Uniform convergence at optimal rate and scale

Noga Alon*, Steve Hanneke†, Shay Moran‡, and Tom Waknine§

October 20, 2026

Abstract

Uniform convergence asks how many random samples suffice for empirical averages to approximate expectations simultaneously over an entire class of functions. It is a central concept in probability and a foundation of statistical learning. For binary-valued classes, the Vapnik–Chervonenkis theorem characterizes uniform convergence by finiteness of the VC dimension. For dimension d, the sharp worst-case sample complexity for additive error ε is of order d/ε^2. For real-valued functions, the analogous combinatorial parameter is the fat-shattering dimension, which depends on an approximation scale γ. The challenge is to achieve error $\gamma + \varepsilon$ with order d/ε^2 samples, matching the sample bound for binary classes, where d is now the fat-shattering dimension at scale γ. Previous results achieved either the optimal scale or the optimal sample complexity, but not both. Our main result closes this gap, establishing the sharp sample complexity at the optimal scale.

Along the way, we use several tools that may be of independent interest. The first concerns ε-nets with margin. A classical ε-net intersects every set of mass at least ε in a prescribed family; we allow the net to intersect a slightly enlarged version of each such set. Remarkably, this relaxation permits bounds independent of the ambient dimension in natural geometric settings. Second, we derive improved covering bounds in both discrete and real-valued settings: we extend Haussler's covering lemma to partial concept classes and use this extension to obtain ℓ_2 covers at the optimal scale for real-valued function classes. Third, we obtain a chaining bound at the optimal scale by aligning the covers so that successive approximations become increasingly close to one another. We also highlight open problems and directions for future research arising from these results.

1 Introduction

Suppose we wish to estimate the expectation of a bounded function f under an unknown distribution. We can draw independent samples and average the values of f on these samples. The Law of Large Numbers guarantees that this average converges to the expectation. Hoeffding's Inequality [Hoe63] makes this quantitative: $O(\varepsilon^{-2}\log(1/\delta))$ samples suffice for error at most ε with confidence $1-\delta$. Suppose now we wish to estimate the expectations of M such functions. A union bound gives the simultaneous guarantee with

$$O\left(\frac{\log M + \log(1/\delta)}{\varepsilon^2}\right)$$

samples. Thus, the number of functions enters only logarithmically.

Uniform convergence asks when a single sample can accurately estimate all expectations in an infinite class $\mathcal{F}$ of functions[1]. The required sample size may depend on $\mathcal{F}$, the desired accuracy, and the confidence, but must work for every underlying distribution. Classes $\mathcal{F}$ for which the empirical averages uniformly approximate the expectations are called *uniform Glivenko–Cantelli classes* [DGZ91].

*Department of Mathematics, Princeton University, Princeton, NJ, USA. Research supported in part by NSF grant DMS-2553988 and by AFOSR grant FA9550261B300.

†Department of Computer Science, Purdue University, West Lafayette, IN, USA. Research supported by grant no. 2024243 from the United States - Israel Binational Science Foundation (BSF).

‡Departments of Mathematics, Computer Science, and Data and Decision Sciences, Technion–Israel Institute of Technology; Google Research.

§Department of Mathematics, Technion–Israel Institute of Technology, Haifa, Israel.

[1]Throughout the paper, we assume that $\mathcal{F}$ satisfies some mild measurability conditions such as [Pol84, Appendix C]

A classical example is estimating a cumulative distribution function (CDF). For independent samples $S = (x_1, \ldots, x_n)$ from a distribution $\mathcal{D}$ on $\mathbb{R}$, the true and empirical CDFs are

$$F_{\mathcal{D}}(t) = \mathop{\mathbb{P}}_{x \sim \mathcal{D}}[x \le t], \qquad \widehat{F}_S(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{\{x_i \le t\}}.$$

These two CDFs are the expectations and sample averages of the threshold functions $f_t(x) = \mathbf{1}_{\{x \le t\}}$. Uniform convergence for this class means that the same sample estimates $F_{\mathcal{D}}(t)$ accurately for all thresholds simultaneously. The Glivenko–Cantelli Theorem, proved in 1933 [Gli33, Can33], guarantees this accuracy with high probability for sufficiently large samples. An entire CDF can thus be estimated without a parametric model. The theorem is therefore sometimes called the *Fundamental Theorem of Mathematical Statistics* [Rén07, p.400].

Binary-valued functions. For classes of binary-valued functions, Vapnik and Chervonenkis [VC71] showed that uniform convergence holds precisely when the VC dimension is finite. This characterization also forms a major part of the Fundamental Theorem of PAC Learning [BEHW89]. Their proof also bounds the number of samples needed. For a class of VC dimension d, their classical symmetrization argument, based on an independent *ghost sample*, gives

$$m(\varepsilon, \delta) = O\left(\frac{d \log(1/\varepsilon) + \log(1/\delta)}{\varepsilon^2}\right),$$

where ε is the approximation error and δ is the allowed failure probability.

The extra logarithmic factor can be removed by combining *chaining* with bounds on the *covering numbers* of the function class. Dudley's chaining method [Dud78], Haussler's covering estimates [Hau95], and Talagrand's bounds for empirical processes [Tal94] yield the optimal worst-case sample complexity

$$\Theta\left(\frac{d + \log(1/\delta)}{\varepsilon^2}\right), \qquad d = \mathtt{VC}(\mathcal{F}).$$

This gives a tight theorem for binary classes, using techniques with applications in probability, functional analysis, and theoretical computer science. Our proof refines the covering and chaining tools, which we recall in Section 2.

Real-valued functions. The binary case corresponds to estimating probabilities of events. More generally, we often want to estimate expectations of real-valued functions. For such classes, VC dimension is replaced by a scale-dependent extension, called the *fat-shattering dimension* [KS94].

Fix a scale $\gamma > 0$. The dimension $\mathtt{fat}_\gamma(\mathcal{F})$ is the largest n for which there are points $x_1, \ldots, x_n$ and thresholds $r_1, \ldots, r_n$ with the following property: for each of the 2^n ways to label the points $x_1, \ldots, x_n$ *high* or *low*, there is a function $f \in \mathcal{F}$ satisfying all the corresponding inequalities simultaneously:

$$f(x_i) \ge r_i + \frac{\gamma}{2} \quad \text{if } x_i \text{ is labeled } \textit{high},$$
$$f(x_i) \le r_i - \frac{\gamma}{2} \quad \text{if } x_i \text{ is labeled } \textit{low}.$$

The thresholds are chosen once, while f may depend on the labeling.

For $\{0,1\}$-valued classes and $\gamma = 1$, this recovers VC dimension: realizing these patterns is exactly realizing every binary labeling. For real-valued classes, the fat-shattering dimension may decrease as the scale increases. Classical work showed that a bounded class satisfies uniform convergence exactly when its fat-shattering dimension is finite at every positive scale [ABDCBH97, BL98]. These classical works reduce the study of uniform convergence to understanding what can be guaranteed at each scale:

If the fat-shattering dimension is finite at a scale γ, how accurately can we estimate the expectations of all functions in the class simultaneously?

Without information at finer scales, we cannot in general guarantee error below γ. The natural target is therefore $\gamma + \varepsilon$, where the additional error ε can be reduced by drawing more samples. Our goal is to achieve this with the same sample complexity as in the binary case. We denote the required sample size, at confidence $1 - \delta$, by $m_\gamma(\varepsilon, \delta)$. If this sample size is finite for every accuracy and confidence, the class satisfies *γ-uniform convergence*.

The classical bounds gave error $C\gamma + \varepsilon$, for a universal constant $C > 1$, with additional logarithmic factors in the sample complexity [ABDCBH97, BL98]. Colomboni, Esposito, and Paudice [CEP25], using bounds of Rudelson and Vershynin [RV06], obtained the optimal binary rate, but with error $C\gamma + \varepsilon$ for a universal constant $C > 1$. The result of [AMM+26] reaches error $\gamma + \varepsilon$, but retains additional logarithmic factors in the sample complexity. Thus, the optimal rate and the optimal scale were available separately.

Our contribution. We obtain both simultaneously:

Theorem 1.1: Main result

Let $\mathcal{F} \subseteq [-1, +1]^{\mathcal{X}}$, let $\gamma \in (0, 1)$, and suppose that $d := \mathtt{fat}_\gamma(\mathcal{F}) < \infty$. Then $\mathcal{F}$ satisfies γ-uniform convergence, with

$$m_\gamma(\varepsilon, \delta) = O\left(\frac{d + \log(1/\delta)}{\varepsilon^2}\right) \qquad \varepsilon, \delta \in (0, 1).$$

Moreover, for every $\gamma \in (0, 1)$ and integer $d \geq 1$, there is a class $\mathcal{F}$ with $\mathtt{fat}_\gamma(\mathcal{F}) = d$ such that, for all sufficiently small $\varepsilon, \delta > 0$,

$$m_\gamma(\varepsilon, \delta) = \Omega\left(\frac{d + \log(1/\delta)}{\varepsilon^2}\right).$$

The theorem gives the optimal binary sample complexity at error $\gamma + \varepsilon$. The matching lower bound shows that this cannot be improved in general using only the fat-shattering dimension at scale γ. Particular classes may admit stronger guarantees when more is known about their behavior at finer scales. Our proof combines covering bounds with chaining at optimal scale; the full proof appears in Section 4.5.

In section 2 we give an self-contained presentation overview of technical tools used to prove our main result. In section 3 we give a proof overview of the main results in the paper. And finally, in section 4 we give formal proofs for all results in the paper.

2 Technical contributions

Before turning to the proof of our uniform-convergence bound, we give a self-contained presentation of the technical tools behind it. Our aim is to explain the ideas in their natural settings and relate them to earlier work, so that the techniques can be understood beyond the present proof. This also brings out results that may be of independent interest, including some that are not used directly in our proof but illustrate the techniques more broadly.

We begin with ε-nets with margin and rounding fractional covers, then present covering bounds for partial binary classes and real-valued classes, and conclude with chaining at optimal scale in Euclidean space. The subsections can be read largely independently. In Section 3, we return to uniform convergence and explain how these tools combine to prove our main result.

2.1 ε-nets with margin and rounding fractional covers

A basic question in combinatorics is how to choose a small set of points that intersects every sufficiently large set in a given family. Let $\mathcal{X}$ be a domain, let $X \subset \mathcal{X}$ be a set of n points, and let $\mathcal{U}$ be a family of subsets of $\mathcal{X}$. A set $E \subseteq \mathcal{X}$ is an ε-net for $\mathcal{U}$ with respect to X if it intersects every $U \in \mathcal{U}$ with

$|U \cap X| \geq \varepsilon n$. The Haussler–Welzl theorem [HW86] guarantees that

$$O\left(\frac{d}{\varepsilon}\log\frac{1}{\varepsilon}\right)$$

points sampled independently and uniformly from X form such a net with constant probability, where d is the VC dimension of the set family $\mathcal{U}$. Importantly, this bound is independent of n.

Halfspaces with margin. Halfspaces in $\mathbb{R}^d$ provide a classical illustration of this theorem. Their VC dimension is $d+1$, so the Haussler–Welzl theorem [HW86] guarantees that, given n points, one can select $O((d/\varepsilon)\log(1/\varepsilon))$ of them to meet every halfspace containing at least εn points. Conversely, for $d \geq 2$, there are point sets for which any such selection requires $\Omega(d/\varepsilon)$ points.[2] The logarithmic gap between these upper and lower bounds has attracted considerable attention: Matoušek, Seidel, and Welzl removed it for $d = 2, 3$ [MSW90], while Pach and Tardos showed that it is necessary already in dimension four [PT11]. Kupavskii, Mustafa, and Pach [KMP16] subsequently established the matching lower bound $\Omega((d/\varepsilon)\log(1/\varepsilon))$ for all $d \geq 4$, settling the dependence on both parameters.[3]

We consider a natural relaxation of ε-nets that allows a *margin*: it suffices to hit each large halfspace after moving its boundary outward by a distance η. Equivalently, place a ball of radius η around each net point and require every large halfspace to intersect one of these balls. This margin viewpoint has been studied through partial concept classes [BCT25, IJN26]. It also appears in work on no-dimensional variants of Carathéodory's, Helly's, and Tverberg's theorems and on weak ε-nets with margin for convex sets; see, e.g., [ABMT20, Iva21, CM22, HPR24, BIP26, IK26].

The following theorem gives a bound independent of the Euclidean dimension. It applies to ℓ_2 geometry in any real Hilbert space, including infinite-dimensional ones.

Theorem 2.1 (Halfspace nets with margin in Hilbert space)**.** *Let X be a set of n points in the unit ball*[4] *of a real Hilbert space $\mathcal{X}$, and let $\varepsilon, \eta \in (0,1)$. Then, there is a set $E \subseteq \mathcal{X}$ of size $O(1/(\varepsilon\eta^2))$ such that every closed halfspace containing at least εn points of X has distance at most η from some point of E.*
The bound is optimal: for all sufficiently small $\varepsilon, \eta > 0$, there is a finite set X in the unit ball of ℓ_2 for which every such net has $\Omega(1/(\varepsilon\eta^2))$ points.

This improves upon a result of Bourneuf, Charbit, and Thomassé [BCT25], who extended the Haussler–Welzl theorem to partial concept classes. Combining their extension with the classical bound of order $1/\eta^2$ on the VC dimension of halfspaces with margin η gives a net of size $O((\varepsilon\eta^2)^{-1}\log(1/\varepsilon))$ in Hilbert space. Theorem 2.1 removes this logarithmic factor and gives the optimal dependence on ε and η. Its full proof appears in Section 4.

It is interesting to compare this bound with the classical bounds without margin. Our bound is linear in $1/\varepsilon$, whereas the classical bound has an extra factor of $\log(1/\varepsilon)$ for $d \geq 4$. Another difference is that our nets are *weak*, or *external*: their points may lie anywhere in the ambient space. They are not obtained by randomly sampling points from X, as in the Haussler–Welzl construction. Indeed, independent uniform sampling from X can still require the logarithmic factor even with a margin, as follows from lower bounds for margin-based generalization; see, e.g., [GKL20, Lemma 11].

Beyond the Hilbert-space setting, Ivanov et al. also study margins for halfspaces in other normed spaces and for balls in L_p spaces [IJN26]. It would be interesting to determine the optimal sizes of margin nets in these settings.

2.1.1 Rounding fractional covers

It is natural to study ε-nets with margin for other geometric families, such as polytopes of bounded complexity or even arbitrary convex sets. In each case, allowing a margin means slightly enlarging the

[2] In dimension one, halfspaces are rays, and the minimum and maximum of X hit every ray that intersects X.
[3] These lower bounds concern strong ε-nets, whose points must belong to X.
[4] For points in a ball of radius R, the same bound holds with margin ηR. Rescaling the lower bound in the proof shows that dependence on the ratio between the radius and the margin is necessary.

set by moving its boundary outward. For balls, this simply increases the radius while keeping the center fixed. As with halfspaces, this relaxation yields dimension-independent bounds for balls. While the result holds in more general spaces, we present it in Hamming space. Both because this is the setting relevant to our proof and because it illustrates the applicability of margin nets in a more combinatorial setting. We use these nets to obtain a fractional extension of the familiar packing–covering relation.

For a finite set A in a metric space (Z, ρ), let $P_r(A)$ be the maximum number of pairwise disjoint closed balls of radius r centered in A, and let $C_r(A)$ be the minimum number of radius-r balls that cover A, with centers anywhere in Z. These are the matching and transversal numbers of the ball family $\{B_r(a) : a \in A\}$: hitting every ball is equivalent to covering its center. The familiar packing–covering inequality is

$$C_{2r}(A) \le P_r(A) \le C_r(A).$$

Indeed, a maximal family of disjoint radius-r balls yields a $2r$-cover by the triangle inequality, while disjoint balls require distinct hitting points.

In a fractional cover, each center receives a nonnegative weight, and the centers within distance r of each $a \in A$ must have total weight at least one. A fractional packing assigns nonnegative weights to the balls $B_r(a)$, with total weight at most one on the balls containing any given point. Linear-programming duality equates the minimum covering weight and maximum packing weight:

$$C_r^*(A) = P_r^*(A).$$

The classical relation therefore gives $C_{2r}(A) \le P_r^*(A) = C_r^*(A)$.

We show that, in Hamming space, a fractional cover can be rounded with a radius increase of only $1+\eta$, at the expense of more centers. We use the normalized Hamming distance d_H on $\{0,1\}^n$, which is the fraction of coordinates on which two vectors differ, and write $B_r(x) = \{y : d_H(x,y) \le r\}$. All packing and covering numbers below use the cube as the ambient space.

Theorem 2.2 (Rounding fractional covers)**.** *Let $A \subseteq \{0,1\}^n$ be nonempty, and let $r, \eta \in (0,1)$. Write $M = C_r^*(A) = P_r^*(A)$ for the common value of the fractional covering and fractional packing numbers at radius r. Then*

$$C_{(1+\eta)r}(A) = O\left(\frac{M}{\eta^2}\log(eM)\right).$$

Thus, for any fixed $\eta > 0$, rounding costs only a logarithmic factor relative to the fractional covering weight, while increasing the radius by a factor of $1+\eta$.

The connection to ε-nets is simple. Normalize the weights of a fractional cover of total weight M to a probability distribution μ on the possible centers. Every ball $B_r(a)$, $a \in A$, then has mass at least $1/M$. A margin net that meets each of these balls after its radius is increased to $(1+\eta)r$ places a point within distance $(1+\eta)r$ of every $a \in A$. Its points therefore form the desired cover. The following theorem, applied with radius r, margin ηr, and mass threshold $1/(2M)$, gives the required bound.[5]

Theorem 2.3 (Nets for Hamming balls with margin)**.** *Let μ be a probability distribution on $\{0,1\}^n$. Fix a radius α and an additive margin η satisfying $0 < \eta \le \alpha$ and $\alpha + \eta \le 1$. For every $\varepsilon \in (0,1)$, there is a set $E \subseteq \{0,1\}^n$ of size*

$$|E| = O\left(\frac{\alpha^2}{\varepsilon\eta^2}\log\frac{e}{\varepsilon}\right)$$

such that every closed Hamming ball of radius α and μ-mass at least ε intersects E after its radius is increased to $\alpha+\eta$.
The bound is optimal up to the logarithmic factor for small radii. For every sufficiently small $\varepsilon, \alpha > 0$ and every $0 < \eta \le \alpha$, there are a dimension n and a distribution μ for which every such net has $\Omega(\alpha^2/(\varepsilon\eta^2))$ points.

[5] If $(1+\eta)r \ge 1$, one ball already covers the entire cube.

The upper bound is independent of n and depends on the radius and margin only through their ratio α/η. The same upper bound extends, uniformly in the dimension, to balls in ℓ_p for every $p \in [1, 2]$, as explained at the end of the proof in Section 4.

A restriction on the radius is necessary for the lower bound: if $\alpha + \eta \geq 1/2$, the two antipodal points $\mathbf{1}$ and $-\mathbf{1}$ meet every enlarged ball. For small radii, a factor of $\log(1/\varepsilon)$ remains between our upper and lower bounds. The lower-bound construction also extends to balls in ℓ_p, with the same dependence on the parameters, when the net is required to be strong. It does not, however, give a corresponding lower bound for weak nets.

2.2 Covering numbers

In the preceding subsection, we studied covers in metric spaces. We now apply this viewpoint to classes of functions on a domain of n points. Each function is identified with the vector of its n values, and a cover consists of finitely many representatives that approximate every function in the class. We consider both binary-valued and real-valued functions. For binary functions, distance is the fraction of domain points on which they disagree; for real-valued functions, we use normalized Euclidean distance.

We first establish covering bounds for partial binary functions, which may be undefined on some inputs, and then use them to cover real-valued classes. This link between partial classes and real-valued learning has been exploited in previous work [BL98, Lon01, AMM+26]. The covering bounds we present build on the ε-net machinery of Subsection 2.1. Combined with the chaining result in Subsection 2.3, these bounds yield our uniform-convergence guarantee, as explained in Section 3.

2.2.1 Covering partial binary classes

Consider a class $\mathcal{F}$ of binary-valued functions on a domain of n points. We wish to approximate every function in $\mathcal{F}$, allowing disagreements on at most an ε-fraction of the domain. Haussler's covering lemma [Hau95] guarantees that this can be done using at most

$$\left(\frac{C}{\varepsilon}\right)^d$$

representatives, where $d = \mathtt{VC}(\mathcal{F})$ and C is a universal constant. Thus, the number of representatives depends only on the VC dimension and the desired accuracy, independently of n.

Haussler obtained the optimal dependence on ε, improving an earlier bound of Dudley [Dud78] by removing a $\log(1/\varepsilon)$ factor from the base. His result is usually stated as a packing bound; the covering form follows by taking a maximal separated subfamily. Chazelle [Cha92] simplified the proof, and Kupavskii and Zhivotovskiy [KZ20] later gave a short proof using the one-inclusion graph algorithm that recovers Haussler's original bound.

We extend this covering bound to functions that may be undefined on some inputs. A *partial binary class* is a family $\mathcal{F} \subseteq \{-1, +1, \star\}^{\mathcal{X}}$, where $\star$ denotes an undefined value. Its VC dimension is the largest size of a set $S \subseteq \mathcal{X}$ on which every *binary* labeling is realized by a member of $\mathcal{F}$ that is defined throughout S. When approximating a partial function, we only count disagreements at points where it is defined. More precisely, for a finite domain $\mathcal{X}$, an *ε-disambiguation* of $\mathcal{F}$ is a class of total binary functions $\mathcal{H} \subseteq \{-1, +1\}^{\mathcal{X}}$ such that, for every $f \in \mathcal{F}$, some $h \in \mathcal{H}$ satisfies

$$\big|\{x \in \mathcal{X} : f(x) \neq \star \text{ and } f(x) \neq h(x)\}\big| \leq \varepsilon|\mathcal{X}|.$$

Thus, each partial function $f \in \mathcal{F}$ can be completed to some $h \in \mathcal{H}$ after flipping its value on at most an ε-fraction of the domain. We prove the following bound.

Theorem 2.4 (Covering partial binary classes)**.** *Let $n \geq 1$. Let $\mathcal{F} \subseteq \{-1, +1, \star\}^n$ be a partial binary class of VC dimension d. For every $\varepsilon \in (0, 1/2]$, there is an ε-disambiguation $\mathcal{H} \subseteq \{-1, +1\}^n$ satisfying*

$$|\mathcal{H}| \leq (1/\varepsilon)^{cd\log(1/\varepsilon)},$$

where $c > 0$ is a universal constant.

This bound is optimal up to constant factors in the exponent: there are universal constants $c_0, \varepsilon_0 > 0$ such that, for every integer $d \geq 1$ and every $\varepsilon \in (0, \varepsilon_0]$, some partial binary class of VC dimension d has the property that every ε-disambiguation $\mathcal{H}$ satisfies

$$|\mathcal{H}| \geq (1/\varepsilon)^{c_0 d \log(1/\varepsilon)}.$$

As in Haussler's lemma, the bound is independent of n, which will be essential when we cover real-valued classes in Subsection 2.2.2.

Our bound has an additional logarithmic factor in the exponent. It would be interesting to determine whether the upper bound can be sharpened to $(C/\varepsilon)^{d \log(1/\varepsilon)}$, with a universal constant in the base, as in Haussler's bound for total classes.

Our proof constructs covers directly from learning rules derived from sample compression. These rules give a fractional cover, which the ε-net theorem for partial concept classes [BCT25] rounds into a finite one. Section 3.2.1 explains this argument.

As for the lower bound, the case of $d = 1$ follows directly from the reduction of Alon, Hanneke, Holzman, and Moran [AHHM22], combined with Pabbaraju's improved construction for the Alon–Saks–Seymour problem [Pab26]. For general d, we take the direct product of d copies of the VC-dimension-one class. Our contribution is to show that the ε-disambiguation complexity grows multiplicatively under this construction, up to constant factors in the exponent. We use a fractional relaxation of exact disambiguation, whose complexity is exactly multiplicative under direct products by linear-programming duality. We then transfer the resulting lower bound to integral ε-disambiguations with only a constant-factor loss in the exponent. The full proof is given in Section 4.

2.2.2 Covering real-valued classes

We now turn to bounded real-valued functions on a domain of n points, which we identify with a family of vectors $V \subseteq [-1, +1]^n$. As in the binary case, we seek a small collection of representatives that approximates every vector in the family. The relevant complexity measure is the fat-shattering dimension, defined in Section 1.

We measure approximation by the normalized ℓ_2 distance

$$\|v - h\|_2 = \left(\frac{1}{n} \sum_{i=1}^{n} |v_i - h_i|^2 \right)^{1/2}.$$

We similarly normalize the ℓ_1 distance by averaging over the coordinates, while the ℓ_∞ distance remains the maximum absolute difference. With this normalization, the usual inequalities are reversed:

$$\|v - h\|_1 \leq \|v - h\|_2 \leq \|v - h\|_\infty.$$

Thus, at the same radius, an ℓ_∞ cover is also an ℓ_2 cover, and an ℓ_2 cover is also an ℓ_1 cover. Table 1 uses these normalized distances.

A cover of V at radius r is a finite set $H \subseteq \mathbb{R}^n$ such that every $v \in V$ is within distance r of some $h \in H$. The representatives need not belong to V. We write $N_2(V, r)$ for the minimum size of such a cover in normalized ℓ_2 distance, and use $N_1(V, r)$ and $N_\infty(V, r)$ for the other two metrics.

Covering bounds in terms of fat-shattering dimension date back to the 1990s [ABDCBH97, BKP97, BL98]. Table 1 summarizes this development. Two issues arise: how small a covering radius can be obtained from $d = \mathtt{fat}_\gamma(V)$, and whether the number of representatives can be bounded independently of n. Classical results give such independent bounds at a radius that is a constant multiple of γ. Below radius $\gamma/2$, no bound independent of n is possible in general, even when $\mathtt{fat}_\gamma(V) = 0$. For example, take the class $(0, \gamma)^n$ of all functions from $[n]$ to $(0, \gamma)$.

More recently, Aiyer et al. [AMM+26] obtained covers at radius $\gamma/2 + \varepsilon$ in the stronger ℓ_∞ metric, which requires approximation within this radius at every coordinate. Their bound depends on n, and some dependence on n is unavoidable for ℓ_∞ covers. We obtain the same radius in normalized ℓ_2 distance with a bound independent of n.

Reference	Metric	Covering radius	Entropy bound	n-free
Alon et al. [ABDCBH97]	ℓ_∞	2γ	$\log N_\infty(V, 2\gamma) = O\left(d \log \frac{en}{\gamma} \log \frac{en}{d\gamma}\right)$	No
Bartlett, Kulkarni, and Posner [BKP97]	ℓ_1	$C\gamma$	$\log N_1(V, C\gamma) = O\left(d \log^2 \frac{e}{\gamma}\right)$	Yes
Bartlett and Long [BL98]	ℓ_1	$\gamma + \varepsilon$	$\log N_1(V, \gamma + \varepsilon) = O\left(d \log \frac{1}{\varepsilon} \log^2(en)\right)$	No
Mendelson and Vershynin [MV03]	ℓ_2	$C\gamma$	$\log N_2(V, C\gamma) = O\left(d \log \frac{e}{\gamma}\right)$	Yes
Rudelson and Vershynin [RV06]	ℓ_∞	$C\gamma/\rho$	$\log N_\infty\left(V, \frac{C\gamma}{\rho}\right) = O\left(d \log \frac{en}{d\gamma} \log^\rho \frac{en}{d}\right)$	No
Aiyer et al. [AMM+26]	ℓ_∞	$\gamma/2 + \varepsilon$	$\log N_\infty\left(V, \frac{\gamma}{2} + \varepsilon\right) = O\left(d \log \frac{1}{\varepsilon} \log^2(en)\right)$	No
This work	ℓ_2	$\gamma/2 + \varepsilon$	$\log N_2\left(V, \frac{\gamma}{2} + \varepsilon\right) = O\left(d \log^3 \frac{e}{\varepsilon}\right)$	Yes

Table 1: Comparison of covering-number bounds for bounded real-valued function classes. Throughout, $d = \mathtt{fat}_\gamma(V)$, $\varepsilon, \rho \in (0,1)$, and $C > 0$ denotes a universal constant. The family $V \subseteq [-1,+1]^n$ is fixed, and $N_p(V,r)$ denotes its covering number in the indicated normalized metric.

Theorem 2.5. *Let $n \geq 1$, let $V \subseteq [-1,+1]^n$, let $\gamma > 0$, and write $d = \mathtt{fat}_\gamma(V) < \infty$. For every $\varepsilon \in (0,1)$,*

$$\log N_2\left(V, \frac{\gamma}{2} + \varepsilon\right) \leq Cd \log^3 \frac{e}{\varepsilon},$$

where $C > 0$ is a universal constant.

The proof constructs real-valued covers by iteratively constructing covers for partial binary classes using Theorem 2.4. We overview this argument in more detail in Section 3.2.1. The full proof appears in Section 4.

It remains open whether our factor $\log^3(1/\varepsilon)$ can be reduced to $\log(1/\varepsilon)$ while retaining radius $\gamma/2 + \varepsilon$. A single logarithm suffices at radius $C\gamma$, by the classical result of Mendelson and Vershynin [MV03], which settled Talagrand's entropy problem.

The bound depends on n only through $d = \mathtt{fat}_\gamma(V)$. This is the covering bound that we combine with the chaining argument in the next subsection to prove our uniform-convergence result.

2.3 Chaining at optimal scale

We now turn to the last part of our technical contribution: a chaining bound at an optimal scale. Chaining is a standard and powerful technique used to transform covering bounds into uniform conver-

gence bounds. We first introduce Rademacher complexity, the intermediate quantity in this approach, then recall classical chaining and explain our contribution.

Rademacher complexity. Consider a family $V \subseteq [-1,+1]^n$ of vectors. Color each coordinate independently red or blue, with equal probability. For any fixed vector, the red and blue sums are equal in expectation. The vector with the largest red-minus-blue discrepancy, however, may depend on the coloring. Rademacher complexity is the expected maximum of this signed discrepancy, normalized by n. Writing $\sigma_i = +1$ for red and $\sigma_i = -1$ for blue, define

$$\mathtt{Rad}(V) = \mathop{\mathbb{E}}_{\sigma}\left[\sup_{v\in V}\frac{1}{n}\sum_{i=1}^{n}\sigma_i v_i\right].$$

For a finite family, the following lemma bounds its Rademacher complexity using the lengths of the vectors and the size of V.

Lemma 2.6 (Massart's lemma, see Lemma 26.8 in [SSBD14])**.** *Let $V \subseteq \mathbb{R}^n$ be finite and nonempty, and suppose that every $v \in V$ has normalized ℓ_2 norm at most R.*[6] *Then*

$$\mathtt{Rad}(V) \le R\sqrt{\frac{2\log|V|}{n}}.$$

From finite covers to chaining. What about infinite families V? A finite cover of radius r approximates every vector in V with error at most r. We can therefore bound $\mathtt{Rad}(V)$ by applying Massart's lemma to the cover and adding r for the approximation error. Chaining refines this argument: it uses covers at successively smaller radii and applies Massart's lemma to the differences between successive approximations. Let H_i be covers of V at decreasing radii r_i, starting with $H_0 = \{0\}$ and $r_0 = 1$. For each $v \in V$, choose $v_i \in H_i$ with $\|v - v_i\|_2 \le r_i$. We write

$$v = (v - v_k) + \sum_{i=1}^{k}(v_i - v_{i-1}).$$

At each level, the possible differences $v_i - v_{i-1}$ form a finite family, with at most $|H_i||H_{i-1}|$ members. Massart's lemma bounds their random imbalance using their lengths and the logarithm of their number. Summing these bounds over the levels, and adding the residual error, gives a chaining bound. A classical form is the following.

Theorem 2.7 (Dudley's chaining bound [SST10])**.** *For every $n \ge 1$, every $V \subseteq [-1,+1]^n$, and every $\alpha \in (0,1)$,*

$$\mathtt{Rad}(V) \le 4\alpha + \frac{10}{\sqrt{n}}\int_0^1 \sqrt{\log N_2(V,\alpha+t)}\,dt.$$

The cutoff α allows us to use covers only at radii above α, at the cost of the leading term 4α. The coefficient in front of the leading term cannot be smaller than 1. Indeed, consider the family $V = \{-\alpha,+\alpha\}^n$. The zero function covers this class at radius α, so the entropy integral vanishes. Yet $\mathtt{Rad}(V) = \alpha$: for each coloring σ, the class contains the vector $\alpha\sigma$. This leaves the question: *can a chaining bound attain the leading term α, using only covers at radii above α?*

This is critical in our application. Here, the covering bounds hold at every radius above $\gamma/2$, so we take $\alpha = \gamma/2$. The symmetrization step that converts Rademacher complexity into uniform convergence introduces another factor of two. The classical bound would therefore give a leading error of 4γ, whereas we need γ. Increasing the sample size does not remove this loss: we need to reduce the coefficient of α from 4 to 1.

[6] Recall that the normalized Euclidean norm is $\|v\|_2 = (n^{-1}\sum_{i=1}^n v_i^2)^{1/2}$.

Aligning the covers. To see the difficulty, recall the successive representatives v_i chosen from covers of radii r_i. When these covers are chosen independently, the triangle inequality gives

$$\|v_i - v_{i-1}\|_2 \le r_i + r_{i-1}.$$

As the radii decrease to $\alpha > 0$, this estimate approaches 2α. Thus, increasingly accurate approximations need not give increasingly small bounds on the successive differences. This is where controlling only the sizes of the covers falls short.

We instead choose the covers jointly, allowing them to be larger in order to align their representatives. Here, alignment means that each vector has a sequence of representatives that retain the prescribed approximation radii, while their successive differences become small as the radii approach α. These smaller differences let us apply Massart's lemma without paying an additional constant multiple of α. We explain the construction, and how we control the increase in cover size, in the proof overview in Section 3.2.2.

Theorem 2.8 (Chaining at optimal scale)**.** *For every $n \ge 1$, every $V \subseteq [-1,+1]^n$, and every $\alpha \in (0,1)$,*

$$\mathtt{Rad}(V) \le \alpha + \frac{32}{\sqrt{n}} \int_0^1 \sqrt{\log N_2(V, \alpha + t^2)}\, dt.$$

The leading term has the optimal coefficient 1. The entropy integral uses radii $\alpha + t^2$ rather than $\alpha + t$; our covering bounds keep it finite. The complete proof appears in Section 4.2.

3 Proof Overview

This section gives an overview of the proof of our uniform-convergence bound, focusing on the main ideas behind the argument. We begin with the lower-bound construction and then turn to the upper bound, where the covering and chaining tools are combined. For the geometric margin-net results in Theorems 2.1 and 2.3, we refer the reader directly to their complete proofs in Section 4.1.

For a class $\mathcal{F} \subseteq [-1,+1]^{\mathcal{X}}$ and a sample $S = (x_1, \dots, x_n)$ from a distribution $\mathcal{D}$, write

$$\mathtt{Rep}_{\mathcal{D}}(\mathcal{F}, S) = \sup_{f \in \mathcal{F}} \left| \mathop{\mathbb{E}}_{x \sim \mathcal{D}} f(x) - \frac{1}{n} \sum_{i=1}^{n} f(x_i) \right|.$$

Thus, $m_\gamma(\varepsilon, \delta)$ is the least sample size m such that, for every $n \ge m$ and every $\mathcal{D}$, this error is at most $\gamma + \varepsilon$ with probability at least $1 - \delta$. We set it to infinity if no such m exists. The class satisfies γ-uniform convergence when this sample complexity is finite for every $\varepsilon > 0$ and $\delta \in (0,1)$.

3.1 Lower bound

The construction separates the two sources of error. We start with a binary class $\mathcal{H}$ of VC dimension d for which uniform convergence requires $\Omega((d + \log(1/\delta))/\varepsilon^2)$ samples: with fewer samples, some distribution gives an estimation error greater than ε with probability greater than δ. We choose $\mathcal{H}$ to be symmetric, meaning that $-h \in \mathcal{H}$ whenever $h \in \mathcal{H}$. Add a coordinate $t \in [0,1]$ and consider the functions

$$f_{h,b}(t,x) = \left(1 - \frac{\gamma}{2}\right) h(x) + \frac{\gamma}{2} b(t), \qquad h \in \mathcal{H},$$

where b ranges over all (measurable) functions from $[0,1]$ to $(-1,+1)$.

The added term can vary by strictly less than γ, so it cannot itself realize the two labels in a fat-shattering pattern at scale γ. The resulting class still has fat-shattering dimension d at this scale. Nevertheless, this term contributes an error of γ: under the uniform distribution on $[0,1]$, a function b can be arbitrarily close to -1 on all sampled points and arbitrarily close to $+1$ everywhere else.

Taking the product with a hard distribution for $\mathcal{H}$ makes these two effects independent. Since h and b can be chosen separately, their worst-case errors add. Achieving error $\gamma + \varepsilon$ therefore requires estimating the binary class to accuracy of order ε. The classical binary lower bound supplies the

required dependence on d, ε, and δ. The complete construction and its analysis appear in the proof of Proposition 4.15 in Section 4.5.

3.2 Upper bound

We first show how the covering and chaining bounds from Section 2 imply the upper bound. We then explain the construction of the covers in Section 3.2.1 and their alignment for chaining in Section 3.2.2.

Let $\mathcal{F} \subseteq [-1,+1]^{\mathcal{X}}$ and write $d = \mathtt{fat}_\gamma(\mathcal{F})$. For a sample $S = (x_1, \dots, x_n)$, consider the family of vectors

$$V_S = \{(f(x_1), \dots, f(x_n)) : f \in \mathcal{F}\} \subseteq [-1,+1]^n.$$

Thus, $\mathtt{fat}_\gamma(V_S) \le d$ and the results of Section 2 apply to every sample S. The covering bound in Theorem 2.5 gives

$$\log N_2(V_S, \gamma/2 + \tau) \le C d \log^3(e/\tau),$$

and chaining at optimal scale (Theorem 2.8), with $\alpha = \gamma/2$, yields

$$\begin{aligned}\mathtt{Rad}(V_S) &\le \frac{\gamma}{2} + \frac{32}{\sqrt{n}} \int_0^1 \sqrt{\log N_2(V_S, \gamma/2 + t^2)}\, dt \\ &\le \frac{\gamma}{2} + C' \sqrt{\frac{d}{n}} \int_0^1 \log^{3/2}(e/t)\, dt = \frac{\gamma}{2} + O\left(\sqrt{\frac{d}{n}}\right).\end{aligned}$$

Since this bound holds for every sample, classical symmetrization [SSBD14, Lemma 26.2] and McDiarmid's concentration inequality [McD89] give, with probability at least $1 - \delta$,

$$\mathtt{Rep}_{\mathcal{D}}(\mathcal{F}, S) \le \gamma + O\left(\sqrt{\frac{d + \log(2/\delta)}{n}}\right).$$

Taking n of order $(d + \log(2/\delta))/\varepsilon^2$ makes this error at most $\gamma + \varepsilon$. The formal argument appears in the proof of Theorem 4.14 in Section 4.5. We now explain the two ingredients in this calculation: constructing covers arbitrarily close to radius $\gamma/2$, and chaining these covers without enlarging the leading error.

3.2.1 Constructing covers

Partial binary classes. We construct approximate disambiguations using the fractional-cover rounding approach from Section 2.1. Let

$$\mathcal{F} \subseteq \{-1, +1, \star\}^{\mathcal{X}}$$

be a partial concept class of VC dimension d. We first construct a distribution over total binary functions under which every $f \in \mathcal{F}$ has a nonnegligible probability of being approximated, and then round this fractional cover into a finite one.

A natural starting point is the learnability of $\mathcal{F}$. Given $f \in \mathcal{F}$, sample sufficiently many points from $\mathcal{X}$ uniformly and apply a learner to the resulting f-labeled sample. With constant probability, the output approximates f on all but an ε-fraction of $\mathcal{X}$. To obtain a distribution that does not depend on f, one might instead sample the points, assign each of them an independent uniformly random label from $\{-1, +1, \star\}$, and apply the learner to the resulting labeled sample. For any fixed f, however, the probability of guessing all labels correctly is 3^{-m}, where m is the number of sampled points. This is too small for the subsequent rounding argument.

The loss comes from guessing the entire labeled sample, although the learner can be reconstructed from much less information. We therefore construct a compression scheme of size

$$O\left(d \log \frac{e}{\varepsilon}\right).$$

Rather than guessing all m labels, we guess only a compression record and the labels of the examples it contains, and then apply the reconstruction map. Controlling the effective number of possible compression records shows that the resulting distribution assigns probability at least

$$\left(\frac{c}{\varepsilon}\right)^{-Cd\log(e/\varepsilon)}$$

to functions that ε-approximate any fixed $f \in \mathcal{F}$. This gives the required fractional disambiguation.

Finally, we round the fractional cover. For each $f \in \mathcal{F}$, let $B_\alpha(f)$ be the set of total binary functions that disagree with f on at most an α-fraction of $\mathcal{X}$, counting only coordinates where f is defined. The $\star$ entries impose no constraints, so these sets are not ordinary Hamming balls and Theorem 2.3 does not apply directly. Nevertheless, the same rounding strategy applies, using the ε-net theorem for partial concept classes of Bourneuf, Charbit, and Thomassé [BCT25].

Increasing the allowed error from α to $\alpha + \eta$ supplies the margin: the associated partial class distinguishes error at most α from error at least $\alpha + \eta$, leaving intermediate errors undefined. This class has VC dimension $O(1/\eta^2)$, so the ε-net theorem for partial concept classes gives a finite cover with error at most $\alpha + \eta$. Taking $\alpha = \eta = \varepsilon/2$ and substituting the fractional-cover bound yields the desired result.

Covering real-valued classes. We now overview the construction of the covers in Theorem 2.5. For $V \subseteq [-1,+1]^n$ with $d = \mathtt{fat}_\gamma(V)$, we seek a cover at radius $\gamma/2 + \varepsilon$ whose size is independent of n. We first recall the standard approach of obtaining covers from packing bounds [Hau95, Dud78], and explain why it does not preserve the optimal scale $\gamma/2$. Start with a packing of M vectors whose pairwise normalized ℓ_2 distances are at least ρ. The idea resembles the Johnson–Lindenstrauss lemma: reduce the dimension while preserving this packing, with a small loss in separation. Here, the reduction simply restricts the vectors to randomly sampled coordinates. For two vectors at normalized ℓ_2 distance at least ρ, sampling $O(1/\rho^2)$ coordinates preserves a constant fraction of their distance with constant probability. Increasing the sample to $O(\rho^{-2}\log M)$ allows concentration and a union bound to preserve the separation of all pairs simultaneously.

We thus retain all M vectors in a packing on only logarithmically many coordinates. Provided the retained separation is sufficiently large relative to γ, known bounds in terms of the fat-shattering dimension limit this projected packing to $M^{o(1)}$ vectors. Since the projected packing still contains M vectors, we obtain $M \le M^{o(1)}$. This inequality fails for sufficiently large M, giving an upper bound on M independent of n.

The difficulty is preserving the radius. A maximal packing by radius-r balls gives a cover at radius $2r$. This route therefore incurs a factor-two loss, whereas we need a radius arbitrarily close to $\gamma/2$. We instead construct the covers directly, adapting the iterative approach of Aiyer et al. [AMM+26].

We now describe one iteration of the cover construction. Suppose the current cover approximates every $v \in V$ within $\gamma/2 + s$ on all but a small fraction of coordinates. We replace each representative g by a small collection that reduces the excess error from s to $s/2$. Different vectors represented by g may require different corrections; our partial-class bound supplies a small list of correction patterns that serves them all. In more detail, for each vector v approximated by g, we record which coordinates of g should be increased or decreased:

$$p_v(i) = \begin{cases} +1, & v_i - g_i \ge \gamma/2, \\ -1, & v_i - g_i \le -\gamma/2, \\ \star, & \text{otherwise.} \end{cases}$$

The undefined coordinates already have error less than $\gamma/2$. For fixed g, these partial vectors form a class of VC dimension at most d: shattering coordinates would fat-shatter them in V. Theorem 2.4 therefore gives a small approximate disambiguation H_g. Replace g by

$$g + \frac{s}{2}h, \qquad h \in H_g,$$

clipping the resulting representatives to $[-1,+1]$.

For each v, choose an h that approximately disambiguates p_v. Where the previous approximation was accurate and h respects the defined label, the correction reduces the error to at most $\gamma/2 + s/2$. On undefined coordinates, either direction satisfies this bound. Thus the new representatives halve the excess error, apart from the coordinates where the previous approximation or the disambiguation was inaccurate.

Starting from $\{0\}$, we repeat this step $O(\log(1/\varepsilon))$ times. We choose the disambiguation errors small enough that the total fraction of inaccurate coordinates remains small. Since all values lie in $[-1, +1]$, these coordinates contribute little to normalized ℓ_2 distance, giving radius $\gamma/2 + \varepsilon$. Each round adds $O(d \log^2(1/\varepsilon))$ to the logarithm of the cover size; the number of rounds supplies the third logarithmic factor:

$$\log N_2(V, \gamma/2 + \varepsilon) = O\big(d \log^3(1/\varepsilon)\big) .$$

3.2.2 Chaining at optimal scale

We now turn to the proof of the chaining theorem (Theorem 2.8). Recall that our goal is to bound the Rademacher complexity using covers at radii approaching α from above, while retaining the optimal leading term α. We follow the chaining argument recalled in Section 2.3: write each vector as its final approximation error plus the successive differences between its representatives, and apply Massart's lemma to those differences. Our modification concerns the choice of representatives. We align them so that successive differences become small even when the approximation radii approach a positive limit α. To see why this is needed, consider covers H_i of $V \subseteq [-1, +1]^n$ at radii r_i decreasing to α. If $h_i(v)$ and $h_{i-1}(v)$ are successive representatives of the same vector v, the triangle inequality gives

$$\|h_i(v) - h_{i-1}(v)\|_2 \le r_i + r_{i-1}.$$

When the radii tend to α, this bound tends to 2α. The following geometric observation allows us to align the representatives while preserving their approximation radii. Suppose that y is the representative already chosen for v at radius R, and x is a representative from the next cover, whose radius is $r < R$. Then

$$v \in B_R(y) \cap B_r(x).$$

The next representative only needs to cover the vectors in this intersection; it need not cover the entire ball $B_r(x)$. This gives us room to move its center toward y while keeping the same radius r.

The following Euclidean fact makes the improvement quantitative (Lemma 4.2): the intersection $B_R(y) \cap B_r(x)$ is contained in a ball $B_r(z)$ whose center satisfies

$$\|z - y\|_2 \le \sqrt{R^2 - r^2}.$$

Unlike $R + r$, this quantity tends to zero when the two radii approach the same positive value; see Figure 1.

Applying this operation successively produces representatives $\pi_i(v)$ that remain within distance r_i of v, while satisfying

$$\|\pi_i(v) - \pi_{i-1}(v)\|_2 \le \sqrt{r_{i-1}^2 - r_i^2}.$$

Thus, although the approximation radii converge only to α, the successive increments converge to zero.

Alignment increases the cover sizes because a center may move to different positions depending on the preceding choices. Choosing $r_i = \alpha + 4^{-i}$ makes the successive differences shrink geometrically, with norms $O(2^{-i})$; the resulting increase in the Massart bounds changes only the universal constant in the entropy term, while the leading approximation error remains α. Stopping after N scales leaves approximation error at most r_N. Applying Massart's lemma to the successive differences and summing their contributions gives

$$\mathtt{Rad}(V) \le r_N + \frac{C}{\sqrt{n}} \sum_{i=1}^{N} 2^{-i} \sqrt{\log |H_i|},$$

where H_i is the original cover at radius r_i, before alignment. Choosing each H_i to have minimum size, bounding the sum by an integral, and letting N tend to infinity gives the desired entropy bound.

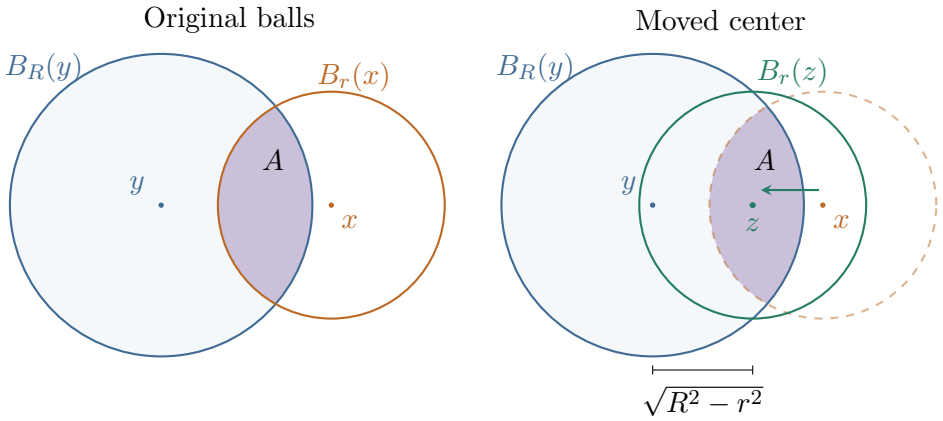


Figure 1: Moving the center while preserving the approximation. The shaded intersection $A = B_R(y) \cap B_r(x)$ remains covered when we replace $B_r(x)$ by $B_r(z)$. The radius stays r, while the new center satisfies $\|z - y\|_2 \le \sqrt{R^2 - r^2}$.

4 Proofs

This section contains the complete proofs of Our results. In section 4.1 we prove the ε-nets results such as Theorems 2.3, and 2.1. In section 4.2 We prove theorem the chaining bound of Theorem 2.8. In section 4.3 We prove results on partial concept classes to derive the covering bound of Theorem 2.4. We apply this result in section 4.4 to derive the covering bounds of Theorem 2.5. Finally in section 4.5 we use the chaining and covering results to prove our main theorem.

Throughout this section, the constants $C, C', c > 0$ are universal and may change from line to line.

4.1 ε-Nets and Fractional Covers with Margin Proofs

The following is Theorem 18 of [BCT25], we repeat the proof here for completeness.

Lemma 4.1 (δ-net for partial concept classes [BCT25])**.** *Let $\mathcal{F} \subseteq \{-1, +1, \star\}^{\mathcal{X}}$ be a partial concept class with $\mathtt{VC}(\mathcal{F}) = d < \infty$. Let $\delta \in (0, 1)$, and let $\mathcal{D}$ be a distribution over $\mathcal{X}$ such that, for every $f \in \mathcal{F}$,*

$$\mathbb{P}_{x \sim \mathcal{D}}\big(f(x) = 1\big) \ge \delta.$$

Then there exists a set $X \subseteq \mathcal{X}$ of size

$$|X| \le C \frac{d}{\delta} \log \frac{e}{\delta} + 1.$$

such that, for every $f \in \mathcal{F}$, there is some $x \in X$ for which $f(x) \neq -1$; that is, $f(x) \in \{1, \star\}$.

Proof. The proof is similar to the one in Theorem 18 of [BCT25], and is given here for completeness. If $d = 0$, the result is clear so we may assume $d > 0$. Assume toward contradiction no such set exists. Let n be chosen later, draw $x_1, \dots, x_n$ independently from $\mathcal{D}$, and independently assign each x_i a uniformly random label $y_i \in \{-1, +1\}$. Let

$$S = \{(x_i, y_i)\}_{i=1}^n,$$

and let A be the event that S is realizable by $\mathcal{F}$.

For every fixed choice of $x_1, \dots, x_n$, Sauer's lemma implies that $\mathcal{F}$ realizes at most

$$\sum_{j=0}^{d} \binom{n}{j} \le \left(\frac{en}{d}\right)^d$$

labelings of these points. Since the labels are uniform,

$$\mathbb{P}(A) \le 2^{-n} \left(\frac{en}{d}\right)^d.$$

We now generate the same random sample in a different order. First choose a random subset $I \subseteq [n]$ by including each index independently with probability $1/2$. The points indexed by I receive label -1, while the remaining points receive label 1. Write

$$X = \{x_i : i \in I\}, \qquad K = |I|.$$

By assumption there exists some $f \in \mathcal{F}$ such that

$$f(x) = -1 \qquad \text{for every } x \in X.$$

Fix such I, f, and note that if $f(x_i) = 1$ for every $i \neq I$, then S is realizable by f so A occurs. For each such point $f(x_i) = 1$ with probability at least δ, hence

$$\mathbb{P}(A \mid X) \geq \delta^{n-K}.$$

Taking expectations

$$\mathbb{P}(A) \geq \mathbb{E}\left[\delta^{n-K}\right] = \frac{1}{2^n} \sum_{k=0}^{n} \binom{n}{k} \delta^{n-k} = \frac{(1+\delta)^n}{2^n}.$$

So we have

$$(1+\delta)^n \leq 2^n \, \mathbb{P}(A) \leq \left(\frac{en}{d}\right)^d.$$

Solving for n in the above, we get that for

$$n = C \frac{d}{\delta} \log \frac{e}{\delta},$$

the above will not hold, from which we deduce the desired result. ■

Theorem 2.3 (Restated)**.** *Let μ be a probability distribution on $\{0,1\}^n$. Fix a radius α and an additive margin η satisfying $0 < \eta \leq \alpha$ and $\alpha + \eta \leq 1$. For every $\varepsilon \in (0,1)$, there is a set $S \subseteq \{0,1\}^n$ of size*

$$|S| = O\left(\frac{\alpha^2}{\varepsilon \eta^2} \log \frac{e}{\varepsilon}\right)$$

such that every closed Hamming ball of radius α and μ-mass at least ε intersects S after its radius is increased to $\alpha + \eta$.

The bound is optimal up to the logarithmic factor for small radii. For every sufficiently small $\varepsilon, \alpha > 0$ and every $0 < \eta \leq \alpha$, there are a dimension n and a distribution μ for which every such net has $\Omega(\alpha^2/(\varepsilon\eta^2))$ points.

Proof of Theorem 2.3. Set $\mathcal{X} = \{0,1\}^n$, and consider the partial concept class $\mathcal{H} = \{h_x : x \in \mathcal{X}\} \subseteq \{-1, +1, \star\}^{\mathcal{X}}$ given by

$$h_x(y) = \begin{cases} +1, & d_H(x,y) \leq \alpha, \\ \star, & \alpha < d_H(x,y) \leq \alpha + \eta, \\ -1, & d_H(x,y) > \alpha + \eta, \end{cases} \qquad x, y \in \mathcal{X}.$$

We first show that

$$\mathtt{VC}(\mathcal{H}) = O\left(\frac{\alpha^2}{\eta^2}\right).$$

It is enough to bound shattered sets whose cardinality is a power of two, since every set of size $m \geq 2$ contains a subset of power-of-two size greater than $m/2$. Suppose, therefore, that $u_1, \ldots, u_d$ are shattered by $\mathcal{H}$, where d is a power of two, and let A be a Hadamard matrix of order d.

Choose $x, y \in \{-1, +1\}^d$ maximizing $\langle x, Ay \rangle$, and set

$$a_{i,j} = A_{i,j} x_i y_j.$$

By shattering, for each $j \in [d]$ there is some $v_j \in \{-1,+1\}^n$ such that

$$d_H(u_i, v_j) \begin{cases} \leq \alpha, & a_{i,j} = +1, \\ \geq \alpha + \eta, & a_{i,j} = -1. \end{cases}$$

For each $i \in [d]$, define u_i' by setting $u_i'(k)$ to be 1 if $u_i(k) = 1$, and -1 otherwise. Define v_j' in the same way. Since

$$d_H(u_i, v_j) = \frac{1}{2} - \frac{1}{2n}\langle u_i', v_j' \rangle,$$

we have

$$2\sum_{i,j} a_{i,j} d_H(u_i, v_j) = \frac{1}{n}\sum_{k=1}^{n}\left(\sum_{i,j} a_{i,j} - \sum_{i,j} a_{i,j} u_i'(k) v_j'(k)\right) \geq 0.$$

Indeed, for each k, replacing x_i by $x_i u_i'(k)$ and y_j by $y_j v_j'(k)$ produces another pair of sign vectors, so the inequality follows from the maximality of x and y.

On the other hand, since A is a Hadamard matrix,

$$\sum_{i,j} a_{i,j} = \langle x, Ay \rangle \leq \|x\|_2 \|Ay\|_2 = d\sqrt{d}.$$

Let P and N denote the numbers of pairs (i,j) for which $a_{i,j} = +1$ and $a_{i,j} = -1$, respectively. Then

$$P + N = d^2, \qquad P - N \leq d\sqrt{d},$$

while the choice of the v_j gives

$$0 \leq \sum_{i,j} a_{i,j} d_H(u_i, v_j) \leq P\alpha - N(\alpha + \eta).$$

Consequently,

$$1 + \frac{\eta}{\alpha} \leq \frac{P}{N} \leq \frac{d^2 + d\sqrt{d}}{d^2 - d\sqrt{d}} = 1 + \frac{2}{\sqrt{d} - 1}.$$

It follows that

$$d \leq \left(1 + \frac{2\alpha}{\eta}\right)^2 \leq 9\frac{\alpha^2}{\eta^2},$$

where the last inequality uses $\eta \leq \alpha$. Since every shattered set contains a power-of-two subset of at least half its size, we conclude that

$$\mathtt{VC}(\mathcal{H}) = O\left(\frac{\alpha^2}{\eta^2}\right).$$

Restrict $\mathcal{H}$ to those h_x for which

$$\mu\big(h_x^{-1}(1)\big) = \mu\big(B_\alpha(x)\big) \geq \varepsilon.$$

Lemma 4.1 now gives a set $S \subseteq \mathcal{X}$ of size

$$|S| = O\left(\frac{\alpha^2}{\eta^2 \varepsilon} \log \frac{e}{\varepsilon}\right)$$

that intersects

$$h_x^{-1}(\{1, \star\}) = B_{\alpha+\eta}(x)$$

for every such x. This proves the upper bound.

For the lower bound, we use the matrix form of Spencer's discrepancy theorem [Spe85]: there is a universal constant C_0 such that, for any collection of at most m vectors in $[-1,+1]^m$, some $z \in \{-1,+1\}^m$ has inner product of absolute value at most $2C_0\sqrt{m}$ with every vector in the collection.

Let $\psi_t : \{0,1\}^m \to \{0,1\}^{tm}$ repeat each coordinate t times. The discrepancy bound implies that, for any set $T \subseteq \{0,1\}^{tm}$ with $|T| \le m$, there is an $x \in \{0,1\}^m$ such that

$$D_H(\psi_t(x), s) \ge t\left(\frac{m}{2} - C_0\sqrt{m}\right) \qquad \text{for every } s \in T,$$

where D_H denotes unnormalized Hamming distance. Indeed, for each $s \in T$, average its signs $2s_j - 1$ within each block of t coordinates to obtain a vector in $[-1,+1]^m$, and apply Spencer's theorem to these vectors. Thus repeating coordinates preserves the lower bound even when the net points do not themselves have repeated coordinates.

We prove the claim for every $0 < \alpha \le 1/32$, every $0 < \eta \le \alpha$, and all sufficiently small $\varepsilon > 0$. Write $q = \alpha/\eta$. First suppose $q \ge q_0$, where q_0 is a sufficiently large universal constant, and set

$$m = \lfloor q^2/16 \rfloor, \qquad r = \left\lfloor \frac{m}{2} - (C_0 + 2)\sqrt{m} \right\rfloor.$$

We choose q_0 large enough that $r \ge m/4$. By the central limit theorem, there is a universal constant $p_0 > 0$ such that

$$2^{-m} \sum_{j=0}^{r} \binom{m}{j} \ge p_0.$$

Let $k = \lfloor p_0/\varepsilon \rfloor$. We place k copies of the repeated cube on the same active coordinates and separate them using additional code coordinates. Repetition lets us provide enough code coordinates while keeping α and η at their prescribed values.

Choose a sufficiently large positive integer t and put

$$N = \left\lceil \frac{tr}{\alpha} \right\rceil, \qquad L = N - tm.$$

Since $r \ge m/4$ and $\alpha \le 1/32$, we have $L \ge 7N/8$. By taking t large enough, a standard random-coding argument gives k words $c_1, \dots, c_k \in \{0,1\}^L$ whose pairwise distances are at least $L/4$. Define

$$\mathcal{F}_i = \{(\psi_t(x), c_i) : x \in \{0,1\}^m\}, \qquad \mathcal{F} = \bigcup_{i=1}^{k} \mathcal{F}_i,$$

and let μ be uniform on $\mathcal{F}$. For every $u \in \mathcal{F}_i$, the ball of normalized radius α around u contains all points in that copy whose underlying m-bit words differ from u in at most r coordinates, since $tr \le \alpha N$. Hence

$$\mu(B_\alpha(u)) \ge \frac{p_0}{k} \ge \varepsilon.$$

In unnormalized distance, the enlarged radius is $R = (\alpha + \eta)N$. The rounding in the definition of N gives

$$\frac{R}{t} < \left(1 + \frac{1}{q}\right)\left(r + \frac{\alpha}{t}\right) < \frac{m}{2} - C_0\sqrt{m},$$

where the last inequality uses $\sqrt{m}/q \le 1/4$ and $\alpha/t \le 1/32$. Also, $R \le N/16$, so distinct copies have distance at least $L/4 > 2R$.

Let S be a net with the required property, and let S_i consist of its points at distance at most R from $\mathcal{F}_i$. The sets S_i are pairwise disjoint, and each must cover $\mathcal{F}_i$ by radius-R balls. If $|S_i| \le m$, applying the repeated-coordinate discrepancy bound to the first tm coordinates of these points produces a point of $\mathcal{F}_i$ farther than R from all of them, a contradiction. Therefore

$$|S| \ge km = \Omega\left(\frac{q^2}{\varepsilon}\right) = \Omega\left(\frac{\alpha^2}{\varepsilon\eta^2}\right).$$

Finally, if $1 \le q < q_0$, take $k = \lfloor 1/\varepsilon \rfloor$ points in a sufficiently large binary cube with pairwise normalized distances at least $1/4$, and put the uniform distribution on them. Each radius-α ball centered at one of these points has mass at least $1/k \ge \varepsilon$. Their enlargements are disjoint because

$2(\alpha+\eta) \le 4\alpha \le 1/8$. Every net therefore needs at least $k = \Omega(1/\varepsilon) = \Omega(q^2/\varepsilon)$ points, since q_0 is a universal constant.

We conclude by explaining the extension to ℓ_p. The only property of the Hamming metric used in the upper bound is

$$\sum_{i,j} a_{i,j} d_H(u_i, v_j) \ge 0.$$

The analogous inequality holds for arbitrary points in L_1. Indeed, every L_1 metric is an integral of cut metrics:

$$\begin{aligned}\sum_{i,j} a_{i,j} \|u_i - v_j\|_1 &= \int_\Omega \int_{\mathbb{R}} \sum_{i,j} a_{i,j} \left| \mathbf{1}_{\{u_i(s) \ge t\}} - \mathbf{1}_{\{v_j(s) \ge t\}} \right| dt \, d\mu(s) \\ &\ge 0.\end{aligned}$$

For each s and t, the inner sum is nonnegative by the same maximality argument used above for binary vectors. The remainder of the proof is therefore unchanged. For $1 < p \le 2$, we use [BDCK66] to embed ℓ_p isometrically, up to a common scaling factor, into L_1. Since this scaling affects the radius and margin equally, the same upper bound follows for every $p \in [1, 2]$.

The lower-bound construction can likewise be viewed inside ℓ_p using the normalized embedding of the Hamming cube, under which

$$\|x - y\|_p = d_H(x, y)^{1/p}.$$

An ℓ_p radius α and additive margin η then correspond to Hamming radius α^p and Hamming margin $(\alpha+\eta)^p - \alpha^p$. Since

$$\frac{\alpha^p}{(\alpha+\eta)^p - \alpha^p} \asymp \frac{\alpha}{\eta} \qquad \text{for } 0 < \eta \le \alpha \text{ and } 1 \le p \le 2,$$

the construction gives the same lower bound

$$\Omega\left(\frac{\alpha^2}{\varepsilon\eta^2}\right)$$

for strong ε-nets in ℓ_p. This argument does not extend to weak nets, whose points may lie outside the embedded Hamming cube. ■

Theorem 2.1 (Restated)**.** *Let A be a set of n points in the unit ball of a real Hilbert space E, and let $\varepsilon, \eta \in (0,1)$. There is a set $S \subseteq E$ of size $O(1/(\varepsilon\eta^2))$ such that every closed halfspace containing at least εn points of A has distance at most η from some point of S.*

The bound is optimal: for all sufficiently small $\varepsilon, \eta > 0$, there is a finite set A in the unit ball of ℓ_2 for which every such net has $\Omega(1/(\varepsilon\eta^2))$ points.

Proof of Theorem 2.1. For the upper bound, let μ be the uniform distribution on A, and let $X \sim \mu$. For a unit vector $u \in E$ and $b \in \mathbb{R}$, write

$$H(u, b) = \{x \in E : \langle u, x \rangle \ge b\}.$$

Consider the second-moment operator $T = \mathbb{E}[X \otimes X]$. This operator has finite rank, since X takes only finitely many values, so the spectral argument below applies even when E is infinite-dimensional. Since A lies in the unit ball,

$$\operatorname{tr}(T) = \mathbb{E}\|X\|^2 \le 1.$$

Let V be the span of the eigenvectors of T whose eigenvalues are larger than $\varepsilon\eta^2/2$, and let P denote the orthogonal projection onto V. Then

$$r := \dim(V) \le \frac{2}{\varepsilon\eta^2}.$$

For every unit vector $u \in E$,

$$\mathbb{E}\left[\langle P^\perp u, X \rangle^2\right] = \langle T P^\perp u, P^\perp u \rangle \le \frac{\varepsilon\eta^2}{2}.$$

Consequently,

$$\mathbb{P}\left(\langle P^{\perp}u, X\rangle > \eta\right) \le \frac{\varepsilon}{2}.$$

If $\mu(H(u,b)) \ge \varepsilon$, there is therefore some $x \in A$ with

$$\langle u, x\rangle \ge b \qquad \text{and} \qquad \langle P^{\perp}u, x\rangle \le \eta.$$

It follows that

$$\langle u, Px\rangle = \langle u, x\rangle - \langle P^{\perp}u, x\rangle \ge b - \eta.$$

Since $\|Px\| \le 1$, the enlarged halfspace $H(u, b-\eta)$ intersects the unit ball of V.

Choose a simplex $\Delta = \operatorname{conv}\{v_0, \dots, v_r\} \subseteq V$ containing the unit ball of V, and set $S = \{v_0, \dots, v_r\}$. A linear functional attains its maximum over a simplex at a vertex. Thus, whenever $H(u, b-\eta)$ intersects the unit ball of V, some v_i satisfies $\langle u, v_i\rangle \ge b - \eta$. As u is a unit vector, the distance from v_i to $H(u,b)$ is at most η. Hence the closed ball of radius η centered at v_i intersects $H(u,b)$, as required. Finally,

$$|S| = r + 1 \le \frac{2}{\varepsilon\eta^2} + 1.$$

The same argument applies to any probability distribution μ supported on the unit ball of a finite-dimensional Euclidean space, with the condition $|A \cap H| \ge \varepsilon n$ replaced by $\mu(H) \ge \varepsilon$.

For the lower bound, set $p = 1/16$ and fix $\varepsilon \in (0, p/2)$ and $\eta \in (0, 1/32)$. Let

$$m = \left\lfloor \frac{1}{256\eta^2} \right\rfloor, \qquad k = \left\lfloor \frac{p}{\varepsilon} \right\rfloor.$$

Then $m = \Theta(\eta^{-2})$ and $\eta \le 1/(16\sqrt{m})$. Let $E = E_1 \oplus \cdots \oplus E_k$ be an orthogonal direct sum of m-dimensional spaces. Choose an orthonormal basis $e_{i1}, \dots, e_{im}$ of each E_i, and take the scaled cube vertices

$$A_i = \left\{ \frac{1}{\sqrt{m}} \sum_{j=1}^{m} \sigma_j e_{ij} : \sigma \in \{-1, +1\}^m \right\}.$$

The set $A = \bigcup_{i=1}^{k} A_i$ consists of $n = k2^m$ points on the unit sphere. Let μ_i be uniform on A_i, so that $\mu = k^{-1}\sum_i \mu_i$ is uniform on A.

We first show that, for every unit vector $z \in E_i$,

$$\mu_i\left(\left\{x : \langle z, x\rangle \ge \frac{1}{4\sqrt{m}}\right\}\right) \ge p.$$

Indeed, if $X_i \sim \mu_i$ and $Z = \sqrt{m}\,\langle z, X_i\rangle$, then Z is symmetric and

$$\mathbb{E}Z^2 = 1, \qquad \mathbb{E}Z^4 = 3 - 2\sum_{j=1}^{m}\langle z, e_{ij}\rangle^4 \le 3.$$

Cauchy–Schwarz gives

$$1 = \mathbb{E}Z^2 \le \frac{1}{16} + \sqrt{3\,\mathbb{P}(|Z| \ge 1/4)}.$$

By symmetry,

$$\mathbb{P}(Z \ge 1/4) = \frac{1}{2}\,\mathbb{P}(|Z| \ge 1/4) \ge \frac{(1 - 1/16)^2}{6} > p.$$

Now let $S \subseteq E$ satisfy the required intersection property, and put $W = \operatorname{span}(S)$. Write P_W for the orthogonal projection onto W. We claim that, for every $i \in [k]$,

$$\sum_{j=1}^{m} \|P_W e_{ij}\|^2 \ge \frac{m}{2}.$$

Otherwise, some unit basis vector $e \in E_i$ satisfies $\|P_W e\|^2 < 1/2$. Define

$$v = \frac{P_{W^\perp} e}{\|P_{W^\perp} e\|}.$$

Then v is a unit vector orthogonal to W, and

$$\|P_{E_i} v\| \geq \langle v, e \rangle = \|P_{W^\perp} e\| > \frac{1}{\sqrt{2}}.$$

Apply the preceding estimate to $z = P_{E_i} v / \|P_{E_i} v\|$. For $x \in A_i$ with $\langle z, x \rangle \geq 1/(4\sqrt{m})$, we have

$$\langle v, x \rangle > \frac{1}{4\sqrt{2m}} > 2\eta.$$

Thus $\mu_i(H(v, 2\eta)) \geq p$ and

$$\mu(H(v, 2\eta)) \geq \frac{p}{k} \geq \varepsilon.$$

Yet every $s \in S$ satisfies $\langle v, s \rangle = 0$, so its distance from $H(v, 2\eta)$ is 2η. No closed ball of radius η centered at a point of S intersects this halfspace, a contradiction.

Finally, the vectors e_{ij} form an orthonormal basis of E, and hence

$$|S| \geq \dim(W) = \operatorname{tr}(P_W) = \sum_{i=1}^{k} \sum_{j=1}^{m} \|P_W e_{ij}\|^2 \geq \frac{km}{2} = \Omega\left(\frac{1}{\varepsilon \eta^2}\right).$$

Finally, embed E isometrically in ℓ_2. Any set of centers in ℓ_2 satisfying the theorem's conclusion can be projected onto E without changing its inner products with vectors in E or increasing its cardinality. The same lower bound therefore holds in ℓ_2 for all sufficiently small $\varepsilon, \eta > 0$. ■

4.2 Chaining at optimal scale

We work with a fixed family $V \subseteq [-1, +1]^n$ and the normalized Euclidean norm. We use $\mathtt{Rad}(V)$ and $N_2(V, r)$ as defined in Sections 2.3 and 2.2.2; the same definition of Rademacher complexity applies to the finite families in $\mathbb{R}^n$ that arise below. We also recall Massart's lemma

Lemma 2.6 (Restated)**.** *Let $V \subseteq \mathbb{R}^n$ be finite and nonempty, and suppose that every $v \in V$ has normalized ℓ_2 norm at most R. Then*

$$\mathtt{Rad}(V) \leq R\sqrt{\frac{2 \log |V|}{n}}.$$

Lemma 4.2. *Let $R > r > 0$ be two positive numbers, and let $B_r(x)$, $B_R(y)$ be two balls in a Euclidean space. Set*

$$A = B_r(x) \cap B_R(y).$$

Then there exists z such that $d(y, z) \leq \sqrt{R^2 - r^2}$, and $A \subseteq B_r(z)$.

Proof. By translation, assume that $y = 0$. If $\|x\| \leq \sqrt{R^2 - r^2}$, take $z = x$. Otherwise, set

$$z = \frac{\sqrt{R^2 - r^2}}{\|x\|} x.$$

Fix $a \in A$. If $\langle z, a - z \rangle \leq 0$, then since $x - z = \lambda z$ for some $\lambda > 0$, we have $\langle x - z, z - a \rangle \geq 0$. Consequently,

$$\|x - a\|^2 = \|x - z\|^2 + \|z - a\|^2 + 2\langle x - z, z - a \rangle \geq \|a - z\|^2.$$

Since $a \in B_r(x)$, this gives $a \in B_r(z)$. If instead $\langle z, a - z \rangle > 0$, then

$$\|a\|^2 = \|z\|^2 + \|a - z\|^2 + 2\langle z, a - z \rangle \geq \|z\|^2 + \|a - z\|^2.$$

Using $\|a\| \leq R$ and $\|z\|^2 = R^2 - r^2$, we again obtain $\|a - z\|^2 \leq r^2$.

■

Lemma 4.3 (Refined Chaining). *Let $V \subseteq [-1,+1]^n$ and $\alpha \in (0,1)$. For every integer $N \geq 1$, there is a finite $(\alpha + 4^{-N})$-cover $\mathcal{G} \subseteq \mathbb{R}^n$ of V satisfying*

$$\mathtt{Rad}(\mathcal{G}) \leq \frac{32}{\sqrt{n}} \int_0^1 \sqrt{\log N_2(V, \alpha + t^2)}\, dt.$$

Proof. Set $\alpha_i = \alpha + 4^{-i}$ for $0 \leq i \leq N$, and let $H_0 = \{0\}$. For $1 \leq i \leq N$, choose an α_i-cover H_i of V of minimum size. For each $v \in V$, fix $h_i(v) \in H_i$ within distance α_i of v.

We construct approximations $\pi_i(v)$ satisfying

$$\|v - \pi_i(v)\|_2 \leq \alpha_i, \qquad \|\pi_i(v) - \pi_{i-1}(v)\|_2 \leq \sqrt{\alpha_{i-1}^2 - \alpha_i^2}.$$

Start with $\pi_0(v) = 0$. Given $\pi_i(v)$, apply Lemma 4.2 to

$$A = B_{\alpha_i}(\pi_i(v)) \cap B_{\alpha_{i+1}}(h_{i+1}(v)).$$

It gives a center $\pi_{i+1}(v)$ such that

$$A \subseteq B_{\alpha_{i+1}}(\pi_{i+1}(v)), \qquad \|\pi_{i+1}(v) - \pi_i(v)\|_2 \leq \sqrt{\alpha_i^2 - \alpha_{i+1}^2}.$$

Since $v \in A$, both required inequalities follow. Fix the choice of the new center for each pair $(\pi_i(v), h_{i+1}(v))$, so it depends only on this pair. Consequently,

$$|\pi_i(V)| \leq \prod_{j=1}^{i} |H_j|.$$

Set $\mathcal{G} = \pi_N(V)$. This is an $(\alpha + 4^{-N})$-cover of V.

To bound its Rademacher complexity, write each representative as a sum of successive differences,

$$\pi_N(v) = \sum_{i=1}^{N} \big(\pi_i(v) - \pi_{i-1}(v)\big),$$

and let $\Delta_i = \{\pi_i(v) - \pi_{i-1}(v) : v \in V\}$. Subadditivity of the supremum gives

$$\mathtt{Rad}(\mathcal{G}) \leq \sum_{i=1}^{N} \mathtt{Rad}(\Delta_i).$$

Every vector in Δ_i has norm at most $\sqrt{\alpha_{i-1}^2 - \alpha_i^2}$, and

$$|\Delta_i| \leq |\pi_i(V)|\, |\pi_{i-1}(V)| \leq \prod_{j=1}^{i} |H_j|^2.$$

Thus Massart's lemma and subadditivity of the square root yield

$$\mathtt{Rad}(\Delta_i) \leq \frac{2\sqrt{\alpha_{i-1}^2 - \alpha_i^2}}{\sqrt{n}} \sum_{j=1}^{i} \sqrt{\log |H_j|}.$$

Since $\alpha < 1$,

$$\sqrt{\alpha_{i-1}^2 - \alpha_i^2} = \sqrt{(4^{1-i} - 4^{-i})(2\alpha + 4^{1-i} + 4^{-i})} \leq 4 \cdot 2^{-i}.$$

Summing in the opposite order and using $\sum_{i=j}^{N} 2^{-i} \leq 2 \cdot 2^{-j}$ gives

$$\begin{aligned} \mathtt{Rad}(\mathcal{G}) &\leq \frac{8}{\sqrt{n}} \sum_{j=1}^{N} \sum_{i=j}^{N} 2^{-i} \sqrt{\log |H_j|} \\ &\leq \frac{32}{\sqrt{n}} \sum_{j=1}^{N} 2^{-j-1} \sqrt{\log N_2(V, \alpha + 4^{-j})}. \end{aligned}$$

By monotonicity of the covering number, each summand is at most the integral of $\sqrt{\log N_2(V, \alpha + t^2)}$ over $[2^{-j-1}, 2^{-j}]$. These intervals are disjoint subsets of $[0,1]$, proving the claim. ∎

Theorem 2.8 (Restated)**.** *For every $n \geq 1$, every $V \subseteq [-1,+1]^n$, and every $\alpha \in (0,1)$,*

$$\mathtt{Rad}(V) \leq \alpha + \frac{32}{\sqrt{n}} \int_0^1 \sqrt{\log N_2(V, \alpha + t^2)}\, dt.$$

Proof of Theorem 2.8. For every integer $N \geq 1$, Lemma 4.3 gives an $(\alpha + 4^{-N})$-cover $\mathcal{G}$ of V with the stated bound on $\mathtt{Rad}(\mathcal{G})$. By Cauchy–Schwarz, replacing a vector by a representative within normalized ℓ_2 distance $\alpha + 4^{-N}$ changes its normalized signed sum by at most $\alpha + 4^{-N}$. Hence

$$\mathtt{Rad}(V) \leq \alpha + 4^{-N} + \mathtt{Rad}(\mathcal{G}) \leq \alpha + 4^{-N} + \frac{32}{\sqrt{n}} \int_0^1 \sqrt{\log N_2(V, \alpha + t^2)}\, dt.$$

Letting $N \to \infty$ proves the theorem. ∎

4.3 Approximate Disambiguation via Compression

Definition 4.4 (Loss for partial labels)**.** For $y \in \{-1,+1,\star\}$ and a total function $h \in \{-1,+1\}^{\mathcal{X}}$, define

$$\ell_{x,y}(h) = \begin{cases} 0, & y = \star, \\ \mathbf{1}_{\{h(x) \neq y\}}, & y \in \{-1,+1\}. \end{cases}$$

Thus, for a partial function $f \in \{-1,+1,\star\}^{\mathcal{X}}$ and a finite set $U \subseteq \mathcal{X}$,

$$L_{U,f}(h) = \frac{1}{|U|} \sum_{x \in U} \ell_{x,f(x)}(h) = \frac{1}{|U|} \big|\{x \in U \cap \mathrm{supp}(f) : h(x) \neq f(x)\}\big|.$$

Similarly, if

$$S = ((x_i, y_i))_{i=1}^m$$

is a labeled sample and $\mathcal{D}$ is a distribution over $\mathcal{X} \times \{-1,+1,\star\}$, set

$$L_S(h) = \frac{1}{m} \sum_{i=1}^m \ell_{x_i,y_i}(h), \qquad L_{\mathcal{D}}(h) = \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \ell_{x,y}(h).$$

Definition 4.5 (ε-compression scheme)**.** Let

$$\mathcal{F} \subseteq \{-1,+1,\star\}^{\mathcal{X}}.$$

An ε-compression scheme of size k for $\mathcal{F}$ consists of a compressor κ and a reconstructor ρ with the following properties.

For every sample

$$S = ((x_i, y_i))_{i=1}^n,$$

the compressor returns a sequence

$$\kappa(S) = I = (i_1, \ldots, i_k) \in [n]^k.$$

Writing

$$S_I = ((x_{i_j}, y_{i_j}))_{j=1}^k,$$

the reconstructor outputs a total function

$$\rho(S_I) \in \{-1,+1\}^{\mathcal{X}}.$$

Moreover, whenever S is realizable by some $f \in \mathcal{F}$, we have

$$L_S\big(\rho(S_I)\big) \leq \varepsilon.$$

For a given n, we say that the scheme has effective size s_n if there exists a set

$$\mathcal{I}_n \subseteq [n]^k, \qquad |\mathcal{I}_n| \leq s_n,$$

such that, for every sample S of size n, there is some $J \in \mathcal{I}_n$ satisfying

$$\rho(S_{\kappa(S)}) = \rho(S_J).$$

Theorem 4.6 (ε-compression for VC classes)**.** *Let $\mathcal{F} \subset \{-1,+1,\star\}^{\mathcal{X}}$ be a partial concept class of VC dimension $d = \mathtt{VC}(\mathcal{F}) \geq 1$.*[7] *Then, for every $\varepsilon \in (0,1)$, $\mathcal{F}$ admits an ε-compression scheme of size*

$$k \leq Cd \log \frac{e}{\varepsilon},$$

where $C > 0$ is a universal constant. Moreover, for every $n > d$, the compression scheme has effective size

$$s_n \leq \left(\frac{en}{d}\right)^{Cd \log \frac{e}{\varepsilon}}.$$

Proof. Fix a finite sample

$$S = \big((x_i, f(x_i))\big)_{i=1}^{n}$$

realizable by some $f \in \mathcal{F}$. We use the partial concept class learner of [AHHM22]. There is a total-valued learner A such that, for every distribution $\mathcal{D}$ supported on S, there is a sequence B of at most Cd examples from S for which

$$L_{\mathcal{D}}(A(B)) \leq \frac{1}{3}.$$

The learner may be taken to be invariant under permutations and repetitions of its input[8].

Strictly speaking, the learner in [AHHM22] is stated for binary-labeled samples. This causes no difficulty here: before applying the learner, we simply remove the $\star$-labeled examples. Since these examples contribute zero loss for every total hypothesis, the same weak-learning guarantee holds with respect to our loss.

By the minimax theorem, there is a distribution μ over such sequences B such that, for every distribution $\mathcal{D}$ supported on S,

$$\mathop{\mathbb{E}}_{B \sim \mu} L_{\mathcal{D}}(A(B)) \leq \frac{1}{3}.$$

In particular, by taking $\mathcal{D}$ to be a point mass, we obtain, for every x_i such that $f(x_i) \neq \star$,

$$\mathop{\mathbb{P}}_{B \sim \mu}\big(A(B)(x_i) \neq f(x_i)\big) \leq \frac{1}{3}.$$

Let

$$T = \left\lceil C \log \frac{e}{\varepsilon} \right\rceil,$$

draw independently

$$B_1, \ldots, B_T \sim \mu,$$

and define

$$h = \mathtt{Maj}\big(A(B_1), \ldots, A(B_T)\big).$$

For every x_i such that $f(x_i) \neq \star$, Hoeffding's inequality gives

$$\begin{aligned}\mathbb{P}\big(h(x_i) \neq f(x_i)\big) &\leq \mathbb{P}\left[\frac{1}{T}\sum_{t=1}^{T} \mathbf{1}_{\{A(B_t)(x_i) \neq f(x_i)\}} \geq \frac{1}{2}\right] \\ &\leq e^{-cT} \leq \varepsilon,\end{aligned}$$

where $c > 0$ is a universal constant. Since $\star$-labeled examples contribute zero loss, it follows that

$$\mathop{\mathbb{E}}_{B_1,\ldots,B_T} L_S(h) \leq \varepsilon.$$

Consequently, there is a deterministic choice of $B_1, \ldots, B_T$ satisfying

$$L_S\left(\mathtt{Maj}\big(A(B_1), \ldots, A(B_T)\big)\right) \leq \varepsilon.$$

[7] If $d = 0$, a fixed total completion of the class gives a scheme with $k = 0$ and $s_n = 1$.

[8] This was not formally proved in [AHHM22] but is clear from the definition of the learner as a one-inclusion-graph based learner.

The compressor chooses one such tuple and records the concatenation of the corresponding index sequences. By padding each block with repetitions if necessary, we may assume that all blocks have the same fixed length. The reconstructor splits the compressed sequence into these blocks, applies A to each block, and returns their majority vote. The total number of recorded examples is at most

$$k \le CdT \le Cd \log \frac{e}{\varepsilon}.$$

It remains to bound the effective size. Since A is invariant under permutations and repetitions, the hypothesis $A(B_t)$ depends only on the set of distinct sample indices appearing in B_t. For $n > d$, the number of possible outputs of a single block is therefore at most

$$\left(\frac{en}{d}\right)^{Cd},$$

after adjusting the universal constant. Since there are

$$T \le C \log \frac{e}{\varepsilon}$$

blocks, the number of possible reconstructed hypotheses is at most

$$s_n \le \left(\frac{en}{d}\right)^{CdT} \le \left(\frac{en}{d}\right)^{Cd \log \frac{e}{\varepsilon}}.$$

■

Theorem 4.7 (Sample complexity for compression-based learners)**.** *Let*

$$\mathcal{F} \subseteq \{-1, +1, \star\}^{\mathcal{X}}$$

admit an α-compression scheme of size k and effective size s_n. Let $\mathcal{D}$ be realizable by $\mathcal{F}$, draw

$$S \sim \mathcal{D}^n,$$

where $n > 2k$, and set

$$m = n - k.$$

Then, with probability at least $1 - \delta$,

$$L_{\mathcal{D}}\big(\rho(S_{\kappa(S)})\big) \le 2\alpha + \sqrt{\frac{4\alpha \log(s_n/\delta)}{m}} + \frac{4 \log(s_n/\delta)}{m}.$$

Proof. Let

$$\mathcal{I}_n = \{I_1, \dots, I_{s_n}\}$$

be the collection in the definition of effective size, and set

$$h_j = \rho(S_{I_j}).$$

Let $J_j \subseteq [n]$ be the set of distinct indices appearing in I_j, and let

$$V_j = [n] \setminus J_j.$$

Since I_j has length k,

$$|V_j| \ge n - k = m.$$

For each fixed j, the function h_j depends only on the examples indexed by J_j. Conditional on these examples, the remaining sample

$$S_{V_j}$$

consists of $|V_j|$ independent samples from $\mathcal{D}$ and is independent of h_j. Therefore, Bernstein's inequality implies that, with probability at least $1 - \delta/s_n$,

$$L_{\mathcal{D}}(h_j) \le L_{S_{V_j}}(h_j) + \sqrt{\frac{2L_{S_{V_j}}(h_j)\log(s_n/\delta)}{|V_j|}} + \frac{4\log(s_n/\delta)}{|V_j|}.$$

A union bound shows that this inequality holds simultaneously for every $j \le s_n$ with probability at least $1 - \delta$.

Write

$$h = \rho(S_{\kappa(S)}).$$

By the definition of effective size, there is some $j \le s_n$ such that

$$h = h_j.$$

Since the compression scheme has empirical error at most α on S,

$$L_S(h_j) \le \alpha.$$

It follows that

$$L_{S_{V_j}}(h_j) \le \frac{n}{|V_j|} L_S(h_j) \le \frac{n}{m}\alpha \le 2\alpha,$$

where the last inequality uses $n > 2k$. Moreover,

$$|V_j| \ge m.$$

Substitution into the Bernstein bound therefore gives

$$L_{\mathcal{D}}(h) \le 2\alpha + \sqrt{\frac{4\alpha\log(s_n/\delta)}{m}} + \frac{4\log(s_n/\delta)}{m}.$$

■

Corollary 4.8 (Sample complexity for VC classes via compression)**.** *Let*

$$\mathcal{F} \subseteq \{-1, +1, \star\}^{\mathcal{X}}$$

be a partial concept class of VC dimension $d \ge 1$. There is a compression-based learner such that, for every $\varepsilon, \delta \in (0,1)$, if

$$n \ge \frac{C}{\varepsilon}\left(d\log^2\frac{e}{\varepsilon} + \log\frac{1}{\delta}\right),$$

then, for every distribution $\mathcal{D}$ realizable by $\mathcal{F}$, with probability at least $1 - \delta$ over $S \sim \mathcal{D}^n$,

$$L_{\mathcal{D}}\big(\rho(S_{\kappa(S)})\big) \le \varepsilon.$$

In particular, for constant confidence, the sample complexity is

$$O\left(\frac{d}{\varepsilon}\log^2\frac{e}{\varepsilon}\right).$$

Proof. Apply Theorem 4.6 with compression error

$$\alpha = \frac{\varepsilon}{8}.$$

This gives a compression scheme of size

$$k \le Cd\log\frac{e}{\varepsilon}.$$

For $n > 2k$, its effective size satisfies

$$\log s_n \le Cd \log \frac{e}{\varepsilon} \log \frac{en}{d}.$$

Consequently,

$$\log \frac{s_n}{\delta} \le Cd \log \frac{e}{\varepsilon} \log \frac{en}{d} + \log \frac{1}{\delta}.$$

We claim that the assumed lower bound on n, with a sufficiently large universal constant, implies

$$\frac{\log(s_n/\delta)}{n} \le c\varepsilon$$

for a sufficiently small universal constant $c > 0$. Write $L = \log(e/\varepsilon)$ and $b = \log(1/\delta)$, and let

$$n_0 = \left\lceil \frac{K}{\varepsilon}(dL^2 + b) \right\rceil.$$

For $n \ge n_0 \ge d$, the function

$$\frac{dL \log(en/d) + b}{n}$$

is decreasing, so it suffices to bound it at n_0. The inequality $\log(1+t) \le t$ gives

$$\log(en_0/d) \le O(L + \log K) + \frac{b}{dL^2}.$$

It follows that $\log(s_n/\delta)/n \le O((\log K)/K)\varepsilon$. Choosing K sufficiently large proves the claim and also ensures $n > 2k$, hence $m = n - k \ge n/2$.

Theorem 4.7 now gives

$$\begin{aligned} L_{\mathcal{D}}\big(\rho(S_{\kappa(S)})\big) &\le 2\alpha + \sqrt{\frac{4\alpha \log(s_n/\delta)}{m}} + \frac{4 \log(s_n/\delta)}{m} \\ &\le \frac{\varepsilon}{4} + \sqrt{C_1 c}\,\varepsilon + C_2 c\varepsilon. \end{aligned}$$

Choosing c sufficiently small makes the right-hand side at most ε, which proves the result. ∎

Lemma 4.9 (Fractional covering for partial function classes)**.** *Let $\mathcal{F} \subset \{-1, +1, \star\}^{\mathcal{X}}$ be a partial concept class over a finite domain $\mathcal{X}$, and assume that $\mathtt{VC}(\mathcal{F}) = d < \infty$. Then, for every $\varepsilon \in (0,1)$, there exists a distribution μ over $\{-1,+1\}^{\mathcal{X}}$ such that, for every $f \in \mathcal{F}$,*

$$\mathop{\mathbb{P}}_{h \sim \mu}\big(L_{\mathcal{X},f}(h) \le \varepsilon\big) \ge \left(\frac{c}{\varepsilon}\right)^{-Cd \log \frac{e}{\varepsilon}},$$

where $C > 0$ and $c > 1$ are universal constants.

Proof. The case $d = 0$ is trivial, so assume that $d \ge 1$. By Theorem 4.6 and Corollary 4.8, after adjusting the universal constants, there is an ε-compression scheme (κ, ρ) for $\mathcal{F}$ of size

$$k \le Cd \log \frac{e}{\varepsilon}$$

such that, for

$$m = \left\lceil \frac{Cd}{\varepsilon} \log^2 \frac{e}{\varepsilon} \right\rceil,$$

the corresponding compression-based learner has loss at most ε with probability at least $1/2$ when applied to an m-sample from the uniform distribution on $\mathcal{X}$. Moreover, the effective size of the compression scheme satisfies

$$s_m \le \left(\frac{em}{d}\right)^{Cd \log \frac{e}{\varepsilon}}.$$

Thus, there is a collection

$$\mathcal{I}_m \subseteq [m]^k, \qquad |\mathcal{I}_m| \le s_m,$$

such that, for every labeled sample S of size m, there is some $I \in \mathcal{I}_m$ for which

$$\rho(S_I) = \rho(S_{\kappa(S)}).$$

We now define the distribution μ. First draw

$$x_1, \dots, x_m$$

independently and uniformly from $\mathcal{X}$. Next, draw

$$I = (i_1, \dots, i_k)$$

uniformly from $\mathcal{I}_m$. Finally, draw

$$y_1, \dots, y_k$$

independently and uniformly from $\{-1, +1, \star\}$, and output

$$h = \rho\big((x_{i_j}, y_j)_{j=1}^k\big).$$

The law of the resulting total function h defines μ and does not depend on the choice of $f \in \mathcal{F}$.

Fix $f \in \mathcal{F}$, and let

$$S = \big((x_i, f(x_i))\big)_{i=1}^m.$$

By the choice of m, with probability at least $1/2$ over the draw of $x_1, \dots, x_m$,

$$L_{\mathcal{X},f}\big(\rho(S_{\kappa(S)})\big) \le \varepsilon.$$

Condition on such a choice of $x_1, \dots, x_m$. By the definition of effective size, there is some $I \in \mathcal{I}_m$ such that

$$\rho(S_I) = \rho(S_{\kappa(S)}).$$

The randomly chosen index list equals one such I with probability at least $1/s_m$. Conditional on this event, the random labels satisfy

$$y_j = f(x_{i_j}) \qquad \text{for every } j \in [k]$$

with probability 3^{-k}. Therefore,

$$\mathbb{P}\left[h = \rho(S_{\kappa(S)}) \,\middle|\, x_1, \dots, x_m\right] \ge \frac{1}{3^k s_m}.$$

It follows that

$$\mathop{\mathbb{P}}_{h\sim\mu}\big(L_{\mathcal{X},f}(h) \le \varepsilon\big) \ge \frac{1}{2 \cdot 3^k s_m}.$$

Using the bounds on k and s_m, we obtain

$$3^k s_m \le \left(\frac{3em}{d}\right)^{Cd\log\frac{e}{\varepsilon}}.$$

Moreover,

$$\frac{m}{d} \le \frac{C}{\varepsilon}\log^2\frac{e}{\varepsilon} \le \frac{c}{\varepsilon^3}$$

Consequently, after adjusting the universal constants

$$\frac{1}{2 \cdot 3^k s_m} \ge \left(\frac{c}{\varepsilon}\right)^{-Cd\log\frac{e}{\varepsilon}}.$$

This proves the result. ■

Lemma 4.10 (From Fractional Cover to Finite Cover)**.** *Let $\mathcal{F} \subseteq \{-1,+1,\star\}^{\mathcal{X}}$ be a partial function class on a finite domain $\mathcal{X}$, and let $\alpha,\delta \in (0,1)$. Assume there is a distribution $\mathcal{D}$ over $\{-1,+1\}^{\mathcal{X}}$ such that, for each $f \in \mathcal{F}$,*

$$\mathop{\mathbb{P}}_{h\sim\mathcal{D}}[L_{\mathcal{X},f}(h) \le \alpha] \ge \delta.$$

Then, for any $\varepsilon \in (0,1)$ there is some $H \subset \{-1,+1\}^{\mathcal{X}}$, of size

$$|H| \le \frac{C}{\varepsilon^2\delta}\log\frac{e}{\delta},$$

Such that for any $f \in \mathcal{F}$, there is some $h \in H$ with

$$L_{\mathcal{X},f}(h) \le \alpha + \varepsilon.$$

Proof. Let $\mathcal{H}$ be the support of $\mathcal{D}$, and for each $f \in \mathcal{F}$ define the partial function $w_f : \mathcal{H} \to \{-1,+1,\star\}$ by

$$w_f(h) = \begin{cases} 1 & L_{\mathcal{X},f}(h) \le \alpha, \\ -1 & L_{\mathcal{X},f}(h) \ge \alpha+\varepsilon, \\ \star & \text{otherwise.} \end{cases}$$

Set $\mathcal{W} = \{w_f \ : \ f \in \mathcal{F}\}$. Let $f \in \mathcal{F}$, $h \in \mathcal{H}$. Treat the $\star$ as 0, and note that under this identification, $|f(x)| - f(x)h(x)$, is 2 if $f(x)$ is defined (nonzero) and not equal $h(x)$, and it is 0 if either $f(x) = h(x)$, or $f(x) = 0$. Summing over this we get

$$L_{\mathcal{X},f}(h) = \frac{1}{2|\mathcal{X}|}\sum_{x\in\mathcal{X}}\big(|f(x)| - f(x)h(x)\big) = \frac{|\mathsf{supp}(f)| - \langle f,h\rangle}{2|\mathcal{X}|},$$

where $\mathsf{supp}(f)$ is the set of points for which $f(x) \neq 0$. Hence the class $\mathcal{W} = \{w_f \ : \ f \in \mathcal{F}\}$ can be embedded as the class of affine classifiers with margin. Formally define

$$\varphi_f = \frac{-f}{2\sqrt{|\mathcal{X}|}}, \quad s_f = \frac{|\mathsf{supp}(f)|}{2|\mathcal{X}|}, \quad x_h = \frac{h}{\sqrt{|\mathcal{X}|}}.$$

Then φ_f, x_h are vectors in the unit ball for any $f \in \mathcal{F}$, $h \in \mathcal{H}$, and we have

$$L_{\mathcal{X},f}(h) = \langle \varphi_f, x_h\rangle + s_f.$$

In particular we have

$$w_f(h) = \begin{cases} 1 & \langle \varphi_f, x_h\rangle + s_f \le \alpha, \\ -1 & \langle \varphi_f, x_h\rangle + s_f \ge \alpha+\varepsilon, \\ \star & \text{otherwise.} \end{cases}$$

Hence, $\mathcal{W}$ can indeed be embedded into the class of affine classifiers with margin ε. By [AHHM22, Proposition 17] we have

$$\mathtt{VC}(\mathcal{W}) \le \frac{C}{\varepsilon^2}.$$

Thus by 4.1, there is some finite set $H \subset \mathcal{H}$, of size at most

$$|H| \le \frac{C}{\varepsilon^2\delta}\log\frac{e}{\delta},$$

such that for any $f \in \mathcal{F}$, there is some $h \in H$, with $w_f(h) \neq -1$, that is

$$L_{\mathcal{X},f}(h) \le \alpha + \varepsilon.$$

This implies the desired result.

■

Theorem 2.4 (Restated). *Let $\mathcal{F} \subseteq \{-1,+1,\star\}^{\mathcal{X}}$ be a partial binary class of VC dimension d on a finite domain $\mathcal{X}$. For every $\varepsilon \in (0,1/2]$, there is an ε-disambiguation $\mathcal{H} \subseteq \{-1,+1\}^{\mathcal{X}}$ satisfying*

$$|\mathcal{H}| \le (1/\varepsilon)^{cd\log(1/\varepsilon)},$$

where $c > 0$ is a universal constant.

This bound is optimal up to constant factors in the exponent: there are universal constants $c_0, \varepsilon_0 > 0$ such that, for every integer $d \ge 1$ and every $\varepsilon \in (0,\varepsilon_0]$, some partial binary class of VC dimension d has the property that every ε-disambiguation $\mathcal{H}$ satisfies

$$|\mathcal{H}| \ge (1/\varepsilon)^{c_0 d\log(1/\varepsilon)}.$$

Proof of Theorem 2.4. Upper bound. If $d = 0$, no point receives both binary labels among the functions in $\mathcal{F}$, so a single total function extends the whole class. Assume henceforth that $d \ge 1$, and put $\rho = \varepsilon/2$. Lemma 4.9 gives a distribution μ over total binary functions such that

$$\mathop{\mathbb{P}}_{h\sim\mu}\big(L_{\mathcal{X},f}(h) \le \rho\big) \ge \delta \quad \text{for every } f \in \mathcal{F}, \qquad \delta = \left(\frac{A}{\rho}\right)^{-Bd\log(e/\rho)},$$

where $A > 1$ and $B > 0$ are universal constants. Apply Lemma 4.10 with approximation error $\alpha = \rho$ and additional error ρ. This gives an ε-disambiguation of size

$$|\mathcal{H}| \le \frac{C_1}{\rho^2\delta}\log\frac{e}{\delta} \le (1/\varepsilon)^{cd\log(1/\varepsilon)}.$$

The second inequality follows from $\log(1/\delta) = O(d\log^2(1/\varepsilon))$, using $\varepsilon \le 1/2$ and $d \ge 1$ to absorb all remaining factors into the universal constant c.

Lower bound. Fix $d \ge 1$ and a sufficiently small $\varepsilon > 0$, and set $n = \lfloor 1/(4\varepsilon)\rfloor$. Combining Pabbaraju's construction [Pab26, Theorem 3 and Appendix B] with the graph-to-class reduction of [AHHM22] gives a partial binary class $\mathcal{F}_n$ on n points of VC dimension one such that every exact disambiguation has size at least $n^{a\log n}$, for a universal $a > 0$. The construction gives such a class for every sufficiently large n.[9]

It remains to amplify this lower bound to VC dimension d, while allowing an ε-fraction of errors.

We do this in three steps: pass to fractional disambiguations, take a direct product, and account for the allowed errors.

1. Fractional disambiguations and their dual. Write $h \succeq f$ if the total function h agrees with the partial function f on every defined input. A fractional exact disambiguation assigns nonnegative weights w_h to total binary functions, subject to

$$\sum_{h\succeq f} w_h \ge 1 \qquad \text{for every } f \in \mathcal{F}_n.$$

Let q_n be the minimum total weight $\sum_h w_h$. We first show that q_n is still large. Normalize optimal weights to a probability distribution. Each $f \in \mathcal{F}_n$ is extended by a draw from this distribution with probability at least $1/q_n$. Hence $k = \lceil q_n(\ln|\mathcal{F}_n| + 1)\rceil$ independent draws fail to extend some member of $\mathcal{F}_n$ with probability at most $|\mathcal{F}_n|e^{-k/q_n} < 1$. Some such sample is therefore an exact disambiguation. Since $q_n \ge 1$ and $|\mathcal{F}_n| \le 3^n$, the one-block lower bound gives

$$q_n \ge \frac{n^{a\log n}}{\ln|\mathcal{F}_n| + 2} \ge \frac{n^{a\log n}}{n\ln 3 + 2} \ge n^{(a/2)\log n}$$

for all sufficiently large n. Thus, passing to the fractional relaxation loses only a constant factor in the exponent.

[9] In Pabbaraju's construction, the domain size n progresses geometrically. Padding with undefined inputs extends the $n^{\Omega(\log n)}$ lower bound to every sufficiently large n.

Linear-programming duality gives nonnegative weights λ_f on the partial functions satisfying

$$\sum_{f \in \mathcal{F}_n} \lambda_f = q_n, \qquad \sum_{f \colon h \succeq f} \lambda_f \leq 1 \quad \text{for every total binary function } h.$$

Consequently, $\nu_n(f) = \lambda_f / q_n$ is a probability distribution on $\mathcal{F}_n$ such that

$$\mathop{\mathbb{P}}_{f \sim \nu_n} [h \succeq f] \leq \frac{1}{q_n} \qquad \text{for every fixed total binary function } h.$$

This is the form of the lower bound that we will use on each block.

2. Taking a direct product. Let $\mathcal{G} = \mathcal{F}_n^{\otimes d}$ be the class on d disjoint blocks of n points obtained by choosing a member of $\mathcal{F}_n$ on each block, with all choices allowed. Its VC dimension is exactly d: one point can be shattered in each block, while no shattered set can contain two points from the same block.

The fractional exact disambiguation number of $\mathcal{G}$ is exactly q_n^d. Indeed, taking products of the one-block weights w_h gives a feasible fractional disambiguation of total weight q_n^d. Taking products of the dual weights λ_f gives a feasible dual solution of the same value. In both cases the constraints factor across the blocks, proving the claimed equality.

3. Allowing approximation errors. Draw $f \in \mathcal{G}$ by choosing its restrictions $f_1, \ldots, f_d$ independently from ν_n, and fix a total binary function h, with restrictions $h_1, \ldots, h_d$. Our choice $n = \lfloor 1/(4\varepsilon) \rfloor$ ensures that an ε-approximation makes at most $\varepsilon dn \leq d/4$ disagreements in total. Every block on which $h_i \not\succeq f_i$ contributes at least one disagreement. Thus, $h_i \succeq f_i$ must hold on at least $\lceil d/2 \rceil$ blocks.

For any specified set of k blocks, independence and the bound from Step 1 give probability at most q_n^{-k} of exact agreement on all of them. Taking $k = \lceil d/2 \rceil$ and a union bound over the possible sets of blocks yields

$$\mathbb{P}\big[h \text{ makes at most } \varepsilon dn \text{ disagreements with } f\big] \leq \binom{d}{\lceil d/2 \rceil} q_n^{-\lceil d/2 \rceil} \leq 2^d q_n^{-d/2}.$$

If $\mathcal{H}$ is an ε-disambiguation of $\mathcal{G}$, every sampled f must be approximated by some $h \in \mathcal{H}$. A final union bound gives $1 \leq |\mathcal{H}|\, 2^d q_n^{-d/2}$, and hence

$$|\mathcal{H}| \geq 2^{-d} q_n^{d/2} \geq 2^{-d} n^{(a/4) d \log n} \geq (1/\varepsilon)^{c_0 d \log(1/\varepsilon)}.$$

The last inequality holds for a universal $c_0 > 0$ and all sufficiently small ε, since $n = \Theta(1/\varepsilon)$. The required upper bound ε_0 on ε is independent of d. ■

4.4 Covering Number Bounds

Recall that H is an (α, ε)-approximate cover of a function class $\mathcal{F}$, over finite domain $\mathcal{X}$, if for every $f \in \mathcal{F}$ there is some $h \in H$ such that

$$|f(x) - h(x)| \leq \alpha$$

for all but at most $\varepsilon |\mathcal{X}|$ points $x \in \mathcal{X}$.

Lemma 4.11. *Let $\mathcal{F}$ be a class of real-valued functions on a finite domain $\mathcal{X}$ of size n. Let $\gamma > 0$, $\varepsilon \in (0,1)$, and suppose that $d = \mathtt{fat}_\gamma(\mathcal{F}) < \infty$. Assume that, for some $\alpha \in [0,1]$ and $r > \gamma$, every $f \in \mathcal{F}$ satisfies $|f(x)| \leq r/2$ on all but at most αn points of $\mathcal{X}$.*

Then $\mathcal{F}$ has an $(\frac{r+\gamma}{4}, \alpha + \varepsilon)$-approximate cover G of size

$$|G| \leq \left(\frac{C}{\varepsilon}\right)^{c \cdot d \log \frac{e}{\varepsilon}},$$

where $C, c > 0$ are universal constants.

Proof. For each $f \in \mathcal{F}$ define the partial function $h_f(x)$ to be 1 if $f(x) \geq \frac{\gamma}{2}$, -1 if $f(x) \leq -\frac{\gamma}{2}$, and undefined if $f(x) \in (\frac{-\gamma}{2}, \frac{\gamma}{2})$. Clearly the induced partial concept class $\mathcal{H}_\mathcal{F} = \{h_f \ : \ f \in \mathcal{F}\}$ has VC dimension at most d. Thus, applying Theorem 2.4 with accuracy $\min\{\varepsilon, 1/2\}$ gives some $H \subset \{-1,+1\}^\mathcal{X}$ of size at most

$$|H| \leq \left(\frac{C}{\varepsilon}\right)^{c \cdot d \log \frac{e}{\varepsilon}},$$

such that for each $h_f \in \mathcal{H}_\mathcal{F}$ there is $h \in H$, with $L_{\mathcal{X},h_f}(h) \leq \varepsilon$. For each such h define $\phi_h : \mathcal{X} \to [-r, r]$ by

$$\phi_h(x) = \begin{cases} \frac{r-\gamma}{4} & h(x) = 1, \\ \frac{\gamma - r}{4} & h(x) = -1. \end{cases}$$

Fix $f \in \mathcal{F}$ and choose $h \in H$ with $L_{\mathcal{X},h_f}(h) \leq \varepsilon$. By assumption there are at most $\alpha|\mathcal{X}|$ points in $\mathcal{X}$, for which $|f(x)| > \frac{r}{2}$. At every other point, if either $h_f(x) = \star$ (so $f(x) \in [\frac{-\gamma}{2}, \frac{\gamma}{2}]$) or $h_f(x) = h(x)$, then

$$|f(x) - \phi_h(x)| \leq \frac{r+\gamma}{4}.$$

For any $f \in \mathcal{F}$ there is some $h \in H$ such that $h_f(x)$ is defined and not equal $h(x)$ for no more than εn points. Thus there are at most $(\alpha+\varepsilon)|\mathcal{X}|$ points for which the above does not hold, and $\{\phi_h \ : \ h \in H\}$ is the desired $(\frac{r+\gamma}{4}, \alpha+\varepsilon)$-approximate cover.

■

Lemma 4.12. *Let $\mathcal{F} \subseteq [-1,+1]^\mathcal{X}$ be a function class over finite domain $\mathcal{X}$ of size n. Let $\gamma > 0$ and $\varepsilon \in (0,1)$ and assume that $d = \mathtt{fat}_\gamma(\mathcal{F})$ is finite. Then there exists $H \subset [-1,+1]^\mathcal{X}$, an $(\frac{\gamma}{2}+\varepsilon, 2\varepsilon \log \frac{e}{\varepsilon})$-approximate cover for $\mathcal{F}$ of size*

$$|H| \leq \left(\frac{C}{\varepsilon}\right)^{c \cdot d \log^2 \frac{e}{\varepsilon}}.$$

Proof. We prove the following slightly more general statement by induction on k: for every $k \geq 0$, the class $\mathcal{F}$ has an

$$\left(\frac{\gamma}{2} + 2^{-k}, k\varepsilon\right)$$

approximate cover of size

$$\left(\frac{C}{\varepsilon}\right)^{cdk \log \frac{e}{\varepsilon}}.$$

Choosing k such that $\frac{1}{2^k} \leq \varepsilon \leq \frac{1}{2^{k-1}}$ then yields the desired bound. For $k = 0$, the zero function gives a cover of radius 1.

Assume the claim holds for k. If $k\varepsilon \geq 1$, a single representative satisfies the claim for $k+1$, so assume $k\varepsilon < 1$. Let H_k be the corresponding cover. Fix $\phi \in H_k$, and let $\mathcal{F}_\phi$ be the set of functions $f \in \mathcal{F}$ which are covered by ϕ, that is, all functions $f \in \mathcal{F}$ satisfying

$$|\{x \in \mathcal{X} : |f(x) - \phi(x)| > \gamma/2 + 2^{-k}\}| \leq k\varepsilon|\mathcal{X}|.$$

Apply Lemma 4.11 to the residual class

$$\mathcal{F}_\phi - \phi = \{f - \phi : f \in \mathcal{F}_\phi\},$$

with

$$\frac{r}{2} = \frac{\gamma}{2} + 2^{-k}, \qquad \alpha = k\varepsilon.$$

Translation by ϕ does not increase the γ-fat-shattering dimension, so the relevant dimension is still at most d.

Lemma 4.11 gives a cover of $\mathcal{F}_\phi - \phi$ with covering error $(k+1)\varepsilon$ and radius

$$\frac{r+\gamma}{4} = \frac{\gamma}{2} + 2^{-(k+1)}.$$

Translating this cover back by adding ϕ, clipping each function above 1 and below -1, so it will be in $[-1,+1]^{\mathcal{X}}$, and doing this for every $\phi \in H_k$, gives the desired cover for the next step. The size is multiplied by at most

$$\left(\frac{C}{\varepsilon}\right)^{c \cdot d \log \frac{e}{\varepsilon}},$$

which proves the induction.

■

Theorem 2.5 (Restated)**.** *Let $n \geq 1$, let $V \subseteq [-1,+1]^n$, let $\gamma > 0$, and write $d = \mathtt{fat}_\gamma(V) < \infty$. For every $\varepsilon \in (0,1)$,*

$$\log N_2\left(V, \frac{\gamma}{2} + \varepsilon\right) \leq C d \log^3 \frac{e}{\varepsilon},$$

where $C > 0$ is a universal constant.

Proof of Theorem 2.5. View V as a class of functions on $[n]$. By Lemma 4.12, for every $\alpha \in (0,1)$ it has a

$$\left(\frac{\gamma}{2} + \alpha, 2\alpha \log \frac{e}{\alpha}\right)$$

approximate cover of size at most

$$\left(\frac{C}{\alpha}\right)^{Cd \log^2(e/\alpha)}.$$

Clipping the representatives to $[-1,+1]^n$ cannot increase any coordinate error. The resulting cover is therefore a normalized ℓ_2 cover at radius

$$\sqrt{\left(\frac{\gamma}{2} + \alpha\right)^2 + 8\alpha \log \frac{e}{\alpha}} \leq \frac{\gamma}{2} + 4\sqrt{\alpha \log \frac{e}{\alpha}}.$$

Choose $\alpha = c\varepsilon^4$, where $c > 0$ is a sufficiently small universal constant. Then $4\sqrt{\alpha \log(e/\alpha)} \leq \varepsilon$, and $\log(e/\alpha) = O(\log(e/\varepsilon))$. Taking the logarithm of the cover size gives

$$\log N_2\left(V, \frac{\gamma}{2} + \varepsilon\right) \leq C' d \log^3 \frac{e}{\varepsilon},$$

as required. ■

4.5 Main Result

The characterization of [AMM+26] states that γ-uniform convergence holds exactly when $\mathtt{fat}_{\gamma'}(\mathcal{F}) < \infty$ for every $\gamma' > \gamma$. Our theorem gives a sharp quantitative guarantee from the stronger assumption $d = \mathtt{fat}_\gamma(\mathcal{F}) < \infty$.

For $S = (x_1, \ldots, x_n)$, write

$$V_S = \{(f(x_1), \ldots, f(x_n)) : f \in \mathcal{F}\}.$$

Then $\mathtt{fat}_\gamma(V_S) \leq \mathtt{fat}_\gamma(\mathcal{F})$, so the vector bounds from the preceding sections apply to every sample.

Define the positive and negative representativeness

$$\mathtt{Rep}^+_{\mathcal{D}}(\mathcal{F}, S) = \sup_{f \in \mathcal{F}} \mathop{\mathbb{E}}_{x \sim \mathcal{D}} f(x) - \frac{1}{n}\sum_{i=1}^n f(x_i), \quad \mathtt{Rep}^-_{\mathcal{D}}(\mathcal{F}, S) = \sup_{f \in \mathcal{F}} \frac{1}{n}\sum_{i=1}^n f(x_i) - \mathop{\mathbb{E}}_{x \sim \mathcal{D}} f(x).$$

Note that $\mathtt{Rep}^+$ is the notion of representativeness defined in [SSBD14]. Its advantage is that it is consistent with our definition of Rademacher complexity by having no absolute value inside the expectation, so we can relate it to the Rademacher complexity.

The following concentration inequality is known as McDiarmid's inequality

Theorem 4.13 (McDiarmid's inequality [McD89]). *Say that $S, S' \in \mathcal{X}^n$ are i-neighbors if $S_j = S'_j$ for all $j \neq i$. Let $f : \mathcal{X}^n \to \mathbb{R}$ be some function such that for every pair of i-neighbors $S, S' \in \mathcal{X}^n$, we have*

$$|f(S) - f(S')| \leq c_i,$$

where $c_1, c_2 \dots c_n$, are some fixed positive constants. Then, for every $\varepsilon > 0$,

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^n}\left[\left|f(S) - \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^n} f(S')\right| \geq \varepsilon\right] \leq 2\exp\left(\frac{-2\varepsilon^2}{\sum_{i=1}^n c_i^2}\right).$$

In particular, if $f(S)$ is one of $\mathtt{Rep}_{\mathcal{D}}(\mathcal{F}, S)$, $\mathtt{Rep}_{\mathcal{D}}{}^{+}(\mathcal{F}, S)$, $\mathtt{Rep}_{\mathcal{D}}{}^{-}(\mathcal{F}, S)$, we may take $c_i = 2/n$ for all $1 \leq i \leq n$, and get

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^n}\left[\left|f(S) - \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^n} f(S')\right| \geq \varepsilon\right] \leq 2\exp\left(\frac{-\varepsilon^2 n}{2}\right).$$

Theorem 4.14. *Let $\mathcal{F} \subset [-1,+1]^{\mathcal{X}}$, let $\gamma \in (0,1)$, and assume that $d = \mathtt{fat}_\gamma(\mathcal{F}) < \infty$ is finite. Then there is a universal constant $C > 0$ such that for every $n \geq 1$ and every $S \in \mathcal{X}^n$,*

$$\mathtt{Rad}(V_S) \leq \frac{\gamma}{2} + C\sqrt{\frac{d}{n}}$$

Consequently, for every $\varepsilon, \delta \in (0,1)$, if

$$n \geq C\frac{d + \log(1/\delta)}{\varepsilon^2},$$

then for every distribution $\mathcal{D}$ over $\mathcal{X}$, with probability at least $1-\delta$ over $S \sim \mathcal{D}^n$,

$$\mathtt{Rep}_{\mathcal{D}}(\mathcal{F}, S) \leq \gamma + \varepsilon.$$

Proof. Fix $S \in \mathcal{X}^n$. Apply Theorem 2.8 to V_S with $\alpha = \gamma/2$ to obtain

$$\mathtt{Rad}(V_S) \leq \frac{\gamma}{2} + \frac{32}{\sqrt{n}}\int_0^1 \sqrt{\log N_2(V_S, \frac{\gamma}{2} + t^2)}dt.$$

By Theorem 2.5 and $\mathtt{fat}_\gamma(V_S) \leq d$, for every $t \in (0,1)$,

$$\log N_2(V_S, \frac{\gamma}{2} + t^2) \leq C \cdot d \cdot \log^3 \frac{e}{t},$$

where $C > 0$ is some universal constant. Hence the integral is bounded

$$\int_0^1 \sqrt{\log N_2(V_S, \frac{\gamma}{2} + t^2)}dt \leq \int_0^1 \sqrt{C \cdot d \cdot \log^3 \frac{e}{t}}dt \leq C'\sqrt{d},$$

where in the last inequality we used convergence of the integral of $\log^{3/2} \frac{e}{t}$ over $[0,1]$. Thus we indeed have

$$\mathtt{Rad}(V_S) \leq \frac{\gamma}{2} + C\sqrt{\frac{d}{n}}.$$

The bound holds for every S. By symmetrization [SSBD14, Lemma 26.2], we have

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^n} \mathtt{Rep}^{+}_{\mathcal{D}}(\mathcal{F}, S) \leq 2 \mathop{\mathbb{E}}_{S\sim\mathcal{D}^n} \mathtt{Rad}(V_S),$$

so, after changing the universal constant,

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^n} \mathtt{Rep}^{+}_{\mathcal{D}}(\mathcal{F}, S) \leq \gamma + C\sqrt{\frac{d}{n}}.$$

Apply the previous bound with

$$n \geq C\frac{d}{\varepsilon^2},$$

to get

$$\mathop{\mathbb{E}}_{S\sim\mathcal{D}^n} \mathtt{Rep}^+_{\mathcal{D}}(\mathcal{F},S) \le \gamma + C\sqrt{\frac{d}{n}} \le \gamma + \frac{\varepsilon}{2}.$$

Using McDiarmid's inequality 4.13, yields

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^n}\left[\mathtt{Rep}^+_{\mathcal{D}}(\mathcal{F},S) > \gamma+\varepsilon\right] \le \mathop{\mathbb{P}}_{S\sim\mathcal{D}^n}\left[\mathtt{Rep}^+_{\mathcal{D}}(\mathcal{F},S) > \mathop{\mathbb{E}}_{S'\sim\mathcal{D}^n}\mathtt{Rep}^+_{\mathcal{D}}(\mathcal{F},S') + \frac{\varepsilon}{2}\right] \le 2\exp(-\frac{\varepsilon^2 n}{8}).$$

Hence if we also have

$$n \ge C\frac{\log 1/\delta}{\varepsilon^2},$$

the above probability is at most $\delta/2$. By replacing $\mathcal{F}$ with $-\mathcal{F}$, we get the same bound for $\mathtt{Rep}^-_{\mathcal{D}}(\mathcal{F},S)$, and since

$$\mathtt{Rep}_{\mathcal{D}}(\mathcal{F},S) = \max\Big(\mathtt{Rep}^-_{\mathcal{D}}(\mathcal{F},S), \mathtt{Rep}^+_{\mathcal{D}}(\mathcal{F},S)\Big),$$

we may use union bound to deduce the desired result. ■

We finish with the matching lower bound.

Proposition 4.15. *There are universal constants $c, \varepsilon_0, \delta_0 > 0$ such that, for every $\gamma \in (0,1)$ and integer $d \ge 1$, there is a class $\mathcal{F} \subseteq [-1,+1]^{\mathcal{X}}$ with $\mathtt{fat}_\gamma(\mathcal{F}) = d$ satisfying the following. For every sufficiently small $\varepsilon, \delta > 0$, and*

$$n \le c\frac{d+\log(1/\delta)}{\varepsilon^2},$$

there is a distribution $\mathcal{D}$ such that

$$\mathop{\mathbb{P}}_{S\sim\mathcal{D}^n}\left[\mathtt{Rep}_{\mathcal{D}}(\mathcal{F},S) > \gamma+\varepsilon\right] > \delta.$$

Proof. Let $Y = [d]\times\{-1,+1\}$ and

$$\mathcal{H} = \{h_a : a \in \{-1,+1\}^d\}, \qquad h_a(j,y) = a_j y.$$

This symmetric class has 2^d functions and VC dimension d. Its uniform-convergence lower bound follows from the standard agnostic-learning lower bound for all binary predictors on $[d]$: their zero-one losses on Y are $(1-h_a)/2$; see [SSBD14, Theorem 6.8 and Chapter 28]. Thus, after adjusting universal constants, for the stated range of ε, δ, n there is a distribution $\mathcal{D}_2$ on Y such that

$$\mathop{\mathbb{P}}_{K\sim\mathcal{D}_2^n}\left[\mathtt{Rep}_{\mathcal{D}_2}(\mathcal{H},K) > 2\varepsilon\right] > \delta.$$

Set $\mathcal{X} = [0,1]\times Y$, let $\mathcal{B}$ be the measurable functions from $[0,1]$ to $(-1,+1)$, and define

$$\mathcal{F} = \{f_{h,b} : h\in\mathcal{H},\ b\in\mathcal{B}\}, \qquad f_{h,b}(t,z) = \left(1-\frac{\gamma}{2}\right)h(z) + \frac{\gamma}{2}b(t).$$

For a fixed h, any two functions $f_{h,b}$ and $f_{h,b'}$ differ by strictly less than γ at every point. Consequently, two different labelings of a γ-fat-shattered set cannot be realized using the same h. A shattered set of size k therefore requires $2^k \le |\mathcal{H}| = 2^d$, proving $\mathtt{fat}_\gamma(\mathcal{F}) \le d$. Conversely, take $b = 0$ and the points $(0,(j,+1))$, for $j\in[d]$. These are γ-fat-shattered with thresholds zero, since $2-\gamma > \gamma$. Hence $\mathtt{fat}_\gamma(\mathcal{F}) = d$.

Let $\mathcal{D}_1$ be uniform on $[0,1]$ and take $\mathcal{D} = \mathcal{D}_1\times\mathcal{D}_2$. For a sample $S = ((t_i,z_i))_{i=1}^n$, write $T = (t_i)_{i=1}^n$ and $K = (z_i)_{i=1}^n$. The symmetry of $\mathcal{H}$ and $\mathcal{B}$, and the independent choices of h and b in the supremum, give

$$\mathtt{Rep}_{\mathcal{D}}(\mathcal{F},S) = \left(1-\frac{\gamma}{2}\right)\mathtt{Rep}_{\mathcal{D}_2}(\mathcal{H},K) + \frac{\gamma}{2}\mathtt{Rep}_{\mathcal{D}_1}(\mathcal{B},T).$$

For every finite T, the supremum defining $\mathtt{Rep}_{\mathcal{D}_1}(\mathcal{B},T)$ is 2: choose b arbitrarily close to -1 on the sampled points and arbitrarily close to $+1$ elsewhere. Thus

$$\mathtt{Rep}_{\mathcal{D}}(\mathcal{F},S) = \gamma + \left(1-\frac{\gamma}{2}\right)\mathtt{Rep}_{\mathcal{D}_2}(\mathcal{H},K).$$

Since $1-\gamma/2 > 1/2$, the lower bound for $\mathcal{H}$ proves the proposition. ■

Theorem 1.1 (Restated). *Let $\mathcal{F} \subseteq [-1,+1]^{\mathcal{X}}$, let $\gamma \in (0,1)$, and suppose that $d := \mathtt{fat}_\gamma(\mathcal{F}) < \infty$. Then $\mathcal{F}$ satisfies γ-uniform convergence, with*

$$m_\gamma(\varepsilon,\delta) = O\left(\frac{d+\log(1/\delta)}{\varepsilon^2}\right) \qquad \varepsilon,\delta \in (0,1).$$

Moreover, for every $\gamma \in (0,1)$ and integer $d \geq 1$, there is a class $\mathcal{F}$ with $\mathtt{fat}_\gamma(\mathcal{F}) = d$ such that, for all sufficiently small $\varepsilon,\delta > 0$,

$$m_\gamma(\varepsilon,\delta) = \Omega\left(\frac{d+\log(1/\delta)}{\varepsilon^2}\right).$$

Proof of Theorem 1.1. The upper bound follows from Theorem 4.14, and the lower bound follows from Proposition 4.15. ■

Use of AI tools. The authors used ChatGPT to assist with writing, reviewing proofs, and developing parts of several arguments. Specifically it suggested applying ε-nets result to a fractional cover, provided Theorem 2.1 and Lemma 4.2, and helped optimize several of the technical results and lower bounds. The authors revised and verified the arguments presented in the paper and take responsibility for its contents.

References

[ABDCBH97] Noga Alon, Shai Ben-David, Nicolò Cesa-Bianchi, and David Haussler. Scale-sensitive dimensions, uniform convergence, and learnability. *J. ACM*, 44(4):615–631, July 1997.

[ABMT20] Karim Adiprasito, Imre Bárány, Nabil H. Mustafa, and Tamás Terpai. Theorems of Carathéodory, Helly, and Tverberg without dimension. *Discrete & Computational Geometry*, 64(2):233–258, 2020.

[AHHM22] 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. IEEE, 2022.

[AMM+26] Shashaank Aiyer, Yishay Mansour, Shay Moran, Han Shao, and Tom Waknine. Scale-sensitive shattering: Learnability and evaluability at optimal scale. *arXiv preprint arXiv:2605.13684*, 2026.

[BCT25] Romain Bourneuf, Pierre Charbit, and Stéphan Thomassé. A dense neighborhood lemma: Applications of partial concept classes to domination and chromatic number. In *2025 IEEE 66th Annual Symposium on Foundations of Computer Science (FOCS)*, pages 1–37, 2025.

[BDCK66] Jean Bretagnolle, Didier Dacunha-Castelle, and Jean-Louis Krivine. Lois stables et espaces L^p. *Annales de l'I.H.P. Probabilités et statistiques*, 2(3):231–259, 1966.

[BEHW89] Anselm Blumer, A. Ehrenfeucht, David Haussler, and Manfred K. Warmuth. Learnability and the vapnik-chervonenkis dimension. *J. ACM*, 36(4):929–965, October 1989.

[BIP26] Polina Barabanshchikova, Grigory Ivanov, and Alexander Polyanskii. Tight colorful no-dimensional Tverberg theorem. *Combinatorica*, 46(3), 2026. Article 25.

[BKP97] Peter L. Bartlett, Sanjeev R. Kulkarni, and Steven R. Posner. Covering numbers for real-valued function classes. *IEEE Transactions on Information Theory*, 43(5):1721–1724, 1997.

[BL98] 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.

[Can33] F. P. Cantelli. Sulla determinazione empirica delle leggi di probabilità. *Giornale dell'Istituto Italiano degli Attuari*, 4:421–424, 1933.

[CEP25] Roberto Colomboni, Emmanuel Esposito, and Andrea Paudice. An improved uniform convergence bound with fat-shattering dimension. *Information Processing Letters*, 188:106539, 2025.

[Cha92] Bernard Chazelle. A note on Haussler's packing lemma, 1992. Unpublished manuscript, Princeton University.

[CM22] Aruni Choudhary and Wolfgang Mulzer. No-dimensional Tverberg theorems and algorithms. *Discrete & Computational Geometry*, 68(4):964–996, 2022.

[DGZ91] Richard M Dudley, Evarist Giné, and Joel Zinn. Uniform and universal glivenko-cantelli classes. *Journal of Theoretical Probability*, 4(3):485–510, 1991.

[Dud78] Richard M Dudley. Central limit theorems for empirical measures. *The Annals of Probability*, pages 899–929, 1978.

[GKL20] Allan Grønlund, Lior Kamma, and Kasper Green Larsen. Near-tight margin-based generalization bounds for support vector machines. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 3779–3788. PMLR, 2020.

[Gli33] V. I. Glivenko. Sulla determinazione empirica delle leggi di probabilità. *Giornale dell'Istituto Italiano degli Attuari*, 4:92–99, 1933.

[Hau95] David Haussler. Sphere packing numbers for subsets of the boolean n-cube with bounded vapnik-chervonenkis dimension. *Journal of Combinatorial Theory, Series A*, 69(2):217–232, 1995.

[Hoe63] Wassily Hoeffding. Probability inequalities for sums of bounded random variables. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[HPR24] Sariel Har-Peled and Eliot W. Robson. No-dimensional Tverberg partitions revisited. In *19th Scandinavian Symposium and Workshops on Algorithm Theory (SWAT 2024)*, volume 294 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 26:1–26:14. Schloss Dagstuhl – Leibniz-Zentrum für Informatik, 2024.

[HW86] David Haussler and Emo Welzl. Epsilon-nets and simplex range queries. In *Proceedings of the second annual symposium on Computational geometry*, pages 61–71, 1986.

[IJN26] Grigory Ivanov, Attila Jung, and Márton Naszódi. The vc dimension of partial concept classes via radon's theorem. *arXiv preprint arXiv:2607.10751*, 2026.

[IK26] Grigory Ivanov and Vladimir Kadets. Optimality of no-dimensional bounds in Banach spaces, 2026. arXiv:2607.04093.

[Iva21] Grigory Ivanov. No-dimension Tverberg's theorem and its corollaries in Banach spaces of type p. *Bulletin of the London Mathematical Society*, 53(2):631–641, 2021.

[KMP16] Andrey Kupavskii, Nabil Mustafa, and János Pach. New lower bounds for ε-nets. In *32nd International Symposium on Computational Geometry (SoCG 2016)*, volume 51 of *Leibniz International Proceedings in Informatics (LIPIcs)*, pages 54:1–54:16, 2016.

[KS94] 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.

[KZ20] Andrey Kupavskii and Nikita Zhivotovskiy. When are ε-nets small? *Journal of Computer and System Sciences*, 110:22–36, 2020.

[Lon01] 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.

[McD89] Colin McDiarmid. On the method of bounded differences. In J. Siemons, editor, *Surveys in Combinatorics, 1989: Invited Papers at the Twelfth British Combinatorial Conference*, London Mathematical Society Lecture Note Series, pages 148–188. Cambridge University Press, Cambridge, 1989.

[MSW90] Jiří Matoušek, Raimund Seidel, and Emo Welzl. How to net a lot with little: Small ε-nets for disks and halfspaces. In *Proceedings of the Sixth Annual Symposium on Computational Geometry*, pages 16–22, 1990.

[MV03] Shahar Mendelson and Roman Vershynin. Entropy and the combinatorial dimension. *Inventiones Mathematicae*, 152(1):37–55, 2003.

[Pab26] Chirag Pabbaraju. Optimal unambiguous DNFs and Alon–Saks–Seymour, 2026. arXiv:2608.02533.

[Pol84] David Pollard. *Convergence of Stochastic Processes*. Springer Series in Statistics. Springer-Verlag, New York, 1984.

[PT11] János Pach and Gábor Tardos. Tight lower bounds for the size of epsilon-nets. In *Proceedings of the twenty-seventh annual symposium on Computational geometry*, pages 458–463, 2011.

[Rén07] Alfréd Rényi. *Probability theory*. Courier Corporation, 2007.

[RV06] Mark Rudelson and Roman Vershynin. Combinatorics of random processes and sections of convex bodies. *Annals of Mathematics*, pages 603–648, 2006.

[Spe85] Joel Spencer. Six standard deviations suffice. *Transactions of the American mathematical society*, 289(2):679–706, 1985.

[SSBD14] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning - From Theory to Algorithms*. Cambridge University Press, 2014.

[SST10] Nathan Srebro, Karthik Sridharan, and Ambuj Tewari. Smoothness, low noise and fast rates. In J. Lafferty, C. Williams, J. Shawe-Taylor, R. Zemel, and A. Culotta, editors, *Advances in Neural Information Processing Systems*, volume 23. Curran Associates, Inc., 2010.

[Tal94] Michel Talagrand. Sharper bounds for Gaussian and empirical processes. *The Annals of Probability*, 22(1):28–76, 1994.

[VC71] V. N. Vapnik and A. Ya. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability & Its Applications*, 16(2):264–280, 1971.