Margin in Abstract Spaces

Yair Ashlagi YAIRASHLAGI@MAIL.TAU.AC.IL
School of Electrical and Computer Engineering, Tel Aviv University.

Roi Livni RLIVNI@TAUEX.TAU.AC.IL
School of Electrical and Computer Engineering, Tel Aviv University.

Shay Moran SMORAN@TECHNION.AC.IL
Departments of Mathematics, Computer Science, and Data and Decision Sciences, Technion and Google Research.

Tom Waknine TOM.WAKNINE@CAMPUS.TECHNION.AC.IL
Departments of Mathematics, Technion.

Editors: Steve Hanneke and Tor Lattimore

Abstract

Margin-based learning, exemplified by linear and kernel methods, is one of the few classical settings where generalization guarantees are independent of the number of parameters. This makes it a central case study in modern highly over-parameterized learning. We ask what minimal mathematical structure underlies this phenomenon.

We begin with a simple margin-based problem in arbitrary metric spaces: concepts are defined by a center point and classify points according to whether their distance lies below r or above R. We show that whenever $R > 3r$, this class is learnable in *any* metric space. Thus, sufficiently large margins make learnability rely only on the triangle inequality, without any linear or analytic structure being necessary. Our first main result extends this phenomenon to concepts defined by bounded linear combinations of distance functions, and reveals a sharp threshold: there exists a universal constant such that whenever the margin is larger than this constant, the class is learnable in every metric space, while below it there exist metric spaces where it is not learnable at all.

We then ask whether margin-based learnability can always be explained via an embedding into a linear space – that is, reduced to linear classification in some Banach space through a kernel-type construction. We answer this negatively by demonstrating a margin learnable class that cannot be embedded into any Banach space in which linear classification with margins is learnable.

1. Introduction

Margin-based learning provides a canonical example of parametric learning in high dimensions due to it being one of the primary settings where generalization does not depend on the number of parameters. Certain hypothesis classes, most notably linear functions, whose learning complexity typically grows with dimensionality, become markedly easier to learn when their effective domain is restricted to points lying a fixed distance away from the decision boundary, i.e., when a margin condition is imposed. Due to the appealing nature of dimensionality-independent generalization, margin assumptions have been a topic for extensive research in the context of halfspaces and kernel methods in Euclidean and Hilbert spaces. Classic maximal margin algorithms, such as the Support Vector Machines, enjoy dimension-independent generalization guarantees when the data satisfies a margin condition. Namely, when some target function realizes the data with a sufficiently large margin Vapnik (1998); Cristianini and Shawe-Taylor (2000); Bartlett and Mendelson (2002).

However, much of the existing work relies on strong geometric assumptions, typically those of Euclidean or, more generally, Hilbert spaces (via kernel methods). This naturally raises the question:

Which basic mathematical properties underlie margin-based learnability, and when can it be reduced to learning embeddings in linear spaces?

We begin by abstracting linear margin-based hypotheses to the minimal geometric setting of metric spaces, and ask whether such weak structure suffices for learnability. Let $\mathcal{X}$ be an arbitrary metric space with distance function $d(\cdot, \cdot)$. A simple margin-based concept class in this setting is defined by points of the space: for a

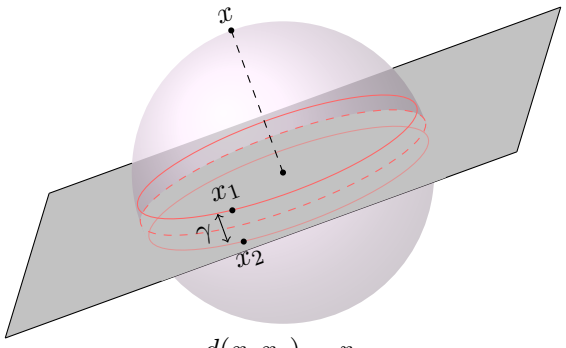
given center $x \in \mathcal{X}$ and parameters $0 \leq r < R$, the corresponding concept labels points within distance at most r from x as positive, and points at distance greater than R from x as negative. Formally, we define

$$d_x(x') = \begin{cases} +1 & d(x, x') \leq r, \\ -1 & d(x, x') > R\,. \end{cases} \tag{1}$$

Points whose distance from x lies in the margin region $(r, R]$ are left unlabeled, and the data distribution is assumed to assign zero probability to this region.

This definition applies to arbitrary metric spaces and relies only on the distance structure. To illustrate how it relates to classical linear classification with margin, consider the special case where $\mathcal{X}$ is the unit ℓ_2-sphere in $\mathbb{R}^d$, scaled to have diameter 1. In this setting, the level sets of the distance function $d(x, \cdot)$ correspond to intersections of affine hyperplanes with the sphere. In particular, a linear classifier with margin γ induces a distance-based labeling of the form above, with parameters $r = \frac{1}{2} - \gamma$ and $R = \frac{1}{2} + \gamma$. Thus, Equation (1) corresponds to the margin structure underlying standard linear classification, while extending naturally to general metric spaces.

Figure 1: Distance based labeling on a sphere



$$d(x, x_1) = r$$
$$d(x, x_2) = R$$

This simple concept class exhibits a sharp threshold behavior: for fixed values r, R such that $R \geq 3r$, the class is learnable regardless of the underlying metric space. In this case, the VC dimension of the class is 1 and even a single pair of points cannot be shattered by it[1]. Indeed, assume by contradiction that there exists a pair $x, x' \in \mathcal{X}$ that are shattered. Then, $\exists x_1 \in \mathcal{X}$ such that $d(x_1, x) \leq r$ and $d(x_1, x') \leq r$. By the triangle inequality, we have $d(x, x') \leq 2r$. On the other hand $\exists x_2 \in \mathcal{X}$ such that $d(x_2, x) < r$ and $d(x_2, x') > R \geq 3r$. By the triangle inequality, we get $d(x, x') > 2r$, a contradiction.

On the other hand, when $R < 3r$, the simple triangle-inequality-based argument no longer applies, and the behavior changes qualitatively: depending on the underlying metric space, the class may be arbitrarily hard to learn, or even unlearnable. Indeed, let $X = A \cup B$ such that $A = \{a_n\}_{n \in \mathbb{N}}$ is a countable set of points and $B = \{b_s\}_{\{s \subseteq A \mid |s| < \infty\}}$ is a set of points referencing finite subsets of A.

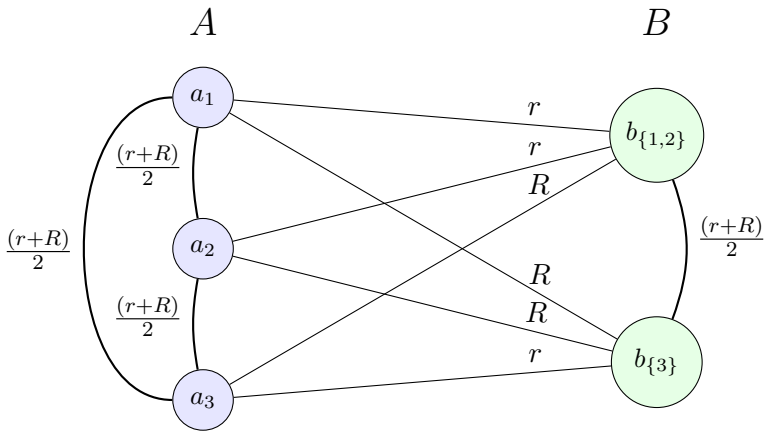
$$\rho(a_i, a_j) = \frac{r + R}{2}, \quad \rho(b_s, b_w) = \frac{r + R}{2}, \quad \rho(a_i, b_s) = \begin{cases} r & a_i \in s \\ R & a_i \notin s \end{cases}$$

For any $R \leq 3r$ this is indeed a metric space, while any finite subset of A is shattered by concepts defined by points in B. Hence the VC dimension of the class is infinite yielding an unlearnable setting.

Simply put, for sufficiently large margins, the learnability of these abstract margin-based classifiers relies only on the triangle inequality and not on any additional linear or geometric structure. This threshold is sharp: for every margin value below it, there exist metric spaces in which the class is not learnable. Thus, while learnability is guaranteed uniformly above the threshold, below it the behavior becomes space-dependent and may range from learnability to complete non-learnability.

1. Note that the VC dimension continues to characterize PAC learnability for such classes of *partial concepts*; this follows from known extensions of PAC learning theory, which we discuss in more detail later

Figure 2: Unlearnable metric space for $R \leq 3r$



In Section 4.1 we extend these ideas to a much richer concept class and identify total boundedness as a key structural property in this setting. When the margin is below the critical threshold, learnability remains possible provided the underlying metric space is totally bounded[2]. Beyond the specific distance-based classes considered here, we show that total boundedness in fact precisely characterizes the learnability of the entire class of Lipschitz functions over the metric space.

Margin-Based Learning in Linear Spaces. Since our analysis of margin-based learning in general metric spaces provides only a partial characterization of margin based learnability, we shift our focus to the more structured setting of linear spaces. A well studied case of margin learning is the one of linear functionals on Hilbert spaces. Although infinite dimensional Hilbert spaces are not totally bounded, linear classification with margin in these spaces provide a canonical example for margin-based learning, a fact that cannot be accounted for by the metric space perspective. This motivates our study of linear spaces, where additional geometric structure plays a central role.

At the same time, linear spaces arise naturally even when dealing with abstract, non-linear learning problems. A central paradigm in learning theory is to embed a classification problem into a linear space, most prominently via kernel methods, so that hypotheses are realized as linear functionals and margin-based generalization guarantees apply. This perspective raises the question of whether such linear embeddings merely provide a convenient analytical framework, or whether they fundamentally characterize margin-based learnability.

These considerations motivate a more focused study of margin problems in the context of Banach spaces, in terms of their role as a potential universal model for margin-based learning via embeddings. We therefore ask the following question:

To what extent is linear classification with margin universal in the context of margin-based learning? For example, can every learnable margin-based class be realized as linear functionals over an appropriate Banach space?

We answer this question negatively and show that margin-based learnability does not, in general, reduce to linear embeddings into Banach spaces. To this end, we construct learnable margin-based classes with learning rates that violate a key property of the sample complexity bounds of learning linear classifiers in Banach spaces: that they must be polynomial in the inverse margin $\frac{1}{\gamma}$. This implies that the learnability of these classes cannot be reduced to learning them linearly in any Banach space.

2. Background & Problem Setup

We begin by introducing several standard notions and the model of γ-learnability which is the main focus of investigation here. Throughout we consider a fixed domain $\mathcal{X}$ and a function class $\mathcal{F}$ of real valued functions from $\mathcal{X}$ to $\mathbb{R}$.

2. Recall that a space is totally bounded if for any $\gamma > 0$ it can be covered by finitely many balls of radius γ.

γ-learnability. For any real-valued function $f : \mathcal{X} \to \mathbb{R}$, a binary labeled dataset $S = \{(x_i, y_i)\}_{i=1}^n$, where $x_i \in \mathcal{X}$ and $y_i \in \{-1, 1\}$, is said to be **γ-realized** by f if $f(x_i) \cdot y_i > \gamma$, for all $(x_i, y_i) \in S$. A distribution $\mathcal{D}$ is **γ-realizable** by $\mathcal{F}$ if there exists an $f \in \mathcal{F}$ that γ-realizes almost surely any i.i.d. dataset sampled from $\mathcal{D}$. A function class $\mathcal{F}$ is said to be **γ-learnable** if there exists a learning algorithm $\mathcal{A}$, mapping a sample S to a function $\mathcal{A}(S) : \mathcal{X} \to \{\pm 1\}$, and a sample-complexity bound $m_{\mathcal{F}_\gamma} : (0,1) \times (0,1) \to \mathbb{N}$ such that for every $\varepsilon, \delta \in (0,1)$ and every distribution $\mathcal{D}$ that is γ-realizable by $\mathcal{F}$, the following holds: If $S \sim \mathcal{D}^m$ with $m \geq m_{\mathcal{F},\gamma}(\varepsilon, \delta)$ then with probability at least $1 - \delta$ over the draw of S,

$$\mathop{\mathbb{P}}_{(x,y)\sim\mathcal{D}}\big[(\mathcal{A}(S)(x)) \cdot y \neq 1\big] \leq \epsilon$$

Partial concept classes We formulate the notion of margin concept class by using partial concept classes, as defined in Long (2001); Alon et al. (2022). Given an input space $\mathcal{X}$, a partial concept is a map $h : \mathcal{X} \to \{-1, 1, \star\}$, where we think of $h(x) = \star$ as meaning that h is undefined on x. A partial concept class $\mathcal{H}$ is a collection of partial concepts. A sample $S = \{(x_i, y_i)\}_{i=1}^n$ is realizable by the partial concept $\mathcal{H}$, if there is some $h \in \mathcal{H}$ such that $h(x_i) = y_i \neq \star$ for all $1 \leq i \leq n$. A distribution $\mathcal{D}$ over $\mathcal{X} \times \{-1, 1\}$ is realizable by $\mathcal{H}$ if any sample drawn from $\mathcal{D}$ is realizable with probability 1. A set $\{x_i\}_{i=1}^n$, is said to be shattered by $\mathcal{H}$ if for any $y \in \{-1, 1\}^n$ the labeled sample $\{(x_i, y_i)\}_{i=1}^n$ is realizable by $\mathcal{H}$, and $\mathtt{VC}(\mathcal{H})$, the VC dimension of $\mathcal{H}$ is the maximal size of a shattered set (or infinite if there are shattered sets of arbitrary size).

A Partial concept class $\mathcal{H}$ is called learnable if there exists a learning algorithm $\mathcal{A}$ and a sample-complexity bound $m_{\mathcal{H}} : (0,1) \times (0,1) \to \mathbb{N}$ such that for every $\varepsilon, \delta \in (0,1)$ and every realizable distribution $\mathcal{D}$ the following holds: If $S \sim \mathcal{D}^m$ with $m \geq m_{\mathcal{H}}(\varepsilon, \delta)$ then with probability at least $1 - \delta$ over the draw of S,

$$\mathop{\mathbb{P}}_{(x,y)\sim\mathcal{D}}\big[(\mathcal{A}(S)(x)) \neq y\big] \leq \epsilon$$

A fundamental result in the theory of PAC learning is that the VC dimension governs the sample complexity of a class. Indeed the results by Alon et al. (2022); Aden-Ali et al. (2023) imply the following characterization

Theorem (Characterization of optimal Sample complexity via VC dimension) *Let $\mathcal{H} \subset \{-1, 1, \star\}^{\mathcal{X}}$ be some partial concept class. Then $\mathcal{H}$ is learnable if and only if $\mathtt{VC}(\mathcal{H})$ is finite, in which case we have the following bound on its optimal sample complexity*

$$m_{\mathcal{H}}(\varepsilon, \delta) = \Theta\left(\frac{\mathtt{VC}(\mathcal{H}) + \log\frac{1}{\delta}}{\varepsilon}\right).$$

For a $\gamma > 0$, and a function $f : \mathcal{X} \to \mathbb{R}$, the margin classifier defined by f is the partial concept class $h_f^\gamma : \mathcal{X} \to \{-1, 1 - \star\}$ defined by

$$h_f^\gamma(x) = \begin{cases} +1 & f(x) \geq \gamma, \\ -1 & f(x) \leq -\gamma, \\ \star & f(x) \in (-\gamma, \gamma). \end{cases}$$

And given a set of functions $\mathcal{F} \subset \mathbb{R}^{\mathcal{X}}$, the induced partial concept class of classifiers is $\mathcal{H}_{\mathcal{F}}^\gamma = \{h_f \ : \ f \in \mathcal{F}\}$. By definition, the margin problem $\mathcal{F}$ is γ-learnable if and only if $\mathcal{H}_{\mathcal{F}}^\gamma$ is learnable, and with the same sample complexity.

γ-fat shattering dimension. We recall the classic definition of the fat-shattering dimension as in Kearns and Schapire (1994); Alon et al. (1997). A set $X \subset \mathcal{X}$ is γ-shattered by a function class $\mathcal{F}$ if there exists a witness threshold function $r : X \to \mathbb{R}$ such that for any labeling $b : X \to \{-1, 1\}$ there exists a function $f_b \in \mathcal{F}$ satisfying:

$$f_b(x) \geq r(x) + \gamma \quad \text{if } b(x) = 1, \quad \text{and} \quad f_b(x) \leq r(x) - \gamma \quad \text{if } b(x) = -1,$$

for all $x \in X$. The **γ-fat shattering dimension** of $\mathcal{F}$, denoted by $\mathtt{fat}_\gamma(\mathcal{F})$, is the cardinality of the largest γ-shattered set $X \subset \mathcal{X}$ (or infinite if there exist shattered sets of arbitrarily large size).

Several restrictive variants of $\mathtt{fat}_\gamma(\mathcal{F})$ exist in the literature. These include confining the threshold function r to be a constant across X, strictly fixing it to 0 (denoted $\mathtt{fat}_\gamma^0(\mathcal{F})$), and replacing the inequality with an equality $f_b(x) = r(x) \pm \gamma$ (see, e.g. Simon (1997); Attias et al. (2023); Alon et al. (1997) for examples of these variants). The following technical result reconciles the numerous definitions and establishes that for many naturally structured function classes, these alternative formulations coincide.

Proposition 2.1 (Characterization of shattering in margin spaces) *Let $\mathcal{X}$ be a set, and let $\mathcal{F}$ be a convex and symmetric class of functions from $\mathcal{X}$ to $\mathbb{R}$; that is:*

1. *$\mathcal{F}$ is closed under convex combinations.*
2. *$f \in \mathcal{F} \implies -f \in \mathcal{F}$.*

Let $\gamma > 0$, and let $S = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$ be a set of n distinct points. Then the following conditions are equivalent:

(1) S is γ-shattered by $\mathcal{F}$: There exists a threshold function $r : S \to \mathbb{R}$ such that for every $s \in \{-1, 1\}^n$, there exists $f \in \mathcal{F}$ satisfying

$$s_i\big(f(x_i) - r(x_i)\big) \geq \gamma \qquad \forall 1 \leq i \leq n.$$

(2) S is γ-shattered at 0 by $\mathcal{F}$ (γ-shattered with threshold $r = 0$): For every $s \in \{-1, 1\}^n$, there exists $f \in \mathcal{F}$ satisfying

$$s_i f(x_i) \geq \gamma \qquad \text{for all } 1 \leq i \leq n.$$

(3) $\mathcal{F}$ contains an n-dimensional γ-cube on S; that is, for every $y \in [-\gamma, \gamma]^n$, there exists $f \in \mathcal{F}$ such that

$$f(x_i) = y_i \qquad \text{for all } 1 \leq i \leq n.$$

(4) For every $\lambda \in \mathbb{R}^n$ satisfying $\sum_{i=1}^n |\lambda_i| = 1$, there exists $f \in \mathcal{F}$ such that

$$\sum_{i=1}^n \lambda_i f(x_i) \geq \gamma.$$

Moreover, the equivalence of conditions (2)–(4) *holds even without the symmetry assumption on $\mathcal{F}$.*

The proof is deferred to Section 5.2 while we turn our attention to the sample complexity of such classes. A consequence of Long (2001); Alon et al. (2022); Aden-Ali et al. (2023) is that the fat-shattering condition with threshold $r = 0$ (condition (2) of Proposition 2.1) precisely characterizes the optimal sample complexity of learning a real-valued function class. Combining this with Proposition 2.1, we conclude that the learnability of convex and symmetric function classes is characterized by $\texttt{fat}_\gamma(\mathcal{F})$.

Corollary 2.2 *Let $\mathcal{F}$ be a convex and symmetric class of functions from $\mathcal{X}$ to $\mathbb{R}$; that is:*

1. *$\mathcal{F}$ is closed under convex combinations.*
2. *$f \in \mathcal{F} \to -f \in \mathcal{F}$.*

Then for any $\gamma > 0$, the following conditions are equivalent

(1) $\mathcal{F}$ is γ-learnable.

(2) $\texttt{fat}_\gamma(\mathcal{F}) < \infty$.

(3) $\texttt{fat}^0_\gamma(\mathcal{F}) < \infty$.

Moreover, for $\texttt{fat}_\gamma(\mathcal{F}) = d$, the optimal sample complexity for learning $\mathcal{F}$ with margin γ satisfies:

$$m_{\mathcal{F},\gamma}(\varepsilon, \delta) = \Theta\left(\frac{d + \log(1/\delta)}{\varepsilon}\right).$$

A few comments regarding the conditions in Proposition 2.1 are in order. Condition (3) is a natural geometric property that connects the learnability of margin spaces to interpolation problems. It also closely relates to the γ-Natarajan dimension and γ-Graph dimension studied by Attias et al. (2023), demonstrating that for the classes under discussion, both are equivalent to the fat-shattering dimension. Condition (4), while less intuitive in an abstract metric setting, turns out to be a useful tool whose geometric meaning becomes more transparent in linear structures. We defer a detailed discussion of this connection to 2, following our Banach space preliminaries.

Metric Spaces. Recall that a domain $\mathcal{X}$, together with a function $d : \mathcal{X}^2 \to \mathbb{R}$ is a metric space if d is symmetric, nonnegative with $d(x, y) = 0$ if and only if $x = y$, and satisfies the traingular inequality, i.e., $d(x, y) \leq d(x, z) + d(z, y)$. For a metric space $\mathcal{X}$, let $\texttt{Lip}_{\mathcal{X}}$ (or simply $\texttt{Lip}$) denote the class of all 1-Lipschitz functions over the metric space. Namely, the set of all functions $f : \mathcal{X} \to \mathbb{R}$ satisfying

$$\forall x, y \in X, \quad |f(x) - f(y)| \leq d(x, y).$$

Within this class we consider the following class of bounded linear combinations of point-distance functions:

$$\mathcal{D}_{\mathcal{X}} := \left\{ \sum_{i=1}^{\infty} a_i \, d_{x_i} \;\middle|\; a_i \in \mathbb{R},\ x_i \in \mathcal{X},\ \sum_{i=1}^{\infty} |a_i| \leq 1 \right\},$$

where each d_{x_i} is defined by $d_{x_i}(x) := d(x_i, x)$. These bounded combinations of distance functions generalize the notion of hyperplanes beyond the unit sphere. For example, consider a subset of $\mathcal{D}_{\mathcal{X}}$ of functions which are differences over $n = 2$ distance functions, i.e. $f(x) = \frac{1}{2}\|x - x_1\| - \frac{1}{2}\|x - x_2\|$ for $x_1, x_2 \in \mathcal{X}$. Each such function encodes a half-space over the space $\mathcal{X} = \mathbb{R}^d$ while its 0-level set corresponds to a hyperplane in $\mathbb{R}^d$.

Banach Spaces. A well studied instance of metric spaces are Banach spaces which are the basis for high-dimensional linear classification. Let $\mathcal{X}$ be a linear space, equipped with a norm $\|\cdot\|$. Recall that $\|\cdot\|$ is a norm if it is non-negative and positive-definite (i.e. $\|x\| = 0$ iff $x = 0$), absolutely homogeneous (i.e. $\|\lambda x\| = |\lambda|\|x\|$ and satisfies the triangular inequality:

$$\|x + y\| \leq \|x\| + \|y\|.$$

The space $\mathcal{X}$ is called a Banach space if it is complete with respect to $\|\cdot\|$; that is, any Cauchy sequence $\{v_n\}_{n=1}^{\infty} \subseteq \mathcal{X}$ converges to a limit in $\mathcal{X}$.

Let $\mathcal{B}$ be a Banach space. In the context of linear classification with margin over $\mathcal{B}$, we assume that the domain is the unit ball $\mathcal{B}_1 = \{x \in \mathcal{B} \ : \ \|x\| \leq 1\}$, and that the corresponding class $\mathcal{F}$ is the set of linear functionals whose *dual-norm* is at most one. Recall that given a norm $\|\cdot\|$ on $\mathcal{B}$ the dual norm is a norm over $\mathcal{B}^\star$, the linear functionals over $\mathcal{B}$, defined as $\|w\| = \sup_{\|x\|=1} w(x)$. The class of 1-bounded linear functions is then denote by $L_{\mathcal{B}} = \{w \in \mathcal{B}^\star \ : \ \|w\|_\star \leq 1\}$, where we consider each linear functional as a function over $\mathcal{B}_1$.

As noted in Theorem 2.1, Condition (4) finds a natural realization in linear spaces where $\mathcal{F} = L_{\mathcal{B}}$. In this setting, a straightforward application of the Hahn-Banach theorem yields

$$\gamma \leq \sup_{f \in \mathcal{F}} \sum \lambda_i f(x_i) = \sup_{f \in \mathcal{F}} f\left(\sum \lambda_i x_i\right) = \left\| \sum \lambda_i x_i \right\|.$$

This dual formulation yields two complementary perspectives. The first is through the geometric lens of the hyperplane separation theorem. A set $\{x_i\}_{i=1}^n$ is γ-shattered (with witness 0) if for every possible binary labeling $y = \{-1, 1\}^n$, the convex hull of $\{y_i x_i\}_{i=1}^n$ is at least γ far away from 0. This mirrors the classical intuition that a shattered set must admit a linear margin from the origin for any arbitrary labeling (see, e.g., Gurvits, 2001). The second perspective interprets shattering as a scale-sensitive generalization of linear independence. While a set of vectors is linearly independent if every non-zero linear combination is non-trivial, Proposition 2.1 implies that $\{x_i\}_{i=1}^n$ is γ-shattered[3] if and only if any (normalized) linear combination is γ-far from zero. We also note that in this case, the linear map $T : \ell_1^n \to \mathcal{X}$ defined by $Te_i = x_i$ shows the existence of a γ-isomorphic copy of ℓ_1^n in $\mathcal{X}$, thus relating fat-shattering to notions in local theory of Banach spaces as observed by Mendelson (2002); Mendelson and Schechtman (2004).

3. Related Work

Our work intersects several foundational streams in statistical learning theory: dimension-independent generalization in normed spaces, metric-space classification, scale-sensitive dimensions, and the limits of kernel-type linear embeddings. Below, we position our contributions within this landscape.

3. Recall that for a Banach space $\mathcal{X}$, we assume the hypothesis class consists of all linear functionals on $\mathcal{X}$ of norm at most one.

3.1. Large-Margin Generalization in Linear and Banach Spaces

The principle of margin-based generalization is a cornerstone of classical learning theory, providing a mechanism for dimension-independent performance in highly over-parameterized models. In Euclidean and Hilbert spaces, the generalization error of linear classifiers can be bounded purely in terms of the margin, bypassing the nominal dimensionality of the space Vapnik (1998); Bartlett (1998); Bartlett and Mendelson (2002); Cortes and Vapnik (1995).

The generalization of this paradigm to more abstract normed and Banach spaces has been explored through the lens of structural functional analysis. Notably, Gurvits (2001) investigated scale-sensitive dimensions for linear functionals, illustrating how margin requirements control capacity via the geometric properties of the underlying space. Building on this, Mendelson (2002); Mendelson and Schechtman (2004) established profound connections between the fat-shattering dimension of linear functionals on a Banach space and its structural properties such as Rademacher type and cotype. Our work directly builds on these insights in our second main result, leveraging Mendelson's asymptotic capacity bounds to prove the impossibility of universally embedding abstract metric margin spaces into learnable Banach structures.

3.2. The Lipschitz Paradigm in Metric Spaces and its Bottlenecks

Moving beyond vector spaces, classical extensions of margin learning to abstract metric spaces have predominantly adopted the framework introduced by von Luxburg and Bousquet (2004). In their setting, the hypothesis class is defined as the *entire class of bounded Lipschitz functions*, where the inverse of the Lipschitz constant behaves as the natural metric analog to a linear margin.

However, a fundamental bottleneck of this approach is that the full Lipschitz class is exceptionally expressive. As studied in von Luxburg and Bousquet (2004); Gottlieb et al. (2014) and extended below, there is a direct connection tying the dimensionality (total boundedness) of a space to the uniform learnability of the class of Lipschitz functions defined on it. This requirement represents a severe restriction; indeed, learning the full Lipschitz class is hard even in standard infinite-dimensional Hilbert or Banach spaces because the space lacks total boundedness, restricting analysis to highly structured domains, such as metric spaces with a low doubling dimension.

3.3. Generalizations of Linear Classifiers to Metric Spaces

In light of the hardness of learning the full Lipschitz class, an independent and highly active line of inquiry has sought to identify and learn restricted sub-classes that act as true structural analogs to linear functions on metric domains. Prior attempts to generalize linear classification to metric spaces predominantly include Kernel Methods and implicit linearization Schölkopf and Smola (2002), learning with general similarity functions Balcan et al. (2008) and nearest-prototype decision boundaries Anthony and Ratsaby (2018).

Rather than targeting the full, unlearnable Lipschitz class or relying on implicit algebraic structures, our work focuses precisely on a well-defined structural generalization of hyperplanes. Our contribution reveals a fundamentally different distribution-free learning landscape than prior work, bypassing requirements such as total boundedness, doubling dimensions, or similarity-goodness axioms.

3.4. Scale-Sensitive Dimensions and Partial Concepts

Methodologically, our distribution-free analysis builds on the fundamental characterization of real-valued learnability established by Kearns and Schapire (1994); Alon et al. (1997); Bartlett and Long (1998). To model margin-based classifiers over abstract spaces rigorously, we formalize this setting using the framework of partial concept classes Long (2001); Alon et al. (2022) and leverage the modern paradigm of optimal sample complexity bounds for partial concepts established by Aden-Ali et al. (2023).

4. Main Results

4.1. Learnability in Metric Spaces

Let $\mathcal{X}$ be a metric space. We begin by studying the learnability of the class $\mathcal{D}_{\mathcal{X}}$ of bounded linear combinations of distance functions, which, as noted earlier, generalizes the notion of half-spaces to general metric spaces.

We show that the class $\mathcal{D}_{\mathcal{X}}$ fulfills a threshold-type phenomenon similar to the one exhibited in the introductory example. When the margin γ is sufficiently large, $\mathcal{D}_{\mathcal{X}}$ is γ-learnable in every metric space - hinging only on the triangle inequality - while smaller margins admit the full spectrum of possibilities, including non-learnability.

Theorem 4.1: A Dichotomy for Distance Functions.

Let $\mathcal{X}$ be a metric space and let $\mathcal{D}_{\mathcal{X}}$ denote the class

$$\mathcal{D}_{\mathcal{X}} := \left\{ \sum_{i=1}^{\infty} a_i \, d_{x_i} \;\middle|\; a_i \in \mathbb{R},\ x_i \in \mathcal{X},\ \sum_{i=1}^{\infty} |a_i| \leq 1 \right\},$$

where $d_x(\cdot) := d(\cdot, x)$. Then the following hold:

1. For every $\gamma \geq \frac{1}{3}$ and every space $\mathcal{X}$ with $\texttt{diam}(\mathcal{X}) \leq 1$, the class $\mathcal{D}_{\mathcal{X}}$ is γ-learnable.
2. For every $\gamma < \frac{1}{3}$, there exists a space $\mathcal{X}$ with $\texttt{diam}(\mathcal{X}) \leq 1$ such that $\mathcal{D}_{\mathcal{X}}$ is not γ-learnable.

By rescaling the metric, the threshold $\frac{1}{3}$ corresponds to a normalized margin of $\gamma/\texttt{diam}(\mathcal{X})$.

The proof is similar to the one from the introduction which shows the dichotomy of the class from Equation (1), both in how learnability follows from the triangle inequality, and in the construction of the non learnable class. For the full proof see Section 5.1.

The proof of the lower bound relies on constructing a metric space that is not *totally bounded*. Recall that a metric space is totally bounded if, for every $\gamma > 0$, there exists a finite cover of $\mathcal{X}$ by balls of radius γ. It is therefore natural to ask whether the dichotomy in the above theorem persists under a total boundedness assumption. Our next result shows that if $\mathcal{X}$ is totally bounded, then $\mathcal{D}_{\mathcal{X}}$ is γ-learnable for *every* $\gamma > 0$, and, moreover, shows that total boundedness is an exact characterization of Lipschitz learnability. While the sufficiency of total boundedness can be inferred from existing sample complexity bounds von Luxburg and Bousquet (2004); Gottlieb et al. (2014), we show that it is also necessary, thereby establishing that total boundedness is equivalent to the learnability of Lipschitz functions.

Lipschitz Functions. Recall that a function $f : \mathcal{X} \to \mathbb{R}$ is *1-Lipschitz* if $|f(x) - f(y)| \leq d(x, y)$ for all $x, y \in \mathcal{X}$, and let $\texttt{Lip}_{\mathcal{X}}$ denote the class of all such functions. Note that $\mathcal{D}_{\mathcal{X}} \subseteq \texttt{Lip}_{\mathcal{X}}$, since every distance function is 1-Lipschitz.

Theorem 4.2: Learnability of Lipschitz Functions.

Let $\mathcal{X}$ be a metric space and let $\texttt{Lip}$ denote the class of all 1-Lipschitz functions over $\mathcal{X}$. Then, the following are equivalent.

1. $\texttt{Lip}$ is γ-learnable for every $\gamma > 0$.
2. $\mathcal{X}$ is totally bounded.

Moreover, $\texttt{fat}_{\gamma}(\texttt{Lip})$ is exactly the 2γ-packing number of $\mathcal{X}$ (i.e., the maximum size of a subset of $\mathcal{X}$ with pairwise distances at least 2γ).

The main insight underlying the proof is that a dataset $S = \{(x_i, y_i)\}_{i=1}^{n}$ is γ-realizable by $\texttt{Lip}$ if and only if the minimum distance between positively and negatively labeled examples is at least 2γ. The full proof is given in Section 5.1.

Previous works mentioned have established learnability guarantees for Lipschitz function classes in terms of covering numbers of the underlying metric space. In particular, von Luxburg and Bousquet (2004); Gottlieb et al. (2014) derived sample complexity bounds for learning $\texttt{Lip}_{\mathcal{X}}$ that depend on the covering numbers of $\mathcal{X}$. Since finite covering numbers imply that $\mathcal{X}$ is totally bounded, these works show that total boundedness is a sufficient condition for learnability. Our contribution therefore is showing that total boundedness is not merely sufficient but also necessary. That is, total boundedness exactly characterizes when the class $\texttt{Lip}_{\mathcal{X}}$ is learnable under a margin assumption. Moreover, our characterization via packing numbers yields tight bounds on the optimal sample complexity.

4.2. Learnability in Banach spaces

We now turn to the question of the universality of linear classification with margin. Linear classifiers play a central role in learning theory, both due to their favorable sample complexity properties and because many abstract learning problems are approached via linear embeddings, for instance through kernel methods.

A notable feature of linear classification is that, unlike the general metric-space setting, margin-based learning in Hilbert spaces is possible for every $\gamma > 0$, despite the fact that such spaces are infinite-dimensional and not totally bounded. This naturally leads to the question of whether general margin-based learning problems can always be reduced to linear classification in normed spaces. More precisely, we ask whether every γ-learnable margin-based concept class can be embedded - possibly via an appropriate kernel - into a Banach space in which linear classification with margin is learnable.

We show that this universality does not hold in general: there exist concept classes that are learnable for every $\gamma > 0$, yet cannot be embedded into any Banach space admitting learnable linear classification with margin. Our argument relies on a characterization of the possible asymptotic dependence of the sample complexity on the margin in Banach spaces, given in the following result by Mendelson (2002); Mendelson and Schechtman (2004):

Theorem 4.3 (Fat-Shattering Dimension Bound for Linear Classifiers) *Mendelson (2002); Mendelson and Schechtman (2004)*
Let $\mathcal{B}$ be a Banach space and assume that $\mathtt{fat}_\gamma(L_\mathcal{B}) < \infty$ for some $\gamma \in (0,1)$. Then $\mathtt{fat}_\gamma(L_\mathcal{B}) < \infty$ for all $\gamma \in (0,1)$ and there is some $p \geq 2$ such that

$$\mathtt{fat}_\gamma(L_\mathcal{B}) = O(\frac{1}{\gamma^p}).$$

It is interesting to note that the converse is also true. That is, for every $p \geq 2$ there exists a Banach space for which $\mathtt{fat}_\gamma(L_\mathcal{B}) = \Theta(\frac{1}{\gamma^p})$. Combining the two, we get a taxonomy of Banach spaces based on the learnability rates of their linear functionals with margins.

4.2.1. EMBEDDING MARGIN PROBLEMS INTO BANACH SPACES

Embedding learning problems into linear spaces is a classical technique that has played a particularly influential role in machine learning, most notably through kernel methods. In this framework, nonlinear concept classes are analyzed via linear predictors in a suitable feature space. Given the central role of kernel methods in explaining margin-based generalization, it is natural to ask to what extent margin-based learnability can - or must - be understood through such linear embeddings.

To formalize this idea, we define an embedding of a margin problem $\mathcal{F}$ over a domain $\mathcal{X}$ into a Banach space $\mathcal{B}$ as a pair of maps $\Phi : \mathcal{X} \to \mathcal{B}_1$ and $\Psi : \mathcal{F} \to \mathcal{B}_1^\star$, where $\mathcal{B}_1$ and $\mathcal{B}_1^\star$ denote the closed unit balls of $\mathcal{B}$ and its dual $\mathcal{B}^\star$, respectively. We require that there exist constants $c_1, c_2 > 0$ such that for all $x \in \mathcal{X}$ and $f \in \mathcal{F}$,

$$c_1\, \Psi(f)\big(\Phi(x)\big) \;\leq\; f(x) \;\leq\; c_2\, \Psi(f)\big(\Phi(x)\big).$$

When $c_1 = c_2 = 1$, this condition reduces to $\Psi(f)(\Phi(x)) = f(x)$, i.e., the class $\mathcal{F}$ is realized exactly as the unit ball of linear functionals on some Banach space $\mathcal{B}$. Allowing c_1 and c_2 to differ relaxes this requirement, permitting controlled distortion when representing $\mathcal{F}$ as a family of linear functionals.

The guiding principle behind this definition is that of reduction of learning problems: an embedding of a margin problem $\mathcal{F}$ into a Banach space $\mathcal{B}$ reduces the task of learning $\mathcal{F}$ to that of learning linear classifiers in $\mathcal{B}$. Indeed, if a labeled sample $\{(x_i, y_i)\}_{i=1}^n$ is γ-realizable by $\mathcal{F}$, then the transformed sample $\{(\Phi(x_i), y_i)\}_{i=1}^n$ is $(C\gamma)$-realizable in $\mathcal{B}$, where $C = 1/\max(c_1, c_2)$, so

$$\mathtt{fat}_\gamma(\mathcal{F}) \leq \dim_{C\gamma}(L_\mathcal{B}).$$

From Theorem 4.3 we know that if a Banach space is γ-learnable for some $\gamma \in (0,1)$, then it is learnable for all $\gamma \in (0,1)$. Thus, if $\mathcal{F}$ admits an embedding into a learnable (for any or all γ) Banach space $\mathcal{B}$, then $\mathcal{F}$ itself is γ-learnable for all margins $\gamma \in (0,1)$.

This naturally raises the converse question: *can every margin problem $\mathcal{F}$ that is learnable for all $\gamma \in (0,1)$ be embedded into a margin-learnable Banach space?* Unfortunately, the answer is negative, as shown by the following result.

Theorem 4.4: Learnable class with no learnable Banach embedding

There exists a symmetric and convex set of functions $\mathcal{F} \subset \mathbb{R}^{\mathbb{N}}$ that is γ-learnable for all $\gamma \in (0,1)$, yet admits no embedding into any learnable Banach space.

The full proof of theorem 4.4, is given in 5.2. The main argument is an application of theorem 4.3, with a construction of a set of functions $\mathcal{F}$ for which $\dim_{\mathcal{F}}(\gamma)$, is not polynomial. Specifically, given a decreasing sequence of positive numbers $\{a_n\}_{n=1}^{\infty}$, define a class $\mathcal{F}$ of functions on the natural numbers by

$$\mathcal{F} = \left\{ f \in (0,1)^{\mathbb{N}} \ : \ |f(n)| \leq a_n \right\}.$$

Clearly $\mathcal{F}$ is symmetric and convex. Additionally, $\mathtt{fat}_\gamma(\mathcal{F})$ is the largest integer n for which $a_n < \gamma$. Consequently, if a_n decreases to 0 slowly enough we have that $\mathtt{fat}_\gamma(\mathcal{F})$ is not bound by any polynomial of $\frac{1}{\gamma}$. Hence $\mathcal{F}$ cannot be embedded into any learnable Banach space $\mathcal{B}$, since by theorem 4.3, any such embedding would enforce a polynomial upper bound on $\mathtt{fat}_\gamma(\mathcal{F})$.

5. Proofs

5.1. Proof of learnability of metric classes

Proof [Proof of Theorem 4.1] First, notice that $\mathcal{D}_{\mathcal{X}}$ is symmetric and closed under convex combinations. Hence by Theorem 2.2 it suffices to show that $\mathtt{fat}^0_\gamma(\mathcal{D}_{\mathcal{X}}) < \infty$. Observe that $\mathcal{D}_{\mathcal{X}}$ can be partitioned into two classes $\mathcal{D}_{\mathcal{X}} = \mathcal{D}^{<}_{\mathcal{X}} \uplus \mathcal{D}^{>}_{\mathcal{X}}$ such that

$$\mathcal{D}^{>}_{\mathcal{X}} := \Big\{ \sum_{i=1}^{n} a_i \, d_{x_i} \in \mathcal{D}_{\mathcal{X}} \mid \sum_{i:a_i \geq 0} a_i \geq \tfrac{1}{2} \Big\}, \quad \mathcal{D}^{<}_{\mathcal{X}} := \Big\{ \sum_{i=1}^{n} a_i \, d_{x_i} \in \mathcal{D}_{\mathcal{X}} \mid \sum_{i:a_i \geq 0} a_i < \tfrac{1}{2} \Big\}.$$

Since the property $\mathtt{fat}^0_\gamma < \infty$ is closed under finite unions (a direct consequence of the scale-sensitive Sauer–Shelah lemma), it suffices to show that both $\mathcal{D}^{>}_{\mathcal{X}}$ and $\mathcal{D}^{<}_{\mathcal{X}}$ have finite γ-fat-shattering dimension at 0. We show this for $\mathcal{D}^{>}_{\mathcal{X}}$. By the symmetric nature of the partition, the bound for $\mathcal{D}^{<}_{\mathcal{X}}$ follows immediately from that of $\mathcal{D}^{>}_{\mathcal{X}}$. Specifically, we show that if $\mathtt{fat}^0_\gamma(\mathcal{D}^{\gamma}_{\mathcal{X}}) \geq 2$ then $\gamma \leq \frac{1}{3}$.

Assume there are some $x, x' \in \mathcal{X}$, which are γ-shattered at 0 by $\mathcal{D}_{\mathcal{X}}{}^{>}$. Then there is some $f \in \mathcal{D}^{>}_{\mathcal{X}}$ such that $f(x) \leq -\gamma$ and $f(x') \leq -\gamma$. By definition, there are points $\{x_i\}_{i=1}^n \subseteq \mathcal{X}$ and scalars $\{a_i\}_{i=1}^n \subseteq \mathbb{R}$, with $\sum |a_i| \leq 1$, and $\sum_{i:a_i>0} a_i \geq \frac{1}{2}$ such that $f(x) = \sum_{i=1}^n a_i d_{x_i}(x) \leq -\gamma$ and $f(x') = \sum_{i=1}^n a_i d_{x_i}(x') \leq -\gamma$. Denote $I = \{1 \leq i \leq n \ : \ a_i \geq 0\}$, and $J = \{1 \leq j \leq n \ : \ a_j < 0\}$ so we have

$$\sum_{i \in I} a_i d(x, x_i) - \sum_{j \in J} |a_j| d(x, x_j) = \sum_{k=1}^{n} a_k d(x, x_k) \leq -\gamma.$$

Which implies

$$\sum_{i \in I} a_i d(x, x_i) \leq \sum_{j \in J} |a_j| d(x, x_j) - \gamma.$$

And the same holds if we replace x with x'. Together with $\sum_{i \in I} a_i \geq \frac{1}{2}$ and the triangle inequality, this implies

$$\begin{aligned}
\frac{1}{2} d(x, x') &\leq \sum_{i \in I} a_i \, d(x, x') \\
&\leq \sum_{i \in I} a_i \big[d(x, x_i) + d(x_i, x') \big] \\
&\leq \Big(\sum_{j \in J} |a_j| \big[d(x, x_j) + d(x_j, x') \big] \Big) - 2\gamma \\
&\leq \frac{1}{2}(1+1) - 2\gamma = 1 - 2\gamma.
\end{aligned}$$

Where in the last inequality we used the assumption that $\mathtt{Diam}(\mathcal{X}) \leq 1$.

Additionally, from γ-fat shattering at 0, we have some $\{x'_i\}_{i=1}^m \subseteq \mathcal{X}$ and $\{a'_i\}_{i=1}^m \subseteq \mathbb{R}$ such that $f'(x) = \sum_{i=1}^m a'_i d(x'_i, x) > \gamma$ and $f'(x') = \sum_{i=1}^m a_i d(x'_i, x') < -\gamma$. From this we get

$$d(x, x') \geq \sum_{i=1}^n a'_i d(x, x') \geq \sum_{i=1}^n a'_i [d(x, x'_i) - d(x', x'_i)] \geq 2\gamma.$$

So our two inequalities gives

$$2\gamma \leq d(x, x') \leq 2 - 4\gamma.$$

Which implies $\gamma \leq \frac{1}{3}$. Define a metric space $(\mathcal{X}, d)$ as follows: Let $X = A \cup B$ such that $A = \{a_n\}_{n\in\mathbb{N}}$ is a countable set of points and $B = B_1 \cup B_2$ where $B_i = \{b^i_s\}_{\{s\subseteq A||s|<\infty\}}$ are a pair of sets of points referencing finite subsets of A.

$$\rho(a_i, a_j) = \frac{2}{3}, \quad \rho(b^i_s, b^j_w) = \frac{2}{3}, \quad \rho(a_i, b^j_s) = \begin{cases} \frac{2}{3} - \gamma & a_i \in s,\ j = 1 \\ \frac{2}{3} + \gamma & a_i \notin s,\ j = 1 \\ \frac{2}{3} + \gamma & a_i \in s,\ j = 2 \\ \frac{2}{3} - \gamma & a_i \notin s,\ j = 2 \end{cases}$$

For any $\gamma \leq 1/3$ this is indeed a metric space with $Diam(\mathcal{X}) \leq 1$. For any finite $A' \subset A$, let $\delta_{A'} := \frac{1}{2} \cdot d_{b^1_{A'}} - \frac{1}{2} \cdot d_{b^2_{A'}} \in \mathcal{D}^\gamma_\mathcal{X}$. Then $\delta|_A = -\gamma \cdot \mathbf{1}_{A'} + \gamma \cdot \mathbf{1}_{A\setminus A'}$. Therefore $\texttt{fat}_\gamma(\mathcal{D}_\mathcal{X}) > n$ for all n, concluding our proof. ■

Proof [Proof of Theorem 4.2 - Learnability of Lipschitz functions] First, notice that `Lip` is symmetric and closed under convex combinations. Hence by Theorem 2.2, γ-learnability of `Lip` is equivalent to the finiteness of $\texttt{fat}^0_\gamma(\texttt{Lip})$.

Let $S = \{(x_i, y_i)\}_{i=1}^n$, be some sample γ-realizable by `Lip`, denote the positive and negative labeled examples by $S^+ = \{(x, y) \in S \ : \ y = 1\}$, $S^- = \{(x, y) \in S \ : \ y = -1\}$, and note that $d(S^+, S^-) \geq 2\gamma$. Conversely, any such sample S for which $d(S^+, S^-) \geq 2\gamma$, is γ-realizable by `Lip`. Indeed define $f \in \texttt{Lip}$ by

$$f(x) = \frac{d(S^-, x) - d(S^+, x)}{2},$$

and note that if $x \in S^+$ then $f(x) = \frac{1}{2} \cdot d(S^-, x) \geq \gamma$, and similarly if $x \in S^-$ then $f(x) \leq -\gamma$. We can also see that $f \in \texttt{Lip}$ since $d(A, y) \leq d(A, x) + d(x, y)$ for any set $A \subset \mathcal{X}$ and points $x, y \in \mathcal{X}$. Thus for any sample S:

$$\textbf{S is } \gamma\textbf{-realizable by } \texttt{Lip} \iff d(S^+, S^-) \geq 2\gamma$$

$1 \to 2$ Assume to the contrary that $\mathcal{X}$ is not totally bounded, i.e. that for every n, there exists a set $S \subseteq \mathcal{X}$ of cardinality n such that $d(x_i, x_j) \geq 2\gamma$ for any pair of different points $x_i, x_j \in S$. Then for any sign pattern $y \in \{\pm 1\}^n$ we have that $d(S^-, S^+) \geq 2\gamma$, where $S^- = \{x_i \in S \ : y_i = -1\}$, and $S^+ = \{x_i \in S \ : \ y_i = 1\}$. Hence $\{(x_i, y_i)\}_{i=1}^n$ is γ-realized by $f \in \texttt{Lip}$ given by

$$f(x) = \frac{d(S^-, x) - d(S^+, x)}{2}.$$

Implying that $\texttt{fat}^0_\gamma(\texttt{Lip}) \geq n$ for all n.

$2 \to 1$ For the other direction, assume by negation that there exists $\gamma > 0$ such that `Lip` is not γ-learnable. Let S be a sample γ-shattered at 0 by `Lip`. Then for any $x_i, x_j \in S$ there is some $f \in \texttt{Lip}$ such that $f(x_i) \geq \gamma$, $f(x_j) \leq -\gamma$, hence we have

$$2\gamma \leq f(x_i) - f(x_j) \leq d(x_i, x_j).$$

Since this is true for sets S of arbitrary size, we conclude that $\mathcal{X}$ is not totally bounded. ■

5.2. Proofs for margin embedding and characterization of shattering results

Proof [Proof for Theorem 4.4 - non embeddability of margin spaces] Define $\mathcal{F} \subset (-1,1)^{\mathbb{N}}$ by

$$\mathcal{F} = \{f \in (-1,1)^{\mathbb{N}} \ : |f(n)| \leq \frac{1}{\log n}\}$$

Clearly $\mathcal{F}$ is symmetric and convex. It is also clear that S is γ-shattered by $\mathcal{F}$ if and only if $\gamma < \frac{1}{\log x}$ for all $x \in S$. Hence we deduce that $\texttt{fat}_\gamma(\mathcal{F}) = \lfloor e^{\frac{1}{\gamma}} \rfloor$, and in particular $\mathcal{F}$ is γ-learnable for all γ. Note that any embedding of $\mathcal{F}$ into a Banach space $\mathcal{B}$ will imply that there is some constant $C > 0$ such that

$$[e^{\frac{1}{\gamma}}] = \texttt{fat}_\gamma(\mathcal{F}) \leq \texttt{fat}_{C\gamma}(L_\mathcal{B}).$$

By Theorem 4.3, this implies that B is not γ learnable for any $\gamma > 0$, as any learnable Banach space will obey the bound $\texttt{fat}_\gamma(L_\mathcal{B}) = O(\frac{1}{\gamma^p})$ for some $p > 0$. ■

Proof [Proof of Proposition 2.1 - Characterization of shattering in margin spaces]

(4) $\implies$ (3) : Assume toward contradiction that (4) holds but (3) does not hold, so there is some $y \in [-\gamma, \gamma]^n$ such that for every $f \in \mathcal{F}$ there is some $i \in [n]$ such that $f(x_i) \neq y_i$. Consider the set

$$\mathcal{F}|_S = \{\big(f(x_1), f(x_2) \dots f(x_n)\big) \ : \ f \in \mathcal{F}\}.$$

Since $\mathcal{F}$ is convex so is $\mathcal{F}|_S \subset \mathbb{R}^n$, and by assumption $y \notin \mathcal{F}|_S$. Hence by the Hahn-Banach theorem there is some vector $\lambda \in \mathbb{R}^n$, and scalar $a \in \mathbb{R}$ such that

$$\sum_{i=1}^n \lambda_i y_i = a.$$

But for any $f \in \mathcal{F}|_S$ we have

$$\sum_{i=1}^n \lambda_i f_i < a.$$

By normalization we may assume $\sum_{i=1}^n |\lambda_i| = 1$. Hence, since we assumed (3), we get that there is some $f \in \mathcal{F}$ such that

$$\sum_{i=1}^n \lambda_i f(x_i) \geq \gamma = \sum_{i=1}^n |\lambda_i|\gamma \geq \sum_{i=1}^n \lambda_i y_i = a$$

In contradiction.

(3) $\implies$ (2) : Obvious.

(2) $\implies$ (4) : Let $\lambda \in \mathbb{R}^n$, be such that $\sum_{i=1}^n |\lambda_i| = 1$. Since we assume that S is shattered by $\mathcal{F}$, for any sign pattern $s \in \{-1,1\}^n$, there is some $f \in \mathcal{F}$ such that $s_i f(x_i) \geq \gamma$ for all $1 \leq i \leq n$. In particular we may find such f for the sign pattern $s_i = \text{sign}(\lambda_i)$ (choosing arbitrarily in the cases where $\lambda_i = 0$). Then for this f we get for all $1 \leq i \leq n$

$$\lambda_i f(x_i) = |\lambda_i| \text{sign}(\lambda_i) f(x_i) \geq |\lambda_i|\gamma.$$

From which we deduce

$$\sum_{i=1}^n \lambda_i f(x_i) \geq \sum_{i=1}^n |\lambda_i|\gamma = \gamma.$$

Which gives the desired result. Finally we show that if (1) and (2) are equivalent when $\mathcal{F}$ is also symmetric. Clearly (2) $\implies$ (1) so we only need to show the converse. Assume we have some $r : S \to \mathbb{R}$ such that for any $s \in \{-1,1\}^n$ there is f_s such that $\big(f_s(x_i) - r(x_i)\big)s_i \geq \gamma$ for all $1 \leq i \leq n$. Then for any such $s \in \{-1,1\}$ define

$$\phi_s = \frac{f_s - f_{-s}}{2}.$$

Where f_{-s} is the function corresponding the labeling $-s \in \{-1,1\}^n$. Since F symmetric we have that $-f_{-s} \in \mathcal{F}$, and since it is convex we have that $\phi_s \in \mathcal{F}$. To conclude the proof we note that for all $1 \leq i \leq n$

$$\phi_s(x_i)s_i = \frac{s_i f_s - s_i f_{-s}}{2} \geq \frac{\gamma + \gamma}{2} = \gamma.$$

■

Acknowledgments

YA and RL are supported by the European Union (ERC, FoG 101116258). Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

SM and TW are supported by Israel PBC-VATAT, by the Technion Center for Machine Learning and Intelligent Systems (MLIS), and by the European Union (ERC, GENERALIZATION, 101039692). Views and opinions expressed are those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

We thank the anonymous reviewers for their insightful suggestions and comments regarding the fat-shattering dimension, which greatly improved the context and technical clarity of our results.

References

Ishaq Aden-Ali, Yeshwanth Cherapanamjeri, Abhishek Shetty, and Nikita Zhivotovskiy. Optimal pac bounds without uniform convergence. In *Proceedings of the 64th IEEE Annual Symposium on Foundations of Computer Science (FOCS)*, 2023.

Noga Alon, Shai Ben-David, Nicolo Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *Journal of the ACM (JACM)*, 44(4):615–631, 1997.

Noga Alon, Steve Hanneke, Ron Holzman, and Shay Moran. A Theory of PAC Learnability of Partial Concept Classes . In *2021 IEEE 62nd Annual Symposium on Foundations of Computer Science (FOCS)*, pages 658–671, Los Alamitos, CA, USA, February 2022. IEEE Computer Society. doi: 10.1109/FOCS52979.2021.00070. URL https://doi.ieeecomputersociety.org/10.1109/FOCS52979.2021.00070.

Martin Anthony and Joel Ratsaby. Large width nearest prototype classification on general distance spaces. *Theoretical Computer Science*, 738:65–79, 2018.

Idan Attias, Steve Hanneke, Alkis Kalavasis, Amin Karbasi, and Grigoris Velegkas. Optimal learners for realizable regression: Pac learning and online learning. *Advances in Neural Information Processing Systems*, 36:44707–44739, 2023.

Maria-Florina Balcan, Avrim Blum, and Nathan Srebro. Improved guarantees for learning via similarity functions. 2008.

Peter L Bartlett. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE transactions on Information Theory*, 44(2):525–536, 1998.

Peter L. Bartlett and Philip M. Long. Prediction, learning, uniform convergence, and scale-sensitive dimensions. *Journal of Computer and System Sciences*, 56(2):174–190, 1998. ISSN 0022-0000. doi: https://doi.org/10.1006/jcss.1997.1557. URL https://www.sciencedirect.com/science/article/pii/S0022000097915579.

Peter L. Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Corinna Cortes and Vladimir Vapnik. Support-vector networks. *Machine learning*, 20(3):273–297, 1995.

Nello Cristianini and John Shawe-Taylor. *An Introduction to Support Vector Machines and Other Kernel-based Learning Methods*. Cambridge University Press, Cambridge, 2000.

Lee-Ad Gottlieb, Aryeh Kontorovich, and Robert Krauthgamer. Efficient classification for metric data. *IEEE Transactions on Information Theory*, 60(9):5750–5759, 2014. doi: 10.1109/TIT.2014.2339840.

Leonid Gurvits. A note on a scale-sensitive dimension of linear bounded functionals in banach spaces. *Theoretical Computer Science*, 261(1):81–90, 2001.

Michael J. Kearns and Robert E. Schapire. Efficient distribution-free learning of probabilistic concepts. *Journal of Computer and System Sciences*, 48(3):464–497, 1994. ISSN 0022-0000. doi: https://doi.org/10.1016/S0022-0000(05)80062-5. URL https://www.sciencedirect.com/science/article/pii/S0022000005800625.

Philip M. Long. On agnostic learning with {0, *, 1}-valued and real-valued hypotheses. In David P. Helmbold and Robert C. Williamson, editors, *Computational Learning Theory, 14th Annual Conference on Computational Learning Theory, COLT 2001 and 5th European Conference on Computational Learning Theory, EuroCOLT 2001, Amsterdam, The Netherlands, July 16-19, 2001, Proceedings*, volume 2111 of *Lecture Notes in Computer Science*, pages 289–302. Springer, 2001. doi: 10.1007/3-540-44581-1_19. URL https://doi.org/10.1007/3-540-44581-1_19.

Shahar Mendelson. Learnability in hilbert spaces with reproducing kernels. *journal of complexity*, 18(1): 152–170, 2002.

Shahar Mendelson and Gideon Schechtman. The shattering dimension of sets of linear functionals. 2004.

Bernhard Schölkopf and Alexander J Smola. *Learning with kernels: support vector machines, regularization, optimization, and beyond*. MIT press, 2002.

Hans Ulrich Simon. Bounds on the number of examples needed for learning functions. *SIAM Journal on Computing*, 26(3):751–763, 1997.

Vladimir N. Vapnik. *Statistical Learning Theory*. Wiley, New York, 1998.

Ulrike von Luxburg and Olivier Bousquet. Distance-based classification with lipschitz functions. *J. Mach. Learn. Res.*, 5:669–695, 2004. URL https://jmlr.org/papers/volume5/luxburg04b/luxburg04b.pdf.