Upper Bounds on Averaged Sampling Numbers for General Model Classes

Rahul Parhi
Department of Electrical and Computer Engineering
University of California, San Diego
La Jolla, CA 92093, USA
Email: rahul@ucsd.edu

Ben Adcock
Department of Mathematics
Simon Fraser University
Burnaby, BC V5A 1S6, Canada
Email: ben_adcock@sfu.ca

***Abstract*—We investigate sampling numbers for essentially arbitrary star-shaped model classes. Although sampling numbers have been determined for a wide-variety of concrete model classes, to the best of our knowledge, abstract characterizations have not received much attention. We employ techniques developed in the context of nonparametric regression and empirical-process theory to derive novel upper bounds on the averaged sampling numbers for general model classes relying only on the knowledge of their entropy numbers. Our formulation provides an abstract characterization for upper bounds of these sampling numbers. Moreover, we show that any interpolator of the data that lies in the model class achieves this bound.**

***Index Terms*—Rademacher complexity, metric entropy, optimal recovery, sampling numbers.**

I. Introduction

Sampling numbers are fundamental objects in approximation theory, numerical analysis, information-based complexity, and machine learning. The determination of such numbers has become a very active area of research, especially in recent years [6], [10], [12], [15]. This has been motivated, in part, by the recent observation in overparameterized machine learning that learning functions that interpolate the training data can still generalize well [3], [4], [5].

Sampling numbers measure the minimal worst-case error (measured in some suitable norm) that can be achieved from a set of n noiseless samples. Conversely, in nonparametric regression and empirical-process theory there has been substantial focus on characterizing the *minimax rate* for a model class $\mathcal{F}$. Here, one assumes noisy data with a fixed and non-negligible variance σ^2 and studies the best possible error in terms of σ as n grows. Abstract characterization of minimax rates have been developed, notably by Yang and Barron [23]. Those results employ techniques such as *metric entropy*.

The research communities interested in sampling numbers and minimax rates are quite disjoint. A by-product of this is that a disjoint set of mathematical tools have been developed to study very similar problems. In this paper we aim to bridge the gap between these two areas/tools. Specifically, we make novel use of tools such as metric entropy to develop new upper bounds on *averaged* $\mathrm{L}^2_\mu(\Omega)$*-sampling numbers* for essentially arbitrary model classes $\mathcal{F}$ with data sites drawn i.i.d. from some probability measure μ on some bounded domain $\Omega \subset \mathbb{R}^d$.[1] By specializing our results to Sobolev model classes, we show that the averaged $\mathrm{L}^2_\mu(\Omega)$-sampling numbers are *no worse* than the $\mathrm{L}^2_\mu(\Omega)$-minimax rates (see Section IV). Furthermore, the algorithm that achieves our predicted rates simply corresponds to finding a minimum-norm interpolator from the model class.

A. Main Contributions

Let $\mathcal{F}$ be a model class that satisfies the compact embedding $\mathcal{F} \subset\subset \mathrm{C}(\Omega)$ (the Banach space of continuous functions on $\mathcal{F}$), $f_0 \in \mathcal{F}$ be a *ground-truth* function and consider the noiseless observations

$$y_i = f_0(x_i), \quad i = 1, \dots, n, \tag{1}$$

where $\{x_i\}_{i=1}^n$ are drawn i.i.d. from μ. The *averaged sampling number* is defined as

$$s_n^{\mathrm{ave}}(\mathcal{F})_{\mathrm{L}^2_\mu} := s_n^{\mathrm{ave}}(\mathcal{F};\mu)_{\mathrm{L}^2_\mu} := \inf_{\widehat{f}} \sup_{f_0 \in \mathcal{F}} \mathbf{E}\|f_0 - \widehat{f}\|_{\mathrm{L}^2_\mu}, \tag{2}$$

where the $\inf$ is taken over all deterministic (measurable) functions of the data $\{(x_i, y_i)\}_{i=1}^n \subset \Omega \times \mathbb{R}$.

Let $N_\varepsilon(\mathcal{F})_{\mathrm{L}^\infty}$ denote the L^∞-covering number of $\mathcal{F}$. Then, the (dyadic) L^∞-entropy number of $\mathcal{F}$ is given by

$$\varepsilon_n(\mathcal{F})_{\mathrm{L}^\infty} = \inf\{\varepsilon > 0 : \ N_\varepsilon(\mathcal{F})_{\mathrm{L}^\infty} \leq 2^n\}. \tag{3}$$

This quantity captures how precisely elements of $\mathcal{F}$ can be specified with n bits. The notion of entropy of a compact set was introduced by Kolmogorov [13] as a way to *quantify* its compactness. The main result of this paper is summarized in the following theorem, which we prove in Section III-A.

Theorem 1. *Let $\mathcal{F} \subset\subset \mathrm{C}(\Omega)$ be a star-shaped model class and suppose that its L^∞-entropy number scales as*

$$\varepsilon_n(\mathcal{F})_{\mathrm{L}^\infty} \asymp n^{-\alpha} \tag{4}$$

for some $\alpha > 0$. Then, the averaged sampling number is upper bounded by

$$s_n^{\mathrm{ave}}(\mathcal{F})_{\mathrm{L}^2_\mu} \lesssim n^{-\frac{\alpha}{2\alpha+1}}, \quad n^{\frac{1}{2\alpha+1}} \gtrsim \log\log n^{\frac{\alpha}{2\alpha+1}}. \tag{5}$$

Moreover, this upper bound is attained by any interpolant $\widehat{f} \in \mathcal{F}$ of the data (1).

[1] We assume that the boundary of Ω is sufficiently regular, e.g., Lipschitz.

This theorem provides an abstract characterization for upper bounds on $s_n^{\mathrm{ave}}(\mathcal{F})_{\mathrm{L}^2_\mu}$. We immediately have the following corollary.

Corollary 2. *When $\mathcal{F} = U(\mathcal{X})$ is the unit ball of some Banach space $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$ then* (5) *is achieved by the solution to the minimum-norm interpolation problem*

$$\min_{f\in\mathcal{X}} \|f\|_{\mathcal{X}} \quad \text{s.t.} \quad f(x_i) = y_i,\ i = 1, \dots, n. \tag{6}$$

Remark 3. We show in Section III that the optimization problem (6) is well-posed, i.e., always admits a minimizer.

Remark 4. We present Theorem 1 with the entropy numbers decaying as in (4) for ease of presentation. The results readily generalize for other decay rates.

B. Relation to Existing Work

As noted, sampling numbers are widely-studied in optimal recovery and information-based complexity [6], [10], [12], [15]. Differing from sampling numbers, in nonparametric regression one considers noisy data

$$y_i = f_0(x_i) + \sigma\eta_i, \quad i = 1, \dots, n, \tag{7}$$

where the noise $\eta_i \sim_{\text{i.i.d.}} \mathcal{N}(0, 1)$ and is independent of the data sites x_i. The *minimax risk* is then defined as

$$m_n(\mathcal{F}; \sigma)_{\mathrm{L}^2_\mu} := \inf_{\widehat{f}} \sup_{f_0\in\mathcal{F}} \mathbf{E}\|f_0 - \widehat{f}\|_{\mathrm{L}^2_\mu}, \tag{8}$$

where the inf is taken over all deterministic (measurable) functions of the noisy data $\{(x_i, y_i)\}_{i=1}^n \subset \Omega \times \mathbb{R}$ and the expectation $\mathbf{E}$ is with respect to the noise $\{\eta_i\}_{i=1}^n$ and, typically, the sample points $\{x_i\}_{i=1}^n$.

In this setting, the noise level σ is assumed to be fixed and strictly positive. Abstract characterizations of the minimax rate for various model classes have been developed (see, e.g., the work of Yang and Barron [23]). These results are typically of the form $(n/\sigma^2)^{-\alpha}$, where α is related to the complexity of the model class. Unfortunately, these kinds of results cannot handle the low-noise or zero-noise regimes since the limit as $\sigma \to 0$ as the limiting quantity to 0, which is certainly not true for any finite number of data n. However, this is precisely the regime of interest in this paper, since the averaged sampling number (2) corresponds precisely to the scenario in which the noise level tends to 0 in nonparametric regression. Therefore, knowledge of the minimax rate of a model class *does not* provide a method to bound $s_n^{\mathrm{ave}}(\mathcal{F})_{\mathrm{L}^2_\mu}$. We also note that recent work [9] characterizes the *noise-level-aware* minimax rates for Besov model classes that simultaneously captures both the usual minimax rate ($\sigma > 0$) and the optimal recovery rate ($\sigma \to 0$) as a function of both n and σ.

In this paper, we provide an upper bound on sampling number for model class $\mathcal{F}$ in terms of their $\mathrm{L}^\infty(\Omega)$ entropy numbers. While sampling numbers have been determined for a wide-variety of model classes (including Sobolev and Besov balls), to the best of our knowledge, there does not exist an abstract characterization based on complexity measures of model classes such as their entropy numbers. Therefore, the main contribution of this paper is to provide a bound for essentially arbitrary model classes based on their entropy numbers (Theorem 1).

Remarkably, by specializing our results to specific model classes (such as Sobolev balls), we are also able to show that the sampling numbers are *no worse* than the minimax rates seen in nonparametric regression. We also remark that the proofs of our main results are based on developments in empirical-process theory, which are typically not used to derive sampling numbers or other approximation-theoretic quantities. Thus, another contribution of this paper is to link ideas from empirical-process theory and approximation theory.

II. Preliminaries

The focus of this paper is on star-shaped model classes $\mathcal{F}$ that satisfy the compact embedding $\mathcal{F} \subset\subset \mathrm{C}(\Omega)$, where $\mathrm{C}(\Omega)$ denotes the Banach space of continuous functions defined on a bounded domain $\Omega \subset \mathbb{R}^d$. This assumption guarantees, in particular, that (i) point evaluations of $f \in \mathcal{F}$ are well-defined and (ii) $\mathcal{F}$ is *uniformly bounded*, i.e.,

$$b := b_{\mathcal{F}} := \sup_{f\in\mathcal{F}} \|f\|_{\mathrm{L}^\infty(\Omega)} := \sup_{f\in\mathcal{F}} \left(\sup_{x\in\Omega} |f(x)|\right) < \infty. \tag{9}$$

Let μ denote a probability measure on Ω and define the usual $\mathrm{L}^2_\mu(\Omega)$-norm of a (measurable) function $f : \Omega \to \mathbb{R}$ as

$$\|f\|_{\mathrm{L}^2_\mu(\Omega)} := \left(\int_{\mathbb{R}^d} |f(x)|^2 \,\mathrm{d}\mu(x)\right)^{1/2}. \tag{10}$$

Consequently, we write $f \in \mathrm{L}^2_\mu(\Omega)$ whenever this norm is finite. Observe that, by assumption, we have $\mathcal{F} \subset \mathrm{L}^2_\mu(\Omega)$. We therefore use the $\mathrm{L}^2_\mu(\Omega)$-norm to measure the error of our estimate of the data-generating function.

Given n samples $\{x_i\}_{i=1}^n \subset \Omega$ drawn i.i.d. from μ, consider the *empirical measure*

$$\mu_n := \frac{1}{n}\sum_{i=1}^n \delta_{x_i}, \tag{11}$$

where δ_{x_i} denotes the Dirac measure centered at x_i. This measure induces the *empirical* L^2*-norm*

$$\|f\|_n := \|f\|_{\mathrm{L}^2_{\mu_n}(\Omega)} = \left(\frac{1}{n}\sum_{i=1}^n |f(x_i)|^2\right)^{1/2}, \tag{12}$$

which is well-defined for any continuous function $f \in \mathrm{C}(\Omega)$, for instance. To that end, we refer to the $\mathrm{L}^2_\mu(\Omega)$-norm as the *population* L^2*-norm*.

Crucial to our analysis is the quantification of the deviation between the population and the empirical L^2-norm with respect to the number of samples n. This is a well-studied problem in empirical-process theory. To that end, let

$$\mathfrak{R}_n(\delta; \mathcal{F}) := \mathbf{E}\left[\sup_{\substack{f\in\mathcal{F} \\ \|f\|_{\mathrm{L}^2_\mu(\Omega)}\le\delta}} \left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\right|\right] \tag{13}$$

denote the *localized Rademacher complexity* of $\mathcal{F}$ and

$$\widehat{\mathfrak{R}}_n(\delta;\mathcal{F}) \coloneqq \mathbf{E}_{\boldsymbol{\varepsilon}}\left[\sup_{\substack{f\in\mathcal{F}\\ \|f\|_n\le\delta}}\left|\frac{1}{n}\sum_{i=1}^n \varepsilon_i f(x_i)\right|\right] \tag{14}$$

denote the *empirical localized Rademacher complexity* (which is a random variable) (see [2], [14] as well as [22, Chapter 14]). In (13) and (14), $\{\varepsilon_i\}_{i=1}^n$ are i.i.d. Rademacher random variables, $\mathbf{E}$ denotes the expectation operator, $\mathbf{E}_{\boldsymbol{\varepsilon}}$ denotes the conditional expectation $\mathbf{E}[\,\cdot \mid \{x_i\}_{i=1}^n]$. The following classical result from empirical-process theory will play a key role in our analysis.

Proposition 5 (adapted from [22, Theorem 14.1 and (14.8)])**.** *Suppose that $\mathcal{F}$ is a star-shaped model class that satisfies the compact embedding $\mathcal{F}\subset\subset \mathrm{C}(\Omega)$. Let δ_n be any positive solution to either of the inequalities*

$$\mathfrak{R}_n(\delta;\mathcal{F}) \le \frac{\delta^2}{b} \quad \text{or} \quad \widehat{\mathfrak{R}}_n(\delta;\mathcal{F}) \le \frac{\delta^2}{b}, \tag{15}$$

where b is as in (9)*. There exist universal constants $c_0, c_1, c_2, c_3 > 0$ such that if $n\delta_n^2 \ge \frac{2}{c_1}\log(4\log(1/\delta_n))$, then*

$$\Big|\|f\|_n - \|f\|_{\mathrm{L}^2_\mu(\Omega)}\Big| \le c_0\delta_n, \quad \text{for all } f\in\mathcal{F}, \tag{16}$$

with probability at least $1 - c_2\mathrm{e}^{-c_3 n\delta_n^2/b^2}$.

III. Main Results

In this section, we will bound $s_\mu(\mathcal{F})_{\mathrm{L}^2_\mu}$ for model classes that satisfy the compact embedding $\mathcal{F}\subset\subset \mathrm{C}(\Omega)$ in terms of their $\mathrm{L}^\infty(\Omega)$ entropy numbers, which are guaranteed to be well-defined thanks to the compact embedding.

Given the L^∞-covering number $N_\varepsilon(\mathcal{F})_{\mathrm{L}^\infty}$ recall that the (dyadic) L^∞-entropy number of $\mathcal{F}$ is given by

$$\varepsilon_n(\mathcal{F})_{\mathrm{L}^\infty} = \inf\{\varepsilon > 0 :\ N_\varepsilon(\mathcal{F})_{\mathrm{L}^\infty} \le 2^n\}. \tag{17}$$

Entropy numbers are well-studied objects in approximation theory. Metric entropies are a related object often studied in empirical-process theory. The L^∞-metric entropy of $\mathcal{F}$ is given by the log covering number, i.e., $\log N_\varepsilon(\mathcal{F})_{\mathrm{L}^\infty}$. Observe that these quantities are "dual" to each other in the sense that

$$\varepsilon_n(\mathcal{F})_{\mathrm{L}^\infty} \asymp n^{-\alpha} \quad\Leftrightarrow\quad \log N_\varepsilon(\mathcal{F})_{\mathrm{L}^\infty} \asymp \left(\frac{1}{\varepsilon}\right)^{\frac{1}{\alpha}}. \tag{18}$$

To upper bound $s_\mu(\mathcal{F})_{\mathrm{L}^2_\mu}$, it suffices to upper bound the expected error for one particular (deterministic) algorithm that constructs an approximation to the data-generating function. More specifically, recall that, given a ground-truth function $f_0\in\mathcal{F}$, we are interested in the construction of an approximation to f_0 from the (noiseless) observations

$$y_i = f_0(x_i), \quad i = 1,\dots,n, \tag{19}$$

where $\{x_i\}_{i=1}^n$ are draw i.i.d. from μ. We shall consider the very simple (deterministic) algorithm of least-squares, which is specified by

$$\min_{f\in\mathcal{F}} \frac{1}{n}\sum_{i=1}^n |y_i - f(x_i)|^2. \tag{20}$$

Observe that this problem is well-posed in the sense that minimizers exist. Indeed, since the embedding of $\mathcal{F}$ into $\mathrm{C}(\Omega)$ is compact, $\mathcal{F}$ is compact with respect to the $\mathrm{C}(\Omega)$-topology. Then, since the point-evaluation functional is continuous on $\mathrm{C}(\Omega)$, this is the minimization of a $\mathrm{C}(\Omega)$-continous function over a $\mathrm{C}(\Omega)$-compact set and so a minimizer exists. Furthermore, it is clear that f_0 is always a minimizer (since the objective evaluated at f_0 is 0). This reveals that the solution set to (20) can be equivalently characterized by

$$S = \{f\in\mathcal{F} :\ f(x_i) = y_i, i = 1,\dots,n\} \tag{21}$$

and is generally not singleton. For our purposes it suffices to consider any minimizer in (21).

When $\mathcal{F} = U(\mathcal{X})$, the unit ball of some Banach space $(\mathcal{X}, \|\cdot\|_{\mathcal{X}})$, an alternative, but compatible, approach to constructing an approximation to f_0 is to consider the minimum-norm interpolation problem

$$\min_{f\in\mathcal{X}} \|f\|_{\mathcal{X}} \quad \text{s.t.} \quad f(x_i) = y_i,\ i = 1,\dots,n. \tag{22}$$

Indeed, any solution to (22) lies in S. This is the typical formulation for learning from data in modern (overparameterized and high-dimensional) settings. Indeed, it has been shown in various scenarios that interpolating functions can still generalize well [3], [4], [5]. Thus, there has been a line of work investigating interpolation learning, which is precisely the formulation of the present paper.

Before stating and proving our main theorem (Theorem 1), we first state and prove the following lemma, which is based on tools from empirical-process theory.

Lemma 6. *Any δ that satisfies*

$$\frac{64}{\sqrt{n}}\int_{\frac{\delta^2}{2b}}^{\delta} \sqrt{\log N_\varepsilon(\mathcal{F})_{\mathrm{L}^\infty}}\,\mathrm{d}\varepsilon \le \frac{\delta^2}{b} \tag{23}$$

satisfies the second inequality in (15)*.*

Proof. Given $\{x_i\}_{i=1}^n \subset \Omega$ drawn i.i.d. from μ and the associated empirical measure μ_n (see (11)), observe that it is always the case that $\|\cdot\|_n = \|\cdot\|_{\mathrm{L}^2_{\mu_n}} \le \|\cdot\|_{\mathrm{L}^\infty}$. Therefore, for any closed—with respect to the topology inherited from $\mathrm{C}(\Omega)$—subset $S\subset\mathcal{F}$, we have that

$$\log N_\varepsilon(S)_{\mathrm{L}^2_{\mu_n}} \le \log N_\varepsilon(S)_{\mathrm{L}^\infty} \le \log N_\varepsilon(\mathcal{F})_{\mathrm{L}^\infty}. \tag{24}$$

Next, consider $B_n(\delta) \coloneqq \{f\in\mathcal{F} :\ \|f\|_n \le \delta\} \subset \mathcal{F}$. By [22, Corollary 14.3], any positive solution to inequality

$$\frac{64}{\sqrt{n}}\int_{\frac{\delta^2}{2b}}^{\delta} \sqrt{\log N_\varepsilon(B_n(\delta))_{\mathrm{L}^2_{\mu_n}}}\,\mathrm{d}\varepsilon \le \frac{\delta^2}{b} \tag{25}$$

satisfies the inequality

$$\widehat{\mathfrak{R}}_n(\delta;\mathcal{F}) \le \frac{\delta^2}{b}, \tag{26}$$

From (24), we see that any positive solution to (23) (which is deterministic) satisfies (26) uniformly over realizations of $\{x_i\}_{i=1}^n$, which completes the proof. □

A. Proof of Theorem 1

Proof. Let $f_0 \in \mathcal{F}$ and suppose that we observe

$$y_i = f_0(x_i),\ i = 1, \dots, n, \tag{27}$$

with $\{x_i\}_{i=1}^N$ drawn i.i.d. from μ. Any solution f to the least-squares problem (20) or the minimum-norm interpolation problem (22) satisfies $f(x_i) = y_i$, $i = 1, \dots, n$. Therefore, $\|f - f_0\|_n = 0$ (almost surely). Therefore, by Proposition 5, for universal constants $c_j > 0$, $j = 0, 1, 2, 3$, we have, for $n\delta_n^2 \geq \frac{2}{c_1}\log(4\log(1/\delta_n))$, that

$$\|f - f_0\|_{\mathrm{L}^2_\mu} \leq c_0 \delta_n \tag{28}$$

with probability at least $1 - c_2 \mathrm{e}^{-c_3 n\delta_n^2/b^2}$, where δ_n is any positive solution to either inequality in (15). By Lemma 6 it suffices to choose δ_n such that it satisfies

$$\frac{64}{\sqrt{n}} \int_{\frac{\delta^2}{2b}}^{\delta} \varepsilon^{-\frac{1}{2\alpha}}\, \mathrm{d}\varepsilon \leq \frac{\delta^2}{b}. \tag{29}$$

This inequality is equivalent to

$$\frac{64}{\sqrt{n}} \cdot \frac{2\alpha}{2\alpha - 1}\left(\delta^{\frac{2\alpha-1}{2\alpha}} - \frac{\delta^{\frac{2\alpha-1}{2\alpha}}}{(2b)^{\frac{2\alpha-1}{2\alpha}}}\right) \leq \frac{\delta^2}{b}. \tag{30}$$

Any $\delta > 0$ that satisfies (30) would also satisfy

$$\frac{64}{\sqrt{n}} \cdot \frac{2\alpha}{2\alpha - 1} \cdot \delta^{\frac{2\alpha-1}{2\alpha}} \leq \frac{\delta^2}{b}. \tag{31}$$

In which case, this implies that

$$\frac{64b}{\sqrt{n}} \cdot \frac{2\alpha}{2\alpha - 1} \leq \delta^{\frac{2\alpha+1}{2\alpha}} \tag{32}$$

and so

$$\delta \geq \left(\frac{128b\alpha}{2\alpha - 1}\right)^{-\frac{2\alpha}{2\alpha+1}} n^{-\frac{\alpha}{2\alpha+1}}. \tag{33}$$

Thus, we can choose δ_n as the right-hand side of (33). For

$$n^{\frac{1}{2\alpha+1}} \gtrsim \log\log n^{\frac{\alpha}{2\alpha+1}}, \tag{34}$$

we have that (28) holds. By integrating the tail probability and observing that the bound is uniform over $f_0 \in \mathcal{F}$, the theorem is proven. □

IV. Discussion and Applications

The Birman-Solomyak theorem [7] says that for Sobolev spaces that satisfy the compact embedding $\mathrm{W}^{s,p}(\Omega) \subset\subset \mathrm{C}(\Omega)$, i.e., $s > d/p$, their unit balls satisfy

$$\varepsilon_n(U(\mathrm{W}^{s,p}(\Omega)))_{\mathrm{L}^\infty} \asymp \varepsilon_n(U(\mathrm{W}^{s,p}(\Omega)))_{\mathrm{L}^2} \asymp n^{-s/d}, \tag{35}$$

where we note that since $\mathrm{W}^{s,p}(\Omega) \subset\subset \mathrm{C}(\Omega)$, it is necessarily the case that $\mathrm{W}^{s,p}(\Omega) \subset\subset \mathrm{L}^2(\Omega)$.

The application of Theorem 1 to these model classes, reveals that $s_n^{\mathrm{ave}}(U(\mathrm{W}^{s,p}(\Omega)))_{\mathrm{L}^2_\mu} \lesssim n^{-\frac{s}{2s+d}}$, for any probability measure μ on Ω, and in particular for the uniform measure. In that case L^2_μ-norm is a constant scaling of the L^2-norm in which case we have that

$$s_n^{\mathrm{ave}}(U(\mathrm{W}^{s,p}(\Omega)))_{\mathrm{L}^2} \lesssim n^{-\frac{s}{2s+d}}, \tag{36}$$

for n sufficiently large. On the other hand, for the problem of nonparametric regression, the L^2-entropy number (35) combined with the abstract characterization of Yang and Barron [23, Proposition 1] reveals that the minimax rate (when μ is the uniform measure) scales as

$$m_n(U(\mathrm{W}^{s,p}(\Omega)); \sigma)_{\mathrm{L}^2} \asymp \left(\frac{n}{\sigma^2}\right)^{-\frac{s}{2s+d}}, \tag{37}$$

for sufficiently large n.

Thus, we see that the averaged sampling number is no worse than the minimax rate. Furthermore, by Corollary 2, the interpolant that achieves (36) can be found by computing the minimum-Sobolev-norm interpolator of the data.

Theorem 1 also can be readily applied to neural network model classes such as the ReLU^k variation spaces on a bounded domain $\Omega \subset \mathbb{R}^d$ [1], [16], [17], [11], [20], which have garnered interest in recent years. In particular, by [19, Theorem 3] combined with a variant of Carl's inequality [8] (see [21, Theorem 10] for the particular variant), the unit ball U_k of the ReLU^k variation spaces satisfies

$$n^{-\frac{1}{2}-\frac{2k+1}{2d}} \lesssim \varepsilon_n(U_k)_{\mathrm{L}^\infty} \lesssim \widetilde{O}(n^{-\frac{1}{2}-\frac{2k+1}{2d}}), \tag{38}$$

where $\widetilde{O}(\cdot)$ hides log factors. Therefore, we have that

$$s_n^{\mathrm{ave}}(U_k)_{\mathrm{L}^2_\mu} \lesssim \widetilde{O}(n^{-\frac{2k+d+1}{2(2k+2d+1)}}) \tag{39}$$

for any probability measure μ. This rate (i) improves the recently reported rate of $n^{-1/4}$ for these model classes in [6, p. 34] and (ii) is no worse than the minimax rates for these model classes in the case when μ is the uniform measure [18]. Furthermore, to the best of our knowledge, these are the fastest known upper bounds on the averaged sampling numbers for these model classes.

V. Conclusion and Future Work

In this paper, we have investigated the averaged $\mathrm{L}^2_\mu(\Omega)$-sampling numbers for general star-shaped model classes. We have abstractly characterized an upper bound on these sampling numbers from knowledge of the L^∞-entropy number of the model class. Our techniques draw on tools developed in the context of minimax rates, which are typically interested in approximating functions from noisy data. Our formulation provides a new link between the tools developed for minimax estimation and the determination of sampling numbers.

This opens the door for a number of follow-up research directions using tools at the intersection of nonparametric regression/empirical-process theory and optimal recovery, further bridging the gap between these two fields. In particular, it would be interesting to understand when the upper bound in Theorem 1 is sharp. In the case of Sobolev model classes with μ as the uniform probability measure on Ω, it has been shown in the recent work of [15] that

$$s_n^{\mathrm{ave}}(U(\mathrm{W}^{s,p}(\Omega)))_{\mathrm{L}^2} \asymp n^{-\frac{s}{d}+\left(\frac{1}{p}-\frac{1}{q}\right)_+}, \tag{40}$$

which reveals that in that case our upper bound is not sharp. Further understanding this gap is a direction of future work.

REFERENCES

[1] F. Bach, "Breaking the curse of dimensionality with convex neural networks," *Journal of Machine Learning Research*, vol. 18, no. 1, pp. 629–681, 2017.

[2] P. L. Bartlett, O. Bousquet, and S. Mendelson, "Local Rademacher complexities," *Annals of Statistics*, vol. 33, no. 4, pp. 1497–1537, 2005.

[3] M. Belkin, "Fit without fear: Remarkable mathematical phenomena of deep learning through the prism of interpolation," *Acta Numerica*, vol. 30, pp. 203–248, 2021.

[4] M. Belkin, D. J. Hsu, and P. Mitra, "Overfitting or perfect fitting? Risk bounds for classification and regression rules that interpolate," *Advances in Neural Information Processing Systems*, vol. 31, 2018.

[5] M. Belkin, A. Rakhlin, and A. B. Tsybakov, "Does data interpolation contradict statistical optimality?" in *International Conference on Artificial Intelligence and Statistics*. PMLR, 2019, pp. 1611–1619.

[6] P. Binev, A. Bonito, R. DeVore, and G. Petrova, "Optimal learning," *Calcolo*, vol. 61, no. 1, p. 15, 2024.

[7] M. S. Birman and M. Z. Solomyak, "Piecewise-polynomial approximations of functions of the classes W_p^α," *Matematicheskii Sbornik*, vol. 115, no. 3, pp. 331–355, 1967.

[8] B. Carl, "Entropy numbers, s-numbers, and eigenvalue problems," *Journal of Functional Analysis*, vol. 41, no. 3, pp. 290–306, 1981.

[9] R. DeVore, R. D. Nowak, R. Parhi, G. Petrova, and J. W. Siegel, "Optimal recovery meets minimax estimation," *arXiv preprint arXiv:2502.17671*, 2025.

[10] M. Dolbeault and A. Cohen, "Optimal pointwise sampling for L^2 approximation," *Journal of Complexity*, vol. 68, p. 101602, 2022.

[11] W. E, C. Ma, and L. Wu, "The Barron space and the flow-induced function spaces for neural network models," *Constructive Approximation*, vol. 55, no. 1, pp. 369–406, 2022.

[12] T. Jahn, T. Ullrich, and F. Voigtlaender, "Sampling numbers of smoothness classes via ℓ^1-minimization," *Journal of Complexity*, vol. 79, p. 101786, 2023.

[13] A. N. Kolmogorov, "On linear dimensionality of topological vector spaces," in *Doklady Akademii Nauk*, vol. 120, no. 2. Russian Academy of Sciences, 1958, pp. 239–241.

[14] V. Koltchinskii, "Local Rademacher complexities and oracle inequalities in risk minimization," *Annals of Statistics*, vol. 34, no. 6, pp. 2593–2656, 2006.

[15] D. Krieg, E. Novak, and M. Sonnleitner, "Recovery of Sobolev functions restricted to iid sampling," *Mathematics of Computation*, vol. 91, no. 338, pp. 2715–2738, 2022.

[16] G. Ongie, R. Willett, D. Soudry, and N. Srebro, "A function space view of bounded norm infinite width ReLU nets: The multivariate case," in *International Conference on Learning Representations*, 2020.

[17] R. Parhi and R. D. Nowak, "Banach space representer theorems for neural networks and ridge splines," *Journal of Machine Learning Research*, vol. 22, no. 43, pp. 1–40, 2021.

[18] R. Parhi and R. D. Nowak, "Near-minimax optimal estimation with shallow ReLU neural networks," *IEEE Transactions on Information Theory*, vol. 69, no. 2, pp. 1125–1140, 2023.

[19] J. W. Siegel, "Optimal approximation of zonoids and uniform approximation by shallow neural networks," *arXiv preprint arXiv:2307.15285*, 2023.

[20] J. W. Siegel and J. Xu, "Characterization of the variation spaces corresponding to shallow neural networks," *Constructive Approximation*, vol. 57, no. 3, pp. 1109–1132, 2023.

[21] J. W. Siegel and J. Xu, "Sharp bounds on the approximation rates, metric entropy, and n-widths of shallow neural networks," *Foundations of Computational Mathematics*, vol. 24, no. 2, pp. 481–537, 2024.

[22] M. J. Wainwright, *High-dimensional statistics: A non-asymptotic viewpoint*. Cambridge University Press, 2019, vol. 48.

[23] Y. Yang and A. Barron, "Information-theoretic determination of minimax rates of convergence," *Annals of Statistics*, pp. 1564–1599, 1999.