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 $\Lambda^{\beta}_{F_A}(x) \leq \Lambda^{\beta}_{\widetilde{F}}(x)$ 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. $\|f|_A\|_{\mathrm{Höl}^\beta} < \infty$ *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 $x \in \Omega$: $\Lambda^{\beta}_{F_A}(x) \leq \Lambda^{\beta}_{f}(x)$. 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 $\sup_{x \in \Omega} R^*_x < \infty \iff \|f\|_{\mathrm{Höl}^\beta} < \infty$.

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

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

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

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

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 $N = \{x_1, \dots, x_{|N|}\}$ 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 $\Pi = \{C_1, \dots, C_{|N|}\}$ 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 $\{0, \frac{\varepsilon'}{2}, \varepsilon', \frac{3\varepsilon'}{2}, \dots, 1\}$.
- 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 $m = 2, \dots, K$ 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 $C_i \in S_m$: $l|_{C_i} \in \{0, \frac{1}{2^m}, \frac{2}{2^m}, \dots, \frac{2^m-2}{2^m}\},\ ,\ u|_{C_i} = l + \frac{1}{2^{m-1}}$.
- In the "remaining" cells $S_{K+1} := \Pi \setminus \bigcup_{j \leq K} S_j$ set for any $C_i \in S_{K+1}$:

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

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 $\{0, \frac{\varepsilon'}{2}, \varepsilon', \dots, 1\}$, 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 $l \leq f \leq u$. To that end, let $f \in \widetilde{\text{Höl}}_L^\beta(\Omega, \mu)$. Denote

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

and notice that $\bigcup\{C_i \in S_1^f\} \subseteq \{x : \Lambda_f^\beta(x) \geq L/\varepsilon'\} \implies \mu(\bigcup\{C_i \in S_1^f\}) \leq \varepsilon'$. 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

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

and notice that $\bigcup\{C_i \in S_m^f\} \subseteq \{x : \Lambda_f^\beta(x) \geq L/(2^{m-1}\varepsilon')\} \implies \mu(\bigcup\{C_i \in S_m^f\}) \leq 2^{m-1}\varepsilon'$. 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

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

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

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

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 $\alpha_i \leq f|_{C_i} \leq \alpha_i + \frac{1}{2^{m-1}}$, 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 $S_{K+1}^f := \Pi \setminus \bigcup_{j \leq K} S_j^f$ 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

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

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

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

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 $\alpha_i \leq f|_{C_i} \leq \alpha_i + \frac{1}{2^K}$, 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

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

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 $\mathrm{Var}(l_{\ell_f}) \leq \|l^2_{\ell_f}\|_{L_1(\mu)} \leq \|l_{\ell_f}\|_{L_1(\mu)}$, 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*

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

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

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

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

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

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

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

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

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 $1 - \frac{Ne^\alpha}{(1+\alpha)^{1+\alpha}}$:

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

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 $1 - \delta/4$:

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

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* $\mathrm{diam}(E)^\beta \le \frac{1}{2}\min_{x \ne x' \in A} \rho(x,x')^\beta$. *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 $2\mathrm{diam}(E)^\beta \le \rho(v_x^*, u_x^*)^\beta \le \rho(v_x^*, x)^\beta + \rho(u_x^*, x)^\beta$, hence

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

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

□

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

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

Lemma B.6 (Ashlagi et al., 2021, Lemma 17)**.** $\|f\|_{\mathrm{H\ddot{o}l}^\beta} \leq \frac{2L}{\gamma}$.

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:

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

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 $\Lambda_f^\beta(x') \leq 2\min_{x\in S_x\cap C}\Lambda_f^\beta(x)$. Thus, under the high probability event described in Lemma B.5 we have

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

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 $\widehat{\Lambda}_{\widehat{f}}^\beta \leq$

$5\log^2(4n/\delta)L$. 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 $\overline{\text{Höl}}^\beta_L(\Omega, \mu)$ 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 $x \in K$:

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

while

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

hence

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

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 $\overline{\text{Höl}}^\beta_L(\Omega)$.

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 $f \in \overline{\mathrm{H\ddot{o}l}}_L^\beta(\Omega)$ 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 $f \notin \widetilde{\mathrm{H\ddot{o}l}}_M^\beta(\Omega,\mu)$ 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 $f \in \widetilde{\mathrm{H\ddot{o}l}}_L^\beta(\Omega)$ for some $L < \infty$. Note that by normalizing the function, the claim holds even for $L = 1$.