilon/2$, we obtain the following realizable sample complexity result.

Theorem 3.4. *For any metric space (Ω,ρ), any $\beta \in (0,1]$ and any $0 < \epsilon < L$, let $\mathcal{D}$ be a distribution over $\Omega \times [0,1]$ realizable by $\widetilde{\mathrm{H\ddot{o}l}}_L^\beta(\Omega,\mu)$. Then there exists $N = N(\beta,\varepsilon,\delta) \in \mathbb{N}$ satisfying*

$$N = \widetilde{O}\left(\frac{\mathcal{N}_\Omega\left(\left(\frac{\varepsilon}{256L\log(1/\varepsilon)}\right)^{1/\beta}\right) + \log(1/\delta)}{\varepsilon}\right)$$

such that as long as $n \ge N$, with probability at least $1-\delta$ over drawing a sample $S \sim \mathcal{D}^n$ it holds that for all $f \in \widetilde{\mathrm{H\ddot{o}l}}_L^\beta(\Omega,\mu)$:

$$L_{\mathcal{D}}(f) \le 1.01 L_S(f) + \varepsilon\,.$$

The same claim holds for the smaller class $\overline{\mathrm{H\ddot{o}l}}_L^\beta(\Omega,\mu)$.

Corollary 3.5 (Doubling metrics)**.** *In most cases of interest, (Ω, ρ) is a doubling metric space of some dimension d,*[3] *e.g. when Ω is a subset of $\mathbb{R}^d$ (or more generally a d-dimensional Banach space). For d-dimensional doubling spaces of finite diameter we have $\mathcal{N}_\Omega(\varepsilon) \lesssim \left(\frac{1}{\varepsilon}\right)^d$ [Gottlieb et al., 2016, Lemma 2.1], which, plugged into Theorem 3.4, yields the simplified sample complexity bound*

$$N = \widetilde{O}\left(\frac{L^{d/\beta}}{\varepsilon^{(d+\beta)/\beta}}\right),$$

or equivalently

$$L_{\mathcal{D}}(f) \leq 1.01 L_S(f) + \widetilde{O}\left(\frac{L^{d/(d+\beta)}}{n^{\beta/(d+\beta)}}\right),$$

up to a constant which depends (exponentially) on d, but is independent of L, n.

Remark 3.6 (Tightness)**.** *The bounds in Theorem 3.1 and Theorem 3.4 are both tight up to logarithmic factors, as we will prove in Section 6.*

4 Realizable learning algorithm

Recall that without knowing μ, the underlying distribution over Ω, we cannot know for sure whether a function f belongs to $\overline{\mathrm{H\ddot{o}l}}_L^\beta(\Omega, \mu)$ (except for the trivial case $f \in \mathrm{H\ddot{o}l}_L^\beta(\Omega)$). This gives rise to the challenge of designing a fully empirical algorithm — since standard empirical risk minimization is not possible. To that end, we provide the following algorithmic result with optimal guarantees (up to logarithmic factors).

Theorem 4.1. *For any metric space (Ω, ρ), any $\beta \in (0, 1]$ and any $0 < \epsilon < L$, let $\mathcal{D}$ be a distribution over $\Omega \times [0, 1]$ realizable by $\overline{\mathrm{H\ddot{o}l}}_L^\beta(\Omega, \mu)$. Then there exists a polynomial time learning algorithm $\mathcal{A}$, which, given a sample $S \sim \mathcal{D}^n$ of size $n \geq N$ for some $N = N(\beta, \varepsilon, \delta) \in \mathbb{N}$ satisfying*

$$N = \widetilde{O}\left(\frac{\mathcal{N}_\Omega\left(\left(\frac{\varepsilon}{256 L \log(1/\varepsilon)}\right)^{1/\beta}\right) + \log(1/\delta)}{\varepsilon}\right),$$

constructs a hypothesis $f = \mathcal{A}(S)$ such that $L_{\mathcal{D}}(f) \leq \varepsilon$ with probability at least $1 - \delta$.

Remark 4.2 (Doubling metrics)**.** *As mentioned in Corollary 3.5, in most cases of interest we have $\mathcal{N}_\Omega(\varepsilon) \lesssim \left(\frac{1}{\varepsilon}\right)^d$. In that case, the algorithm above has sample complexity*

$$N = \widetilde{O}\left(\frac{L^{d/\beta}}{\varepsilon^{(d+\beta)/\beta}}\right),$$

or equivalently

$$L_{\mathcal{D}}(f) = \widetilde{O}\left(\frac{L^{d/(d+\beta)}}{n^{\beta/(d+\beta)}}\right),$$

up to a constant which depends (exponentially) on d, but is independent of L, n.

Remark 4.3 (Computational complexity)**.** *The algorithm constructed in Theorem 4.1 involves a one-time preprocessing step after which $f(x)$ can be evaluated at any given $x \in \Omega$ in $O(n^2)$ time. We note that the computation at inference time matches that of (classic) Lipschitz/Hölder regression (e.g. Gottlieb et al., 2017). Furthermore, the computational complexity of the preprocessing step is similar to that in Ashlagi et al. [2021, Theorem 7] for the average Lipschitz case, where it is shown to run in time $\widetilde{O}(n^2)$. The complexity analysis of our prepossessing step is entirely analogous to theirs, and we forgo repeating it here.*

We will now outline the proof of Theorem 4.1, which appears in Appendix B.3. The key idea is to analyze a natural fully-empirical quantity that will serve as an estimator of the true unknown average

[3] Namely, any ball of radius $r > 0$ can be covered by 2^d balls of radius $r/2$.

smoothness. To that end, given a sample $S = (X_i, Y_i)_{i=1}^n \sim \mathcal{D}^n$ and a function $f : \Omega \to [0,1]$, consider the following quantity which can be established directly from the data:

$$\widehat{\Lambda}_f^\beta := \frac{1}{n}\sum_{i=1}^n \sup_{X_j \neq X_i} \frac{|f(X_i) - f(X_j)|}{\rho(X_i, X_j)^\beta} .$$

Namely, this is the empirical average smoothness with respect to the sampled points. It would suit us well if the empirical average smoothness of a function did not greatly exceed its true average smoothness, with high probability. The fact something like this turns out to be true is somewhat surprising and may be of independent interest:

Proposition B.1. *(Informal) Let $f^* : \Omega \to [0,1]$. Then with high probability $\widehat{\Lambda}_{f^*}^\beta \lesssim \overline{\Lambda}_{f^*}^\beta$.*

The proposition above implies that restricting to the sample, and letting $\widehat{f}(X_i) := Y_i$ yields a function over $\{X_1, \ldots, X_n\}$ which is empirically average-smooth over the sample (with high probability). We then turn to show that any such function can be approximately extended to the whole space, in a way that guarantees its average smoothness with respect to the *underlying distribution*.

Proposition B.3. *(Informal) Let $\widehat{f} : \{X_1, \ldots, X_n\} \to [0,1]$ where $(X_i)_{i=1}^n \sim \mu^n$. Then it is possible to construct $f : \Omega \to [0,1]$ such that with high probability $f(X_i) \approx \widehat{f}(X_i)$ for all $i \in [N]$, and $\overline{\Lambda}_f^\beta(\mu) \lesssim \widehat{\Lambda}_{\widehat{f}}^\beta$.*

We will now sketch the procedure described in Proposition B.3, which serves as the main challenge in proving Theorem 4.1. Roughly speaking, the algorithm sorts the sampled points with respect to their relative slope to one another. Then, it discards a fraction of the sampled points with largest relative slope, which can be thought of as "outliers". Then, the algorithm proceeds to extend the function in a smooth fashion among the remaining "well-behaved" samples. A careful probabilistic analysis shows that disregarding just the right amount of samples induces small error, while being average-smooth with high probability.

Overall this procedure yields a function $f : \Omega \to [0,1]$ which is an approximate empirical-minimizer (since $f(X_i) \approx \widehat{f}(X_i) = Y_i$), while guaranteed to be averagely-smooth with respect to the unknown distribution. Thus we can apply the uniform convergence of Theorem 3.4, proving Theorem 4.1.

5 Agnostic learning algorithm

Noticeably, up to this point, both the uniform convergence result we derived (Theorem 3.4) as well as the algorithmic result (Theorem 4.1) are tailored for the realizable regression setting. Inspired by a recent result of Hopkins et al. [2022] that showed a reduction from agnostic learning to realizable learning, we provide an algorithm for agnostic (i.e. noisy) regression of average-smooth functions. It is worth noting that the following algorithm does not require any prior assumption on the noise model, unlike many nonparametric regression methods, due to our distribution free analysis.

Theorem 5.1. *There exists a learning algorithm $\mathcal{A}$ such that for any metric space (Ω, ρ), any $\beta \in (0,1]$, $0 < \epsilon < L$, and any distribution $\mathcal{D}$ over $\Omega \times [0,1]$, given a sample $S \sim \mathcal{D}^n$ of size $n \geq N$ for some $N = N(\beta, \varepsilon, \delta)$ satisfying*

$$N = \widetilde{O}\left(\frac{\mathcal{N}_\Omega\left(\left(\frac{\varepsilon}{640L\log(1/\varepsilon)}\right)^{1/\beta}\right) + \log(1/\delta)}{\varepsilon^2}\right),$$

the algorithm constructs a hypothesis $f = \mathcal{A}(S)$ such that $L_{\mathcal{D}}(f) \leq \inf_{f^* \in \overline{\mathrm{H\ddot{o}l}}_L^\beta(\Omega,\mu)} L_{\mathcal{D}}(f^*) + \epsilon$ with probability at least $1 - \delta$.

Remark 5.2 (Doubling metrics)**.** *As mentioned in Corollary 3.5, in most cases of interest we have $\mathcal{N}_\Omega(\varepsilon) \lesssim \left(\frac{1}{\varepsilon}\right)^d$. In that case, the algorithm above has sample complexity*

$$N = \widetilde{O}\left(\frac{L^{d/\beta}}{\varepsilon^{(d+2\beta)/\beta}}\right),$$

or equivalently

$$L_{\mathcal{D}}(f) = \inf_{f^* \in \overline{\mathrm{H\ddot{o}l}}_L^{\beta}(\Omega,\mu)} L_{\mathcal{D}}(f^*) + \widetilde{O}\left(\frac{L^{d/(d+2\beta)}}{n^{\beta/(d+2\beta)}}\right) ,$$

up to a constant which depends (exponentially) on d, but is independent of L, n.

Though our agnostic algorithm is similar in spirit to that obtained by the reduction of Hopkins et al. [2022], our analysis is self-contained and crucially relies on the bracketing bound given by Theorem 3.1, as well as analyzing the empirical smoothness estimator as provided by Proposition B.1. We also note that unlike our algorithm for realizable learning, the agnostic algorithm is not computationally efficient. This seems to be inherent for such reductions, and we do not know whether this blow-up in running time can be avoided or not.

We will now describe the proof of Theorem 5.1 which appears in Appendix B.4. Given a sample S of size n, consider dividing it into two sub-samples $S_1 \cup S_2 = S$ of size $n/2$ each. We first use S_1 in order to construct an empirical ϵ-net $h_1, \ldots, h_N : S_1 \to [0,1]$, namely a set of functions which are sufficiently empirically smooth over the sample, yet far away enough from one another when averaged over the sample. Recalling that bracketing numbers upper bound covering numbers (Eq. (1)), and since Theorem 3.1 holds true for every measure (in particular for the empirical measure), we can bound $\log N \lesssim \mathcal{N}_{\Omega}((\epsilon/L)^{1/\beta})$. Moreover, using Proposition B.1 we know that $f^* := \arg\min_{f \in \overline{\mathrm{H\ddot{o}l}}_L^{\beta}(\Omega,\mu)} L_{\mathcal{D}}(f)$ is likely to be $\widetilde{O}(L)$ average-smooth over S_1, so there must exist some h_j with ϵ excess empirical loss (since f^* is in the class we are ϵ-covering). Thus running the realizable algorithm of Theorem 4.1 over all $\{h_1, \ldots, h_N\}$, producing $f_1, \ldots, f_N : \Omega \to [0,1]$, yields at least one function which has both small excess empirical error, while being smooth with respect to the underlying distribution. Finally, running ERM over $\{f_1, \ldots, f_N\}$ with respect to the fresh sample S_2 reveals such a good candidate function within $\frac{\log(N)+\log(1/\delta)}{\epsilon^2}$ samples by applying standard uniform convergence for finite classes (i.e. Hoeffding's inequality with the union bound).

6 Lower bound

We now turn to show that the bounds proved in Theorem 3.1, Theorem 3.4 and Theorem 4.1 are all tight up to logarithmic factors. In fact, since the bracketing entropy bound of Theorem 3.1 implies the generalization bound of Theorem 3.4 and the latter implies the sample complexity in Theorem 4.1, it is enough to show that the latter is nearly optimal.

Theorem 6.1. *For any $\beta \in (0,1]$, $\varepsilon \in (0,1)$ any metric space (Ω, ρ) and $L \geq \frac{8}{\mathrm{diam}(\Omega)}$, there exists a distribution $\mathcal{D}$ over $\Omega \times [0,1]$ which is realizable by $\overline{\mathrm{H\ddot{o}l}}_L^{\beta}(\Omega)$ such that any learning algorithm that produces $f = A(S)$ with $L_{\mathcal{D}}(f) \leq \varepsilon$ with constant probability, must have sample complexity*

$$n = \Omega\left(\frac{\mathcal{N}_{\Omega}((\varepsilon/L)^{1/\beta})}{\varepsilon}\right) .$$

Remark 6.2 (Typical case)**.** *In most cases of interest it holds that $\mathcal{N}_{\Omega}(\varepsilon) \gtrsim \left(\frac{1}{\varepsilon}\right)^d$ for some constant d, e.g. when Ω is a subset of non-empty interior in $\mathbb{R}^d$ (or more generally in any d-dimensional Banach space).*[4] *That being the case, Theorem 6.1 yields the simplified sample complexity lower bound of*

$$n = \Omega\left(\frac{L^{d/\beta}}{\varepsilon^{(d+\beta)/\beta}}\right)$$

Equivalently, we obtain an excess risk lower bound of

$$L_{\mathcal{D}}(f) = \Omega\left(\frac{L^{d/(d+\beta)}}{n^{\beta/(d+\beta)}}\right) .$$

We will now provide a proof sketch of Theorem 6.1, while the full proof appears in Appendix B.5. Suppose $K \subset \Omega$ is a $(\varepsilon/L)^{1/\beta}$-net of *most* of Ω, yet $x_0 \in \Omega$ is some "isolated" point at constant

[4] Note that assuming a subset has nonempty interior implies that it cannot be isometrically embedded to a lower dimensional space. Hence, this d encapsulates the "true" intrinsic metric dimension.

distance away from K (we show that such x_0, K always exist). Let μ be the measure that assigns $1-\varepsilon$ probability mass to x_0, while the rest of the probability mass is distributed uniformly over K. Now consider a (random) function that independently assigns either 0 or 1 to each point in K uniformly, and is constant over x_0. Since points in K are $(\varepsilon/L)^{1/\beta}$ away from one another, the local β-slope at each point in K is roughly $1/((\varepsilon/L)^{1/\beta})^\beta = L/\varepsilon$, while the slope at x_0 is small since it is far enough from other points. Averaging over the space with respect to μ, we see that the function is $\mu(K) \cdot L/\varepsilon = L$ average-Hölder. Now, we imitate the standard lower bound proof for VC classes over K: Since any point in K is sampled with probability $\varepsilon/|K|$, any learning algorithm with much fewer than $|K|/\varepsilon \approx \mathcal{N}_\Omega((\varepsilon/L)^{1/\beta})/\varepsilon$ examples will guess wrong a large portion of K, suffering L_1-loss of at least order of $\mu(K) = \varepsilon$.

7 Illustrative examples

Having established the control that average-Hölder smoothness has on generalization, we illustrate the vast possible gap between the average smoothness and it's "worst-case" classic counterpart. Indeed, in the examples we provide, the gap is infinite. Moreover, we also show that classes of average-Hölder smoothness are significantly richer than the previously studied average-Lipschitz, motivating the more general Hölder framework considered in this work. Finally, it is illuminating to notice that both claims to follow actually consist of the same simple function $f(x) = \mathbf{1}[x > \frac{1}{2}]$ though with respect to different distributions, emphasizing the crucial role of the underlying distribution in terms of establishing the function classes.

Claim 7.1. *For any $L > 0, \beta \in (0,1)$, there exist $f : \Omega \to [0,1]$ and a probability measure μ such that*

- *f is average-Hölder: $f \in \overline{\mathrm{Höl}}^\beta_L(\Omega, \mu)$.*
- *f is not Hölder with any finite Hölder constant: For all $M > 0$: $f \notin \mathrm{Höl}^\beta_M(\Omega)$.*
- *f is not (even) weakly-average-Lipschitz with any finite modulus: For all $M > 0$: $f \notin \widetilde{\mathrm{Lip}}_M(\Omega, \mu)$.*

Thus, $\overline{\mathrm{Höl}}^\beta_L(\Omega, \mu) \not\subset \bigcup_{M=0}^\infty \left(\mathrm{Höl}^\beta_M(\Omega) \cup \widetilde{\mathrm{Lip}}_M(\Omega, \mu)\right)$.

Claim 7.2. *For any $L > 0, \beta \in (0,1)$, there exist $f : \Omega \to [0,1]$ and a probability measure μ such that*

- *f is weakly-average-Hölder: $f \in \widetilde{\mathrm{Höl}}^\beta_L(\Omega, \mu)$.*
- *f is not strongly-average-Hölder with any finite modulus: For all $M > 0$: $f \notin \overline{\mathrm{Höl}}^\beta_M(\Omega)$.*
- *f is not (even) weakly-average-Lipschitz with any finite modulus: For all $M > 0$: $f \notin \widetilde{\mathrm{Lip}}_M(\Omega, \mu)$.*

Thus, $\widetilde{\mathrm{Höl}}^\beta_L(\Omega, \mu) \not\subset \bigcup_{M=0}^\infty \left(\overline{\mathrm{Höl}}^\beta_M(\Omega, \mu) \cup \widetilde{\mathrm{Lip}}_M(\Omega, \mu)\right)$.

We prove both of the claims above in Appendix B.6.

8 Discussion

In this work, we have defined a notion of an average-Hölder smoothness, extending the average-Lipschitz one introduced by Ashlagi et al. [2021]. Using proof techniques based on bracketing numbers, we have established the minimax rate for average-smoothness classes in the realizable setting with respect to the L_1 risk up to logarithmic factors, and have provided a nontrivial learning algorithm that attains this nearly-optimal learning rate. Moreover, we have also provided yet another learning algorithm for the agnostic setting. All of these results improve upon previously known rates even in the special case of average-Lipschitz classes.

A few notes are in order. First, the choice of focusing on L_1 risk as opposed to general L_p losses is merely a matter of conciseness, as to avoid introducing additional parameters. Indeed, the only place throughout the proofs which we use the L_1 loss is in the proof of Proposition 3.2, where we show that the loss-class $\mathcal{L}_{\mathcal{F}} := \{x \mapsto |f(x) - f^*(x)| : f \in \mathcal{F}\}$ satisfies

$$\mathcal{N}_{[]}(\mathcal{L}_{\mathcal{F}}, L_1(\mu), \alpha) \leq \mathcal{N}_{[]}(\mathcal{F}, L_1(\mu), \alpha) \ .$$

It is easy to show via essentially the same proof that for any $p \in [1, \infty)$, the L_p-composed loss-class satisfies $\mathcal{N}_{[]}(\mathcal{L}_{\mathcal{F}}, L_1(\mu), \alpha) \leq \mathcal{N}_{[]}(\mathcal{F}, L_1(\mu), \alpha^{1/p})$, and the remaining proofs can be invoked verbatim. This yields a realizable sample complexity (in the typical, d-dimensional case) of order $N = \widetilde{O}\left(\frac{L^{d/p\beta}}{\varepsilon^{(d+p\beta)/p\beta}}\right)$, or equivalently L_p-risk decay rate of $L_{\mathcal{D}}(f) = \widetilde{O}\left(\frac{L^{d/(d+p\beta)}}{n^{p\beta/(d+p\beta)}}\right)$ which are also easily translatable to their corresponding agnostic rates.

Focusing again on L_1 minimax rates of average-Hölder classes, it is interesting to compare them to the minimax rates of "classic" (i.e., worst-case) Hölder classes. Schreuder [2020] has shown the minimax risk to be of order $n^{-\beta/d}$, whereas we showed the average-smooth case has the slightly worse rate of $n^{-\beta/(d+\beta)}$ (which cannot be improved, due to our matching lower bound). However, comparing the rates alone is rather misleading, since both risks are multiplied by a factor depending on their corresponding Hölder constant, which can be considerably smaller in the average-case result. Still, it is interesting to note that in the asymptotic regime there is a marginal advantage in case the learned function is worst-case Hölder, as opposed to Hölder on average.

Our work leaves open several questions. A relatively straightforward one is to compute the minimax rates and construct an optimal algorithm for the *classification* setting, which is not addressed by this paper. Moreover, there is a slight mismatch between our established upper and lower bounds in the agnostic setting, ranging between $\widetilde{O}(n^{-\beta/(d+2\beta)})$ and $\Omega(n^{-\beta/(d+\beta)})$. Closing this gap is an interesting problem which we leave for future work.

References

Mikhail S. Agranovich. *Sobolev spaces, their generalizations and elliptic problems in smooth and Lipschitz domains*. Springer Monographs in Mathematics. Springer, Cham, 2015. ISBN 978-3-319-14647-8; 978-3-319-14648-5. doi: 10.1007/978-3-319-14648-5. URL https://doi.org/10.1007/978-3-319-14648-5. Revised translation of the 2013 Russian original.

Martin Anthony and Peter L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, Cambridge, 1999. ISBN 0-521-57353-X. doi: 10.1017/CBO9780511624216. URL http://dx.doi.org/10.1017/CBO9780511624216.

Jürgen Appell, Józef Banaś, and Nelson Merentes. *Bounded variation and around*, volume 17 of *De Gruyter Series in Nonlinear Analysis and Applications*. De Gruyter, Berlin, 2014. ISBN 978-3-11-026507-1; 978-3-11-026511-8.

Yair Ashlagi, Lee-Ad Gottlieb, and Aryeh Kontorovich. Functions with average smoothness: structure, algorithms, and learning. In *Conference on Learning Theory*, pages 186–236. PMLR, 2021.

Peter L. Bartlett, Sanjeev R. Kulkarni, and S. E. Posner. Covering numbers for real-valued function classes. *IEEE Trans. Information Theory*, 43(5):1721–1724, 1997. doi: 10.1109/18.623181. URL https://doi.org/10.1109/18.623181.

Kamalika Chaudhuri and Sanjoy Dasgupta. Rates of convergence for nearest neighbor classification. In *NIPS*, 2014.

David L. Donoho and Iain M. Johnstone. Minimax estimation via wavelet shrinkage. *Ann. Statist.*, 26 (3):879–921, 1998. ISSN 0090-5364. doi: 10.1214/aos/1024691081. URL https://doi.org/10.1214/aos/1024691081.

Evarist Giné and Richard Nickl. *Mathematical foundations of infinite-dimensional statistical models*. Cambridge university press, 2021.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Adaptive metric dimensionality reduction. *Theoretical Computer Science*, 620:105–118, 2016.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Efficient regression in metric spaces via approximate lipschitz extension. *IEEE Transactions on Information Theory*, 63(8):4838–4849, 2017. doi: 10.1109/TIT.2017.2713820.

László Györfi, Michael Kohler, Adam Krzyżak, and Harro Walk. *A distribution-free theory of nonparametric regression*. Springer Series in Statistics. Springer-Verlag, New York, 2002. ISBN 0-387-95441-4. doi: 10.1007/b97848. URL http://dx.doi.org/10.1007/b97848.

Max Hopkins, Daniel M Kane, Shachar Lovett, and Gaurav Mahajan. Realizable learning is all you need. In *Conference on Learning Theory*, pages 3015–3069. PMLR, 2022.

Michael J Kearns and Umesh Vazirani. *An introduction to computational learning theory*. MIT press, 1994.

Samory Kpotufe, Ruth Urner, and Shai Ben-David. Hierarchical label queries with data-dependent partitions. In Peter Grünwald, Elad Hazan, and Satyen Kale, editors, *Proceedings of The 28th Conference on Learning Theory, COLT 2015, Paris, France, July 3-6, 2015*, volume 40 of *JMLR Workshop and Conference Proceedings*, pages 1176–1189. JMLR.org, 2015. URL http://proceedings.mlr.press/v40/Kpotufe15.html.

L. Kuipers and H. Niederreiter. *Uniform distribution of sequences*. Wiley-Interscience [John Wiley & Sons], New York-London-Sydney, 1974. Pure and Applied Mathematics.

Philip M. Long. Efficient algorithms for learning functions with bounded variation. *Inf. Comput.*, 188 (1):99–115, 2004. doi: 10.1016/S0890-5401(03)00164-0. URL https://doi.org/10.1016/S0890-5401(03)00164-0.

Yu. V. Malykhin. Averaged modulus of continuity and bracket compactness. *Mat. Zametki*, 87(3): 468–471, 2010. ISSN 0025-567X. doi: 10.1134/S0001434610030181. URL https://doi.org/10.1134/S0001434610030181.

Richard Nickl and Benedikt M. Pötscher. Bracketing metric entropy rates and empirical central limit theorems for function classes of Besov- and Sobolev-type. *J. Theoret. Probab.*, 20(2):177–199, 2007. ISSN 0894-9840. doi: 10.1007/s10959-007-0058-1. URL https://doi.org/10.1007/s10959-007-0058-1.

Harald Niederreiter and Denis Talay, editors. *Monte Carlo and quasi-Monte Carlo methods 2004*, 2006. Springer-Verlag, Berlin. ISBN 978-3-540-25541-3; 3-540-25541-9. doi: 10.1007/3-540-31186-6. URL https://doi.org/10.1007/3-540-31186-6.

Adam M. Oberman. An explicit solution of the Lipschitz extension problem. *Proc. Amer. Math. Soc.*, 136(12):4329–4338, 2008. ISSN 0002-9939. doi: 10.1090/S0002-9939-08-09457-4. URL https://doi.org/10.1090/S0002-9939-08-09457-4.

Nicolas Schreuder. Bounding the expectation of the supremum of empirical processes indexed by hölder classes. *Mathematical Methods of Statistics*, 29(1):76–86, 2020.

Blagovest Sendov and Vasil A. Popov. *The averaged moduli of smoothness*. Pure and Applied Mathematics (New York). John Wiley & Sons, Ltd., Chichester, 1988. ISBN 0-471-91952-7. Applications in numerical methods and approximation, A Wiley-Interscience Publication.

Alexandre B. Tsybakov. *Introduction to Nonparametric Estimation*. Springer Publishing Company, Incorporated, 1st edition, 2008. ISBN 0387790519.

Ruth Urner and Shai Ben-David. Probabilistic Lipschitzness: A niceness assumption for deterministic labels. In *Learning Faster from Easy Data WorkshopNIPS*, 2013.

Ruth Urner, Sharon Wullf, and Shai Ben-David. PLAL: Cluster-based active learning. In *Conference on Learning Theory*, 2013.

Aad W. van der Vaart and Jon A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

Roman Vershynin. *High-dimensional probability: An introduction with applications in data science*, volume 47. Cambridge university press, 2018.

A Minimal β-slope Hölder extension

In this section we describe a procedure that extends Hölder functions in an optimally smoothest manner at every point, as it will serve as a crucial ingredient in our proofs. That is, given a subset of a metric space $A \subset \Omega$ and a function $f : \Omega \to [0,1]$, it produces $F_A : \Omega \to [0,1]$ such that

1. It extends $f|_A$: $F_A|_A = f|_A$.
2. For any $\widetilde{F} : \Omega \to [0,1]$ that extends $f|_A$, it holds that $\Lambda^\beta_{F_A}(x) \le \Lambda^\beta_{\widetilde{F}}(x)$ for all $x \in \Omega$.

Such a procedure was described for Lipschitz extensions (namely when $\beta = 1$) in Ashlagi et al. [2021]. The purpose of this section is to generalize this procedure to any Hölder exponent.

Throughout this section we fix $\beta \in (0,1]$, $\emptyset \neq A \subset \Omega$ and $f : \Omega \to [0,1]$, and will always assume the following.

Assumption A.1. $\|f|_A\|_{\mathrm{H\ddot{o}l}^\beta} < \infty$ *and* $\mathrm{diam}(A) < \infty$.

Keeping in mind that the case we are really interested in is when A is finite (i.e. a sample), the conditions above are trivially satisfied. Nonetheless, everything we will present continues to hold in this more general setting. For $u, v \in A$ we introduce the following notation:

$$\begin{aligned} R_x(u,v) &:= \frac{f(v) - f(u)}{\rho(x,v)^\beta + \rho(x,u)^\beta} \,, \\ F_x(u,v) &:= f(u) + R_x(u,v)\rho(x,u)^\beta \,, \\ R^*_x &:= \sup_{u,v \in A} R_x(u,v) \,, \\ W_x(\varepsilon) &:= \{(u,v) \in A \times A : R_x(u,v) > R^*_x - \varepsilon\} \,, \quad 0 < \varepsilon < R^*_x \\ \Phi_x(\varepsilon) &:= \{F_x(u,v) : (u,v) \in W_x(\varepsilon)\} \,. \end{aligned}$$

Definition A.2. *We define the β-pointwise minimal slope extension (β-PMSE) to be the function $F_A : \Omega \to \mathbb{R}$ satisfying*

$$F_A(x) := \lim_{\varepsilon \to 0^+} \Phi_x(\varepsilon) \,.$$

In the degenerate case in which $f(u) = f(v)$ for all $u, v \in A$, define $F_A(x) := f(u)$ for some (and hence any) $u \in A$.

Theorem A.3. *Let $\emptyset \neq A \subset \Omega$, $f : \Omega \to [0,1]$, such that Assumption A.1 holds. Then $F_A : \Omega \to [0,1]$ is well defined, and satisfies for any $x \in \Omega$: $\Lambda^\beta_{F_A}(x) \le \Lambda^\beta_f(x)$. Furthermore, if A is finite, then $F_A(x)$ can be computed for any $x \in \Omega$ within $O(|A|^2)$ arithmetic operations.*

Remark A.4. *When $R_x(\cdot,\cdot)$ has a unique maximizer $(u^*_x, v^*_x) \in A \times A$, the definition of F_A simplifies to*

$$F_A(x) = f(u^*_x) + \frac{\rho(x,u^*_x)^\beta}{\rho(x,u^*_x)^\beta + \rho(x,v^*_x)^\beta}(f(v^*_x) - f(u^*_x)) \,. \tag{3}$$

We conclude that under Assumption A.1, we can assume without loss of generality that for each $x \in \Omega$ there is such a unique maximizer (since the function is well defined, thus does not depend on the choice of the maximizer). Furthermore, this readily shows that when A is finite, we can compute $F_A(x)$ for any $x \in \Omega$ within $O(|A|^2)$ arithmetic operations — simply by finding this maximizer.

Proof. (of Theorem A.3)

We will assume that there exist $u, v \in A$ such that $f(u) \neq f(v)$, since the degenerate (constant extension) case is trivial to verify. This assumption implies that $R^*_x > 0$. It is also easy to verify that $\sup_{x \in \Omega} R^*_x < \infty \iff \|f\|_{\mathrm{H\ddot{o}l}^\beta} < \infty$.

Lemma A.5. *F_A is well defined. Namely, under Assumption A.1 the limit $\lim_{\varepsilon \to 0^+} \Phi_x(\varepsilon) \in [0,1]$ exists.*

Proof. Fix $x \in \Omega$ (we will omit the x subscripts from now on). Let $\varepsilon < R^*/2$, $(u,v), (u',v') \in W(\varepsilon)$. Note that $R(u,v) > 0$ and that $F(u,v) = f(v) - R(u,v)\rho(x,v)^\beta$. Hence

$$f(u) \le F(u,v) \le f(v) \,, \tag{4}$$

and the same clearly holds if we replace (u,v) by (u',v'). Assume without loss of generality that $F(u,v) \le F(u',v')$, hence $f(u) \le F(u,v) \le F(u',v') \le f(v')$. We get

$$\begin{aligned}
R(u',v') + \varepsilon &> R^* \\
&\ge \frac{f(v') - f(u)}{\rho(x,v')^\beta + \rho(x,u)^\beta} \\
&= \frac{f(v') - F(u',v') + F(u,v) - f(u)}{\rho(x,v')^\beta + \rho(x,u)^\beta} + \frac{F(u',v') - F(u,v)}{\rho(x,v')^\beta + \rho(x,u)^\beta} \\
&= \frac{R(u',v')\rho(x,v')^\beta + R(u,v)\rho(x,u)^\beta}{\rho(x,v')^\beta + \rho(x,u)^\beta} + \frac{F(u',v') - F(u,v)}{\rho(x,v')^\beta + \rho(x,u)^\beta} \\
&\ge \frac{R(u',v')\rho(x,v')^\beta + (R(u',v') - \varepsilon)\rho(x,u)^\beta}{\rho(x,v')^\beta + \rho(x,u)^\beta} + \frac{F(u',v') - F(u,v)}{2\mathrm{diam}(A)^\beta} \\
&\ge R(u',v') - \varepsilon + \frac{F(u',v') - F(u,v)}{2\mathrm{diam}(A)^\beta} \\
&\implies |F_x(u,v) - F_x(u',v')| \le 4\varepsilon\,\mathrm{diam}(A)^\beta\,.
\end{aligned}$$

We conclude that if $\mathrm{diam}(A) < \infty$ then $\lim_{\varepsilon \to 0^+} \Phi_x(\varepsilon)$ indeed exists.

□

It remains to prove the optimality of the β-slope. Throughout the proof we will denote for any $u \ne v \in \Omega$:

$$S(u,v) := \frac{|F_A(u) - F_A(v)|}{\rho(u,v)^\beta}\,,$$

and for any point $x \in \Omega$, subset $B \subset \Omega$ and function $g : \Omega \to [0,1]$ we let

$$\Lambda_g^\beta(x,B) := \sup_{y \in B \setminus \{x\}} \frac{|g(x) - g(y)|}{\rho(x,y)^\beta}\,.$$

The proof is split into three claims.

Claim I. $\forall x \in \Omega \setminus A :\ \Lambda_{F_A}^\beta(x,A) \le \Lambda_f^\beta(x,A)$.

Let $x \in \Omega \setminus A$, and let $(u^*,v^*) \in A \times A$ be its associated maximizer of R_x. Recall Eq. (4) from which we can deduce that $F_A(u^*) \le F_A(x) \le F_A(v^*)$. Also note that a simple rearrangement based on Eq. (3) (and the fact that f and F_A agree on A) shows that $S(u^*,x) = R_x(u^*,v^*) = S(x,v^*)$. Furthermore, we claim that $\Lambda_{F_A}^\beta(x,A) := \sup_{y \in A \setminus \{x\}} S(x,y) = S(x,u^*)$. If this were not true then we would have $S(x,y) > S(x,u^*) = S(x,v^*)$ for some $y \in A \setminus \{x,u^*,v^*\}$. Using the mediant inequality, if $f(y) \ge f(x)$ this implies

$$R_x(u^*,y) = \frac{f(y) - f(u^*)}{\rho(x,y)^\beta + \rho(x,u^*)^\beta} = \frac{F_A(y) - F_A(x) + F_A(x) - F_A(u^*)}{\rho(x,y)^\beta + \rho(x,u^*)^\beta} > S(x,u^*) = R_x(u^*,v^*)\,,$$

while if $f(y) < f(x)$ then

$$R_x(y,v^*) = \frac{f(v^*) - f(y)}{\rho(x,v^*)^\beta + \rho(x,y)^\beta} = \frac{F_A(v^*) - F_A(x) + F_A(x) - F_A(y)}{\rho(x,v^*)^\beta + \rho(x,y)^\beta} > S(x,v^*) = R_x(u^*,v^*)\,,$$

both contradicting the maximizing property of (u^*,v^*) - so indeed $\Lambda_{F_A}^\beta(x,A) = S(x,u^*) = S(x,v^*)$. In particular, we see that if $F_A(x) \ge f(x)$ then

$$\Lambda_f^\beta(x,A) = \sup_{y \in A \setminus \{x\}} \frac{|f(y) - f(x)|}{\rho(y,x)^\beta} \ge \frac{f(v^*) - f(x)}{\rho(v^*,x)^\beta} \ge \frac{F_A(v^*) - F_A(x)}{\rho(v^*,x)^\beta} = S(x,v^*) = \Lambda_{F_A}^\beta(x,A)\,,$$

while if $F_A(x) < f(x)$ then

$$\Lambda_f^\beta(x,A) = \sup_{y \in A \setminus \{x\}} \frac{|f(x) - f(u)|}{\rho(x,y)^\beta} \ge \frac{f(x) - f(u^*)}{\rho(x,u^*)^\beta} > \frac{F_A(x) - F_A(u^*)}{\rho(x,u^*)^\beta} = S(x,u^*) = \Lambda_{F_A}^\beta(x,A)\,,$$

proving Claim I in either case.

Claim II. $\forall x \in \Omega \setminus A:\ \Lambda^{\beta}_{F_A}(x, \Omega \setminus A) \leq \Lambda^{\beta}_{F_A}(x, A)$, in particular $\Lambda^{\beta}_{F_A}(x, \Omega) = \Lambda^{\beta}_{F_A}(x, A)$.

It suffices to show that for any $x, y \in \Omega \setminus A$:

$$S(x, y) \leq \min\{\Lambda^{\beta}_{F_A}(x, A), \Lambda^{\beta}_{F_A}(y, A)\} ,$$

since taking the supremum of the left hand side over $y \in \Omega \setminus A$ shows the claim. Let $(u^*_x, v^*_x), (u^*_y, v^*_y)$ the associated maximizers of R_x, R_y respectively, and note that by definition we have

$$\Lambda^{\beta}_{F_A}(x, A) = \sup_{z \in A \setminus \{x\}} S(x, z) \geq \max\{S(x, u^*_y), S(x, v^*_y)\} . \tag{5}$$

We assume without loss of generality that $\Lambda^{\beta}_{F_A}(x, A) \leq \Lambda^{\beta}_{F_A}(y, A)$, and recall that by Eq. (4) we can deduce that $F_A(u^*_x) \leq F_A(x) \leq F_A(v^*_x)$ and $F_A(u^*_y) \leq F_A(y) \leq F_A(v^*_y)$. Now suppose by contradiction that $S(x, y) > \Lambda^{\beta}_{F_A}(x, A)$. If $F_A(x) \leq F_A(y)$ then

$$\begin{aligned} F_A(v^*_y) &= F_A(x) + \rho(x, y)^{\beta} S(x, y) + \rho(y, v^*_y)^{\beta} \Lambda^{\beta}_{F_A}(y, A) \\ &> F_A(x) + \rho(x, y)^{\beta} \Lambda^{\beta}_{F_A}(x, A) + \rho(y, v^*_y)^{\beta} \Lambda^{\beta}_{F_A}(x, A) \\ &\geq F_A(x) + \rho(x, v^*_y)^{\beta} \Lambda^{\beta}_{F_A}(x, A) , \end{aligned}$$

thus $S(x, v^*_y) = \frac{|F_A(x) - F_A(v^*_y)|}{\rho(x, v^*_y)^{\beta}} > \Lambda^{\beta}_{F_A}(x, A)$ which contradicts Eq. (5). On the other hand, if $F_A(x) > F_A(y)$ then

$$\begin{aligned} F_A(x) &= F_A(u^*_y) + \rho(u^*_y, y)^{\beta} \Lambda^{\beta}_{F_A}(y, A) + \rho(y, x)^{\beta} S(x, y) \\ &> F_A(u^*_y) + \rho(u^*_y, y)^{\beta} \Lambda^{\beta}_{F_A}(x, A) + \rho(y, x)^{\beta} \Lambda^{\beta}_{F_A}(x, A) \\ &\geq F_A(u^*_y) + \rho(u^*_y, x)^{\beta} \Lambda^{\beta}_{F_A}(x, A) , \end{aligned}$$

thus $S(x, u^*_y) = \frac{|F_A(x) - F_A(u^*_y)|}{\rho(x, u^*_y)^{\beta}} > \Lambda^{\beta}_{F_A}(x, A)$ which contradicts Eq. (5), and proves claim Claim II.

Claim III. $\forall x \in A:\ \Lambda^{\beta}_{F_A}(x, \Omega) = \Lambda^{\beta}_{F_A}(x, A) \leq \Lambda^{\beta}_{f}(x, \Omega)$.

Let $x \in A$. Assume towards contradiction that there exists $y \notin A$ such that

$$\Lambda_{F_A}(x, \Omega) \geq S(x, y) > \Lambda^{\beta}_{F_A}(x, A) .$$

We denote by $(u^*_y, v^*_y) \in A \times A$ the maximizer of $R_y(\cdot, \cdot)$. Recall that since $x \in A$, in the proof of Claim I we showed that $S(x, y) \leq S(y, u^*_y) = S(y, v^*_y)$. If $F_A(x) \leq F_A(y) \leq F_A(v^*_y)$ then

$$\begin{aligned} S(x, v^*_y) = \frac{F_A(v^*_y) - F_A(x)}{\rho(v^*_y, x)^{\beta}} &\geq \frac{F_A(v^*_y) - F_A(y) + F_A(y) - F_A(x)}{\rho(v^*_y, y)^{\beta} + \rho(x, y)^{\beta}} \\ &\geq \min\{S(y, v^*_y), S(x, y)\} = S(x, y) > \Lambda^{\beta}_{F_A}(x, A) , \end{aligned}$$

while on the other hand if $F_A(x) > F_A(y) \geq F_A(u^*_y)$ then

$$\begin{aligned} S(x, u^*_y) = \frac{F_A(x) - F_A(u^*_y)}{\rho(x, u^*_y)^{\beta}} &\geq \frac{F_A(x) - F_A(y) + F_A(y) - F_A(u^*_y)}{\rho(x, y)^{\beta} + \rho(u^*_y, y)^{\beta}} \\ &\geq \min\{S(x, y), S(y, u^*_y)\} = S(x, y) > \Lambda^{\beta}_{F_A}(x, A) , \end{aligned}$$

where in both calculations we used the mediant inequality. Both inequalities above contradict the definition of $\Lambda^{\beta}_{F_A}(x, A)$, thus proving Claim III.

Combining the ingredients. We are now ready to finish the proof. For $x \in \Omega$, if $x \in A$ then Claim III provides the desired inequality. Otherwise, if $x \in \Omega \setminus A$ then

$$\Lambda^{\beta}_{F_A}(x, \Omega) \overset{\text{Claim II}}{=} \Lambda^{\beta}_{F_A}(x, A) \overset{\text{Claim I}}{\leq} \Lambda^{\beta}_{f}(x, A) \leq \Lambda^{\beta}_{f}(x, \Omega) .$$

□

B Proofs

B.1 Proof of Theorem 3.1

We start by stating a strengthened version of the triangle inequality (also known as the "snowflake" triangle inequality) which we will use later on. For any $\beta \in (0,1],\ x \neq y, z \in \Omega$:

$$\rho(x,y)^\beta \leq \rho(x,z)^\beta + \rho(z,y)^\beta \ . \tag{6}$$

Indeed, this follows from

$$\begin{aligned}\frac{\rho(x,z)^\beta + \rho(z,y)^\beta}{\rho(x,y)^\beta} &\geq \frac{\rho(x,z)^\beta + \rho(z,y)^\beta}{(\rho(x,z)+\rho(z,y))^\beta} = \left(\frac{\rho(x,z)}{\rho(x,z)+\rho(z,y)}\right)^\beta + \left(\frac{\rho(z,y)}{\rho(x,z)+\rho(z,y)}\right)^\beta \\ &\geq \left(\frac{\rho(x,z)}{\rho(x,z)+\rho(z,y)}\right) + \left(\frac{\rho(z,y)}{\rho(x,z)+\rho(z,y)}\right) = 1 \ .\end{aligned}$$

Let $0 < \varepsilon < \frac{1}{4}$, denote $K := \lceil \log_2(1/\varepsilon) \rceil,\ \varepsilon' := \frac{1}{(K+1)2^K}$ and note that

$$\varepsilon' \geq \frac{1}{(\log_2(1/\varepsilon)+2)\,2^{\log_2(1/\varepsilon)+1}} = \frac{\varepsilon}{2\,(\log_2(1/\varepsilon)+2)} \geq \frac{\varepsilon}{4\log_2(1/\varepsilon)} \ . \tag{7}$$

Let $N = \{x_1, \ldots, x_{|N|}\}$ be a $\left(\frac{\varepsilon'}{32L}\right)^{1/\beta}$-net of Ω of size $|N| = \mathcal{N}_\Omega\left(\left(\frac{\varepsilon'}{32L}\right)^{1/\beta}\right)$, and let $\Pi = \{C_1, \ldots, C_{|N|}\}$ be its induced Voronoi partition. We define $\mathcal{B} = \{[l_j, u_j]\}_{j\in J} \subset [0,1]^\Omega \times [0,1]^\Omega$ to be the pairs of functions constructed as follows:

- l, u are both constant over every cell $C_i \in \Pi$, and map each cell to a value in $\{0, \frac{\varepsilon'}{2}, \varepsilon', \frac{3\varepsilon'}{2}, \ldots, 1\}$.
- Choose some cells $S_1 \subset \Pi$ such that $\mu(\bigcup_{C_i \in S_1} C_i) \leq \varepsilon'$ and set for any $C_i \in S_1 :\ l|_{C_i} = 0,\ u|_{C_i} = 1$.
- For $m = 2, \ldots, K$ choose some "unchosen" cells $S_m \subset \Pi \setminus \bigcup_{j<m} S_j$ such that $\mu(\bigcup_{C_i \in S_m} C_i) \leq 2^{m-1}\varepsilon'$ and set for any $C_i \in S_m :\ l|_{C_i} \in \{0, \frac{1}{2^m}, \frac{2}{2^m}, \ldots, \frac{2^m-2}{2^m}\},\ ,\ u|_{C_i} = l + \frac{1}{2^{m-1}}$.
- In the "remaining" cells $S_{K+1} := \Pi \setminus \bigcup_{j\leq K} S_j$ set for any $C_i \in S_{K+1}$:

$$l|_{C_i} \in \left\{0, \frac{1}{2^{K+1}}, \frac{2}{2^{K+1}}, \ldots, \frac{2^{K+1}-2}{2^{K+1}}\right\},\ u|_{C_i} = l + \frac{1}{2^K} \ .$$

Notice that for any $[l,u] \in \mathcal{B}$ we have

$$\begin{aligned}\|l-u\|_{L^1(\mu)} &= \sum_{C_i\in\Pi}\int_{C_i} |l(x)-u(x)|d\mu(x) = \sum_{m=1}^{K+1}\sum_{C_i\in S_m}\int_{C_i} |l(x)-u(x)|d\mu(x) \\ &= \sum_{m=1}^{K+1}\sum_{C_i\in S_m}\int_{C_i} \frac{1}{2^{m-1}}d\mu(x) = \sum_{m=1}^{K+1}\frac{1}{2^{m-1}}\sum_{C_i\in S_m}\mu(C_i) \\ &= \sum_{m=1}^{K+1}\frac{2^{m-1}\varepsilon'}{2^{m-1}} = \varepsilon'(K+1) = \frac{1}{2^K} \leq \varepsilon \ .\end{aligned}$$

Furthermore, we can bound $|\mathcal{B}|$ by noticing that any such l is defined by its values over $|N|$ cells who all belong to $\{0, \frac{\varepsilon'}{2}, \varepsilon', \ldots, 1\}$, and that once l is fixed then any associated u has at most $K+1$ possible values over each cell since it equals $l + \frac{1}{2^{m-1}}$ for some $m \in [K+1]$. Thus

$$|\mathcal{B}| \leq (K+1)\left(\frac{8}{\varepsilon'}\right)^{|N|} \leq \log_2\left(\frac{1}{\varepsilon}\right)\cdot\left(\frac{16\log_2(1/\varepsilon)}{\varepsilon}\right)^{\mathcal{N}\left(\left(\frac{\varepsilon}{128L\log(1/\varepsilon)}\right)^{1/\beta}\right)} \ ,$$

where the last inequality uses Eq. (7) and definition of K. In order to finish the proof, in remains to show that $\mathcal{B}$ indeed cover $\widetilde{\text{Höl}}_L^\beta(\Omega,\mu)$ as brackets. Namely, that for any $f \in \widetilde{\text{Höl}}_L^\beta(\Omega,\mu)$ there exist $[l,u] \in \mathcal{B}$ such that $l \le f \le u$. To that end, let $f \in \widetilde{\text{Höl}}_L^\beta(\Omega,\mu)$. Denote

$$S_1^f := \left\{ C_i \in \Pi \ : \ \forall x \in C_i : \Lambda_f^\beta(x) \ge \frac{L}{\varepsilon'} \right\}$$

and notice that $\bigcup\{C_i \in S_1^f\} \subseteq \{x : \Lambda_f^\beta(x) \ge L/\varepsilon'\} \implies \mu(\bigcup\{C_i \in S_1^f\}) \le \varepsilon'$. Hence we can pick $[l,u] \in \mathcal{B}$ such that $(l|_{C_i}, u|_{C_i}) \equiv (0,1)$ for any $C_i \in S_1^f$ (serving as S_1 in their construction). Clearly any such l,u bound f over these cells. Furthermore, for $m = 2,\ldots,K$ we denote

$$S_m^f := \left\{ C_i \in \Pi \setminus \bigcup_{j<m} S_j^f \ : \ \forall x \in C_i : \Lambda_f^\beta(x) \ge \frac{L}{2^{m-1}\varepsilon'} \right\} ,$$

and notice that $\bigcup\{C_i \in S_m^f\} \subseteq \{x : \Lambda_f^\beta(x) \ge L/(2^{m-1}\varepsilon')\} \implies \mu(\bigcup\{C_i \in S_m^f\}) \le 2^{m-1}\varepsilon'$. Consequently we can let S_m^f serve as S_m in the construction of $[l,u] \in \mathcal{B}$, assuming we will show such a choice can serve as a bracket of f over such cells. Indeed, for any $x \in C_i$ we have

$$|f(x) - f(z_i)| \le \Lambda_f^\beta(z_i) \cdot \rho(x,z_i)^\beta \overset{Eq.\ (6)}{\le} \frac{L}{2^{m-2}\varepsilon'} \cdot \frac{2\varepsilon'}{32L} = \frac{1}{2^{m+2}},$$

which by the triangle inequality shows in particular that for any $x,y \in C_i$:

$$|f(x) - f(y)| \le |f(x) - f(z_i)| + |f(z_i) - f(y)| \le \frac{1}{2^{m+1}} = \frac{1}{4 \cdot 2^{m-1}} .$$

So clearly there exists $\alpha_i \in \{0, \frac{1}{2^m}, \frac{2}{2^m}, \ldots, \frac{2^m-2}{2^m}\}$ such that $\alpha_i \le f|_{C_i} \le \alpha_i + \frac{1}{2^{m-1}}$, and by setting $l|_{C_i}, u|_{C_i} = (\alpha_i, \alpha_i + \frac{1}{2^{m-1}})$ for any $C_i \in S_m^f$ we ensure the bracketing property. Finally, for any of the remaining cells $S_{K+1}^f := \Pi \setminus \bigcup_{j \le K} S_j^f$ we get by construction that $\exists z_i \in C_i : \Lambda_f^\beta(z_i) < \frac{L}{2^K \varepsilon'}$ (otherwise they would satisfy the condition for some previously constructed S_m^f). Hence for any $x \in C_i$ we have

$$|f(x) - f(z_i)| \le \Lambda_f^\beta(z_i) \cdot \rho(x,z_i)^\beta \overset{Eq.\ (6)}{\le} \frac{L}{2^K \varepsilon'} \cdot \frac{2\varepsilon'}{32L} = \frac{1}{2^{K+4}},$$

which by the triangle inequality shows that for any $x,y \in C_i$:

$$|f(x) - f(y)| \le \frac{1}{2^{K+3}} = \frac{1}{8 \cdot 2^K} .$$

So as before, there clearly exists $\alpha_i \in \{0, \frac{1}{2^{K+1}}, \frac{2}{2^{K+1}}, \ldots, \frac{2^{K+1}-2}{2^{K+1}}\}$ such that $\alpha_i \le f|_{C_i} \le \alpha_i + \frac{1}{2^K}$, and by setting $l|_{C_i}, u|_{C_i} = (\alpha_i, \alpha_i + \frac{1}{2^K})$ for any $C_i \in S_m^f$ we ensure the bracketing property over all of Ω, which finishes the proof.

B.2 Proof of Proposition 3.2

Recalling that the realizability assumption ensures a "perfect" predictor $f^* \in \mathcal{F}$, we start by introducing the loss class $\mathcal{L}_\mathcal{F} \subset [0,1]^\Omega$:

$$\mathcal{L}_\mathcal{F} = \{\ell_f(x) := |f(x) - f^*(x)| : f \in \mathcal{F}\} .$$

Fix $\alpha > 0$. We observe that $\mathcal{L}_\mathcal{F}$ is no larger than $\mathcal{F}$ in terms of bracketing entropy, namely

$$\mathcal{N}_{[]}(\mathcal{L}_\mathcal{F}, L_1(\mu), \alpha) \le \mathcal{N}_{[]}(\mathcal{F}, L_1(\mu), \alpha) . \tag{8}$$

Indeed, suppose we are given an α-bracketing of $\mathcal{F}$ denoted by $\mathcal{B}_\alpha$, and denote for any $f \in \mathcal{F}$ by $[l_f, u_f] \in \mathcal{B}_\alpha$ its associated bracket. Then any $\ell_f \in \mathcal{L}_\mathcal{F}$ is inside the bracket $[l_{\ell_f}, u_{\ell_f}]$ where

$$\begin{aligned} l_{\ell_f} &:= \max\{0\,,\, \min\{l_f - f^*, f^* - u_f\}\} , \\ u_{\ell_f} &:= \min\{1\,,\, \max\{u_f - f^*, f^* - l_f\}\} . \end{aligned}$$

It is straightforward to verify that $\|u_{\ell_f} - l_{\ell_f}\|_{L_1(\mu)} \leq \|u_f - l_f\|_{L_1(\mu)} \leq \alpha$, and clearly the set of all such brackets is of size at most $|\mathcal{B}_\alpha|$, yielding Eq. (8).

Now notice that for any $f \in \mathcal{F}$:

$$L_{\mathcal{D}}(f) - 1.01 L_S(f) = \|\ell_f\|_{L_1(\mu)} - 1.01\|\ell_f\|_{L_1(\mu_n)} \leq \alpha + \|l_{\ell_f}\|_{L_1(\mu)} - 1.01\|l_{\ell_f}\|_{L_1(\mu_n)} ,$$

hence

$$\sup_{f\in\mathcal{F}} \left(L_{\mathcal{D}}(f) - 1.01 L_S(f)\right) \leq \alpha + \max_{l_{\ell_f}} (\|l_{\ell_f}\|_{L_1(\mu)} - 1.01\|l_{\ell_f}\|_{L_1(\mu_n)}) . \tag{9}$$

In order to bound the right hand side, fix some l_{ℓ_f}, and note that $\mathrm{Var}(l_{\ell_f}) \leq \|l_{\ell_f}^2\|_{L_1(\mu)} \leq \|l_{\ell_f}\|_{L_1(\mu)}$, since $l_{\ell_f}(x) \in [0,1]$. Thus by Bernstein's inequality and the AM-GM inequality we get that with probability at least $1-\gamma$:

$$\begin{aligned} \|l_{\ell_f}\|_{L_1(\mu)} - \|l_{\ell_f}\|_{L_1(\mu_n)} &\leq \frac{\log(1/\gamma)}{n} + \sqrt{\frac{2\|l_{\ell_f}\|_{L_1(\mu)}\log(1/\gamma)}{n}} \\ &\leq \frac{202\log(1/\gamma)}{n} + \frac{1}{101}\|l_{\ell_f}\|_{L_1(\mu)} \\ \implies \|l_{\ell_f}\|_{L_1(\mu)} - 1.01\|l_{\ell_f}\|_{L_1(\mu_n)} &\leq \frac{205\log(1/\gamma)}{n} . \end{aligned}$$

Setting $\gamma = \delta/\mathcal{N}_{[]}(\mathcal{F}, L_1(\mu), \alpha)$ and taking a union bound over l_{ℓ_f} whose number is bounded due to Eq. (8), we see that with probability $1-\delta$:

$$\max_{l_{\ell_f}} (\|l_{\ell_f}\|_{L_1(\mu)} - 1.01\|l_{\ell_f}\|_{L_1(\mu_n)}) \leq \frac{205\log \mathcal{N}_{[]}(\mathcal{F}, L_1(\mu), \alpha) + 205\log(1/\delta)}{n} .$$

Plugging this back into Eq. (9), and minimizing over $\alpha > 0$ finishes the proof.

B.3 Proof of Theorem 4.1

Proposition B.1. *Let $f : \Omega \to [0,1]$. Then with probability at least $1-\delta/2$ over drawing a sample it holds that*

$$\widehat{\Lambda}_f^\beta \leq 4\log^2(4n/\delta)\overline{\Lambda}_f^\beta(\mu) + \frac{4\log^2(4n/\delta)}{n} .$$

Corollary B.2. *If $\mathcal{D}$ is realizable by $\overline{\mathrm{Höl}}_L^\beta(\Omega,\mu)$, then for $f^* : \Omega \to [0,1]$ such that $L_{\mathcal{D}}(f^*) = 0$ it holds with probability at least $1-\delta/2$: $\widehat{\Lambda}_{f^*}^\beta \leq 5\log^2(4n/\delta)L$. Hence, $\widehat{f}(X_i) := f^*(X_i) = Y_i$ satisfies $L_S(\widehat{f}) = 0$ and $\widehat{\Lambda}_{\widehat{f}}^\beta \leq 5\log^2(4n/\delta)L$.*

Proof. (of Proposition B.1) Fix $f : \Omega \to [0,1]$. Given a sample $(X_i)_{i=1}^n \sim \mu^n$ which induces an empirical measure μ_n, we get

$$\widehat{\Lambda}_f^\beta \leq \frac{1}{n}\sum_{i=1}^n \sup_{z\neq X_i} \frac{|f(X_i) - f(z)|}{\rho(X_i,z)^\beta} = \mathop{\mathbb{E}}_{X\sim\mu_n}[\Lambda_f^\beta(X)] \leq 2\log(n)\mathbb{W}_{X\sim\mu_n}[\Lambda_f^\beta(X)] , \tag{10}$$

where the last inequality follows from the reversed strong-weak mean inequality for uniform measures. We will now show that with high probability $\mathbb{W}_{X\sim\mu_n}[\Lambda_f^\beta(X)] \lesssim \mathbb{W}_{X\sim\mu}[\Lambda_f^\beta(X)] = \widetilde{\Lambda}_f^\beta$. To that end, we denote for any $t > 0$: $M_f(t) := \{x : \Lambda_f^\beta(x) \geq t\} \subset \Omega$, let $K := \widetilde{\Lambda}_f^\beta(\mu)$, $N := \lceil 2\log(4n/\delta)\log\log(4n/\delta)\rceil$ and note that

$$\begin{aligned} \mathbb{W}_{X\sim\mu_n}[\Lambda_f^\beta(X)] &= \sup_{t>0} t\mu_n(M_f(t)) \qquad (11) \\ &\leq \sup_{0<t\leq K} t\mu_n(M_f(t)) + 2\max_{j\in\{0,1,\ldots,N-1\}} 2^j K\mu_n(M_f(2^j K)) + \sup_{t\geq 2^N K} t\mu_n(M_f(t)) . \end{aligned}$$

We will bound all three summands above. We easily bound the first term by

$$\sup_{0<t\leq K} t\mu_n(M_f(t)) \leq K\cdot 1 = \widetilde{\Lambda}_f^\beta(\mu) . \tag{12}$$

For the second term, denote for any $t > 0$ by $M_f^+(t) \supset M_f(t)$ a containing set for which $\frac{1}{n} \leq \mu(M_f^+(t)) \leq \mu(M_f(t)) + \frac{1}{n}$. We can always assume without loss of generality that such a set exists.[5] By the multiplicative Chernoff bound we have for any $t, \alpha > 0$:

$$\Pr_S \left[\mu_n(M_f^+(t)) \geq (1+\alpha)\mu(M_f^+(t)) \right] \leq \frac{e^\alpha}{(1+\alpha)^{1+\alpha}} ,$$

hence by the union bound we get with probability at least $1 - \frac{Ne^\alpha}{(1+\alpha)^{1+\alpha}}$:

$$\begin{aligned}
\max_{j \in \{0,1,\ldots,N-1\}} 2^j K \mu_n(M_f(2^j K)) &\leq \max_{j \in \{0,1,\ldots,N-1\}} 2^j K \mu_n(M_f^+(2^j K)) \\
&\leq (1+\alpha) \max_{j \in \{0,1,\ldots,N-1\}} 2^j L \mu(M_f^+(2^j K)) \\
&\leq (1+\alpha) \max_{j \in \{0,1,\ldots,N-1\}} 2^j K \left(\mu(M_f(2^j K)) + \frac{1}{n} \right) \\
&\leq (1+\alpha) \widetilde{\Lambda}_f^\beta(\mu) + \frac{1+\alpha}{n} .
\end{aligned}$$

Letting $\alpha = \log(4n/\delta) - 1$, by our choice of $N = \lceil 2\log(4n/\delta)\log\log(4n/\delta) \rceil$ we get that with probability at least $1 - \delta/4$:

$$2 \max_{j \in \{0,1,\ldots,N-1\}} 2^j K \mu_n(M_f(2^j K)) \leq 2\log(4n/\delta)\widetilde{\Lambda}_f^\beta(\mu) + \frac{2\log(4n/\delta)}{n} . \tag{13}$$

In order to bound the last term in Eq. (11), we observe that the empirical measure satisfies for any $A \subset \Omega$: $\mu_n(A) < \frac{1}{n} \iff \mu_n(A) = 0$, and that $M_f(s) \subset M_f(t)$ for $s > t$. Furthermore, by definition of $K = \widetilde{\Lambda}_f^\beta(\mu)$ we have $\mu(M_f(t)) \leq \frac{K}{t}$, hence by Markov's inequality

$$\Pr_S \left[\sup_{s \geq t} \mu_n(M_f(s)) \neq 0 \right] \leq \Pr_S \left[\mu_n(M_f(t)) \neq 0 \right] = \Pr_S \left[\mu_n(M_f(t)) \geq \frac{1}{n} \right] \leq \frac{nK}{t} .$$

For $t := 2^N K$ yields $\Pr_S \left[\sup_{s \geq 2^n K} \mu_n(M_f(s)) \neq 0 \right] \leq \frac{n}{2^N} \leq \frac{\delta}{4}$. Combining this with Eq. (12), Eq. (13) and plugging back into Eq. (11), we get that with probability at least $1 - \delta/2$:

$$\mathbb{W}_{X \sim \mu_n}[\Lambda_f^\beta(X)] \leq (1+2\log(4n/\delta))\widetilde{\Lambda}_f^\beta(\mu) + \frac{2\log(4n/\delta)}{n} \leq (1+2\log(4n/\delta))\overline{\Lambda}_f^\beta(\mu) + \frac{2\log(4n/\delta)}{n} .$$

Recalling Eq. (10), we get overall that

$$\widehat{\Lambda}_f^\beta \leq 2\log(n) \left[(1+2\log(4n/\delta))\overline{\Lambda}_f^\beta(\mu) + \frac{2\log(4n/\delta)}{n} \right] .$$

Simplifying the expression above finishes the proof.

□

Proposition B.3. *Under the same setting, for any $\gamma > 0$ there exists an algorithm that given a sample $S \sim \mathcal{D}^n$ and any function $\widehat{f} : S \to [0,1]$, provided that $n \geq N$ for $N = \widetilde{O}\left(\frac{\mathcal{N}_\Omega(\gamma) + \log(1/\delta)}{\gamma} \right)$, constructs a function $f : \Omega \to [0,1]$ such that with probability at least $1 - \delta/2$:*

- $\|f - \widehat{f}\|_{L_1(\mu_n)} \leq \gamma(1 + 2\widehat{\Lambda}_{\widehat{f}}^\beta)$. *In particular* $L_S(f) \leq L_S(\widehat{f}) + \gamma(1 + 2\widehat{\Lambda}_{\widehat{f}}^\beta)$.
- $\overline{\Lambda}_f^\beta(\mu) \leq 5\widehat{\Lambda}_{\widehat{f}}^\beta$.

[5] Such a set does not exist only in the case of atoms $x_0 \in \Omega$ with large probability mass $\mu(x_0)$. If that is the case, consider a "copy" metric space $\widetilde{\Omega}$ with x_0 split into two points $x_1, x_2 \in \widetilde{\Omega}$ at distance ε apart and each of mass $\mu(x_0)/2$. Any function $f : \Omega \to \mathbb{R}$ is extended to $\widetilde{f} : \widetilde{\Omega} \to \mathbb{R}$ via $\widetilde{f}(x_1) = \widetilde{f}(x_2) = f(x_0)$. Repeating the split if necessary and taking $\varepsilon \downarrow 0$, we obtain a space $\widetilde{\Omega}$ with all of the relevant properties of Ω but no atoms of large mass.

Proof. Throughout the proof, we denote for any point $x \in \Omega$, subset $B \subset \Omega$ and function $g : B \to [0,1]$:

$$\Lambda_g^\beta(x, B) := \sup_{y \in B \setminus \{x\}} \frac{|g(x) - g(y)|}{\rho(x,y)^\beta} .$$

Give the sample $S = (X_i, Y_i)_{i=1}^n$, we denote $S_x = (X_i)_{i=1}^n$. Let $\gamma > 0$. The algorithm constructs $f : \Omega \to [0,1]$ as follows:

1. Let $S_x(\gamma) \subset S_x$ consist of the $\lfloor \gamma n \rfloor$ points whose $\Lambda_{\widehat{f}}(\cdot, S_x)$ values are the largest (with ties broken arbitrarily), and $S'_x(\gamma) := S_x \setminus S_x(\gamma)$ be the rest.
2. Let $A \subset S'_x(\gamma)$ be a $\gamma^{1/\beta}$-net of $S'_x(\gamma)$.
3. Define $f : \Omega \to [0,1]$ to be the β-PMSE extension of $\widehat{f}$ from A to Ω as defined in Definition A.2 (and analyzed throughout Appendix A).

We will prove that f satisfies both requirements. For the first requirement, since $f|_A = \widehat{f}|_A$ and $S_x = S'_x(\gamma) \uplus S_x(\gamma)$ we have

$$\|f - \widehat{f}\|_{L_1(\mu_n)} := \frac{1}{n}\sum_{i=1}^n |f(x_i) - g(x_i)| = \frac{1}{n} \sum_{x \in S_x(\gamma) \setminus A} |f(x) - \widehat{f}(x)| + \frac{1}{n} \sum_{x \in S'_x(\gamma) \setminus A} |f(x) - \widehat{f}(x)| .$$

The first summand above is bounded by γ since $0 \le f, \widehat{f} \le 1 \implies |f(x) - \widehat{f}(x)| \le 1$ and $|S_x(\gamma)| \le \gamma n$. In order to bound the second term, we denote by $N_A : S'_x(\gamma) \to A$ to be the mapping of each element to its nearest neighbor in the net, and note that $\rho(x, N_A(x)) \le \gamma^{1/\beta}$. Then

$$\begin{aligned}
\frac{1}{n} \sum_{x \in S'_x(\gamma) \setminus A} |f(x) - \widehat{f}(x)| &\le \frac{1}{n} \sum_{x \in S'_x(\gamma) \setminus A} \frac{\gamma}{\rho(x, N_A(x))^\beta} |f(x) - \widehat{f}(x)| \\
&\le \frac{\gamma}{n} \sum_{x \in S'_x(\gamma) \setminus A} \frac{|f(x) - \widehat{f}(N_A(x))| + |\widehat{f}(N_A(x)) - \widehat{f}(x)|}{\rho(x, N_A(x))^\beta} \\
&= \frac{\gamma}{n} \sum_{x \in S'_x(\gamma) \setminus A} \frac{|f(x) - f(N_A(x))|}{\rho(x, N_A(x))^\beta} + \frac{|\widehat{f}(N_A(x)) - \widehat{f}(x)|}{\rho(x, N_A(x))^\beta} \\
&\le \frac{\gamma}{n} \sum_{x \in S'_x(\gamma) \setminus A} \Lambda_f^\beta(x, A) + \Lambda_{\widehat{f}}^\beta(x, A) \\
[\text{Theorem A.3}] &\le \frac{2\gamma}{n} \sum_{x \in S'_x(\gamma) \setminus A} \Lambda_{\widehat{f}}^\beta(x, A) \\
&\le 2\gamma L .
\end{aligned}$$

So overall we get $\|f - \widehat{f}\|_{L_1(\mu_n)} \le \gamma + 2\gamma L = \gamma(1 + 2L)$ as claimed in the first bullet.

We move on to prove the second bullet. Let $U \subset \Omega$ be a $\frac{\gamma^{1/\beta}}{4}$-net of Ω, Π be its induced Voronoi partition and let $m := |\Pi| \le \mathcal{N}_\Omega(\gamma^{1/\beta}/4)$. Let Consider the following partition of Π into "light" and "heavy" cells:

$$\Pi_l := \{C \in \Pi \,:\, \mu_n(C) < n\gamma/m\} , \quad \Pi_h := \Pi \setminus \Pi_l .$$

We will now state three lemmas required for the proof, two of which are due to [Ashlagi et al., 2021].

Lemma B.4. *Suppose $A \subset \Omega$ and that $f : \Omega \to [0,1]$ is the β-PMSE extension of some function from A to Ω. Let $E \subset \Omega$ such that $\operatorname{diam}(E)^\beta \le \frac{1}{2} \min_{x \ne x' \in A} \rho(x, x')^\beta$. Then $\sup_{x,x' \in E} \frac{\Lambda_f^\beta(x)}{\Lambda_f^\beta(x')} \le 2$.*

Proof. Let $u_x^*, v_x^* \in A$ be the pair of points which satisfy $\Lambda_f^\beta(x) = \frac{f(v_x^*) - f(u_x^*)}{\rho(v_x^*, x)^\beta + \rho(u_x^*, x)^\beta}$. By assumption on E, we know that $2\operatorname{diam}(E)^\beta \le \rho(v_x^*, u_x^*)^\beta \le \rho(v_x^*, x)^\beta + \rho(u_x^*, x)^\beta$, hence

$\rho(v_x^*, x)^\beta + \rho(u_x^*, x)^\beta + 2\mathrm{diam}(E)^\beta \le 2(\rho(v_x^*, x)^\beta + \rho(u_x^*, x)^\beta)$. We get

$$\begin{aligned}\Lambda_f^\beta(x') &\ge \frac{f(v_x^*) - f(u_x^*)}{\rho(v_x^*, x')^\beta + \rho(u_x^*, x')^\beta} \\ &\ge \frac{f(v_x^*) - f(u_x^*)}{\rho(v_x^*, x)^\beta + \mathrm{diam}(E)^\beta + \rho(u_x^*, x)^\beta + \mathrm{diam}(E)^\beta} \\ &\ge \frac{f(v_x^*) - f(u_x^*)}{2(\rho(v_x^*, x)^\beta + \rho(u_x^*, x)^\beta)} = \frac{1}{2}\Lambda_f^\beta(x) \,.\end{aligned}$$

□

Lemma B.5 (Ashlagi et al., 2021, Lemma 16). *If $n\gamma^2 \ge m$, then*

$$\begin{aligned}\Pr_{S\sim\mathcal{D}^n}\left[\min_{C\in\Pi_h}\frac{\mu_n(C)}{\mu(C)} > \frac{1}{2}\right] &\ge 1 - m\exp(-n\gamma/4m) \,, \\ \Pr_{S\sim\mathcal{D}^n}\left[\max_{C\in\Pi_h}\frac{\mu_n(C)}{\mu(C)} < 2\right] &\ge 1 - m\exp(-n\gamma/3m) \,, \\ \Pr_{S\sim\mathcal{D}^n}\left[\sum_{C\in\Pi_l}\mu(C) < 2\gamma\right] &\ge 1 - \exp\left(-n(\gamma - \sqrt{m/n})^2/2\right) \,.\end{aligned}$$

Lemma B.6 (Ashlagi et al., 2021, Lemma 17). $\|f\|_{\mathrm{Höl}^\beta} \le \frac{2L}{\gamma}$.

Equipped with the three lemmas, we calculate

$$\overline{\Lambda}_f^\beta(\mu) = \int_\Omega \Lambda_f^\beta(x)d\mu = \sum_{C\in\Pi_l}\int_C \Lambda_f^\beta(x)d\mu + \sum_{C\in\Pi_h}\int_C \Lambda_f^\beta(x)d\mu \,. \tag{14}$$

The first summand above is bounded with high probability using Lemma B.5 and Lemma B.6, since under the event described in Lemma B.5 we have:

$$\begin{aligned}\sum_{C\in\Pi_l}\int_C \Lambda_f^\beta(x)d\mu &\le \sum_{C\in\Pi_l}\int_C \frac{2L}{\gamma}d\mu = \frac{2L}{\gamma}\sum_{C\in\Pi_l}\mu(C) \\ &\le \frac{2L}{\gamma}\cdot 2q = \frac{L}{4} \,.\end{aligned}$$

In order to bound the second term in Eq. (14), let $C \in \Pi$, $x' \in C$ and note that by applying Lemma B.4 to $E := S_x \cap C$ we get that $\Lambda_f^\beta(x') \le 2\min_{x\in S_x\cap C}\Lambda_f^\beta(x)$. Thus, under the high probability event described in Lemma B.5 we have

$$\begin{aligned}\sum_{C\in\Pi_h}\int_C \Lambda_f^\beta(x)d\mu &\le \sum_{C\in\Pi_h}\int_C 2\min_{x\in S_x\cap C}\Lambda_f^\beta(x)d\mu = 2\sum_{C\in\Pi_h}\min_{x\in S_x\cap C}\Lambda_f^\beta(x)\mu(C) \\ &\le 4\sum_{C\in\Pi_h}\min_{x\in S_x\cap C}\Lambda_f^\beta(x)\mu_n(C) = \frac{4}{n}\sum_{C\in\Pi_h}\sum_{x'\in S_x\cap C}\min_{x\in S_x\cap C}\Lambda_f^\beta(x) \\ &\le \frac{4}{n}\sum_{C\in\Pi_h}\sum_{x'\in S_x\cap C}\Lambda_f^\beta(x') \le \frac{4}{n}\sum_{x'\in S_x}\Lambda_f^\beta(x') \le 4L \,,\end{aligned}$$

where the last inequality is due to the extension property of Theorem A.3. Overall, plugging these bounds into Eq. (14) and using the union bound to ensure all required events to hold simultaneously, we see that the desired second bullet holds holds with probability at least $1 - m\exp(-n\gamma/4m) - \exp\left(-n(\gamma - \sqrt{m/n})^2/2\right)$. A straightforward computation shows that by our assumption on n being large enough, this probability exceeds $1 - \delta/2$ as required.

□

We are now ready to finish the proof of Theorem 4.1. Let $\gamma > 0$. By Corollary B.2, we can construct $\widehat{f} : S \to [0,1]$ such that with probability at least $1 - \delta/2$: $L_S(\widehat{f}) = 0$ and $\widehat{\Lambda}_{\widehat{f}}^\beta \le$

$5\log^2(4n/\delta)L$. Assuming n is appropriately large, we further apply Proposition B.3 in order to obtain $f:\Omega\to[0,1]$ such that with probability at least $1-\delta/2$: $f\in\overline{\text{Höl}}^{\beta}_{25\log^2(4n/\delta)L}(\Omega)$ and also $L_S(f)\le L_S(\widehat{f})+\gamma(1+2L)=\gamma(1+2L)$. By the union bound, we get that with probability at least $1-\delta$:

$$\begin{aligned}L_{\mathcal{D}}(f) &= 1.01L_S(f)+(L_{\mathcal{D}}(f)-1.01L_S(f))\\ &\le \gamma(1+2L)+\sup_{f\in\overline{\text{Höl}}^{\beta}_{25\log^2(4n/\delta)L}(\Omega)}(L_{\mathcal{D}}(f)-1.01L_S(f))\,.\\ &\overset{(*)}{\le} \frac{\varepsilon}{2}+\frac{\varepsilon}{2}=\varepsilon\,,\end{aligned}$$

where $(*)$ is justified by setting $\gamma=\Theta(\varepsilon/L)$ and applying Theorem 3.4 for appropriately large n.

B.4 Proof of Theorem 5.1

Given a sample $S=(X_i,Y_i)_{i=1}^n\sim\mathcal{D}^n$, denote the empirically smooth class

$$\widehat{\text{Höl}}:=\left\{f:\{X_1,\dots,X_{\lfloor n/2\rfloor}\}\to[0,1]\,:\,\widehat{\Lambda}^{\beta}_f\le 5\log^2(4n/\delta)L\right\}\,.$$

Consider the following procedure:

1. (*Empirical cover*) Construct $h_1,\dots,h_N\in\widehat{\text{Höl}}$ for maximal N such that $\forall i\neq j\in[N]:\ \|h_i-h_j\|_{L_1(\mu_n)}\ge\frac{\epsilon}{4}$.
2. (*Run realizable algorithm on cover*) For any $j\in[N]$, execute the realizable algorithm $\mathcal{A}_{\text{realizable}}$ of Theorem 4.1 on the "relabeled" dataset $(X_i,h_j(X_i))_{i=1}^{\lfloor n/2\rfloor}$, and obtain $f_j:\Omega\to[0,1]$.
3. (*ERM*) Return $\arg\min_{f_1,\dots,f_N}\sum_{i=\lfloor n/2\rfloor+1}^{n}|f_j(X_i)-Y_i|$.

We will now prove that the algorithm above satisfies the theorem. Let $f^*\in\arg\min_{f\in\overline{\text{Höl}}^{\beta}_L(\Omega,\mu)}L_{\mathcal{D}}(f)$,[6] and note that by Proposition B.1 (as explained in Corollary B.1) we have $f^*\in\widehat{\text{Höl}}$ with probability at least $1-\delta/2$. By construction, $h_1,\dots,h_N$ is a maximal $\frac{\epsilon}{4}$-packing of $\widehat{\text{Höl}}$, which is known to imply that it is also a $\frac{\epsilon}{4}$-net [Vershynin, 2018, Lemma 4.2.8] with respect to the metric $L_1(\mu_n)$. In particular, this implies that there exists $j^*\in[N]$ such that

$$\|f^*-h_{j^*}\|_{L_1(\mu_n)}\le\frac{\epsilon}{4}\implies L_S(h_{j^*})\le L_S(f^*)+\frac{\epsilon}{4}\,.$$

Further note for any $j\in[N]:\ h_j\in\widehat{\text{Höl}}$, so our realizable algorithm (as manifested in Proposition B.3 for $\gamma=\Theta(\epsilon/L)$) when fed the "smoothed" labels $(X_i,h_j(X_i))_{i=1}^{\lfloor n/2\rfloor}$ will produce f_j such that $L_S(f_j)\le L_S(h_j)\le\frac{\epsilon}{4}$ and $\overline{\Lambda}^{\beta}_{f_j}(\mu)\le 5\widehat{\Lambda}^{\beta}_{h_j}\le 25\log^2(4n/\delta)L$. In particular

$$L_S(f_{j^*})\le L_S(h_{j^*})+\frac{\epsilon}{4}\le L_S(f^*)+\frac{\epsilon}{2}\,.$$

Finally, by Eq. (1) and Theorem 3.1 (which holds for any measure, in particular for the empirical measure μ_n)

$$\begin{aligned}\log N &\le \log\mathcal{N}_{\widehat{\text{Höl}}}(\epsilon/2)\\ &\le \log\mathcal{N}_{[\,]}(\widehat{\text{Höl}},L_1(\mu_n),\epsilon)\\ &\le \log\mathcal{N}_{\Omega}\left(\left(\frac{\varepsilon}{640\log^2(4n/\delta)L\log(1/\varepsilon)}\right)^{1/\beta}\right)\cdot\log\left(\frac{16\log_2(1/\varepsilon)}{\varepsilon}\right)\,.\end{aligned}$$

Hence, by a standard Chernoff-Hoeffding bound over the finite class $\{f_1,\dots,f_N\}$, step (3) of the algorithm yields $\frac{\epsilon}{2}$ excess risk as long as $\frac{n}{2}=\Omega\left(\frac{\log(N)+\log(1/\delta)}{\epsilon^2}\right)$.

[6] We assume without loss of generality that the infimum is obtained. Otherwise we can take a function whose loss is arbitrarily close enough to the optimal value and continue with the proof verbatim.

B.5 Proof of Theorem 6.1

We start by providing a simple structural result which we will use for our lower bound construction, showing that in any metric space there exists a sufficiently isolated point from a large enough subset.

Lemma B.7. *There exists a point $x_0 \in \Omega$ and a subset $K \subset \Omega$ such that*

- $\forall x \in K : \rho(x_0, x) \geq \frac{\mathrm{diam}(\Omega)}{4}$.
- $\forall x \neq y \in K : \rho(x, y) \geq (\varepsilon/L)^{1/\beta}$.
- $|K| = \left\lfloor \frac{\mathcal{N}_\Omega((\varepsilon/L)^{1/\beta})}{2} \right\rfloor$.

Proof. Denote $D := \mathrm{diam}(\Omega)$, let x_0, x_1 be two points such that $\rho(x_0, x_1) > D/2$, and let $\Pi = \{C_0, C_1\}$ be a Voronoi partition of Ω induced by $\{x_0, x_1\}$. For $\gamma > 0$, let N_γ be a maximal γ-packing of Ω. By the pigeonhole principle there must exist a cell $C_i \in \Pi$ such that $|C_i \cap N_\gamma| \geq |N_\gamma|/2$, which we assume without loss of generality to be C_1. Now note that any $x \in C_1$ satisfies $\rho(x, x_0) \geq \frac{1}{2}\rho(x, x_0) + \frac{1}{2}\rho(x, x_1) \geq \frac{1}{2}\rho(x_0, x_1) > D/4$. Finally, set $\gamma := \varepsilon^{1/\beta}$ and let $K \subset C_1 \cap N_\gamma$ be any subset of size $\left\lfloor \frac{\mathcal{N}_\Omega((\varepsilon/L)^{1/\beta})}{2} \right\rfloor$. □

Given x_0, K from the lemma above, we denote $\overline{K} = \{x_0\} \cup K$ and define the distribution μ over Ω supported on $\overline{K}$ such that $\mu(x_0) = 1 - \frac{\varepsilon}{2}$ and $\mu(x) = \frac{\varepsilon}{2|K|}$ for all $x \in K$. From now on, the proof is similar to a classic lower bound strategy for VC classes in the realizable case (e.g. Kearns and Vazirani, 1994, Proof of Theorem 3.5). To that end, it is enough to provide a distribution over functions in $\overline{\mathrm{Höl}}_L^\beta(\Omega, \mu)$ such that with constant probability any algorithm must suffer significant loss for some function supported by the distribution.

We define such a distribution over functions $\overline{f} : \overline{K} \to \{0, 1\}$ as follows: $\Pr[\overline{f}(x_0) = 0] = 1$, while for any $x \in K : \Pr[\overline{f}(x) = 0] = \Pr[\overline{f}(x) = 1] = \frac{1}{2}$ independently of other points. We will now show that any such $\overline{f} : \overline{K} \to \{0, 1\}$ is average Hölder smooth with respect to μ. Indeed, for every $x \in K$:

$$\Lambda_{\overline{f}}^\beta(x) = \sup_{x' \in \overline{K} \setminus \{x\}} \frac{|\overline{f}(x) - \overline{f}(x')|}{\rho(x, x')^\beta} \leq \frac{1}{\varepsilon/L} = \frac{L}{\varepsilon} ,$$

while

$$\Lambda_{\overline{f}}^\beta(x_0) = \sup_{x' \in \overline{K} \setminus \{x_0\}} \frac{|\overline{f}(x_0) - \overline{f}(x')|}{\rho(x_0, x')^\beta} \leq \frac{1}{\mathrm{diam}(\Omega)/4} = \frac{4}{\mathrm{diam}(\Omega)} ,$$

hence

$$\overline{\Lambda}_{\overline{f}}^\beta(x) = \mu(x_0)\Lambda_{\overline{f}}^\beta(x_0) + \sum_{x \in K} \mu(x)\Lambda_{\overline{f}}^\beta(x) \leq \frac{4}{D} + \frac{L}{2} \leq L .$$

Finally, we define the (random) function $f^* : \Omega \to [0, 1]$ to be the β-PMSE extension of $\overline{f}$ from $\overline{K}$ to Ω as defined in Definition A.2, and note that f^* satisfies the required smoothness assumption. Setting $\mathcal{D}$ over $\Omega \times [0, 1]$ to have marginal μ and $Y = f^*(X)$, we ensure that $\mathcal{D}$ is indeed realizable by $\overline{\mathrm{Höl}}_L^\beta(\Omega)$.

Now assume A is a learning algorithm which is given a sample S of size $|S| \leq \frac{|K|}{4\varepsilon}$ and produces $A(S) : \Omega \to [0, 1]$. We call a point $x \in K$ "misclassified" by the algorithm if $|A(S)(x) - f^*(x)| \geq \frac{1}{2}$, and denote the set of misclassified points by $M \subset K$. Recalling that $\forall x \in K : \Pr[f(x) = 0] = \Pr[f(x) = 1] = \frac{1}{2}$ independently, and that $\mu(x) = \frac{\varepsilon}{2|K|}$, we observe that with probability at least $\frac{1}{2}$ the algorithm will misclassify more than $|K|/2$ points.[7] Thus, we get that with probability at least $\frac{1}{2}$:

$$L_{\mathcal{D}}(A(S)) = \mathop{\mathbb{E}}_{X \sim \mu}[|A(S)(X) - f^*(X)|] \geq \sum_{x \in M} \mu(x) \cdot |A(S)(x) - f^*(x)| \geq \frac{|K|}{2} \cdot \frac{\varepsilon}{2|K|} \cdot \frac{1}{2} = \frac{\varepsilon}{8} .$$

[7] Indeed, denoting $C = K \setminus M$ we see that $\Pr[|C| \geq |K|/8] \leq \frac{8}{|K|} \cdot \mathbb{E}[|C|] = \frac{8}{|K|} \cdot \frac{|S|}{2} \cdot \mu(K) \leq \frac{8}{|K|} \cdot \frac{|K|}{8\varepsilon} \cdot \frac{\varepsilon}{2} = \frac{1}{2}$.

By rescaling ε, we see that in order to obtain $L_{\mathcal{D}}(A(S)) \leq \varepsilon$ the sample size must be of size

$$\Omega\left(\frac{|K|}{\varepsilon}\right) = \Omega\left(\frac{\mathcal{N}_{\Omega}((\varepsilon/L)^{1/\beta})}{\varepsilon}\right) .$$

B.6 Proofs from Section 7

Proof of Claim 7.1. Let $\beta \in (0,1)$. Consider the unit segment $\Omega = [0,1]$ with the standard metric, equipped with the probability measure μ with density $\frac{d\mu}{dx} = \frac{1}{Z}|x - \frac{1}{2}|^{\frac{\beta-1}{2}}$ (where $Z = \int_0^1 |x - \frac{1}{2}|^{\frac{\beta-1}{2}} < \infty$ is a normalizing constant). We examine the function $f(x) = \mathbf{1}[x > \frac{1}{2}]$ which is clearly not Hölder continuous since it is discontinuous. Furthermore,

$$\mu(\{x : \Lambda_f^1(x) \geq t\}) = \mu\left(\left\{\left|x - \frac{1}{2}\right| \leq \frac{1}{t}\right\}\right) = \frac{2}{Z}\int_0^{1/t} x^{\frac{\beta-1}{2}} dx \asymp t^{-\frac{\beta+1}{2}}$$
$$\implies \widetilde{\Lambda}_f^1 = \sup_{t>0} t \cdot \mu(\{x : \Lambda_f^1(x) \geq t\}) \asymp \sup_{t>0} t^{\frac{1-\beta}{2}} = \infty ,$$

hence $f \notin \widetilde{\mathrm{Lip}}_M(\Omega, \mu)$ for all $M > 0$. On the other hand, $\Lambda_f^\beta(x) = \frac{1}{|x-\frac{1}{2}|^\beta}$ so

$$\overline{\Lambda}_f^\beta = \int_0^1 \Lambda_f^\beta(x) d\mu = \frac{1}{Z}\int_0^1 \frac{|x - \frac{1}{2}|^{\frac{\beta-1}{2}}}{|x - \frac{1}{2}|^\beta} dx = \frac{1}{Z}\int_0^1 \frac{1}{|x - \frac{1}{2}|^{\frac{\beta+1}{2}}} dx \overset{(\beta<1)}{<} \infty ,$$

thus $f \in \overline{\mathrm{Höl}}_L^\beta(\Omega)$ for some $L < \infty$. Note that by normalizing the function, the claim holds even for $L = 1$.

Proof of Claim 7.2. Let $\beta \in (0,1)$. Consider the unit segment $\Omega = [0,1]$ with the standard metric, equipped with the probability measure μ with density $\frac{d\mu}{dx} = \frac{1}{Z}|x - \frac{1}{2}|^{\beta-1}$ (where $Z = \int_0^1 |x - \frac{1}{2}|^{\beta-1} dx < \infty$ is a normalizing constant). We examine the function $f(x) = \mathbf{1}[x > \frac{1}{2}]$. Note that for any $x \neq \frac{1}{2}$: $\Lambda_f^1(x) = \frac{1}{|x-\frac{1}{2}|}$, hence

$$\mu(\{x : \Lambda_f^1(x) \geq t\}) = \mu\left(\left\{x : |x - \frac{1}{2}| \leq \frac{1}{t}\right\}\right) = \frac{2}{Z}\int_0^{1/t} x^{\beta-1} dx \asymp t^{-\beta} .$$

This shows that

$$\widetilde{\Lambda}_f^1 = \sup_{t>0} t \cdot \mu(\{x : \Lambda_f^1(x) \geq t\}) \asymp \sup_{t>0} t^{1-\beta} = \infty ,$$

hence $f \notin \widetilde{\mathrm{Lip}}_M(\Omega, \mu)$ for all $M > 0$. Furthermore, for $x \neq \frac{1}{2}$: $\Lambda_f^\beta(x) = \frac{1}{|x-\frac{1}{2}|^\beta}$ so

$$\overline{\Lambda}_f^\beta = \int_0^1 \frac{1}{|x - \frac{1}{2}|^\beta} d\mu = \frac{1}{Z}\int_0^1 \frac{1}{|x - \frac{1}{2}|} dx = \infty ,$$

hence $f \notin \widetilde{\mathrm{Höl}}_M^\beta(\Omega, \mu)$ for all $M > 0$. On the other hand

$$\mu(\{x : \Lambda_f^\beta(x) \geq t\}) = \mu(\{|x - \frac{1}{2}| \leq t^{-1/\beta}\}) = \frac{2}{Z}\int_0^{t^{-1/\beta}} x^{\beta-1} dx \asymp t^{-1}$$
$$\implies \widetilde{\Lambda}_f^\beta = \sup_{t>0} t \cdot \mu(\{x : \Lambda_f^\beta(x) \geq t\}) < \infty ,$$

thus $f \in \widetilde{\mathrm{Höl}}_L^\beta(\Omega)$ for some $L < \infty$. Note that by normalizing the function, the claim holds even for $L = 1$.