arXiv preprint. Under review.

Which Spaces can be Embedded in $\mathcal{L}_p$-type Reproducing Kernel Banach Space? A Characterization via Metric Entropy

Yiping Lu[1], **Daozhe Lin**[2] **and Qiang Du**[2]

[1] *Industrial Engineering & Management Sciences, Northwestern University*
[2] *Department of Applied Physics & Applied Mathematics, Columbia University*

This manuscript was compile on September 1, 2024

Abstract

In this paper, we establish a novel connection between the metric entropy growth and the embeddability of function spaces into reproducing kernel Hilbert/Banach spaces. Metric entropy characterizes the information complexity of function spaces and has implications for their approximability and learnability. Classical results show that embedding a function space into a reproducing kernel Hilbert space (RKHS) implies a bound on its metric entropy growth. Surprisingly, we prove a **converse**: a bound on the metric entropy growth of a function space allows its embedding to a $\mathcal{L}_p-$type Reproducing Kernel Banach Space (RKBS). This shows that the $\mathcal{L}_p-$type RKBS provides a broad modeling framework for learnable function classes with controlled metric entropies.Our results shed new light on the power and limitations of kernel methods for learning complex function spaces.

Keywords: *Metric Entropy, Reproducing Kernel Banach Space*

Corresponding author: Yiping Lu *E-mail address:* yiping.lu@northwestern.edu

1. Introduction

Learning a function from its finite samples is a fundamental science problem. A recent emerging trend in machine learning is to use Reproducing Kernel Hilbrt/Banach Spaces (RKHSs/RKBSs) [41, 51, 29, 53, 21] as a powerful framework for studying the theoretical properties of neural networks[5, 48, 42, 40, 6] and other machine learning models. The RKBS framework offers a principled approach to numerical implementable parametric representation via the representer theorem[46, 47, 33], characterizing the hypothesis spaces induced by neural networks[14, 24, 36] and study the generalization properties [1, 3, 8]. The Reproducing Kernel Banach Space (RKBS) framework offers a flexible and general approach to characterize complex machine learning estimators. However, most of the construction and statistical analysis in the literature focuses on and is based on the structure of $\mathcal{L}_p$-type RKBS, i.e., the feature space is specifically embedded into an $\mathcal{L}_p$ space. In this paper, we aim to answer the following questions for general machine learning problems:

*Can $\mathcal{L}_p-$type Reproducing Kernel Banach Spaces offer a general enough framework for machine learning studies? Which spaces can be embedded into a $\mathcal{L}_p-$**type** ?*

Surprisingly, we provide an affirmative answer to the previous questions. We demonstrate that every function class learnable with a polynomial number of data points with respect to the excess risk can be embedded into a $\mathcal{L}_p$-type Reproducing Kernel Banach space. This result indicates that $\mathcal{L}_p$-type Reproducing Kernel Banach spaces constitute a powerful and expressive model class for machine learning tasks.

To show this, we link the learnability and metric entropy [27] with the embedding to the reproducing Kernel Banach Space. Metric entropy quantifies the number of balls of a certain radius required to cover the hypothesis class. A smaller number of balls implies a simpler hypothesis class, which in turn suggests better generalization performance. Conversely, a larger number of balls indicates a more complex hypothesis class, potentially leading to over-fitting or poor generalization. Classical results show that embedding a function space into a reproducing kernel Hilbert space implies a polynomial bound on its metric entropy growth [43, 45].

Our main result demonstrates that if the growth rate of a Banach hypothesis space's metric entropy can be bounded by a polynomial function of the radius of the balls, then the hypothesis space can be embedded into a $\mathcal{L}_p$-type Reproducing Kernel Banach space for some $1 \le p \le 2$. This result indicates that if a function space can be learned with a polynomially large dataset with respect to the learning error, then it can be embedded into a p-norm Reproducing Kernel Banach Space. Thus, Reproducing Kernel Banach Spaces provide a powerful theoretical model for studying learnable datasets.

1.1. Related Works

Reproducing Kernel Hilbert Space and Reproducing Kernel Banach Space A Reproducing kernel Banach space (RKBS) is a space of functions on a given set Ω on which point evaluations are continuous linear functionals. For example, the space of $\mathbb{R}$-valued, bounded continuous functions $C^0(\Omega)$ on some metric space Ω is also a Reproducing Kernel Banach Space. Finally, the space $\ell_\infty(\Omega)$ of all bounded functions $f : \Omega \to \mathbb{R}$ equipped with the supremum norm is also a Reproducing Kernel Banach Space. A formal definition is given below.

Definition 1. *A reproducing kernel Banach space $\mathcal{B}$ on a prescribed nonempty set X is a Banach space of certain functions on X such that every point evaluation functional δ_x, $x \in X$ on B is continuous, that is, there exists a positive constant C_x such that*

$$|\delta_x(f)| = |f(x)| \le C_x \|f\|_{\mathcal{B}} \text{ for all } f \in B.$$

Note that in all RKBS $\mathcal{B}$ on Ω norm-convergence implies pointwise convergence, that is, if $(f_n) \subset \mathcal{B}$ is a sequence converging to some $f \in \mathcal{B}$ in the sense of $\|f_n - f\|_{\mathcal{B}} \to 0$, then $f_n(x) \to f(x)$ for all $x \in \Omega$. Note that in the special case with the norm $\|\cdot\|_{\mathcal{B}}$ being induced by an inner product, the space is called a Reproducing Kernel Hilbert Space (RKHS).

Compared to Hilbert spaces, Banach spaces possess much richer geometric structures, which are potentially useful for developing learning algorithms. For example, in some applications, a norm from a Banach space is invoked without being induced from an inner product. It is known that minimizing about the ℓ_p norm in $\mathbb{R}^d$ leads to sparsity of the minimizer when p is close to 1.

As in the case of RKHS, a feature map (which is the Reproducing kernel in Hilbert space) can also be introduced as an appropriate measurement of similarities between elements in the domain of the function. To see this, [53, 29, 5] provides a way to construct the Reproducing Kernel Banach Spaces via feature map. In this construction, the reproducing kernels naturally represents the similarity of two elements in the feature space.

Construction of a Reproducing Kernel Banach Space

For a Banach space $\mathcal{W}$, let $[\cdot,\cdot]_{\mathcal{W}} : \mathcal{W}' \times \mathcal{W} \to \mathbb{R}$ be its duality pairing. Suppose there exist an nonempty set Ω and a corresponding feature mappings $\Phi : \Omega \to \mathcal{W}'$,. We can construct a Reproducing Kernel Banach Space as

$$\mathcal{B} := \{f_v(x) := [\Phi(x), v]_{\mathcal{W}} : v \in \mathcal{W}, x \in \Omega\}$$

with norm$\|f_v\|_{\mathcal{B}} := \inf\{\|v\|_{\mathcal{W}} : v \in \mathcal{W}$ with $f = [\Phi(\cdot), v]_{\mathcal{W}}\}$.

In [5], the relation between the feature map construction and the RKBS has been established in the following theorem.

Theorem 1 (Proposition 3.3[5]). *A space $\mathcal{B}$ of function on Ω is a RKBS if and only if there is a Banach space $\mathcal{W}$ and a feature map $\Phi : \Omega \to \mathcal{W}'$ such that $\mathcal{B}$ is constructed by the method above.*

As discussed in [5], the feature maps are generally not unique, and the relation between the Banach space W and the RKBS $\mathcal{B}$ is presented in the following technique remark:

Remark. *The RKBS $\mathcal{B}$ is isometrically isomorphic to the quotient space $\mathcal{W}/\mathcal{N}$, where*

$$\mathcal{N} = \{v \in \mathcal{W} : f_v = 0\}$$

$\mathcal{L}^p$-type Reproducig Kernel Banach Space For a probability measure space $(\Omega, \mathcal{M}, \mu)$, the space $\mathcal{L}_p(\mu)$ for $1 \leq p < \infty$ is defined as $\mathcal{L}_p(\mu) = \left\{ f : \Omega \to \mathbb{R} \;\middle|\; f \text{ is measurable and } \int_X |f|^p \, d\mu < \infty \right\}$. It is known that, under proper assumptions, the Reproducing Kernel Hilbert Space [45] can be characterized in two equivalent feature spaces: ℓ_2 and $\mathcal{L}_2(\mu)$.

In this paper, our focus lies in the generalization of the $\mathcal{L}_2$ characterization of the RKHS to the RKBS, i.e., the $\mathcal{L}_p$-type Reproducing Kernel Banach space, defined as follows:

Definition 2 ($\mathcal{L}_p-$type Reproducing Kernel Banach Space). *If the feature space $\mathcal{W}$ is given by $\mathcal{W} = \mathcal{L}_p(\mu)$ for some measure μ, then we call the constructed Reproducing Kernel Banach Space as $\mathcal{L}_p-$type.*

Example 1 (Reproducing Kernel Hilbert Space). *$\mathcal{L}_2-$type Reproducing Kernel Banach Space is a Reproducing Kernel Hilbert Space.*

Example 2 (Barron Space [26, 4, 36, 32, 50]). *Barron space is used to characterize the approximation properties of shallow neural networks from the point of view of non-linear dictionary approximation. Let $\mathcal{X}$ be a Banach space and $\mathbb{D} \subset \mathcal{X}$ be a uniformly bounded dictionary, i.e. $\mathbb{D}$ is a subset such that $\sup_{h \in \mathbb{D}} \|h\|_{\mathcal{X}} = K_{\mathbb{D}} < \infty$. Barron space is concerned with approximating a target function f by non-linear n-term dictionary expansions, i.e. by an element of the set $\Sigma(\mathbb{D}) = \left\{\sum_{j=1}^n a_j h_j : h_j \in \mathbb{D}\right\}$. The approximation is non-linear since the elements h_j in the expansion will depend upon the target function f. It is often also important to have some control over the coefficients a_j. For this purpose, we introduce the sets*

$$\Sigma_M^p(\mathbb{D}) = \left\{ \sum_{j=1}^n a_j h_j : h_j \in \mathbb{D}, n \in \mathbb{N}, \left(\sum_{i=1}^n |a_i|^p \right)^{\frac{1}{p}} \leq M \right\}$$

[42] showed that the Barron space $\Sigma_M^1(\mathbb{D})$ can be represented as a $\mathcal{L}_1-$type RKBS. Furthermore, we will show later on that $\Sigma_M^1(\mathbb{D})$ can be embeded into a Reproducing Kernel Hilbert Space with a weak assumption on the dictionary.

Learnability and Metric Entropy The metric entropy [27, 49, 23] indicates how precisely we can specify elements in a function class given a finite mount of bits information and it is closely related to the approximation by stable non-linear methods [16]. Metric entropy is quantified as the log of the covering number, which counts the minimum number of balls of a certain radius needed to cover the space. In information theory, metric entropy is the natural characterization of the complexity of a function class. [7, 19, 2] showed that a concept class is learnable with respect to a fixed data distribution if and only if the concept class is finitely coverable (i.e., there exists a finite ϵ cover for every $\epsilon > 0$) with respect to the distribution. In this paper, we extend this result to concept classes that can be learned with a polynomially large dataset with respect to the learning error. We demonstrate that the growth speed of the metric entropy of such concept classes can also be polynomially bounded.

1.2. Contribution

In this paper, we aim to establish connections between $\mathcal{L}_p$-type RKBS and function classes that can be learned efficiently with a polynomially large dataset with respect to the learning error. Specifically, it is shown that such classes have metric entropies enjoys a power law relationship with the covering radius and can be embedded into an $\mathcal{L}_p$-type reproducing kernel Banach space (RKBS). Classical results indicate that the ability to embed a hypothesis space into a reproducing kernel Hilbert space (RKHS) implies a metric entropy decay rate (Steinwart, 2000), which in turn suggests learnability. Our novel contribution is establishing a converse connection between the metric entropy and the type of a Banach space. We demonstrate that concept classes whose metric entropy can be polynomially bounded lead to the embedding into $\mathcal{L}_p$-type RKBSs. These results highlight the generality of using $\mathcal{L}_p$-type RKBSs as prototypes for learnable function classes and are particularly useful because bounding the metric entropy of a function class is often straightforward. Several illustrative examples are provided in Section 4.

2. Preliminary

Type and Cotype of a Banach Space The type and cotype of a Banach space are classification s of Banach spaces through probability theory. They measure how far a Banach space is from a Hilbert space. The idea of type and cotype emerged from the work of J. Hoffmann-Jorgensen, S. Kwapien, B. Maurey and G. Pisier in the early 1970's. The type of a Banach space is defined as follows

Definition 3 (Banach Space of Type-p). *A Banach space $\mathcal{B}$ is of type p for $p \in [1, 2]$ if there exist a finite constant $C \geq 1$ such that for any integer n and all finite sequences $(x_i)_{i=1}^n \in \mathcal{B}^n$ we have*

$$\left(\mathbb{E} \left\| \sum_{i=1}^n \varepsilon_i x_i \right\|_{\mathcal{B}}^p \right)^{\frac{1}{p}} \leq C \left(\sum_{i=1}^n \|x_i\|_{\mathcal{B}}^p \right)^{\frac{1}{p}}$$

where ε is a sequence of independent Rademacher random variables, i.e., $P(\varepsilon_i = -1) = P(\varepsilon_i = 1) = \frac{1}{2}$ and $\mathbb{E}[\varepsilon_i \varepsilon_m] = 0$ for $i \neq m$

and $Var[\varepsilon_i] = 1$. *The sharpest constant C is called type p constant and denoted as* $T_p(\mathcal{B})$.

Definition 4 (Banach Space of Cotype-q). *A Banach space $\mathcal{B}$ is of cotype q for $q \in [2, \infty]$ if there exist a finite constant $C \geq 1$ such that*

$$\left(\mathbb{E}\left\|\sum_{i=1}^{n} \varepsilon_i x_i\right\|_{\mathcal{B}}^{q}\right)^{\frac{1}{q}} \geq \frac{1}{C}\left(\sum_{i=1}^{n} \|x_i\|_{\mathcal{B}}^{q}\right)^{\frac{1}{q}},$$

if $2 \leq q < \infty$ for any integer n and all finite sequences $(x_i)_{i=1}^n \in \mathcal{B}^n$. The sharpest constant C is called cotype q constant and denoted as $C_q(\mathcal{B})$.

The previous work [37] utilizes the following Kwapien's Theorem to charaterize whether there exists a RKHS H with a bounded kernel such that certain Banach space $E \subset H$. As a result, it was shown that typical classes of function spaces described by the smoothness have a strong dependence on the underlying dimension: the smoothness s required for the space E needs to grow *proportionally* to the underlying dimension in order to allow for the embedding to a RKHS H.

Theorem 2 (Kwapien's Theorem [28, 52]). *For a Banach space E, $id : E \to E$ being Type 2 and Cotype 2 is equivalent to E being isomorphic to a Hilbert Space*

The relation of the type of a Banach space and $\mathcal{L}_p$ can be characterized by the following Theorem:

Theorem 3 (Lemma 11.18 in [17], corollary of Pietsch Domination Theorem and Maurey-Pisier Theorem). *Consider type$-p$ $(1 < p \leq 2)$ Banach Space $\mathcal{X}$ which is a closed subspace of $\mathcal{L}_1(\mu)$ for some measure μ, then for any $1 < r < p$ there exists isomorphic embedding $u : \mathcal{X} \to L_r(\nu)$ (isomorphic to a subspace of $\mathcal{L}_r(\nu)$) for some probability ν.*

Covering Number and Metric Entropy Covering number and metric entropy measure the size of the hypotheses space on which we work. For many machine learning problems, a natural way to measure the size of the set is via the number of balls of a fixed radius $\delta > 0$ required to cover the set.

Definition 5 ($\delta-$Covering Number for metric space $(\mathcal{X}, d)$ [49]). *Consider a metric space $(\mathcal{X}, d)$ where d is the metric for space $\mathcal{X}$. Let $\delta \geq 0$. A δ-covering or δ-net of metric space $(\mathcal{X}, d)$ is a set of elements of $\mathcal{X}$ given by $\{\theta_1, \dots, \theta_N\} \subseteq \mathcal{X}$ where $N = N(\delta)$, such that for any $\theta \in X$, there exists $i \in [N]$ such that $d(\theta, \theta_i) \leq \delta$. The δ-covering number of $(\mathcal{X}, d)$, denoted as $N(\delta, \mathcal{X}, d)$, is the smallest cardinality of all δ-covering.*

We can define a related measure—more useful for constructing our lower bounds—of size that is related to the number of disjoint balls of radius $\delta > 0$ that can be placed into the set

Definition 6 (δ-Packing numbers for metric space $(\mathcal{X}, d)$). *A δ-packing of $(\mathcal{X}, d)$ is a set of elements of $\mathcal{X}$ given by $\{\theta_1, \dots, \theta_M\} \subseteq \mathcal{X}$ where $M = M(\delta)$, such that for all $i \neq j$, $d(\theta_i, \theta_j) > \delta$. The δ-packing number of $(\mathcal{X}, d)$, denoted as $M(\delta, \mathcal{X}, d)$, is the largest cardinality of all δ-packing set.*

The following lemma showed that the packing and covering numbers of a set are in fact closely related:

Lemma 1 (Lemma 4.3.8 [18]). *For any $\delta > 0$, $M(2\delta, \mathcal{X}, d) \leq N(\delta, \mathcal{X}, d) \leq M(\delta, \mathcal{X}, d)$*

The metric entropy, which is defined as log of the covering number, indicate how precisely we can specify elements in a function class given fixed bits of information.

Definition 7. *The metric entropy of $(\mathcal{X}, d)$ is defined as $\log N(\delta, \mathcal{X}, d)$.*

3. Main Results

In recent literature, reproducing kernel Banach spaces (RKBS) have been gaining interest for the analysis of neural networks. Moreover, RKBS also offers a versatile and comprehensive framework for characterizing complex machine learning estimators. However, the majority of the constructions and statistical analyses in the literature are concentrated on and based on the structure of $\mathcal{L}_p$-type RKBS, specifically embedding the feature space into an $\mathcal{L}_p$ space. However, *we still do not know whether $\mathcal{L}_p$-type RKBS is a flexible enough modeling*. In this paper, we consider the following questions:

> **Question.** Given a RKBS E of functions from $\Omega \to \mathbb{R}$, does there exist an $\mathcal{L}_p-$type RKBS $\mathcal{B}_p$ on X with the embeddings $E \hookrightarrow \mathcal{B}_p \hookrightarrow F = \mathcal{L}_\infty(\Omega)$, where $\mathcal{L}_\infty(\Omega)$ denotes the space of all the pointwise bounded function on Ω.

Recently, the question was studied in [37] for the case $p = 2$. The authors showed that there exists no Reproducing Kernel Hilbert Space $\mathcal{H}$ with a bounded kernel such that the space of all bounded, continuous functions from Ω to $\mathbb{R}$ satisfies $C(\Omega) \subset \mathcal{H}$. At the same time, the smoothness required for the space E needs to grow *proportionally* to the underlying dimension in order to allow for embedding into an intermediate RKHS $\mathcal{H}$.

In the literature, one way to describe the "size" of a RKBS is by means of denseness in a surrounding space F and universal consistency can be established for kernel-based learning algorithms if universal kernels are used, [44, 45]. However, universal consistency does not mean that the problem can be efficiently learned. To precisely approximate arbitrary continuous functions, having a large RKHS norm is sufficient but may lead to a large sample complexity requirement [9, 22].

Surprisingly, we show the following connection between the sample complexity and the embedding to $\mathcal{L}_p-$type RKBS :

All the polynomially learnable RKBS can be embeded to a $\mathcal{L}_p-$type RKBS.

We first demonstrate the relationship between metric entropy and embedding in the following theorem, and subsequently establish the connection between metric entropy and sample complexity in Section **??**. The significance of this result lies in the fact that estimating metric entropy is considerably more straightforward in practice than finding the embedding. For instance, the metric entropies of all classical Sobolev and Besov finite balls in $\mathcal{L}_p$ or Sobolev spaces are well-known.

> **Theorem 4.** *Given a bounded domain $\Omega \in \mathcal{R}^d$, a RKBS E of functions on Ω, and $F = \ell_\infty(\Omega)$ on Ω, which means the embedding $id : E \to F$ is a compact embedding. If the growth of metric entropy can be bounded via*
>
> $$\mathcal{E}_E^F(\delta) := \log N(\delta, \{x \in E : \|x\|_E \leq 1\}, \|\cdot\|_F) \leq \delta^{-p}, p \geq 2.$$
>
> *Then for any $s > p$, there exist a $\mathcal{L}_s-$type RKBS $\mathcal{B}_s$, such that*
>
> $$E \hookrightarrow \mathcal{B}_s \hookrightarrow F.$$

Related Work A series of earlier works [10, 34, 11, 12, 13] providedthe metric entropy control of the convex hull in a type-p Banach space which showed that a type-p Banach space always has metric entorpy control. [25] showed that a Banach space is of weak type p if and only if it is of entropy type p' with

$1/p' + 1/p = 1$. All type-p Banach space is weak type-p [35]. Thus our work showed a stronger result than [25].

3.1. Proof Sketch

A sketch of the proof of metric entorpy bound to embedding is given below.

1. We first bound the Rademacher norm $\mathbb{E}_{\epsilon_i}\frac{1}{n}\left\|\sum_{i=1}^n \epsilon_i x_i\right\|_F$ of the Banach space E via generalizations of the Massart's lemma and Dudley's chaining theorem to general Banach space.
2. We provide a novel lemma which shows that type of a Banach space can be inferred from the estimation of Rademacher norm $\mathbb{E}_{\epsilon_i}\frac{1}{n}\left\|\sum_{i=1}^n \epsilon_i x_i\right\|_F$
3. Using the isomorphism between the Banach space $\hat{E} = (E, \|\cdot\|_F)$ and subspace of $\mathcal{L}_{s'}(\mu)$ to construct the feature mapping of the $\mathcal{L}_s$−type RKBS.

To be more specific,given $p > 2$, for any $s > p$, our proof takes on the following pathway:

Metric entropy control δ^{-p} → Type s' → $\mathcal{L}_s$-type RKBS

where $1 < s', p' < 2$ such that $1/s + 1/s' = 1/p + 1/p' = 1$. The detailed proof can be found in the appendix.

Metric Entropy Bound leads to bound of the Rademacher norm We generalize the Dudley's Chaining Theorem to abstract Banach space, so that we can show a $n^{-\frac{1}{p}}$ decay of the Rademacher norm $\mathbb{E}_{\epsilon_i}\frac{1}{n}\left\|\sum_{i=1}^n \epsilon_i x_i\right\|_F$ based on the assumption that $\log \mathcal{E}_E^F(\delta)$.

Theorem 5 (Dudley's Chaining for Abstract Banach Space). *Given two Banach Spaces $(E, \|\cdot\|_E)$ and $(F, \|\cdot\|_F)$, an upper bound on the Rademacher norm can be showed by a Dudley's chaining argument as follows:*

$$\mathbb{E}_{\epsilon_i} \sup_{\substack{x_1,\cdots,x_n\in E\\ \|x_1\|_E\le 1,\|x_2\|_E\le 1,\\ \cdots,\|x_n\|_E\le 1}} \frac{1}{n}\left\|\sum_{i=1}^n \epsilon_i x_i\right\|_F \le C\inf_\alpha\left\{\alpha + \frac{6}{\sqrt{n}}\int_\alpha^2 \sqrt{\mathcal{E}_E^F(\delta)}\,d\delta\right\},$$

holds for all $0 < \alpha < 1$, where: ϵ_i are independent Rademacher variables, taking values in $\{-1, +1\}$ with equal probability.

According to Theorem 5, if the entropy number $\mathcal{E}_E^F(\delta) \le \delta^{-p}$ for some $p > 2$, we can have

$$\begin{aligned}\mathbb{E}_{\epsilon_i}\frac{1}{n}\left\|\sum_{i=1}^n \epsilon_i x_i\right\|_F &\lesssim n^{-\frac{1}{p}} + \frac{1}{\sqrt{n}}\int_{n^{-\frac{1}{p}}}^1 \sqrt{\delta^{-p}}d\delta \quad (\text{Take } \alpha = n^{-\frac{1}{p}})\\ &\lesssim n^{-\frac{1}{p}} \quad (\text{The integral is of } O(n^{-\frac{1}{p}}))\end{aligned} \tag{1}$$

for all $\|x_i\|_E \le 1$.

Proof of Theorem 5. We first extend Massart's lemma to Banach space.

Lemma 2 (Generalized Massart's Lemma in Banach Space). *Let $\mathcal{B}$ be banach space and $A \subset \mathcal{B}$ be a finite set with $r = \max_{a\in A}\|a\|_{\mathcal{B}}$, then*

$$\mathbb{E}\left[\sup_{a\in A}\left\|\sum_{i=1}^m \sigma_i a_i\right\|_{\mathcal{B}}\right] \le r\sqrt{2\log|A|}$$

where $|A|$ denotes the cardinality of A, σ_i's are Rademacher random variables (which are independent and identically distributed random variables taking values $\{-1, 1\}$ with equal probability) and a_i are components of vector a.

Proof. Here's a proof of the Massart's Lemma. It basically follows from Hoeffding's Lemma.

$$\begin{aligned}\exp\left(\lambda\mathbb{E}\left[\sup_{a\in A}\left\|\sum_{i=1}^m \sigma_i a_i\right\|_{\mathcal{B}}\right]\right) &\le \mathbb{E}\exp\left(\left[\sup_{a\in A}\left\|\sum_{i=1}^m \lambda\sigma_i a_i\right\|_{\mathcal{B}}\right]\right)\\ &\quad(\text{Jensen's for } \lambda > 0)\\ &\le \mathbb{E}\left[\sum_{a\in A}\exp\left(\left\|\sum_{i=1}^m \lambda\sigma_i a_i\right\|_{\mathcal{B}}\right)\right]\\ &\le \sum_{a\in A}\mathbb{E}\left[\exp\left(\left\|\sum_{i=1}^m \lambda\sigma_i a_i\right\|_{\mathcal{B}}\right)\right]\\ &\quad(\text{as } \sigma_i\text{'s are i.i.d.})\\ &\le \sum_{a\in A}\prod_{i=1}^m \mathbb{E}\left[\exp\left(\|\lambda\sigma_i a_i\|_{\mathcal{B}}\right)\right]\\ &\quad(\text{by Traingular Inequality})\\ &\le \sum_{a\in A}\exp\left(\frac{m\lambda^2 r^2}{2}\right)\\ &\quad(\text{Using Hoeffding's Lemma})\\ &= |A|\exp\left(\frac{m\lambda^2 r^2}{2}\right)\end{aligned}$$

Applying the logarithm operator to the inequality and multiplying by $\frac{1}{\lambda}$

$$\frac{1}{\lambda}\log\left(\exp\left(\lambda\mathbb{E}\left[\sup_{a\in A}\left\|\sum_{i=1}^m \sigma_i a_i\right\|_{\mathcal{B}}\right]\right)\right) \le \frac{1}{\lambda}\log\left(|A|\exp\left(\frac{m\lambda^2 r^2}{2}\right)\right)$$

$$\mathbb{E}\left[\sup_{a\in A}\left\|\sum_{i=1}^m \sigma_i a_i\right\|_{\mathcal{B}}\right] \le \frac{\log|A|}{\lambda} + \frac{m\lambda r^2}{2}$$

Set value of $\lambda = \sqrt{\frac{2\log|A|}{mr^2}}$ above to obtain

$$\mathbb{E}\left[\sup_{a\in A}\left\|\sum_{i=1}^m \sigma_i a_i\right\|_{\mathcal{B}}\right] \le r\sqrt{2\log|A|}$$

□

To prove the Dudley's Chaining Theorem 5 for abstract Banach spaces, we start by the most crude ϵ-cover for our function class. To simplify the notation we denote:

$$N_\delta := N(\delta, \{x \in E : \|x\|_E \le 1\}, \|\cdot\|_F)$$

For any $0 < \alpha < 1$, we can set $\epsilon_0 = 2^m\alpha$, where m is choosed properly such that $\epsilon_0 \ge \sup_{i=1,\cdots,n}\|x_i\|_E$ and note that we have the covering net $\mathcal{N}_{\epsilon_0} = \{g_0\}$ for $g_0 = 0$ which implies $N_{\epsilon_0} = 1$.

Next , define the sequence of epsilon covers $\mathcal{N}_{\epsilon_j}$ by setting $\epsilon_j = 2^{-j}\epsilon_0 = 2^{m-j}\alpha$ for $j = 0, ..., m$. By definition, $\forall x \in E, \|x\|_E \le 1$, we can find $g_j(x) \in \mathcal{N}_{\epsilon_j}$ that such that $\|x - g_j(x)\|_F \le \epsilon$. Therefore we can write the telescopic sum

$$x = x - g_m + \sum_{j=1}^m g_j(x) - g_{j-1}(x).$$

By triangle inequality, for any x we have $\|g_j(x) - g_{j-1}(x)\|_F \le \|g_j(x) - x\|_F + \|x - g_{j-1}(x)\|_F \le \epsilon_j + \epsilon_{j-1} = 3\epsilon_j$. Thus,

$$\mathbb{E}_{\epsilon_i} \sup_{\substack{x_1,\cdots,x_n\in E\\ \|x_1\|_E\le 1,\|x_2\|_E\le 1,\\ \cdots,\|x_n\|_E\le 1}} \frac{1}{n}\left\|\sum_{i=1}^n \epsilon_i x_i\right\|_F \le \mathbb{E}\frac{1}{n}\left[\sup_{\substack{x_1,\cdots,x_n\in E\\ \|x_1\|_E\le 1,\|x_2\|_E\le 1,\\ \cdots,\|x_n\|_E\le 1}} \left\|\sum_{i=1}^n \epsilon_i(x - g_m(x))\right\|_F\right.$$

$$
\begin{aligned}
&+ \sum_{j=1}^{m} \left\| \sum_{i=1}^{n} \epsilon_i (g_j(x_i) - g_{j-1}(x_i)) \right\|_F \Bigg] \\
&\le \frac{1}{n} \cdot n\epsilon_m + \mathbb{E}\frac{1}{n} \left[\sup_{\substack{x_1,\cdots,x_n \in E \\ \|x_1\|_E \le 1, \|x_2\|_E \le 1, \\ \cdots, \|x_n\|_E \le 1}} \sum_{j=1}^{m} \left\| \sum_{i=1}^{n} \epsilon_i (g_j(x_i) - g_{j-1}(x_i)) \right\|_F \right] \\
&\le \epsilon_m + \mathbb{E}\frac{1}{n} \left[\sum_{j=1}^{m} \sup_{\substack{x_1,\cdots,x_n \in E \\ \|x_1\|_E \le 1, \|x_2\|_E \le 1, \\ \cdots, \|x_n\|_E \le 1}} \left\| \sum_{i=1}^{n} \epsilon_i (g_j(x_i) - g_{j-1}(x_i)) \right\|_F \right] \\
&\qquad\qquad (\text{by } \sup \textstyle\sum \le \sum \sup) \\
&\le \alpha + \mathbb{E}\frac{1}{n} \left[\sum_{j=1}^{m} \sup_{\substack{y_1,\cdots,y_n \in E \\ \|y_1\|_E \le 3\epsilon_j, \|y_2\|_E \le 3\epsilon_j, \\ \cdots, \|y_n\|_E \le 3\epsilon_j}} \left\| \sum_{i=1}^{n} \epsilon_i y_i \right\|_F \right] \\
&\le \alpha + \sum_{j=1}^{m} \frac{3\epsilon_j}{n} \sqrt{2n \log \left|\mathcal{N}_{\epsilon_j}\right|^2} \qquad (\text{by Massart's lemma}) \\
&\le \alpha + \frac{6}{\sqrt{n}} \sum_{j=1}^{m} (\epsilon_j - \epsilon_{j+1}) \sqrt{\log \left|\mathcal{N}_{\epsilon_j}\right|} \le \alpha + \frac{6}{\sqrt{n}} \int_{\epsilon_m}^{\epsilon_0} \sqrt{\log |\mathcal{N}_t|} dt. \\
&\le \alpha + \frac{6}{\sqrt{n}} \int_{\alpha}^{D} \sqrt{\log |\mathcal{N}_t|} dt.
\end{aligned}
$$

where we take $D = 2 \sup_{i=1,\cdots,n} \|x_i\|_E$ and therefore $D > \epsilon_0$. □

From the Bounded Rademacher norm to the Type of the Banach Space We now present a novel lemma which shows that the previous estimation of the Rademacher norm $\mathbb{E}_{\epsilon_i} \frac{1}{n} \left\| \sum_{i=1}^{n} \epsilon_i x_i \right\|_F$ can imply the type of the Banach space.

Lemma 3 (Techinque Contribution: *From bounded Rademacher norm to type of the Banach space*)**.** *Given two Banach spaces $(E, \|\cdot\|_E)$ and $(F, \|\cdot\|_F)$ on X where we have the embedding $E \hookrightarrow F$, if for $1 \le p' \le 2$, the following inequality*

$$\mathbb{E}_{\epsilon_i} \left\| \sum_{i=1}^{n} \epsilon_i x_i \right\|_F \lesssim n^{\frac{1}{p'}}$$

holds when $\|x_i\|_E \le 1, i = 1, \cdots, n, \forall n \in \mathbb{N}$, then $\hat{E} = (E, \|\cdot\|_F)$ is of the type s', for each $1 \le s' < p'$.

Proof.

Lemma 4 (Kahane-Khintchine Inequality)**.** *If $(E, \|\cdot\|_E)$ is any normed space and $x_1, \cdots, x_n \in E$, then*

$$\mathbb{E}\|\sum_{i=1}^{n} \epsilon_i x_i\|_E \le \left(\mathbb{E}\|\sum_{i=1}^{n} \epsilon_i x_i\|_E^p \right)^{1/p} \le K_P \mathbb{E}\|\sum_{i=1}^{n} \epsilon_i x_i\|_E$$

Using the previous lemma, we are ready to prove the Lemma 3

We first prove for all $\|x_i\|_E \le 1, i = 1, \cdots, N, \forall N \in \mathbb{N}$, the inequality holds. By the embedding $E \hookrightarrow F$, we have $\|x_i\|_F \le c\|x_i\|_E \le c$ for some constant $c > 0$, WLOG we can assume $c = 1$. In the following proof, we will fix an $m \in \mathbb{N}$. For $j, k = 0, 1, 2, ...$ define the two sets

$$U_j = \left\{ i : \|x_i\|_F \in \left(\frac{1}{2^{j+1}}, \frac{1}{2^j} \right] \right\} \quad \text{and} \quad V_k = \{ j : |U_j| \in (m^{k-1}, m^k] \}$$

Fix a k and a $j \in V_k$. We will perform a calculation as above, but now taking advantage of the assumption that $s < q$, which buys us a bit of room that will come in handy later. Let $\tau = s^{-1} - q^{-1} > 0$. By the fact that $|U_j| \le m^k$, $\|2^k x_i\|_F \le 1$ and using Lemma 4

$$\left(\mathbb{E} \left\| \sum_{i \in U_j} \epsilon_i x_i \right\|_F^s \right)^{\frac{1}{s}} \lesssim \mathbb{E} \left\| \sum_{i \in U_j} \epsilon_i x_i \right\|_F = \frac{1}{2^k} \mathbb{E} \left\| \sum_{i \in U_j} \epsilon_i \left(2^k x_i\right) \right\|_F \le \frac{1}{2^j} m^{k/q} = \frac{1}{2^j} m^{k/s} \tag{2}$$

For each j define $f_j : \{-1, 1\}^n \to \mathbb{R}$ by $f_j(\epsilon) = \left\| \sum_{i \in U_j} \epsilon_i x_i \right\|_F$. Then we have

$$
\begin{aligned}
\left(\mathbb{E} \left\| \sum_{i=1}^{n} \epsilon_i x_i \right\|_F^s \right)^{\frac{1}{s}} &= \left(\mathbb{E} \left\| \sum_{j=0}^{\infty} \sum_{i \in U_j} \epsilon_i x_i \right\|_F^s \right)^{\frac{1}{s}} \le \left(\mathbb{E}_\epsilon \sum_{j=0}^{\infty} f_j(\epsilon)^s \right)^{\frac{1}{s}} && (3) \\
&\le \sum_{j=0}^{\infty} \left(\mathbb{E}_\epsilon f_j(\epsilon)^s \right)^{\frac{1}{s}} = \sum_{k=0}^{\infty} \sum_{j \in V_k} \left(\mathbb{E}_\epsilon \left\| \sum_{i \in U_j} \epsilon_i x_i \right\|_F^s \right)^{\frac{1}{s}} && (4) \\
&\le \sum_{k=0}^{\infty} \sum_{j \in V_k} m^{-\tau k} \frac{m^{k/s}}{2^j} \le \left(\sum_{k=0}^{\infty} m^{-\tau k} \right) \max_k \left\{ m^{k/s} \sum_{j \in V_k} \frac{1}{2^j} \right\} && (5) \\
&\le m^{\frac{1}{s}} \max_k \left\{ |U_j|^{\frac{1}{s}} \sum_{j \in V_k} \frac{1}{2^j} \right\} \qquad (\text{for } |U_j| \ge m^{k-1}) && (6) \\
&\le 2m^{\frac{1}{s}} \max_k \max_{j \in V_k} \left\{ |U_j|^{\frac{1}{s}} \frac{1}{2^j} : j \in V_k \right\} && (7) \\
&\qquad \left(\text{for } \sum_{j \in V_k} \frac{1}{2^j} \le \sum_{j \ge \min(V_k)} \frac{1}{2^j} \le \frac{2}{2^{\min(V_k)}} \right) && (8) \\
&\le 2m^{\frac{1}{s}} \max_j \left\{ |U_j|^{\frac{1}{s}} \frac{1}{2^j} : j = 0, 1, 2, ... \right\} && (9) \\
&\le 4m^{\frac{1}{s}} \max_j \left\{ \left(\sum_{i \in U_j} \|x_i\|_F^s \right)^{\frac{1}{s}} : j = 0, 1, 2, ... \right\} && (10) \\
&\qquad \left(\text{for } \frac{1}{2^{j+1}} \le \|x_i\| \text{ holds } \forall i \in U_j \right) && (11) \\
&\le 4m^{\frac{1}{s}} \left(\sum_{i=1}^{n} \|x_i\|_F^s \right)^{\frac{1}{s}}. && (12)
\end{aligned}
$$

Gathering up the implicit factors in the above inequalities and noting that they all only depend on r and τ gives the result. Now we prove for all $x_i \in E, i = 1, \cdots, n, \forall n \in \mathbb{E}$ and any $s < q$, we have

$$\left(\mathbb{E} \left\| \sum_{i=1}^{n} \epsilon_i x_i \right\|_F^s \right)^{\frac{1}{s}} \lesssim \left(\sum_{i=1}^{n} \|x_i\|_F^s \right)^{\frac{1}{s}}. \tag{13}$$

Let $\tilde{x}_i = \frac{x_i}{\max_{i=1}^{n}\{\|x_i\|_E\}}$, then $\|\tilde{x}_i\|_E \le 1$ holds for all $i = 1, \cdots, n$. From the previous proof we have

$$\left(\mathbb{E} \left\| \sum_{i=1}^{n} \epsilon_i \tilde{x}_i \right\|_F^s \right)^{\frac{1}{s}} \lesssim \left(\sum_{i=1}^{n} \|\tilde{x}_i\|_F^s \right)^{\frac{1}{s}},$$

which will leads to the complete proof of (3). □

Embedding to $\mathcal{L}_s$-type RKBS From the previous lemma, the space $\hat{E} = (E, \|\cdot\|_F)$ is type s' for all $1 \le s' \le p'$. By Theorem 3, $\hat{E}$ is isometric to a closed subspace $\mathcal{W}_{s'}$ of $\mathcal{L}_{s'}(\mu)$ for $1 \le s' \le p'$. Now fixing an $1 \le s' \le p'$, we can construct the $\mathcal{L}_s$-type RKBS

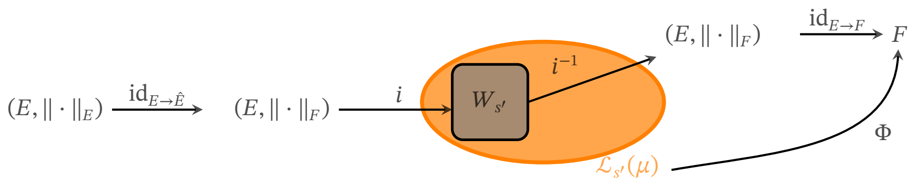

Figure 1. In our paper , we first use the type of a RKBS to build an isomorphic mapping to a subspace of $\mathcal{L}_p(\mu)$ for some probability measure μ. Then we construct the $\mathcal{L}_p$-type RKBS via an extension maps from $\mathcal{L}_p(\mu)$ to F.

$\mathcal{B}_s$ using the mapping relation in Figure 1. We firstly use Hahn-Banach continuous extension theorem to extend $id_{\hat{E}\to F}\circ i^{-1}$ to a continuous linear functional Φ from $\mathcal{L}_{s'}(\mu) \to F$ such that $\Phi|_{W_{s'}} = i^{-1}\circ id_{\hat{E}\to F}|_{W_{s'}}$. We define the feature map via $\phi : \Omega \to \mathcal{L}_s(\mu)$ by

$$\phi(x) := \Phi^*\delta_x^F$$

where $\delta_x^F \in F'$ denotes the evaluation functional at x acting on F and $\Phi^* : F' \to \mathcal{L}_{s'}(\mu)$ is the adjoint of operator that is uniquely determined by

$$[f, \Phi h]_F = [\Phi^* f, h]_{\mathcal{L}_s(\mu)}, \qquad \text{for all } f \in F', h \in \mathcal{L}_s(\mu).$$

Then we have for any $e \in \hat{E}$

$$\begin{aligned}[][\phi(x), i(e)]_{\mathcal{L}_s(\mu)} &= [\Phi^*\delta_x^F, i(e)]_{\mathcal{L}_s(\mu)} = [\delta_x^F, \Phi i(e)]_F \\ &\overset{(1)}{=} [\delta_x^F, id_{\hat{E}\to F}(e)]_F = id_{\hat{E}\to F}(e)(x),\end{aligned} \tag{14}$$

where (1) is based on the fact that $\Phi i(e) = (id_{\hat{E}\to F}\circ i^{-1})(i(e)) = id_{\hat{E}\to F}(e)$. Now we define the RKBS,

$$\mathcal{B}_s := \left\{ f_v(x) := [\phi(x), v]_{\mathcal{L}_{s'}} : v \in W_{s'}, x \in \Omega \right\}$$

then we can show that $E \hookrightarrow \mathcal{B}_s \hookrightarrow F$. The detailed proof is left to the Appendix A.

4. Applications

Spaces of (Generalized) Mixed Smoothness The Besov space is a considerably general function space including the Hölder space and Sobolev space, and especially can capture spatial inhomogeneity of smoothness.

Definition 8 (Besov Space [20], Definition 2.2.1). *Let $0 \le s < \infty$, $1 \le p \le \infty$, and $1 \le q \le \infty$, with $q = 1$ in case $s = 0$. For $f \in L^p(\mathbb{R}^d, \lambda)$ define*

$$\|f\|_{s,p,q} := \left(\sum_{k=0}^{\infty} 2^{ksq} \|\mathcal{F}^{-1}(\phi_k \mathcal{F} f)\|_p \right)^{1/q}$$

where ϕ_0 is a complex-valued C^∞-function on $\mathbb{R}^d$ with $\phi_0(x) = 1$ if $\|x\| \le 1$ and $\phi_0(x) = 0$ if $\|x\| \ge 3/2$. Define $\phi_1(x) = \phi_0(x/2) - \phi_0(x)$ and $\phi_k(x) = \phi_1(2^{-k+1}x)$ for $k \in \mathbb{N}$. (ϕ_k form a dyadic resolution of unity) and $\mathcal{F}$ denote the Fourier transform acting on this space (with scaling constant $(2\pi)^{-d/2}$). We further define

$$B_{pq}^s\left(\mathbb{R}^d, \langle x\rangle^\beta\right) := \left\{ f : \left\| f \cdot \langle x\rangle^\beta \right\|_{s,p,q} < \infty \right\}$$

where $\langle x\rangle^\beta = \frac{1}{(1+x^2)^\beta}$ is the polynomial weighting function parameterized by $\beta \in \mathbb{R}_+$.

Remark. *Let $S'(\mathbb{R}^d)$ denote the space of complex tempered distributions on $\mathbb{R}^d$. Since any $f \in L^p(\mathbb{R}^d)$ gives rise to an element of $S'(\mathbb{R}^d)$, the quantity $\mathcal{F}^{-1}(\phi_k\mathcal{F}f)$ is well-defined (for any k) as an element of $S'(\mathbb{R}^d)$. Moreover $\mathcal{F}^{-1}(\phi_k\mathcal{F}g)$ is an entire analytic function on $\mathbb{R}^d$ for any $g \in S'(\mathbb{R}^d)$ and any k by the Paley-Wiener-Schwartz theorem.*

Theorem 6 (Metric Entropy of Besov Space [31]). *Let $1 \le p \le \infty$, $1 \le q \le \infty$, $\beta \in \mathbb{R}_+$, and $s - \frac{d}{p} > 0$. Suppose E is a (non-empty) bounded subset of $B_{pq}^s(\mathbb{R}^d, \{x\}^\beta)$. For $\beta > s - \frac{d}{p}$ we have*

$$\log N(\delta, E, \|\cdot\|_\infty) \lesssim \delta^{-d/s}.$$

> **Corollary 1.** *Let $\frac{d}{2} \le p' \le p \le \infty$, $1 \le q \le \infty$, $\frac{d}{p'} < s$ and $\beta > s - \frac{d}{p}$. Then $B_{pq}^s\left(\mathbb{R}^d, \langle x\rangle^\beta\right)$ can be embedded into an $\mathcal{L}_{\frac{p'}{p'-1}}$–-type RKBS.*

Remark. *Our Corollary 1 covers the results in [37, Section 4.2, 4.3 and 4.5] for embedding to Reproducing Kernel Hilbert Space by taking $p' = 2$.*

Barron Space Barron space is used to characterize the function space represented by two-layer neural networks and comment belief is Barron space is larger than any Reproducing Kernel Hilbert Space. For example, [32] has showed that Barron space is not isometric to a Reproducing Kernel Hilbert Space because Barron space violates the parallelogram law. However Barron space still can be embeded into a Reproducing Kernel Hilbert Space using our theory. [37, Section 4.4] showed similar property for a special kind of dictionary. via the metric entropy estimation of convex hull. To show this, we utilize the metric entropy of convex hull in Banach space [13, 12], which is the technique used widely in estimating the metric entropy of Barron space / Integral Reproducing Kernel Banach Spaces [39, 38, 30].

Theorem 7 (Convex Hull Metric Entropy [13, 12]). *Let $A \subset X$ be a precompact subset of the unit ball of a Banach space X of type p, $p > 1$, with the property that there are constants $\rho, \alpha > 0$ such that*

$$\mathcal{N}(\delta, A, \|\cdot\|_X) \le \rho\delta^{-\alpha}$$

Then there exists a positive constant $c_{p,\alpha,\rho}$ such that for the dyadic entropy numbers of the convex hull we have the asymptotically optimal estimate

$$\log \mathcal{N}(\delta, \overline{\mathrm{co}(A)}, \|\cdot\|_X) \le c_{p,\alpha,\rho}\delta^{-(1-(1/p))-\alpha} \quad \text{for } n = 1, 2, \ldots$$

Since every Banach space is type-1 (from the triangular inequality), we can have the following coro llary.

> **Corollary 2.** *If the dictionary space $\mathcal{D}$ satisfies $\mathcal{N}(\delta, \mathcal{D}, \|\cdot\|_\infty) \le \rho\delta^{-p+\epsilon}$ for some constant $p > 2, \epsilon > 0, \rho > 0$, then Barron space $\Sigma_M^1(\mathbb{D})$ can be embeded into $\mathcal{L}_{\frac{p}{p-1}}$-type RKBS.*

Remark. *[37] showed that if the dictionary has a positive decomposition then the Barron space can be embeded to a Reproducing Kernel Hilbert space. Our condition provides a new class of conditions which utilize the smoothness of the dictionary [38, 39].*

[15] provide the metric entropy estimate of q−hull in type-p Banach space, which help us to embed to Reproducing Kernel Banach space.

Lemma 5 (Metric Etnropy of q−hull in Type−p Banach Space[15]). *Let $K \subset \mathcal{X}$ be a precompact subset of the unit ball of a Banach space $\mathcal{X}$ of type p ($p > 1$), if $N(\delta, K, \|\cdot\|_{\mathcal{X}}) = O(\delta^{-\alpha+\epsilon})$ with $\alpha > 0, \epsilon > 0$ and $\beta \in \mathbb{R}$, then we have*

$$\log N(\epsilon, H_q(K), \|\cdot\|_{\mathcal{X}}) = O\left(\epsilon^{-\frac{\alpha pq}{pq+\alpha(p-q)}}\right).$$

where $H_q(K) := \overline{\left\{\sum_{i=1}^n c_i x_i \mid x_i \in K, 1 \le i \le n, n \in \mathbb{N}, \sum_{i=1}^n |c_i|^q \le 1\right\}}.$

Based on the result in [15], we have the following corollary

Corollary 3. *If the dictionary space $\mathcal{D}$ satisfies $\mathcal{N}(\delta, \mathcal{D}, \|\cdot\|_\infty) \le \rho\delta^{-p+\epsilon}$ for some constant $p > 2, \epsilon > 0, \rho > 0$, then the Barron space $\Sigma_M^q(\mathbb{D})$ can be embeded into $\mathcal{L}_{\frac{pq}{2pq-q-p}}$-type RKBS.*

References

[1] Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in neural information processing systems*, 32, 2019.

[2] Dana Angluin and Philip Laird. Learning from noisy examples. *Machine learning*, 2:343–370, 1988.

[3] Sanjeev Arora, Simon Du, Wei Hu, Zhiyuan Li, and Ruosong Wang. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. In *International Conference on Machine Learning*, pages 322–332. PMLR, 2019.

[4] Francis Bach. Breaking the curse of dimensionality with convex neural networks. *Journal of Machine Learning Research*, 18(19):1–53, 2017.

[5] Francesca Bartolucci, Ernesto De Vito, Lorenzo Rosasco, and Stefano Vigogna. Understanding neural networks with reproducing kernel banach spaces. *Applied and Computational Harmonic Analysis*, 62:194–236, 2023.

[6] Francesca Bartolucci, Ernesto De Vito, Lorenzo Rosasco, and Stefano Vigogna. Neural reproducing kernel banach spaces and representer theorems for deep networks. *arXiv preprint arXiv:2403.08750*, 2024.

[7] Gyora M Benedek and Alon Itai. Learnability with respect to fixed distributions. *Theoretical Computer Science*, 86(2):377–389, 1991.

[8] Yuan Cao and Quanquan Gu. Generalization bounds of stochastic gradient descent for wide and deep neural networks. *Advances in neural information processing systems*, 32, 2019.

[9] Andrea Caponnetto and Ernesto De Vito. Optimal rates for the regularized least-squares algorithm. *Foundations of Computational Mathematics*, 7:331–368, 2007.

[10] Bernd Carl. Inequalities of bernstein-jackson-type and the degree of compactness of operators in banach spaces. In *Annales de l'institut Fourier*, volume 35, pages 79–118, 1985.

[11] Bernd Carl. Metric entropy of convex hulls in hilbert spaces. *Bulletin of the London Mathematical Society*, 29(4):452–458, 1997.

[12] Bernd Carl, Aicke Hinrichs, and Philipp Rudolph. Entropy numbers of convex hulls in banach spaces and applications. *Journal of Complexity*, 30(5):555–587, 2014.

[13] Bernd Carl, Ioanna Kyrezi, and Alain Pajor. Metric entropy of convex hulls in banach spaces. *Journal of the London Mathematical Society*, 60(3):871–896, 1999.

[14] Lin Chen and Sheng Xu. Deep neural tangent kernel and laplace kernel have the same rkhs. *arXiv preprint arXiv:2009.10683*, 2020.

[15] James Cockreham, Fuchang Gao, and Yuhong Yang. Metric entropy of q-hulls in banach spaces of type-p. *Proceedings of the American Mathematical Society*, 145(12):5205–5214, 2017.

[16] Albert Cohen, Ronald DeVore, Guergana Petrova, and Przemyslaw Wojtaszczyk. Optimal stable nonlinear approximation. *Foundations of Computational Mathematics*, 22(3):607–648, 2022.

[17] Joe Diestel, Hans Jarchow, and Andrew Tonge. *Absolutely summing operators*. Number 43. Cambridge university press, 1995.

[18] John Duchi. Lecture notes for statistics 311/electrical engineering 377. *URL: https://stanford. edu/class/stats311/Lectures/full notes. pdf. Last visited on*, 2:23, 2016.

[19] Richard M Dudley, Sanjeev R Kulkarni, Thomas Richardson, and Ofer Zeitouni. A metric entropy bound is not sufficient for learnability. *IEEE Transactions on Information Theory*, 40(3):883–885, 1994.

[20] David Eric Edmunds and Hans Triebel. Function spaces, entropy numbers, differential operators. *(No Title)*, 1996.

[21] Gregory E Fasshauer, Fred J Hickernell, and Qi Ye. Solving support vector machines in reproducing kernel banach spaces with positive definite functions. *Applied and Computational Harmonic Analysis*, 38(1):115–139, 2015.

[22] Simon Fischer and Ingo Steinwart. Sobolev norm learning rates for regularized least-squares algorithms. *Journal of Machine Learning Research*, 21(205):1–38, 2020.

[23] Sara A Geer. *Empirical Processes in M-estimation*, volume 6. Cambridge university press, 2000.

[24] Amnon Geifman, Abhay Yadav, Yoni Kasten, Meirav Galun, David Jacobs, and Basri Ronen. On the similarity between the laplace and neural tangent kernels. *Advances in Neural Information Processing Systems*, 33:1451–1461, 2020.

[25] Marius Junge and Martin Defant. Some estimates on entropy numbers. *Israel Journal of Mathematics*, 84(3):417–433, 1993.

[26] Jason M Klusowski and Andrew R Barron. Risk bounds for high-dimensional ridge function combinations including neural networks. *arXiv preprint arXiv:1607.01434*, 2016.

[27] Andrei Nikolaevich Kolmogorov. On linear dimensionality of topological vector spaces. In *Doklady Akademii Nauk*, volume 120, pages 239–241. Russian Academy of Sciences, 1958.

[28] Stanisław Kwapień. Isomorphic characterizations of inner product spaces by orthogonal series with vector valued coefficients. *Studia mathematica*, 44(6):583–595, 1972.

[29] Rong Rong Lin, Hai Zhang Zhang, and Jun Zhang. On reproducing kernel banach spaces: Generic definitions and unified framework of constructions. *Acta Mathematica Sinica, English Series*, 38(8):1459–1483, 2022.

[30] Fanghui Liu, Leello Dadi, and Volkan Cevher. Learning with norm constrained, over-parameterized, two-layer neural networks. *arXiv preprint arXiv:2404.18769*, 2024.

[31] Richard Nickl and Benedikt M Pötscher. Bracketing metric entropy rates and empirical central limit theorems for function classes of besov-and sobolev-type. *Journal of Theoretical Probability*, 20:177–199, 2007.

[32] Greg Ongie, Rebecca Willett, Daniel Soudry, and Nathan Srebro. A function space view of bounded norm infinite width relu nets: The multivariate case. *arXiv preprint arXiv:1910.01635*, 2019.

[33] Rahul Parhi and Robert D Nowak. Banach space representer theorems for neural networks and ridge splines. *Journal of Machine Learning Research*, 22(43):1–40, 2021.

[34] Gilles Pisier. Remarques sur un résultat non publié de b. maurey. *Séminaire d'Analyse fonctionnelle (dit" Maurey-Schwartz")*, pages 1–12, 1981.

[35] Gilles Pisier. *The volume of convex bodies and Banach space geometry*, volume 94. Cambridge University Press, 1999.

[36] Pedro Savarese, Itay Evron, Daniel Soudry, and Nathan Srebro. How do infinite width bounded norm networks look in function space? In *Conference on Learning Theory*, pages 2667–2690. PMLR, 2019.

[37] Max Schölpple and Ingo Steinwart. Which spaces can be embedded in reproducing kernel hilbert spaces? *arXiv preprint arXiv:2312.14711*, 2023.

[38] Lei Shi, Yun-Long Feng, and Ding-Xuan Zhou. Concentration estimates for learning with ℓ_1-regularizer and data dependent hypothesis spaces. *Applied and Computational Harmonic Analysis*, 31(2):286–302, 2011.

[39] Lei Shi, Xiaolin Huang, Yunlong Feng, and Johan AK Suykens. Sparse kernel regression with coefficient-based ℓ_q- regularization. *Journal of Machine Learning Research*, 20(161):1–44, 2019.

[40] Alistair Shilton, Sunil Gupta, Santu Rana, and Svetha Venkatesh. Gradient descent in neural networks as sequential learning in reproducing kernel banach space. In *International Conference on Machine Learning*, pages 31435–31488. PMLR, 2023.

[41] Guohui Song, Haizhang Zhang, and Fred J Hickernell. Reproducing kernel banach spaces with the ℓ_1 norm. *Applied and Computational Harmonic Analysis*, 34(1):96–116, 2013.

[42] Len Spek, Tjeerd Jan Heeringa, Felix Schwenninger, and Christoph Brune. Duality for neural networks through reproducing kernel banach spaces. *arXiv preprint arXiv:2211.05020*, 2022.

[43] Ingo Steinwart. *Entropy of C (K)-valued operators and some applications*. PhD thesis, 2000.

[44] Ingo Steinwart. On the influence of the kernel on the consistency of support vector machines. *Journal of machine learning research*, 2(Nov):67–93, 2001.

[45] Ingo Steinwart and Andreas Christmann. *Support vector machines*. Springer Science & Business Media, 2008.

[46] Michael Unser. A representer theorem for deep neural networks. *Journal of Machine Learning Research*, 20(110):1–30, 2019.

[47] Michael Unser. A unifying representer theorem for inverse problems and machine learning. *Foundations of Computational Mathematics*, 21(4):941–960, 2021.

[48] Michael Unser. From kernel methods to neural networks: A unifying variational formulation. *Foundations of Computational Mathematics*, pages 1–40, 2023.

[49] Aad W Van der Vaart. *Asymptotic statistics*, volume 3. Cambridge university press, 2000.

[50] E Weinan, Chao Ma, and Lei Wu. Barron spaces and the compositional function spaces for neural network models. *arXiv preprint arXiv:1906.08039*, 2019.

[51] Yuesheng Xu and Qi Ye. Generalized mercer kernels and reproducing kernel banach spaces. *arXiv preprint arXiv:1412.8663*, 2014.

[52] Yasuo Yamasaki. A simple proof of kwapien's theorem. *Publications of the Research Institute for Mathematical Sciences*, 20(6):1247–1251, 1984.

[53] Haizhang Zhang, Yuesheng Xu, and Jun Zhang. Reproducing kernel banach spaces for machine learning. *Journal of Machine Learning Research*, 10(12), 2009.

A. Proof of the Main Theorem

In this section, we present the proof of our main theorem

> Given a bounded domain $\Omega \in \mathcal{R}^d$, a RKBS E of functions on Ω, and $F = \ell_\infty(\Omega)$ on Ω , which means the embedding $id : E \to F$ is a compact embedding. If the growth of metric entropy can be bounded via
>
> $$\mathcal{E}_E^F(\delta) := \log N(\delta, \{x \in E : \|x\|_E \leq 1\}, \|\cdot\|_F) \leq \delta^{-p}, p \geq 2.$$
>
> Then for any $s > p$, there exist a $\mathcal{L}_s$−type RKBS $\mathcal{B}_s$, such that
> $$E \hookrightarrow \mathcal{B}_s \hookrightarrow F\,.$$

Proof. First of all, according to Theorem 5, if the entropy number $\log N(\delta, E, \|\cdot\|_F) \leq \delta^{-p}$ for some $p > 2$, we can have

$$\begin{aligned}\mathbb{E}_{\epsilon_i}\frac{1}{n}\left\|\sum_{i=1}^n \epsilon_i x_i\right\|_F &\lesssim n^{-\frac{1}{p}} + \frac{1}{\sqrt{n}}\int_{n^{-\frac{1}{p}}}^1 \sqrt{\delta^{-p}}d\delta \qquad (\text{Take } \alpha = n^{-\frac{1}{p}})\\ &\lesssim n^{-\frac{1}{p}} \qquad (\text{The integral is of } O(n^{-\frac{1}{p}}))\\ \Rightarrow \mathbb{E}_{\epsilon_i}\left\|\sum_{i=1}^n \epsilon_i x_i\right\|_F &\lesssim n^{\frac{1}{p'}}\end{aligned} \tag{15}$$

for all $\|x_i\|_E \leq 1$.

Therefore by our technique Lemma 3, we can conclude that the space $\hat{E} = (E, \|\cdot\|_F)$ is of type s' for any $1 < s' < p'$. Recall that E is an RKBS, so it is a closed subspace of $F = \ell_\infty(X)$. Therefore E is a closed subspace of $L_1(X)$ because $\ell_\infty(X)$ embeds continuously to L_1, so is $\hat{E}$. Consequently, $\hat{E}$ is a closed subspace of $L_1(\nu)$, where ν is the uniform distribution on X. By Theorem 3, $\hat{E}$ is isometric to a subspace of $L_{s'}(\mu)$ for some measure μ for any $1 < s' < p'$.

By the induction above the following embedding holds

$$E = (E, \|\cdot\|_E) \overset{id_{E\to\hat{E}}}{\hookrightarrow} \hat{E} = (E, \|\cdot\|_F) \overset{id_{\hat{E}\to F}}{\hookrightarrow} F = (F, \|\cdot\|_F)$$

Therefore,
We have the following embedding

$$E = (E, \|\cdot\|_E) \overset{id_{E\to\hat{E}}}{\hookrightarrow} \hat{E} = (E, \|\cdot\|_F) \overset{id_{\hat{E}\to F}}{\hookrightarrow} F = (F, \|\cdot\|_F)$$

and $\hat{E}$ is isometric to $W_{s'}$, a closed subspace of $L_{s'}(\mu)$ by the isometric mapping i.

First, for any $x \in \Omega$, we denote $\delta_x^F \in F'$ as the evaluation functional at x acting on F. Then we consider the following linear functional for any $w \in W_{s'}$:

$$[\hat{\phi}(x), w]_{W_{s'}} := [\delta_x^F, (id_{\hat{E}\to F}\circ i^{-1})(w)]_F$$

Since $W_{s'}$ is a subspace of $L_{s'}(\mu)$, by Hahn-Banach continuous extension theorem, we can extend this mapping $\hat{\phi}(x) : W_{s'} \to \mathbb{R}$ to a continuous linear functional $\phi : L_{s'}(\mu) \to \mathbb{R}$. Then we have for any $e \in \hat{E}$, $i(e) \in W_{s'}$

$$[\phi(x), i(e)]_{\mathcal{L}_{s'}(\mu)} = [\hat{\phi}(x), i(e)]_{W_{s'}} = [\delta_x^F, (id_{\hat{E}\to F}\circ i^{-1})(i(e))]_F = [\delta_x^F, id_{\hat{E}\to F}(e)]_F = id_{\hat{E}\to F}(e)(x),$$

where (1) is based on the fact that $\Phi i(e) = (id_{\hat{E}\to F}\circ i^{-1})(i(e)) = id_{\hat{E}\to F}(e)$. Now we define the $\mathcal{L}_s$-typ RKBS,

$$\mathcal{B}_s := \left\{f_v(x) := [\phi(x), v]_{\mathcal{L}_{s'}} : v \in \mathcal{W}_{s'}, x \in \Omega\right\}$$

with norm

$$\|f_v\|_{\mathcal{B}_s} := \inf\{\|v\|_W : v \in \mathcal{W}_{s'} \text{ with } f_v = [\phi(\cdot), v]_{\mathcal{L}_{s'}}\}.$$

Next we show the embedding $E \hookrightarrow \mathcal{B}_s \hookrightarrow F$. Noticing that $\mathcal{B}_s$ is an RKBS, so $\mathcal{B}_s \hookrightarrow F$, we only need to show the first embedding. Since $E \hookrightarrow \hat{E}$ and $\hat{E}$ is isometric to $\mathcal{W}_{s'}$, we will prove this by showing that $\mathcal{B}_s$ is the image of the mapping $id_{\hat{E}\to F}\circ i^{-1}$ on $\mathcal{W}_{s'}$. Noticing that for all $v \in \mathcal{W}_{s'}$, $f_v = [\phi(\cdot), v]_{\mathcal{L}_{s'}}\} = id_{\hat{E}\to F}\circ i^{-1}(v) \in \mathcal{B}_s$. Conversely, for any $f \in \mathcal{B}_s$, one can find a $v \in \mathcal{W}_{s'}$ such that $f = [\phi(\cdot), v]_{\mathcal{L}_{s'}}\}$ by definition. Therefore $\mathcal{B}_s$ is the image of the mapping $id_{\hat{E}\to F}\circ i^{-1}$ on $\mathcal{W}_{s'}$.

Now, since $E = \left(id_{\hat{E}\to F}\circ i^{-1}\right)\circ i \circ id_{E\to\hat{E}}E$, therefore E is a subset of the image of the mapping $id_{\hat{E}\to F}\circ i^{-1}$ on $\mathcal{W}_{s'}$, then we can conclude that $E \hookrightarrow \mathcal{B}_s$. □

Upon further review, we recognize that our initial approach contained some inaccuracies. We appreciate the opportunity to clarify our position. What we intended to convey is that consider the following linear functional for any $w \in W_{s'}$, where $W_{s'}$ is a closed subspace of $L_{s'}(\mu)$:

$$[\hat{\phi}(x), w]_{W_{s'}} := [\delta_x^F, (id_{\hat{E}\to F}\circ i^{-1})(w)]_F$$

Since $W_{s'}$ is a subspace of $L_{s'}(\mu)$, by Hahn-Banach continuous extension theorem, we can extend this mapping $\hat{\phi}(x) : W_{s'} \to \mathbb{R}$ to a continuous linear functional $\phi : L_{s'}(\mu) \to \mathbb{R}$. Noticing that in this case we only apply Hahn-Banach Theorem on linear functional and then we can proceed with our proof of the Main result further since we have for any $e \in \hat{E}$, $i(e) \in W_{s'}$

$$[\phi(x), i(e)]_{\mathcal{L}_{s'}(\mu)} = [\hat{\phi}(x), i(e)]_{W_{s'}} = [\delta_x^F, (id_{\hat{E}\to F}\circ i^{-1})(i(e))]_F = [\delta_x^F, id_{\hat{E}\to F}(e)]_F = i$$

This refined statement more accurately reflects our stance on the matter. We apologize for any confusion our previous communication may have caused and are committed to providing clear and accurate information moving forward.