Proceedings of Machine Learning Research vol 237:1–13, 2024 35th International Conference on Algorithmic Learning Theory

Efficient Agnostic Learning with Average Smoothness

Steve Hanneke STEVE.HANNEKE@GMAIL.COM
Purdue University

Aryeh Kontorovich KARYEH@CS.BGU.AC.IL
Ben-Gurion University of the Negev

Guy Kornowski GUY.KORNOWSKI@WEIZMANN.AC.IL
Weizmann Institute of Science

Editors: Claire Vernade and Daniel Hsu

Abstract

We study distribution-free nonparametric regression following a notion of average smoothness initiated by Ashlagi et al. (2021), which measures the "effective" smoothness of a function with respect to an arbitrary unknown underlying distribution. While the recent work of Hanneke et al. (2023) established tight uniform convergence bounds for average-smooth functions in the realizable case and provided a computationally efficient realizable learning algorithm, both of these results currently lack analogs in the general agnostic (i.e. noisy) case.

In this work, we fully close these gaps. First, we provide a distribution-free uniform convergence bound for average-smoothness classes in the agnostic setting. Second, we match the derived sample complexity with a computationally efficient agnostic learning algorithm. Our results, which are stated in terms of the intrinsic geometry of the data and hold over any totally bounded metric space, show that the guarantees recently obtained for realizable learning of average-smooth functions transfer to the agnostic setting. At the heart of our proof, we establish the uniform convergence rate of a function class in terms of its bracketing entropy, which may be of independent interest.

Keywords: agnostic learning, average smoothness, bracketing numbers, generalization, metric space

1. Introduction

Numerous frameworks in learning theory and statistics formalize the intuitive insight that "smooth functions are easier to learn than rough ones" (Györfi et al., 2002; Tsybakov, 2008; Giné and Nickl, 2021). The various measures of smoothness that were studied in a statistical context include the popular Lipschitz or Hölder seminorms; the bounded variation norm (Long, 2004); Sobolev, Sobolev-Slobodetskii and Besov norms (Nickl and Pötscher, 2007); averaged modulus of continuity (Sendov and Popov, 1988; Malykhin, 2010); and probabilistic Lipschitzness in the context of classification (Urner and Ben-David, 2013; Urner et al., 2013; Kpotufe et al., 2015).

In particular, a recent line of work (Ashlagi et al., 2021; Hanneke et al., 2023) studied a notion of *average smoothness* with respect to an arbitrary measure. Informally, the average smoothness is defined by considering the "local" Hölder (or Lipschitz) smoothness of a function at each point of the instance space, averaged with respect to the marginal distribution over the space; see Figure 1 for a simple illustration, and Section 2.1 for a formal definition. The main conclusion of the aforementioned works is that it is possible to guarantee statistical generalization solely in terms of the average smoothness for any underlying measure, effectively replacing the classic Hölder (or Lipschitz) constant with a much tighter distribution-dependent quantity. In particular, Hanneke

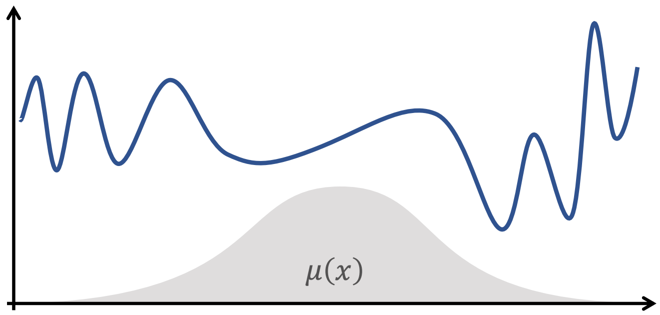

Figure 1: Illustration of a function and a measure μ exhibiting a large gap between "worst-case" smoothness (occurring in low density regions) and average-smoothness with respect to μ; figure taken from (Hanneke et al., 2023).

et al. (2023) proved a uniform convergence bound for the class of on-average-smooth functions in the realizable (i.e. noiseless) case, and complemented this result with an efficient realizable learning algorithm. With regard to the the general case of agnostic learning, the results of Hanneke et al. had some limitations. In particular, the general reduction from agnostic to realizable learning (Hopkins et al., 2022) deployed therein left two unfulfilled desiderata. From a statistical perspective, it remained open whether a function class with bounded average smoothness under some distribution μ is μ-Glivenko-Cantelli, namely that the excess risk decays uniformly over the class; only the existence of *some* returned predictor with small excess risk was established. On the computational side, the agnostic algorithm is highly inefficient: its runtime complexity is exponential in the sample size, in contrast with the polynomial-time realizable algorithm.

1.1. Our Contributions.

In this paper we study distribution-free agnostic learning of average-smooth functions, and address the issues raised above. Our main contributions can be summarized as follows:

- **Agnostic uniform convergence (Theorem 4 and Theorem 6).** We provide a distribution-free uniform convergence bound for the class of average-smooth functions in the agnostic setting (Theorem 6). This bound actually follows from a more general result, in which we bound the uniform convergence in terms of the bracketing entropy of the class (Theorem 4). The latter is widely applicable and may be of independent interest.

- **Efficient agnostic algorithm (Theorem 8).** We present a polynomial time algorithm for agnostic learning of on-average-smooth functions. The resulting sample complexity matches the aforementioned uniform convergence bound, which also matches that of the exponential-time agnostic algorithm of Hanneke et al. (2023). Furthermore, the algorithm's running time matches that of their efficient realizable learning algorithm.

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

Metric notions. 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. A metric space (Ω, ρ) is said to be *doubling* with constant $D \in \mathbb{N}$ if every ball $B \subset \Omega$ of radius r verifies $\mathcal{N}_B(r/2) \le D$. The *doubling dimension* is defined as $\inf_{D\in\mathbb{N}} \log_2 D$, where the infimum runs over all valid doubling constants for (Ω, ρ).

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 every 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 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 $\mathcal{N}_{\mathcal{F}}(\varepsilon) \le \mathcal{N}_{[]}(\mathcal{F}, \|\cdot\|, 2\varepsilon)$, 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).*

2.1. Average smoothness (Ashlagi et al., 2021; Hanneke et al., 2023).

The definition of average smoothness closely follows that given by Hanneke et al. (2023). 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

$$\|f\|_{\mathrm{Höl}^\beta} := \sup_{x\in\Omega} \Lambda_f^\beta(x) < \infty\ ;$$

the latter is known as the 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 $X \sim \mu$. We define

$$\overline{\Lambda}_f^\beta(\mu) := \mathop{\mathbb{E}}_{X\sim\mu}\left[\Lambda_f^\beta(X)\right] .$$

Notably

$$\overline{\Lambda}_f^\beta(\mu) \le \|f\|_{\mathrm{Höl}^\beta} , \tag{1}$$

where the gap can be *infinitely* large, as demonstrated by Hanneke et al. (2023). The notion of average smoothness induces the corresponding function class (alongside the classic "worst-case" one):

$$\mathrm{Höl}_H^\beta(\Omega) := \left\{f:\Omega\to[0,1]\ :\ \|f\|_{\mathrm{Höl}^\beta}\le H\right\},$$
$$\overline{\mathrm{Höl}}_H^\beta(\Omega,\mu) := \left\{f:\Omega\to[0,1]\ :\ \overline{\Lambda}_f^\beta(\mu)\le H\right\}.$$

We occasionally omit μ when it is clear from context. Note that for any measure μ :

$$\mathrm{Höl}_H^\beta(\Omega) \subset \overline{\mathrm{Höl}}_H^\beta(\Omega,\mu)$$

due to Eq. (1), where the containment is strict in general. The special case of $\beta = 1$ recovers the average-Lipschitz function class $\mathrm{Lip}_H(\Omega) \subset \overline{\mathrm{Lip}}_H(\Omega,\mu)$ studied by Ashlagi et al. (2021), while the general case $\beta \in (0,1]$ was studied by Hanneke et al. (2023).

In particular, we will now recall one of the main results of Hanneke et al. (2023) which establishes a bound on the bracketing entropy of average-smoothness classes. Crucially, the bound does not depend on μ, which allows to obtain distribution-free generalization guarantees.

Theorem 2 (Hanneke et al., 2023, Theorem 1) *For any metric probability space* (Ω,ρ,μ)*, any* $\beta\in(0,1]$ *and any* $0<\varepsilon<H$ *:*

$$\log\mathcal{N}_{[]}(\overline{\mathrm{Höl}}_H^\beta(\Omega,\mu),L_1(\mu),\varepsilon)\le\mathcal{N}_\Omega\left(\left(\frac{\varepsilon}{128H\log(1/\varepsilon)}\right)^{1/\beta}\right)\cdot\log\left(\frac{16\log_2(1/\varepsilon)}{\varepsilon}\right) .$$

Remark 3 (Weak average) *Hanneke et al. (2023) also considered the even larger space of functions which are weakly-average-smooth, namely such that* $\sup_{t>0} t\cdot\mu(x:\Lambda_f^\beta(x)\ge t)\le H$*. Note that this class is indeed larger than* $\overline{\mathrm{Höl}}_H^\beta(\Omega,\mu)$ *by Markov's inequality. The bracket entropy bound in Theorem 2 was actually proven for this even larger class. Consequently, all the uniform convergence results to appear in the next section also hold for this larger class. We choose to focus on the class* $\overline{\mathrm{Höl}}_H^\beta(\Omega,\mu)$ *throughout this paper for ease of presentation.*

3. Generalization bounds

Our first goal is to establish a uniform convergence result for the class $\overline{\mathrm{Höl}}_H^\beta(\Omega,\mu)$, which holds regardless of the distribution $\mathcal{D}$ whose marginal is μ (in particular, the bound does not depend on μ). Notably, a bound of this sort was previously established by Hanneke et al. (2023) only for $\mathcal{D}$ that are realizable by the function class, namely for which there exists an $f^*\in\overline{\mathrm{Höl}}_H^\beta(\Omega,\mu)$ with $L_{\mathcal{D}}(f^*)=0$.

In order to leverage Theorem 2 towards establishing an agnostic risk bound, we prove what is apparently a novel uniform deviation bound in terms of bracketing numbers:

Theorem 4 *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 $\Omega \times [0,1]$ 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 $f \in \mathcal{F},\ \alpha \geq 0$:*

$$|L_{\mathcal{D}}(f) - L_S(f)| \leq \alpha + O\left(\sqrt{\frac{\log \mathcal{N}_{[]}(\mathcal{F}, L_1(\mu), \alpha) + \log(1/\delta)}{n}}\right) .$$

Remark 5 (Other losses) *The proof of Theorem 4 is the only place throughout the paper that relies on the considered risk being with respect to the L_1 loss. In particular, in Eq. (5) we prove an analog of the contraction lemma (cf. Mohri et al., 2018, Lemma 5.7) for bracketing entropies with respect to the L_1 loss. This statement holds with essentially the same proof under mild assumption on the loss, e.g. as long as the loss $\ell(f(x), y)$ is symmetric with respect to exchanging its variables, monotone and Lipschitz with respect to $|f(x) - y|$ (with an incurred dependence on the Lipschitz constant). In particular, since the functions discussed in this paper are bounded, the results are readily extendable to L_p losses for any $p \in [1, \infty)$ (naturally yielding p-dependent rates due to the p-dependent Lipschitz constant).*

In our case of interest, plugging the bracket entropy bound for average smoothness classes from Theorem 2 into the uniform deviation bound in Theorem 4 yields the following:

Theorem 6 *For any metric space (Ω, ρ) and distribution $\mathcal{D}$ with marginal μ as above, it holds with probability at least $1-\delta$ over drawing a sample $S \sim \mathcal{D}^n$ that for all $f \in \overline{\text{Höl}}_H^\beta(\Omega, \mu),\ \alpha \geq 0$:*

$$|L_{\mathcal{D}}(f) - L_S(f)| = \alpha + \widetilde{O}\left(\sqrt{\frac{\mathcal{N}_\Omega\left(\left(\frac{\alpha}{128H\log(1/\alpha)}\right)^{1/\beta}\right) + \log(1/\delta)}{n}}\right) .$$

Remark 7 (Doubling metrics) *In most cases of interest, (Ω, ρ) is a doubling metric space of some dimension d,[1] 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 by plugging into Theorem 6 and optimizing over $\alpha \geq 0$ yields the simplified generalization bound*

$$\sup_{f \in \overline{\text{Höl}}_H^\beta(\Omega,\mu)} |L_{\mathcal{D}}(f) - L_{\mathcal{D}}(f)| = \widetilde{O}\left(\frac{H^{d/(d+2\beta)}}{n^{\beta/(d+2\beta)}}\right) .$$

Equivalently, $\sup_{f \in \overline{\text{Höl}}_H^\beta(\Omega,\mu)} |L_{\mathcal{D}}(f) - L_S(f)| \leq \varepsilon$ whenever $n \geq N$ for

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

up to a constant that depends (exponentially) on d, but is independent of H, ε.

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

4. Efficient agnostic learning algorithm

Having established the sample complexity required for controlling the excess risk uniformly over average-smooth functions, we turn to seeking an efficient agnostic regression algorithm that attains this sample complexity. We note that this is a nontrivial task due to the nature of the class $\overline{\mathrm{H\ddot{o}l}}_H^\beta(\Omega, \mu)$, which is *unknown to the learner*. Indeed, without knowledge of the underlying distribution, the learner cannot evaluate a candidate function's average smoothness with respect to the given distribution, thus a naive empirical-risk-minimization approach over the function class is inapplicable. The key to designing an average-smooth regression algorithm is the analysis of the *empirical smoothness* induced by the sample, namely the quantity

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

for any function $f : \Omega \to [0,1]$. Hanneke et al. (2023) proved a tail bound for the empirical smoothness in terms of the true average smoothness. Subsequently, their agnostic algorithm is a certain exhaustive search procedure over the space of empirically-smooth functions, and thus highly inefficient. In particular, the runtime of the algorithm is exponential in the sample size, which provided ample motivation to seek an efficient one.

In the following theorem we provide a polynomial-time algorithm that matches the same sample complexity, thus closing the exponential gap.

Theorem 8 *There is a polynomial time algorithm $\mathcal{A}$ such that for any metric space (Ω, ρ), any $\beta \in (0,1]$, $0 < \varepsilon < H$, and any distribution $\mathcal{D}$ over $\Omega \times [0,1]$, given a sample $S \sim \mathcal{D}^n$ of size $n \geq N$ where $N = N(\beta, \varepsilon, \delta)$ satisfies*

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

the algorithm constructs a hypothesis $f = \mathcal{A}(S)$ such that

$$L_{\mathcal{D}}(f) \leq \inf_{f^* \in \overline{\mathrm{H\ddot{o}l}}_H^\beta(\Omega,\mu)} L_{\mathcal{D}}(f^*) + \varepsilon$$

with probability at least $1 - \delta$.

Remark 9 (Doubling metrics) *As previously discussed, in most cases of interest we have $\mathcal{N}_\Omega(\varepsilon) \lesssim \left(\frac{1}{\varepsilon}\right)^d$ for some dimension $d \in \mathbb{N}$. That being the case, Theorem 8 yields the simplified sample complexity bound*

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

or equivalently

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

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

Remark 10 (Computational complexity) *The algorithm described in Theorem 8 involves a single preprocessing step with runtime $\widetilde{O}(n^{2\omega})$ where $\omega \approx 2.37$ is the current matrix multiplication exponent, 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).*

We will now outline the proof of Theorem 8, which appears in Section 5.2 along the full description of the algorithm. Denoting the Bayes-optimal risk by $L^* = \inf_{f^* \in \overline{\mathrm{H\ddot{o}l}}^{\beta}_{H}(\Omega,\mu)} L_{\mathcal{D}}(f^*)$, we assume without loss of generality (by a standard approximation argument) that the infimum is obtained, and let $f^* \in \overline{\mathrm{H\ddot{o}l}}^{\beta}_{H}(\Omega,\mu)$ be a function with $L_{\mathcal{D}}(f^*) = L^*$. Given a sample $(X_i, Y_i)_{i=1}^n \sim \mathcal{D}^n$, the algorithm first constructs labels $(\widehat{f}(X_i))_{i=1}^n$ such that

$$L_S(\widehat{f}) \leq L^* + \frac{\varepsilon}{3} \tag{2}$$

and

$$\widehat{\Lambda}^{\beta}_{\widehat{f}} \lesssim L \,. \tag{3}$$

We show that such a "relabeling" is obtainable by solving a linear program which minimizes the empirical error under the empirical smoothness constraint. This program is feasible since f^* satisfies both conditions with high probability. Indeed, f^* satisfies Eq. (2) by Theorem 6 (for large enough sample size), while Eq. (3) follows from the aforementioned tail bound of empirical smoothness. With these labels in hand, we invoke an approximate-extension procedure due to Hanneke et al. (2023) that extends $\widehat{f}$ to $f : \Omega \to [0,1]$ satisfying $L_S(f) \leq L_S(\widehat{f}) + \frac{\varepsilon}{3}$ and $\overline{\Lambda}^{\beta}_{f}(\mu) \lesssim \widehat{\Lambda}^{\beta}_{\widehat{f}}$ with high probability. Combining the latter property with Eq. (3) yields $\overline{\Lambda}^{\beta}_{f}(\mu) \lesssim H$. Thus, we have overall obtained some f in the average-smooth class (with a slightly inflated average-smoothness parameter) whose empirical risk is bounded according to Eq. (2) by

$$L_S(f) \leq L_S(\widehat{f}) + \frac{\varepsilon}{3} \leq L^* + \frac{2\varepsilon}{3} \,.$$

Finally, invoking Theorem 6 we conclude that the smooth-on-average f has small excess risk, resulting in

$$L_{\mathcal{D}}(f) \leq L_S(f) + \frac{\varepsilon}{3} \leq L^* + \varepsilon$$

with high probability, whenever the sample is large enough.

5. Proofs

5.1. Proof of Theorem 4

We start by denoting the loss class $\mathcal{L}_{\mathcal{F}} \subseteq [0,1]^{\Omega \times [0,1]}$:

$$\mathcal{L}_{\mathcal{F}} := \{\ell_f(x,y) := |f(x) - y| : f \in \mathcal{F}\} \,. \tag{4}$$

We will show that for any $\alpha > 0$, the bracketing entropy of $\mathcal{L}_{\mathcal{F}}$ is no larger than that of $\mathcal{F}$, namely

$$\mathcal{N}_{[]}(\mathcal{L}_{\mathcal{F}}, L_1(\mu), \alpha) \leq \mathcal{N}_{[]}(\mathcal{F}, L_1(\mu), \alpha) \,. \tag{5}$$

To that end, fix $\alpha > 0$, let $\mathcal{B}_\alpha$ be a minimal α-bracket of $\mathcal{F}$, and denote for any $f \in \mathcal{F}$ by $[f_L, f_U] \in \mathcal{B}_\alpha$ its associated bracket. For $\ell_f \in \mathcal{L}_\mathcal{F}$ as defined in Eq. (4), we define the bracket $[(\ell_f)_L, (\ell_f)_U]$ as

$$((\ell_f)_L(x,y), (\ell_f)_U)(x,y)) := \begin{cases} (0, f_U(x) - f_L(x)), & \text{if } f_L(x) \leq y \leq f_U(x) \\ (f_L(x) - y, f_U(x) - y), & \text{if } y < f_L(x) \\ (y - f_U(x), y - f_L(x)), & \text{if } y > f_U(x) \,. \end{cases}$$

Notice that this is indeed a valid bracket, since $f_L(x) \leq f(x) \leq f_U(x)$ implies that for any $(x, y) \in \Omega \times [0, 1]$:

$$\begin{aligned} (\ell_f)_L(x,y) &= \begin{cases} 0, & \text{if } f_L(x) \leq y \leq f_U(x) \\ f_L(x) - y, & \text{if } y < f_L(x) \\ y - f_U(x), & \text{if } y > f_U(x) \end{cases} \\ &\leq \begin{cases} |f(x) - y|, & \text{if } f_L(x) \leq y \leq f_U(x) \\ f(x) - y, & \text{if } y < f_L(x) \\ y - f(x), & \text{if } y > f_U(x) \end{cases} \\ &= |f(x) - y| = \ell_f(x,y) \,, \end{aligned}$$

and similarly

$$\begin{aligned} (\ell_f)_U(x,y) &= \begin{cases} f_U(x) - f_L(x), & \text{if } f_L(x) \leq y \leq f_U(x) \\ f_U(x) - y, & \text{if } y < f_L(x) \\ y - f_L(x), & \text{if } y > f_U(x) \end{cases} \\ &\geq \begin{cases} |f(x) - y|, & \text{if } f_L(x) \leq y \leq f_U(x) \\ f(x) - y, & \text{if } y < f_L(x) \\ y - f(x), & \text{if } y > f_U(x) \end{cases} \\ &= |f(x) - y| = \ell_f(x,y) \,. \end{aligned}$$

Moreover, by construction we see that for any $(x, y) \in \Omega \times [0, 1] : (\ell_f)_U(x, y) - (\ell_f)_L(x, y) = f_U(x) - f_L(x)$, hence

$$\|(\ell_f)_U - (\ell_f)_L\|_{L_1(\mathcal{D})} \leq \|f_U - f_L\|_{L_1(\mu)} \leq \alpha \,,$$

showing we indeed constructed an α-bracket. As it is clearly of size at most $|\mathcal{B}_\alpha|$, we proved Eq. (5).

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

$$\begin{aligned} L_\mathcal{D}(f) - L_S(f) &= \|\ell_f\|_{L_1(\mathcal{D})} - \|\ell_f\|_{L_1(\mathcal{D}_n)} \\ &\leq \|\ell_f - (\ell_f)_L\|_{L_1(\mathcal{D})} + \|(\ell_f)_L\|_{L_1(\mathcal{D})} - \|\ell_f\|_{L_1(\mathcal{D}_n)} \\ &\leq \|(\ell_f)_U - (\ell_f)_L\|_{L_1(\mathcal{D})} + \|(\ell_f)_L\|_{L_1(\mathcal{D})} - \|\ell_f\|_{L_1(\mathcal{D}_n)} \\ &\leq \alpha + \|(\ell_f)_L\|_{L_1(\mathcal{D})} - \|\ell_f\|_{L_1(\mathcal{D}_n)} \\ &\leq \alpha + \|(\ell_f)_L\|_{L_1(\mathcal{D})} - \|(\ell_f)_L\|_{L_1(\mathcal{D}_n)} \,, \end{aligned}$$

hence

$$\sup_{f\in\mathcal{F}} \left(L_{\mathcal{D}}(f) - L_S(f)\right) \leq \alpha + \max_{(\ell_f)_L} \left(\|(\ell_f)_L\|_{L_1(\mathcal{D})} - \|(\ell_f)_L\|_{L_1(\mathcal{D}_n)}\right) .$$

Similarly, we also have

$$\sup_{f\in\mathcal{F}} \left(L_S(f) - L_{\mathcal{D}}(f)\right) \leq \alpha + \max_{(\ell_f)_U} \left(\|(\ell_f)_U\|_{L_1(\mathcal{D}_n)} - \|(\ell_f)_U\|_{L_1(\mathcal{D})}\right) ,$$

thus overall

$$\begin{aligned}\sup_{f\in\mathcal{F}} |L_{\mathcal{D}}(f) - L_S(f)| \leq \alpha &+ \max_{(\ell_f)_L} \left(\|(\ell_f)_L\|_{L_1(\mathcal{D})} - \|(\ell_f)_L\|_{L_1(\mathcal{D}_n)}\right) \\ &+ \max_{(\ell_f)_U} \left(\|(\ell_f)_U\|_{L_1(\mathcal{D}_n)} - \|(\ell_f)_U\|_{L_1(\mathcal{D})}\right) .\end{aligned}$$

In order to bound the right-hand side, all that is left is a standard application of Hoeffding's inequality with a union bound over the finite bracket class, whose size is bounded by $\mathcal{N}_{[\,]}(\mathcal{F}, L_1(\mu), \alpha)$ due to Eq. (5). Minimizing over $\alpha > 0$ completes the proof.

5.2. Proof of Theorem 8

Algorithm 1 Approximate extension

1: **Input:** Sample $S = (X_i)_{i=1}^n$, labels $(\widehat{f}(X_i))_{i=1}^n$, exponent $\beta \in (0, 1]$, accuracy parameter $\gamma > 0$.
2: **Preprocessing:**
3: Sort $(X_1, \dots, X_n)$ according to

$$w(X_i) = \max_{j\neq i} \frac{\left|\widehat{f}(X_i) - \widehat{f}(X_j)\right|}{\rho(X_i, X_j)^\beta} .$$

4: Let $S' \subset \{X_1, \dots, X_n\}$ consist of the $n - \lfloor \gamma n \rfloor$ points with smallest $w(X_i)$ value.
5: Let $A \subset S'$ be a $\gamma^{1/\beta}$ net of S'.
6: **Inference:**
7: For any $x \in \Omega$, compute

$$(u^*, v^*) = \operatorname*{arg\,max}_{(u,v)\in A\times A} \frac{\widehat{f}(v) - \widehat{f}(u)}{\rho(x, u)^\beta + \rho(x, v)^\beta}$$

and set

$$f(x) := \widehat{f}(u^*) + \frac{\rho(x, u^*)^\beta}{\rho(x, u^*)^\beta + \rho(x, v^*)^\beta} (\widehat{f}(v^*) - \widehat{f}(u^*)) .$$

We will state two propositions due to Hanneke et al. (2023) which, together with Theorem 6, will serve as the main components of the proof.

Proposition 11 (Hanneke et al., 2023) *Let $f : \Omega \to [0, 1]$ and μ be any distribution over Ω. Then with probability at least $1 - \delta$ over drawing a sample $(X_i)_{i=1}^n \sim \mu^n$ it holds that*

$$\widehat{\Lambda}_f^\beta \leq 5\log^2(2n/\delta)\overline{\Lambda}_f^\beta(\mu) \, .$$

Proposition 12 (Hanneke et al., 2023) *Algorithm 1 is an algorithm with $\widetilde{O}(n^2)$ preprocessing time and $O(n^2)$ inference time, that given any $\gamma > 0$, 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 $1 - \delta$:

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

We are now ready to prove Theorem 8. Let $\delta' = \frac{\delta}{3}$, and fix $\alpha, \gamma > 0$ to be determined later. Denote $L^* = \inf_{f\in\overline{\text{Höl}}_H^\beta(\Omega,\mu)} L_{\mathcal{D}}(f)$, and let $f^* \in \overline{\text{Höl}}_H^\beta(\Omega, \mu)$ be such that $L_{\mathcal{D}}(f^*) \leq L^* + \alpha$. We will now describe two desirable events that hold with high probability over drawing the sample $S \sim \mathcal{D}^n$, which we will condition on throughout the rest of the proof. Consider the event in which

$$\widehat{\Lambda}_{f^*}^\beta \leq \widehat{H} := 5\log^2(2n/\delta')H \, , \tag{6}$$

and note that this event holds with probability at least $1 - \delta'$ according to Proposition 11. Further consider the event in which for all $f \in \overline{\text{Höl}}_{5\widehat{H}}^\beta(\Omega, \mu)$:

$$|L_{\mathcal{D}}(f) - L_S(f)| = \alpha + \widetilde{O}\left(\sqrt{\frac{\mathcal{N}_\Omega\left(\left(\frac{\alpha}{640\widehat{H}\log(1/\alpha)}\right)^{1/\beta}\right) + \log(1/\delta')}{n}}\right) , \tag{7}$$

and note that this event holds with probability at least $1 - \delta'$ according to Theorem 6. In particular, since $f^* \in \overline{\text{Höl}}_H^\beta(\Omega, \mu) \subset \overline{\text{Höl}}_{5\widehat{H}}^\beta(\Omega, \mu)$, we get that as long as

$$n = \widetilde{\Omega}\left(\frac{\mathcal{N}_\Omega\left(\left(\frac{\alpha}{640\widehat{H}\log(1/\alpha)}\right)^{1/\beta}\right) + \log(1/\delta')}{\alpha^2}\right) , \tag{8}$$

it holds that

$$L_S(f^*) \leq L_{\mathcal{D}}(f^*) + 2\alpha \leq L^* + 3\alpha \, . \tag{9}$$

Thus, by solving the following feasible linear program over the variables $\left(\widehat{f}(X_i), \mathrm{err}_i, \widehat{H}_i\right)_{i=1}^n$:

$$\begin{aligned}
\text{Minimize} \quad & \sum_{i=1}^n \mathrm{err}_i \\
\text{subject to} \quad & \mathrm{err}_i \geq \left|\widehat{f}(X_i) - Y_i\right| && \forall i \in [n] && (10)\\
& 0 \leq \widehat{f}(X_i) \leq 1 && \forall i \in [n] \\
& \frac{1}{n}\sum_{i=1}^n \widehat{H}_i \leq 5\widehat{H} && \forall i \in [n] && (11)\\
& |\widehat{f}(X_i) - \widehat{f}(X_j)| \leq \widehat{H}_i \cdot \rho(X_i, X_j)^\beta && \forall i,j \in [n] : X_i \neq X_j && (12)
\end{aligned}$$

it is possible to find $(\widehat{f}(X_1), \ldots, \widehat{f}(X_n))$ so that

$$L_S(\widehat{f}) = \sum_{i=1}^n \left|\widehat{f}(X_i) - Y_i\right| \overset{\text{Eq. (10)}}{\leq} \sum_{i=1}^n \mathrm{err}_i \overset{(\star)}{\leq} L^* + 4\alpha$$

and

$$\widehat{\Lambda}^\beta_{\widehat{f}} = \frac{1}{n}\sum_{i=1}^n \max_{X_j \neq X_i} \frac{|\widehat{f}(X_i) - \widehat{f}(X_j)|}{\rho(X_i, X_j)^\beta} \overset{\text{Eq. (12)}}{\leq} \frac{1}{n}\sum_{i=1}^n \widehat{H}_i \overset{\text{Eq. (11)}}{\leq} 5\widehat{H} \ ,$$

within polynomial time. Indeed, the feasibility is observed by considering the variable assignment

$$\widehat{f}(X_i) = f^*(X_i) \tag{13}$$

$$\mathrm{err}_i = |f^*(X_i) - Y_i| \tag{14}$$

$$\widehat{H}_i = \max_{X_j \neq X_i} \frac{|f^*(X_i) - f^*(X_j)|}{\rho(X_i, X_j)^\beta} \ , \tag{15}$$

since Eqs. (13) and (14) imply Eq. (10); Eqs. (13) and (15) imply Eq. (12); and Eq. (6) implies Eq. (11). Moreover, $(\star)$ follows as long as the program is solved up to accuracy at most α due to Eq. (9). The runtime required for solving the program with $O(n^2)$ constraints up to accuracy at most α is bounded, according to the currently best known complexity of linear programming, by $\widetilde{O}\left((n^2)^\omega\right) = \widetilde{O}\left(n^{2\omega}\right)$ where $\omega \approx 2.37$ is the current matrix multiplication exponent (Cohen et al., 2021).

With such $\widehat{f}$ in hand, we can apply Algortithm 1 in order to obtain $f : \Omega \to [0,1]$, whose guaranteed by Proposition 12 to satisfy with probability at least $1 - \delta'$:

$$L_S(f) \leq L_S(\widehat{f}) + \gamma(1 + 2\widehat{\Lambda}^\beta_{\widehat{f}}) \leq L^* + 4\alpha + \gamma(1 + 5\widehat{H})$$

and

$$\overline{\Lambda}^\beta_f(\mu) \leq 5\widehat{\Lambda}^\beta_{\widehat{f}} \leq 5\widehat{H} = 25\log^2(2n/\delta')H \ .$$

By Eq. (7) and Eq. (8), the latter property ensures that

$$L_{\mathcal{D}}(f) \leq L_S(f) + 2\alpha \leq L^* + 6\alpha + \gamma(1 + 5\widehat{H}) \ .$$

Setting

$$\alpha = \frac{\varepsilon}{12}\,, \quad \gamma = \frac{\varepsilon}{2+10\widehat{H}} = \widetilde{\Theta}\left(\frac{\varepsilon}{H}\right)\,,$$

and applying the union bound yields

$$L_{\mathcal{D}}(f) \leq L^* + \varepsilon$$

with probability at least $1 - 3\delta' = 1 - \delta$.

Acknowledgments

AK is partially supported by the Israel Science Foundation (grant No. 1602/19), an Amazon Research Award, and the Ben-Gurion University Data Science Research Center. GK is supported by an Azrieli Foundation graduate fellowship.

References

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.

Michael B Cohen, Yin Tat Lee, and Zhao Song. Solving linear programs in the current matrix multiplication time. *Journal of the ACM (JACM)*, 68(1):1–39, 2021.

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.

Steve Hanneke, Aryeh Kontorovich, and Guy Kornowski. Near-optimal learning with average hölder smoothness. In *Neural Information Processing Systems*, 2023.

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.

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.

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.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

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.

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 Wulff, 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.