Bounding the Expectation of the Supremum of Empirical Processes Indexed by Hölder Classes

N. Schreuder*

CREST, ENSAE, Institut Polytechnique de Paris, 91120 Palaiseau, France

Abstract—In this note, we provide upper bounds on the expectation of the supremum of empirical processes indexed by Hölder classes of any smoothness and for any distribution supported on a bounded set in $\mathbb{R}^d$. These results can alternatively be seen as non-asymptotic risk bounds, when the unknown distribution is estimated by its empirical counterpart, based on n independent observations, and the error of estimation is quantified by integral probability metrics (IPM). In particular, IPM indexed by Hölder classes are considered and the corresponding rates are derived. These results interpolate between two well-known extreme cases: the rate $n^{-1/d}$ corresponding to the Wasserstein-1 distance (the least smooth case) and the fast rate $n^{-1/2}$ corresponding to very smooth functions (for instance, functions from a RKHS defined by a bounded kernel).

1. INTRODUCTION

In many problems of mathematical statistics and learning theory, a crucial step is to understand how well the empirical distribution of a sample approximates the underlying true distribution. The theory of empirical processes is devoted to this question. There are many papers and books treating this and related problems, both from asymptotic and nonasymptotic points of view; see, for instance, del Barrio et al. [5], van der Vaart and Wellner [20]. Among many remarkable achievements of the theory of empirical processes, there are two results that have been particularly often evoked and used in the recent literature in statistics and machine learning.

To quickly present these two results, let us give some details on the framework. It is assumed that n independent copies $X_1, \ldots, X_n$ of a random variable X taking its values in the d-dimensional hypercube $[0,1]^d$ are observed. The aforementioned two results characterize the order of magnitude of supremum of the empirical process $\mathbb{X}_n(f) = \frac{1}{n} \sum_{i=1} f(X_i) - \mathbb{E}[f(X)]$ over some class of functions $\mathcal{F}$. More precisely, the first result established by [6] states that $\sup_{f \in \mathsf{Lip(1)}} \mathbb{X}_n(f)$ is of order $O(n^{-1/d})$, where $\mathsf{Lip(1)}$ is the set of all the Lipschitz-continuous functions with Lipschitz constant 1. The second result [3, Lemma 1], tells us that if $\mathcal{F}$ contains functions that are smooth enough, for instance functions that are in a finite ball of a RKHS defined by a bounded kernel, then $\sup_{f \in \mathcal{F}} \mathbb{X}_n(f)$ is of order $O(n^{-1/2})$, i.e., the same order as in the case when $\mathcal{F}$ contains only one function.

The main result of this note provides an interpolation between the two aforementioned results. Roughly speaking, it shows that if $\mathcal{F}$ is the class of functions defined on $[0,1]^d$ that are Hölder-continuous for a given constant L and a given order $\alpha > 0$, then the supremum of the empirical process over $\mathcal{F}$ is of order $O(n^{-(\frac{\alpha}{d} \wedge \frac{1}{2})})$ with an additional slowly varying factor $\log n$ when $\alpha = d/2$. Clearly, when $\alpha = 1$ this coincides with the result from [6], while for $\alpha \geqslant d/2$ we get the fast and dimension-free rate $n^{-1/2}$, up to a log factor.

*E-mail: nicolas.schreuder@ensae.fr

The rest of this note is organized as follows. We complete this introduction by providing all the important notations used throughout this note. Section 2 is devoted to presenting and formally defining Hölder classes and Integral Probability Metrics (IPM). In Section 3, we expose some important concepts and results from empirical process theory needed for our proofs. We end this note by stating our main theorem in Section 4. Some extensions are mentioned in Section 5. The proofs are postponed to the Appendix.

Notations

A multi-index $\mathbf{k}$ is a vector with integer coordinates $(k_1, \ldots, k_d)$. We write $|\mathbf{k}| = \sum_{i=1}^{d} k_i$. For a given multi-index $\mathbf{k} = (k_1, \ldots, k_d)$, we define the differential operator

$$D^{\mathbf{k}} = \frac{\partial^{|\mathbf{k}|}}{\partial x_1^{k_1} \ldots \partial x_d^{k_d}}.$$

For any positive real number x, $\lfloor x \rfloor$ denotes the largest integer strictly smaller than x. We let $\mathcal{X}$ be a convex bounded set in $\mathbb{R}^d$ with non-empty interior. We assume that all the functions and function classes considered in this note are supported on the bounded set $\mathcal{X}$. For any integer k, we denote by $C^k(\mathcal{X}, \mathbb{R})$ the class of real-valued functions with domain $\mathcal{X}$ which are k-times differentiable with continuous k-th differentials. For any real-valued bounded function f on $\mathcal{X}$, we let $||f||_\infty := \sup_{x \in \mathcal{X}} |f(x)| \in [0, +\infty)$. Note that we can consider the essential supremum instead of the supremum over $\mathcal{X}$ in which case our results would hold almost surely. We let $|| \cdot ||$ denote some norm on $\mathbb{R}^d$. We denote by $\sigma_1, \ldots, \sigma_n$ i.i.d. Rademacher random variables, i.e., discrete random variables such that $\mathbb{P}(\sigma_1 = 1) = \mathbb{P}(\sigma_1 = -1) = 1/2$ which are independent of any other source of randomness. We use the convention $1/0 = +\infty$.

2. A PRIMER ON HÖLDER CLASSES AND INTEGRAL PROBABILITY METRICS

In this section we define Hölder classes of functions and integral probability metrics. We then discuss some properties of these notions and highlight their role in statistics and statistical learning theory.

2.1. Hölder Classes

A central problem in nonparametric statistics is to estimate a function belonging to an infinite-dimensional space (e.g., density estimation, regression function estimation, hazard function estimation), see Tsybakov [19] for an introduction to the topic of nonparametric estimation. To obtain nontrivial rates of convergence, some kind of regularity is assumed on the function of interest. It can be expressed as conditions on the function itself, on its derivatives, on the coefficients of the function in a given basis, etc. Hölder classes are one of the most common classes considered in the nonparametric estimation literature, they form a natural extension of Lipschitz-continuous functions and can be formalised with the following simple conditions. For any real number $\alpha > 0$, we define the Hölder norm of smoothness α of a $\lfloor \alpha \rfloor$-times differentiable function f as

$$||f||_{\mathcal{H}^\alpha} := \max_{|k| \leqslant \lfloor \alpha \rfloor} ||D^k f||_\infty + \max_{|k| = \lfloor \alpha \rfloor} \sup_{x \neq y} \frac{|D^k f(x) - D^k f(y)|}{||x - y||^{\alpha - \lfloor \alpha \rfloor}}.$$

The Hölder ball of smoothness α and radius $L > 0$, denoted by $\mathcal{H}^\alpha(L)$, is then defined as the class of $\lfloor \alpha \rfloor$-times continuously differentiable functions with Hölder norm bounded by the radius L:

$$\mathcal{H}^\alpha(L) = \left\{ f \in C^{\lfloor \alpha \rfloor}(\mathcal{X}, \mathbb{R}) \mid ||f||_{\mathcal{H}^\alpha} \leqslant L \right\}.$$

2.2. Integral Probability Metrics

The class $\mathcal{H}^1(1)$ of 1-Lipschitz functions has received a lot of attention in the optimal transport literature; see [13] for an overview of the topic of mathematical optimal transport. This interest comes from the Kantorovitch duality, which implies that the Wasserstein-1 distance (also known as the earth mover's distance) can be expressed, for any probability measures P, Q, as a supremum of some functional over 1-Lipschitz functions:

$$W_1(P,Q) = \sup_{f \in \mathcal{H}^1(1)} |\mathbb{E}_{X \sim P} f(X) - \mathbb{E}_{Y \sim Q} f(Y)|.$$

More generally, for a given class $\mathcal{F}$ of bounded functions, one can define a pseudo-metric on the space of probability measures, the integral probability metric (IPM) induced by the class $\mathcal{F}$, as

$$d_{\mathcal{F}}(P,Q) = \sup_{f \in \mathcal{F}} |\mathbb{E}_{X \sim P} f(X) - \mathbb{E}_{Y \sim Q} f(Y)|.$$

The literature on IPM has recently been boosted by the advent of adversarial generative models [1, 8]. A reason for this is that an IPM can be seen as an adversarial loss: to compare two probability distributions, it seeks for the function which discriminates the most the two distributions in expectation. Initially studied by the deep learning community, impressive empirical results obtained by adversarial generative models on several tasks such as image generation led statisticians to study it theoretically [3, 4, 10] (see also Sriperumbudur et al. [18] for statistical results on IPM in a general framework). Since, as pointed out earlier, Lipschitz functions are also Hölder, one can wonder what happens for IPM indexed by general Hölder classes. Such IPM already appeared in the literature: Scetbon et al. [14] showed that α-Hölder IPM with smoothness $\alpha \leqslant 1$ correspond to the cost of a generalized optimal transport problem.

To further motivate our study, let us consider the abstract problem of minimum distance estimation: for a given probability measure P, find a distribution Q in a given set of probability measures $\mathcal{Q}$ such that Q is close to P under the metric $d_{\mathcal{F}}$:

$$\min_{Q \in \mathcal{Q}} d_{\mathcal{F}}(Q, P). \tag{1}$$

For example, when $\mathcal{F}$ is taken to be the class of 1-Lipschitz function, this problem is known as minimum Kantorovitch estimation [2]. In statistics, the probability P is usually unknown and one is only given i.i.d. samples $X_1, \ldots, X_n$ from the probability distribution P. A natural strategy is then to employ the empirical distribution $P_n = 1/n \sum_{i=1}^{n} \delta_{X_i}$ as a proxy for the theoretical distribution and instead of (1) solve the problem:

$$\min_{Q \in \mathcal{Q}} d_{\mathcal{F}}(Q, P_n). \tag{2}$$

Since the triangle inequality yields

$$|d_{\mathcal{F}}(Q, P) - d_{\mathcal{F}}(Q, P_n)| \leqslant d_{\mathcal{F}}(P, P_n) = \sup_{f \in \mathcal{F}} \left| \frac{1}{n} \sum_{i=1}^{n} f(X_i) - \mathbb{E} f(X) \right|,$$

one question of interest is to measure how fast the empirical measure approximates the true measure under the IPM $d_{\mathcal{F}}$. If the rates are fast, we do not loose much by considering the empirical problem (2) instead of the theoretical one of (1). However if the rates are slow, one cannot expect the distances of the solutions to the measure P to be close. We will see in the next section that the latter expression corresponds to the supremum of the empirical process indexed by the class $\mathcal{F}$, it will enable us to leverage the rich literature on empirical processes to obtain rates of convergence for $d_{\mathcal{F}}(P, P_n)$.

3. EMPIRICAL PROCESSES, METRIC ENTROPY AND DUDLEY'S BOUNDS

This section provides a short account of the notions and tools from the theory of empirical processes which are necessary for stating and establishing the main result.

3.1. Empirical Processes

Empirical process are ubiquitous in statistical learning theory, we refer the reader to [7, 9] for a general presentation of results on empirical processes and their link with statistics and learning theory. For clarity, we begin by recalling the definition of an empirical process.

Definition 1. *Let $\mathcal{F}$ be a class of real-valued functions $f\colon \mathcal{X} \to \mathbb{R}$, where $(\mathcal{X}, \mathcal{A}, P)$ is a probability space. Let X be a random point in $\mathcal{X}$ distributed according to the distribution P and let $X_1, \ldots, X_n$ be independent copies of X. The random process $\big(\mathbb{X}_n(f)\big)_{f\in\mathcal{F}}$ defined by*

$$\mathbb{X}_n(f) := \frac{1}{n}\sum_{i=1}^{n} f(X_i) - \mathbb{E}f(X),$$

is called an empirical process indexed by $\mathcal{F}$.

In our case, we are interested in controlling the (expectation of the) supremum of an empirical process, a common case in the literature. Most of the time, the first step to apply for achieving this goal is to "symmetrize" the empirical process as allowed by the following lemma. Let $\widehat{R}_n(\mathcal{F})$ be the empirical Rademacher complexity of function class $\mathcal{F}$, defined as

$$\widehat{R}_n(\mathcal{F}) = \mathbb{E}\left[\sup_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^{n} \sigma_i f(X_i) \,\Big|\, X_1, \ldots, X_n\right].$$

Lemma 1 (Symmetrization). *For any class $\mathcal{F}$ of P-integrable functions,*

$$\mathbb{E}\left[\sup_{f\in\mathcal{F}} |\mathbb{X}_n(f)|\right] \leqslant 2\mathbb{E}\big[\widehat{R}_n(\mathcal{F})\big].$$

The advantage of Rademacher processes is that, regardless of the distribution of the random variable X and the function class $\mathcal{F}$, for a fixed sample $X_1, \ldots, X_n$, the random variable $\sum_{i=1}^{n} \sigma_i f(X_i)$ has a sub-Gaussian behavior, in the following sense.

Definition 2 (Sub-Gaussian behavior). *A centered random variable Y has a sub-Gaussian behavior if there exists a positive constant σ such that*

$$\mathbb{E}e^{\lambda Y} \leqslant e^{\lambda^2 \sigma^2/2}, \quad \forall \lambda \in \mathbb{R}.$$

In that case, we define the sub-Gaussian norm[1] of Y as

$$\|Y\|_{\psi_2} = \inf\left\{t > 0 : \mathbb{E}e^{Y^2/t^2} \leqslant 2\right\}.$$

Having a sub-Gaussian behavior essentially means to be at least as concentrated as a Gaussian random variable around its mean. Our definition is equivalent to the tail inequalities

$$\mathbb{P}(|Y| > t) \leqslant 2e^{-t^2/(2\sigma^2)}, \quad \forall t > 0.$$

This type of behavior will be crucial to obtain the main result of this note. Indeed, as we will see, the behavior of the supremum of an empirical process (and more generally a stochastic process) which has sub-Gaussian increments exclusively depends on the topology of the space by which the process is indexed.

---

[1] See [21, Section 2.5] for the link between definitions of sub-Gaussian random variables (bound on moment-generating function, tail inequalities, ...) and the Orlicz norm ψ_2.

3.2. Metric Entropy

Let (T, d) be a totally bounded metric space, i.e., for every real number $\varepsilon > 0$, there exists a finite collection of open balls of radius ε whose union contains T. We give a formal definition of such finite collections.

Definition 3. *Given $\varepsilon > 0$, a subset $T_\varepsilon \subset T$ is called an ε-cover of T if for every $t \in T$, there exists $s \in T_\varepsilon$ such that $d(s, t) \leqslant \varepsilon$.*

Note that adding any point to an ε-cover still yields an ε-cover. Thus we can look for ε-covers of a set with smallest cardinality, which we call covering number.

Definition 4. *The ε-covering number of T, denoted by $\mathcal{N}(T, d, \varepsilon)$, is the cardinality of the smallest ε-cover of T, that is*

$$\mathcal{N}(T, d, \varepsilon) := \min\left\{|T_\varepsilon| : T_\varepsilon \text{ is an } \varepsilon\text{-cover of } T\right\}.$$

The metric entropy of T is given by the logarithm of the ε-covering number.

Remark 1. *A totally bounded metric space (T, d) is pre-compact in the sense that its closure is compact. The metric entropy (or entropic numbers) of (T, d) can then be seen as some measure of compactness of the space. Indeed, $\mathcal{N}(T, d, \varepsilon)$ quantifies precisely how many balls of radius ε are needed to cover the whole space T.*

Entropic numbers for Hölder classes are known and can be found in e.g., Shiryayev [15], van der Vaart and Wellner [20].

Theorem 1 (Theorem 2.7.3 in [20]). *Let $\mathcal{X}$ be a bounded, convex subset of $\mathbb{R}^d$ with nonempty interior. There exists a constant $K_{\alpha, d}$ depending only on α and d such that, for every $\varepsilon > 0$,*

$$\log \mathcal{N}(\mathcal{H}^\alpha(1), ||\cdot||_\infty, \varepsilon) \leqslant K_{\alpha, d} \lambda_d(\mathcal{X}^1) \varepsilon^{-d/\alpha},$$

where λ_d is the d-dimensional Lebesgue measure and $\mathcal{X}^1$ is the 1-blowup of $\mathcal{X}$: $\mathcal{X}^1 = \{y : \inf_{x \in \mathcal{X}} ||y - x|| < 1\}$.

3.3. Dudley's Bound and Its Refined Version

We now present classic results which show the link between the topology of the indexing set and the behavior of the supremum of the corresponding empirical process. Following [21, Definition 8.1.1], for $K \geqslant 0$, we say that a random process $(X_t)_{t \in T}$ on a metric space (T, d) has K-sub-Gaussian increments if

$$||X_t - X_s||_{\psi_2} \leqslant Kd(t, s), \quad \text{for all} \quad t, s \in T.$$

Theorem 2 (Dudley's inequality). *Let $(X_t)_{t \in T}$ be a mean-zero random process on a metric space (T, d) with K-sub-Gaussian increments. Then*

$$\mathbb{E}\left[\sup_{t \in T} X_t\right] \leqslant CK \int\limits_0^{+\infty} \sqrt{\log \mathcal{N}(T, d, \varepsilon)}\, d\varepsilon$$

for some universal constant $C > 0$.

One drawback of Dudley's bound is that the integral on the right hand side may diverge if the metric entropy of T tends to infinity at a very fast rate when $\varepsilon \to 0$. For example, when the metric entropy is upper bounded by $\varepsilon^{-\gamma}$, as it was seen to be the case with $\gamma = d/\alpha$ for α-Hölder-smooth d-variate functions, the integral converges if and only if $\gamma < 2$.

An improvement of Dudley's bound in the case where the process X_t is a Rademacher average indexed by a class of functions $\mathcal{F}$—circumventing the problem of divergence of the integral—was proposed by [17, Lemma A.3] (see also Srebro and Sridharan [16]). Before stating the theorem, let us recall the definition of the $L_2(P_n)$ norm of a function f:

$$||f||_{L_2(P_n)}^2 = \int\limits_{\mathcal{X}} f^2 dP_n = \frac{1}{n} \sum_{i=1}^n f(X_i)^2.$$

Theorem 3. *Let $\mathcal{F} \subset \{f : \mathcal{X} \to \mathbb{R}\}$ be any class of measurable functions containing the uniformly zero function and let $S_n(\mathcal{F}) = \sup_{f \in \mathcal{F}} ||f||_{L_2(P_n)}$. We have*

$$\widehat{R}_n(\mathcal{F}) \leqslant \inf_{\tau > 0} \left\{ 4\tau + \frac{12}{\sqrt{n}} \int_{\tau}^{S_n(\mathcal{F})} \sqrt{\log \mathcal{N}(\mathcal{F}, L_2(P_n), \varepsilon)} d\varepsilon \right\}.$$

Note that the refined Dudley bound gives an upper bound on the empirical Rademacher process and depends on the metric entropy with respect to the empirical norm $L_2(P_n)$. The following simple lemma shows that the $L_2(P_n)$-norm can be replaced by the supremum-norm in the refined Dudley bound.

Lemma 2. *Let $\mathcal{F}$ be any class of bounded functions defined on $\mathcal{X}$. For any sample $X_1, \dots, X_n$, let $\mathcal{F}_{|X_1,\dots,X_n}$ be the subset of $\mathbb{R}^n$ defined by*

$$\mathcal{F}_{|X_1,\dots,X_n} = \big\{ u \in \mathbb{R}^n : \exists f \in \mathcal{F} \text{ such that } u_i = f(X_i) \text{ for all } i = 1, \dots, n \big\}.$$

For any $\varepsilon > 0$, we have

$$\mathcal{N}(\mathcal{F}, L_2(P_n), \varepsilon) \leqslant \mathcal{N}(\mathcal{F}_{|X_1,\dots,X_n}, || \cdot ||_\infty, \varepsilon) \leqslant \mathcal{N}(\mathcal{F}, || \cdot ||_\infty, \varepsilon).$$

Proof. Let $\{u_1, \dots, u_M\}$ be a minimal ε-net for $\mathcal{F}_{|X_1,\dots,X_n}$ with respect to the supremum norm. Let $f_1, \dots, f_M \in \mathcal{F}$ be such that $\big(f_j(X_1), \dots, f_j(X_n)\big) = u_j$ for every $j = 1, \dots, M$. Then, for any $f \in \mathcal{F}$, there exists an index $j \in [M]$ such that $\max_i |f(X_i) - (u_j)_i| = \max_i |f(X_i) - f_j(X_i)| \leqslant \varepsilon$. Since for any function f in $\mathcal{F}$,

$$||f - f_j||^2_{L_2(P_n)} = \frac{1}{n} \sum_{i=1}^{n} (f(X_i) - f_j(X_i))^2 \leqslant ||f - f_j||^2_\infty,$$

$\{f_1, \dots, f_M\}$ is an ε-net for $\mathcal{F}$ with respect to the empirical L_2 norm. This proves the first inequality. Let now $f_1, \dots, f_M$ be an ε-net of $(\mathcal{F}, || \cdot ||_\infty)$. One readily checks that $u_1, \dots, u_M$ defined by $u_j = (f_j(X_1), \dots, f_j(X_n))$ is an ε-net of $\mathcal{F}_{|X_1,\dots,X_n}$. This completes the proof. $\square$

4. MAIN RESULT

We are now in a position to state the main theorem which gives, for an IPM defined by a Hölder class, the rate of convergence of the empirical measure towards its theoretical counterpart.

Theorem 4. *Let $\mathcal{X} \subset \mathbb{R}^d$ be a convex bounded set with non-empty interior. Let $\mathcal{H}^\alpha(L)$ be the Hölder class of α-smooth functions supported on the set $\mathcal{X}$ and with Hölder norm bounded by L. For any probability distribution P supported on $\mathcal{X}$, denoting by P_n the empirical measure associated to i.i.d. samples $X_1, \dots, X_n \sim P$, we have,*

$$\mathbb{E}\big[d_{\mathcal{H}^\alpha(L)}(P_n, P)\big] = \mathbb{E}\left[\sup_{h \in \mathcal{H}^\alpha(L)} |\mathbb{X}_n(h)|\right] \leqslant cL \begin{cases} n^{-\alpha/d} & \text{if } \alpha < d/2 \\ n^{-1/2} \ln(n) & \text{if } \alpha = d/2 \\ n^{-1/2} & \text{if } \alpha > d/2, \end{cases}$$

where c is a constant depending only on d, $\lambda_d(\mathcal{X}^1)$ and α.

We notice two different regimes: for highly smooth functions ($\alpha > d/2$), the rate of convergence does not depend on the smoothness α nor on the dimension d and corresponds to the usual parametric rate of convergence (note that it also matches the rate known for the Maximum Mean Discrepancy metric, which is an IPM indexed by the unit ball of a RKHS with bounded kernel [3]). For less regular Hölder functions ($\alpha < d/2$), the rate of convergence depends both on the smoothness and on the dimension in a typical curse of dimensionality behavior. These two regimes coincide, up to a logarithmic factor, at their smoothness boundary $\alpha = d/2$: we have a continuous transition in terms of the exponent of the sample size. Interestingly the rates we obtain interpolate between the $n^{-1/d}$ rate known for Wasserstein-1 distance [22] when considering $\mathcal{H}^1(1)$ and the $n^{-1/2}$ rate for Maximum Mean Discrepancy when considering Hölder classes with enough smoothness.

Finally, let us mention that the formulation of Theorem 4 given above aims at characterizing the behaviour of the expected error in the asymptotic setting of large samples. This result follows from the following finite sample upper bound (proved in Section 6.2):

$$\mathbb{E}\left[d_{\mathcal{H}^\alpha(L)}(P_n, P)\right] \leqslant 12L \begin{cases} \left(\frac{K\lambda}{n}\right)^{\alpha/d}\left[\frac{d}{d-2\alpha} \wedge \left(1 + 0.5\log(\frac{n}{9K\lambda})\right)\right] & \text{if } \alpha < d/2 \\ \left(\frac{K\lambda}{n}\right)^{1/2}\left[\frac{2\alpha}{2\alpha-d} \wedge \left(1 + \frac{\alpha}{d}\log(\frac{n}{9K\lambda})\right)\right] & \text{if } \alpha \geqslant d/2, \end{cases}$$

where $\lambda := \lambda_d(\mathcal{X}^1)$ and $K = K_{\alpha,d}$ is the constant depending only on α and d borrowed from Theorem 1.

5. SOME EXTENSIONS

A slightly less precise but more general result can be obtained for any bounded class whose entropy grows polynomially in $1/\varepsilon$; see also Rakhlin et al. [12, Theorem 2], where this condition naturally arises. Such an extension can be stated as follows.

Theorem 5. *Let $\mathcal{X} \subset \mathbb{R}^d$ be a convex bounded set with non-empty interior. Let $\mathcal{H}$ be a bounded class of functions supported on the set $\mathcal{X}$. Assume that the entropy of the class grows polynomially, i.e., there exist positive real numbers p and A such that*

$$\forall \varepsilon > 0, \quad \log \mathcal{N}(\mathcal{H}, ||\cdot||_\infty, \varepsilon) \leqslant A\varepsilon^{-p}.$$

Then, for any probability distribution P supported on $\mathcal{X}$, denoting by P_n the empirical measure associated to i.i.d. samples $X_1, \ldots, X_n \sim P$, we have,

$$\mathbb{E}\left[d_{\mathcal{H}}(P_n, P)\right] = \mathbb{E}\left[\sup_{h\in\mathcal{H}} \left|\mathbb{X}_n(h)\right|\right] \leqslant c \begin{cases} n^{-1/p} & \text{if } p > 2 \\ n^{-1/2}\ln(n) & \text{if } p = 2 \\ n^{-1/2} & \text{if } p < 2, \end{cases}$$

where c is a constant.

The proof of the extension is exactly the same as the proof of Theorem 4 up to constants. In this note we have seen Hölder classes as examples of classes with polynomial growth of the entropy but there are many other such classes. To illustrate this we give the example of Sobolev classes which, in some cases, are more general than Hölder classes. For a positive integer s and a real number $1 \leqslant p \leqslant +\infty$, define the Sobolev space $\mathcal{W}_p^s(r)$ with radius $r > 0$ as

$$\mathcal{W}_p^s(r) := \left\{ f \in C^s(\mathcal{X}, \mathbb{R}) : \sum_{|k| \leqslant s} ||D^k f||_p \leqslant r \right\}.$$

Note that for any positive integer s and for any positive radius L, there exist radii r and r' such that

$$\mathcal{W}_\infty^s(r) \subset \mathcal{H}^s(L) \subset \mathcal{W}_\infty^{s-1}(r').$$

A consequence of [11, Corollary 1] is that for any positive integer $s > 0$, and real number p such that $d/s < p \leqslant +\infty$, the entropy of a Sobolev class grows polynomially as

$$\log \mathcal{N}(\mathcal{W}_p^s(L), ||\cdot||_\infty, \varepsilon) \leqslant A\varepsilon^{-d/s}$$

for some positive constant A. Thus Theorem 5 holds for this class. Finally we point out that such bounds on the entropy hold for more general spaces such as some Besov spaces. We refer the reader to Nickl and Pötscher [11] for more details.

APPENDIX

PROOFS

This section contains the proofs of the main results, Theorems 3 and 4, stated in the main body of the note.

A.1. Proof of Theorem 3

The proof of Theorem 3 can be found in [16]. We add it here for completeness.

Let $\gamma_0 = S_n(\mathcal{F}) = \sup_{f\in\mathcal{F}} \|f\|_{L_2(P_n)}$. Define $\gamma_j = 2^{-j}\gamma_0$, for every integer $j \in \mathbb{N}$, and let T_j be a minimal γ_j-cover of $\mathcal{F}$ with respect to $L_2(P_n)$. For any function $f \in \mathcal{F}$, we denote by $\widehat{f}_j$ an element of T_j which is an γ_j approximation of f. For any positive integer N we can decompose the function f as

$$
f = f - \widehat{f}_N + \sum_{j=1}^{N}(\widehat{f}_j - \widehat{f}_{j-1}),
$$

where $\widehat{f}_0 = 0 \in \mathcal{F}$. Hence, for any positive integer N, we have

$$
\widehat{R}_n(\mathcal{F}) = \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i\left(f(X_i) - \widehat{f}_N(X_i) + \sum_{j=1}^{N}(\widehat{f}_j(X_i) - \widehat{f}_{j-1}(X_i))\right)\right]
$$

$$
\leqslant \frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i(f(X_i) - \widehat{f}_N(X_i))\right] + \sum_{j=1}^{N}\frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i(\widehat{f}_j(X_i) - \widehat{f}_{j-1}(X_i))\right]
$$

$$
\leqslant \frac{1}{n}\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}|(f(X_i) - \widehat{f}_N(X_i))| + \sum_{j=1}^{N}\frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i(\widehat{f}_j(X_i) - \widehat{f}_{j-1}(X_i))\right]
$$

$$
= \sup_{f\in\mathcal{F}}\|f - \widehat{f}_N\|_{L_2(P_n)} + \sum_{j=1}^{N}\frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i(\widehat{f}_j(X_i) - \widehat{f}_{j-1}(X_i))\right]
$$

$$
\leqslant \gamma_N + \sum_{j=1}^{N}\frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i(\widehat{f}_j(X_i) - \widehat{f}_{j-1}(X_i))\right].
$$

For any positive integer j, the triangle inequality gives

$$
\|\widehat{f}_j - \widehat{f}_{j-1}\|_{L_2(P_n)} \leqslant \|\widehat{f}_j - f\|_{L_2(P_n)} + \|f - \widehat{f}_{j-1}\|_{L_2(P_n)} \leqslant \gamma_j + \gamma_{j-1} = 3\gamma_j. \qquad (A.1)
$$

We need the following classic lemma which controls the expectation of a Rademacher average over a finite set.[2]

Lemma A.1 (Massart's finite class lemma). *Let $\mathcal{X}$ be a finite subset of $\mathbb{R}^n$ and let $\sigma_1,\ldots,\sigma_n$ be independent Rademacher random variables. Denote the radius of $\mathcal{X}$ by $R = \sup_{x\in\mathcal{X}}\|x\|$. Then, we have,*

$$
\mathbb{E}\left[\sup_{x\in\mathcal{X}}\frac{1}{n}\sum_{i=1}^{n}\sigma_i x_i\right] \leqslant R\frac{\sqrt{2\log|\mathcal{X}|}}{n}.
$$

Applying this lemma to $\mathcal{X}_j = \left\{(\widehat{f}_j(X_i) - \widehat{f}_{j-1}(X_i))_{i=1}^{n} \in \mathbb{R}^n : f \in \mathcal{F}\right\}$ for any $j = 1,\ldots,n$ and using (3), we get

$$
\sum_{j=1}^{N}\frac{1}{n}\mathbb{E}_\sigma\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{n}\sigma_i(\widehat{f}_j(X_i) - \widehat{f}_{j-1}(X_i))\right] \leqslant \sum_{j=1}^{N}3\gamma_j\frac{\sqrt{2\log(|T_j|\cdot|T_{j-1}|)}}{n}
$$

Therefore we have

$$
\widehat{R}_n(\mathcal{F}) \leqslant \gamma_N + \sum_{j=1}^{N}3\gamma_j\frac{\sqrt{2\log(|T_j|\cdot|T_{j-1}|)}}{n} \leqslant \gamma_N + \frac{6}{n}\sum_{j=1}^{N}\gamma_j\sqrt{\log|T_j|}
$$

---

[2] We refer the reader to https://ttic.uchicago.edu/tewari/lectures/lecture10.pdf for a simple proof of this lemma.

$$= \gamma_N + \frac{12}{n} \sum_{j=1}^{N} (\gamma_j - \gamma_{j+1}) \sqrt{\log |T_j|} = \gamma_N + \frac{12}{n} \sum_{j=1}^{N} (\gamma_j - \gamma_{j+1}) \sqrt{\log \mathcal{N}(\mathcal{F}, L_2(P_n), \gamma_j)}$$

$$\leqslant \gamma_N + \frac{12}{n} \int_{\gamma_{N+1}}^{\gamma_0} \sqrt{\log \mathcal{N}(\mathcal{F}, L_2(P_n), \varepsilon)} d\varepsilon.$$

For any $\tau > 0$, pick $N = \sup\{j : \gamma_j > 2\tau\}$. Then $\gamma_N = 2\gamma_{N+1} \leqslant 4\tau$ and $\gamma_{N+1} = \gamma_N/2 \geqslant \tau$. Hence, we conclude that

$$\widehat{R}_n(\mathcal{F}) \leqslant 4\tau + \frac{12}{\sqrt{n}} \int_{\tau}^{\gamma_0} \sqrt{\log \mathcal{N}(\mathcal{F}, L_2(P_n), \varepsilon)} d\varepsilon.$$

Since τ can take any positive value we can take the infimum over all positive τ and this concludes the proof.

A.2. Proof of Theorem 4

Without loss of generality, we prove the theorem in the case $L = 1$. The general case will follow by homogeneity. For simplicity we write $\mathcal{H}^\alpha = \mathcal{H}^\alpha(1)$, $Ph = \int_{\mathcal{X}} h \, dP$ and $P_n h = \int_{\mathcal{X}} h \, dP_n$. A symmetrization argument (Lemma 1) gives

$$\mathbb{E}\left[\sup_{h \in \mathcal{H}^\alpha} |Ph - P_n h| \right] \leqslant 2\mathbb{E}\left[\widehat{R}_n(\mathcal{H}^\alpha) \right],$$

where the empirical Rademacher process $\widehat{R}_n(\mathcal{H}^\alpha)$ is given by

$$\widehat{R}_n(\mathcal{H}^\alpha) = \frac{1}{n} \mathbb{E}\left[\sup_{h \in \mathcal{H}^\alpha} \sum_{i=1}^{n} \sigma_i h(X_i) \, \Big| \, X_1, \ldots, X_n \right].$$

Noting that, for any $h \in \mathcal{H}^\alpha$,

$$P_n h^2 := \frac{1}{n} \sum_{i=1}^{n} h^2(X_i) \leqslant ||h^2||_\infty \leqslant 1,$$

the improved Dudley bound (Theorem 3) coupled with Lemma 2 yields,

$$\mathbb{E}\left[\sup_{h \in \mathcal{H}^\alpha} |P_n h - Ph| \right] \leqslant \inf_{\tau > 0} \left(4\tau + \frac{12}{\sqrt{n}} \int_{\tau}^{1} \sqrt{\log \mathcal{N}(\mathcal{H}^\alpha, || \cdot ||_\infty, \varepsilon)} d\varepsilon \right)$$

$$\leqslant \inf_{\tau > 0} \left(4\tau + \frac{12\sqrt{K\lambda_d(\mathcal{X}^1)}}{\sqrt{n}} \int_{\tau}^{1} \varepsilon^{-d/2\alpha} d\varepsilon \right).$$

Applying Lemma A.2 with $\beta = \frac{d}{2\alpha}$ and $a = 3\sqrt{\frac{K\lambda}{n}}$ where $K = K_{\alpha,d}$ is the constant depending only on α and d borrowed from Theorem 1 and $\lambda := \lambda_d(\mathcal{X}^1)$, we get

$$\mathbb{E}\left[\sup_{h \in \mathcal{H}^\alpha} |P_n h - Ph| \right] \leqslant 12 \begin{cases} \left(\frac{K\lambda}{n}\right)^{\alpha/d} \left[\frac{d}{d-2\alpha} \wedge (1 + 0.5 \log(\frac{n}{9K\lambda})) \right] & \text{if } \alpha < d/2 \\ \left(\frac{K\lambda}{n}\right)^{1/2} \left[\frac{2\alpha}{2\alpha-d} \wedge (1 + \frac{\alpha}{d} \log(\frac{n}{9K\lambda})) \right] & \text{if } \alpha \geqslant d/2. \end{cases} \tag{A.2}$$

The proof is finished since the upper bound stated in Theorem 4 is a direct consequence of (A.2)

A.3. Additional Lemma

The following lemma enables to obtain an upper bound on Dudley's refined bound (Theorem 3) for any bounded class whose entropy grows polynomially in $1/\varepsilon$.

Lemma A.2. *For any real positive numbers a and β, it holds*

$$\min_{0 \leqslant \tau \leqslant 1} \left(\tau + a \int_\tau^1 \varepsilon^{-\beta} d\varepsilon \right) \leqslant (a^{1/\beta} \vee a) \left[\left(\frac{\beta \vee 1}{|\beta - 1|} \right)) \wedge \left(1 + \frac{\log(1/a)}{\beta \vee 1} \right) \right].$$

Proof. Let a and β be real positive numbers. Define the function

$$f : [0, 1] \to \mathbb{R},$$

$$\tau \mapsto \tau + a \int_\tau^1 \varepsilon^{-\beta} d\varepsilon.$$

One can easily check that

$$f^* := \min_{0 \leqslant \tau \leqslant 1} f(\tau) = \begin{cases} 1 & \text{if } a > 1 \\ a^{1/\beta} + \frac{a}{1-\beta}(1 - a^{1/\beta-1}) & \text{if } a < 1. \end{cases} \tag{A.3}$$

In the case $a < 1$, using the fact that $1 - x^\alpha \leqslant \log(x^{-\alpha})$ for any $\alpha > 0$ and $x \in (0, 1]$, we have

$$f^* \leqslant (a^{1/\beta} \vee a) \left[\left(\frac{\beta \vee 1}{|\beta - 1|} \right)) \wedge \left(1 + \frac{\log(1/a)}{\beta \vee 1} \right) \right]. \tag{3}$$

Finally, since the RHS of (A.3) is greater than 1 for any $a > 1$, (A.3) holds for any positive real a and this concludes the proof. $\square$

ACKNOWLEDGMENTS

REFERENCES

1. M. Arjovsky, S. Chintala, and L. Bottou, *Wasserstein Generative Adversarial Networks*, Ed. by D. Precup and Y. W. Teh, *Proceedings of the 34th International Conference on Machine Learning*, Vol. 70 of Proceedings of Machine Learning Research, International Convention Centre, Sydney, Australia, PMLR (2017), pp. 214–223.
2. F. Bassetti, A. Bodini, and E. Regazzini, "On minimum Kantorovich distance estimators," Statistics and probability letters **76** (12), 1298–1302 (2006).
3. F.-X. Briol, A. Barp, A. B. Duncan, and M. Girolami, Statistical inference for generative models with maximum mean discrepancy (2019). arXiv preprint arXiv:1906.05944.
4. M. Chen, W. Liao, H. Zha, and T. Zhao, Statistical guarantees of generative adversarial networks for distribution estimation (2020). arXiv preprint arXiv:2002.03938.
5. E. del Barrio, P. Deheuvels, and S. van de Geer, Lectures on empirical processes. EMS Series of Lectures in Mathematics. European Mathematical Society (EMS), Zürich. Theory and statistical applications, With a preface by Juan A. Cuesta Albertos and Carlos Matrán (2007).
6. R. M. Dudley, "The speed of mean Glivenko-Cantelli convergence," Ann. Math. Statist. **40**, 40–50 (1968).
7. E. Giné and R. Nickl, *Mathematical Foundations of Infinite-Dimensional Statistical Models*, Vol. 40 (Cambridge University Press, 2016).
8. I. Goodfellow, J. Pouget-Abadie, M. Mirza, B. Xu, D. Warde-Farley, S. Ozair, A. Courville, and Y. Bengio, "Generative adversarial nets," In Advances in Neural Information Processing Systems, 2672–2680 (2014).
9. V. Koltchinskii, *Oracle Inequalities in Empirical Risk Minimization and Sparse Recovery Problems: Ecole d'Eté de Probabilités de Saint-Flour XXXVIII-2008*, Vol. 2033 (Springer Science and Business Media, 2011).
10. T. Liang, On how well generative adversarial networks learn densities: Nonparametric and parametric results (2018). arXiv preprint arXiv:1811.03179.

11. R. Nickl and B. M. Pötscher, "Bracketing metric entropy rates and empirical central limit theorems for function classes of besov-and sobolev-type," Journal of Theoretical Probability **20** (2), 177−199 (2007).
12. A. Rakhlin, K. Sridharan, and A. B. Tsybakov, "Empirical entropy, minimax regret and minimax risk," Bernoulli **23** (2), 789−824 (2017).
13. F. Santambrogio, "Optimal transport for applied mathematicians," Birkäuser, NY **55** (58−63), 94 (2015).
14. M. Scetbon, L. Meunier, J. Atif, and M. Cuturi, "Equitable and optimal transport with multiple agents," in *Proceedings of the 24th International Conference on Artificial Intelligence and Statistics,* Vol. 130, *Proceedings of Machine Learning Research* (2021), pp. 2035−2043; arXiv: 2006.07260 (2020).
15. A. Shiryayev, *Selected Works of AN Kolmogorov, Vol. III, Information Theory and the Theory of Algorithms* (Springer, 1993), Vol. 27.
16. N. Srebro and K. Sridharan, Note on refined Dudley integral covering number bound. Unpublished results (2010). http://ttic. uchicago. edu/karthik/dudley. pdf
17. N. Srebro, K. Sridharan, and A. Tewari, "Smoothness, low noise and fast rates," in Advances in neural information processing systems, 2199−2207 (2010).
18. B. K. Sriperumbudur, K. Fukumizu, A. Gretton, B. Schölkopf, G. R. Lanckriet, et al., "On the empirical estimation of integral probability metrics," Electronic Journal of Statistics **6**, 1550−1599 (2012).
19. A. B. Tsybakov, *Introduction to Nonparametric Estimation* (Springer Science and Business Media, 2008).
20. A. W. van der Vaart and J. A. Wellner, *Weak Convergence and Empirical Processes* (Springer Series in Statistics. Springer-Verlag, New York. With applications to statistics, 1996).
21. R. Vershynin, *High-dimensional probability: An introduction with applications in data science* (Cambridge university press, 2018), vol. 47.
22. J. Weed, F. Bach, et al. "Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance," Bernoulli **25** (4A), 2620−2648 (2019).