Scale-Sensitive Shattering: Learnability and Evaluability at Optimal Scale

Shashaank Aiyer*
University of Maryland
saiyer1@umd.edu

Yishay Mansour
Tel Aviv University and Google Research
mansour.yishay@gmail.com

Shay Moran
Technion and Google Research
smoran@technion.ac.il

Han Shao
University of Maryland
hanshao@umd.edu

Tom Waknine
Technion
tom.waknine@campus.technion.ac.il

Abstract

We study the optimal scales at which real-valued function classes exhibit uniform convergence and learnability. Our main result is a scale-sensitive generalization of the fundamental theorem of PAC learning: for every bounded real-valued class $\mathcal{F}$ and every $\gamma > 0$, uniform convergence at scale γ, agnostic learnability at scale $\gamma/2$, and finiteness of the fat-shattering dimension at every scale $\gamma' > \gamma$ are equivalent. This extends Long's (COLT '01) equivalence between agnostic learnability at scale $\gamma/2$ and fat-shattering at scale γ by adding uniform convergence at the corresponding optimal scale. Moreover, if $d = \mathrm{fat}_\gamma(\mathcal{F})$, our sample complexity is $\widetilde{O}(d/\varepsilon^2)$, improving Long's $\widetilde{O}(d/\varepsilon^3)$ bound and matching the optimal dependence on ε up to logarithmic factors. The key technical ingredient is a direct bound on empirical ℓ_∞ covering numbers, avoiding the standard detour through packing numbers. As a consequence, we obtain sharp asymptotic metric-entropy bounds in terms of the fat-shattering scale γ: an $O(\log^2 n)$ bound holds already at scale $\gamma/2$, while an $O(\log n)$ bound holds at scale 2γ. We further show that the $O(\log^2 n)$ bound is sometimes tight. These results settle the precise scale relationship between empirical covering numbers and fat-shattering dimension, and resolve questions raised and studied by Alon et al. (JACM '97), Anthony and Bartlett (Cambridge Univ. Press '99), and Rudelson and Vershynin (Ann. of Math. '06). As an application, we establish a sharp dichotomy for bounded integral probability metrics: every such IPM is either estimable or cannot be weakly evaluated within any multiplicative factor $c < 3$, while 3-weak evaluability always holds, resolving an open question from Aiyer et al. (ICML '26). We also highlight several open questions on quantitative sample complexity and evaluability.

1 Introduction

One of the central results in statistical learning theory is the fundamental theorem of PAC learning, which establishes an equivalence between three different notions: agnostic learnability, uniform

*Authors are ordered alphabetically.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

convergence, and finiteness of the VC dimension. In the binary classification setting, this theorem provides a complete characterization of when a class of functions can be learned from data in a distribution-free manner, and serves as a cornerstone of learning theory [VC71, BEHW89].

A natural extension of this framework considers real-valued function classes $\mathcal{F}$, where predictions are no longer binary but continuous. In this setting the VC dimension is replaced by a scale-sensitive extension called the fat-shattering dimension (defined below). This notion was introduced in learning theory by Kearns and Schapire [KS94], and also appears in approximation theory (see Lorentz [Lor86], and Tikhomirov [Tik60] who attributes it to earlier work of Kolmogorov).

Understanding the relationship between learnability, fat-shattering, and uniform convergence in the real-valued setting has been a long-standing goal; see, in particular, the detailed discussion in Anthony and Bartlett [AB99, Section 12.7]. The classical no-free-lunch argument shows that if the γ-fat-shattering dimension is unbounded, then the class is not learnable at scale $\gamma/2$ and does not satisfy uniform convergence at scale γ, reflecting a scale-sensitive version of the no-free-lunch principle. In the opposite direction, Alon et al. [ABDCBH97] showed that finite γ-fat-shattering dimension implies uniform convergence, and hence learnability, at scale $c\gamma$ for some universal constant $c \geq 10$. Bartlett and Long [BL98] improved this implication to every constant $c > 2$, and Long [Lon01] later showed that the learnability implication holds at the optimal scale: finite γ-fat-shattering dimension suffices for agnostic learnability at every scale strictly larger than $\gamma/2$.

The sample-complexity bounds were even less complete. At larger scales, Bartlett and Long obtained the standard ε^{-2} dependence on the excess error parameter ε, up to logarithmic factors; these logarithmic losses were later removed by [RV06, CEP25] via sharp asymptotic entropy bounds. At the optimal learnability scale, however, the only available bound was Long's, with the weaker dependence $\widetilde{O}(\varepsilon^{-3})$. Thus, Long's result closed the qualitative scale gap for learnability, but left open both the corresponding optimal-scale question for uniform convergence and the sharp quantitative dependence for learning at that scale.

Our first main result closes the remaining optimal-scale question for uniform convergence and improves the quantitative dependence for learning to near-optimal. To state the result, we briefly recall the relevant notions. Let $\mathcal{X}$ be a domain, let $\mathcal{F} \subset [-R, R]^{\mathcal{X}}$ be a class of bounded real-valued functions [2] over $\mathcal{X}$, and let $\gamma > 0$. We say that $\mathcal{F}$ satisfies γ-uniform convergence if for every $\varepsilon, \delta > 0$ there exists m such that for every distribution q over $\mathcal{X}$, with probability at least $1 - \delta$ over a sample $S \sim q^n$ of size $n \geq m$,

$$\sup_{f \in \mathcal{F}} \left| \mathbb{E}_{x \sim q}[f(x)] - \frac{1}{n} \sum_{x \in S} f(x) \right| \leq \gamma + \varepsilon.$$

We say that $\mathcal{F}$ is γ-learnable if there exists a learner $\mathcal{A}$ (that is, a mapping from samples to hypotheses) such that for every $\varepsilon, \delta > 0$ there exists m such that for every distribution q over $\mathcal{X} \times [-R, R]$, with probability at least $1 - \delta$ over $S \sim q^n$ of size $n \geq m$,

$$\mathop{\mathbb{E}}_{(x,y) \sim q} \left[\left| \mathcal{A}_S(x) - y \right| \right] \leq \inf_{f \in \mathcal{F}} \mathop{\mathbb{E}}_{(x,y) \sim q} \left[|f(x) - y| \right] + \gamma + \varepsilon,$$

where $\mathcal{A}_S$ denotes the hypothesis output by $\mathcal{A}$ on input sample S. Finally, the γ-fat-shattering dimension of $\mathcal{F}$, denoted $\mathrm{fat}_\gamma(\mathcal{F})$, is the largest integer d for which there exist points $x_1, \ldots, x_d \in \mathcal{X}$ and thresholds $r_1, \ldots, r_d \in \mathbb{R}$ such that for every $b \in \{\pm 1\}^d$ there exists $f \in \mathcal{F}$ satisfying

$$f(x_i) \geq r_i + \tfrac{\gamma}{2} \text{ if } b_i = +1, \quad \text{and} \quad f(x_i) \leq r_i - \tfrac{\gamma}{2} \text{ if } b_i = -1.$$

If for arbitrarily large d there exist points $x_1, \ldots, x_d$ and thresholds $r_1, \ldots, r_d$ satisfying the above, we say that $\mathrm{fat}_\gamma(\mathcal{F}) = \infty$.

When $\mathcal{F}$ consists of $\{0, 1\}$-valued functions and the scale $\gamma < 1$, the fat-shattering dimension coincides with the VC dimension, and these notions reduce to their classical binary counterparts. Thus, the real-valued framework can be viewed as a strict generalization of the binary setting.

Theorem 1. *Let $\mathcal{F} \subset [-R, R]^{\mathcal{X}}$ be a bounded function class and let $\gamma > 0$. Then, the following are equivalent:*

1. *$\mathcal{F}$ is $\frac{\gamma}{2}$-learnable;*

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

2. *$\mathcal{F}$ satisfies γ-uniform convergence;*

3. *for every $\gamma' > \gamma$, the γ'-fat-shattering dimension of $\mathcal{F}$ is finite.*

The proof appears in Section A.2. The implications $(1) \Rightarrow (3)$ and $(2) \Rightarrow (3)$ were established by [BL98]. The main new implication is $(3) \Rightarrow (2)$: finite fat-shattering dimension at the corresponding scale yields uniform convergence. Since uniform convergence implies agnostic learnability by the ERM principle, this also recovers Long's learnability implication [Lon01] ($(3) \Rightarrow (1)$), with sharper sample-complexity bounds matching the optimal dependence on the excess error parameter up to logarithmic factors. Removing these logarithmic factors remains open.

Covering Numbers. The derivation of uniform convergence from finite combinatorial dimension follows the classical approach of Vapnik and Chervonenkis. The argument proceeds via symmetrization (also known as double sampling), which reduces the problem to controlling the behavior of the class on a finite sample. In the binary case, the behavior of the class is controlled using the Sauer-Shelah Lemma [Sau72], which bounds the number of distinct labelings induced by the class on a sample of size n. This combinatorial bound allows one to replace an infinite union bound with a finite one.

In the real-valued setting, the role of the Sauer–Shelah Lemma is played by covering numbers. Concretely, for $\mathcal{F} \subset [-R, R]^{\mathcal{X}}$, the empirical ℓ_∞ covering number $N(\mathcal{F}, \gamma, n)$ is the smallest integer N for which for any sequence $x_1, \dots, x_n$, there exist functions $g_1, \dots, g_N : \mathcal{X} \to \mathbb{R}$ such that for every $f \in \mathcal{F}$, there exists $i \in [N]$ satisfying $\max_{j \in [n]} |f(x_j) - g_i(x_j)| \le \gamma$. The quantity

$$H(\mathcal{F}, \gamma, n) := \log N(\mathcal{F}, \gamma, n)$$

is called the (empirical) metric entropy, and serves as the appropriate analogue of the combinatorial growth function. A classical result of Dudley et al. [DGZ91] shows that sublinear entropy growth, namely $H(\mathcal{F}, \gamma/2, n) = o(n)$, implies γ-uniform convergence.

A long line of work has sought to bound this entropy in terms of the fat-shattering dimension. Alon et al. [ABDCBH97] showed that if $\mathrm{fat}_\gamma(\mathcal{F}) < \infty$, then

$$H(\mathcal{F}, 2\gamma, n) = O(\log^2 n)$$

In addition to establishing this bound, they explicitly asked whether the $\log^2 n$ dependence is inherent, or whether it can be improved to $O(\log n)$. Subsequent work of Rudelson and Vershynin [RV06] showed that if one allows the scale to increase further, then the entropy bound can be improved: for every $\varepsilon > 0$,

$$H\big(\mathcal{F}, \tfrac{c\gamma}{\varepsilon}, n\big) = O\big(\log^{1+\varepsilon} n\big).$$

They further asked whether the exponent of the logarithm can be reduced all the way to 1, i.e., whether an $O(\log n)$ bound is achievable.

These results are obtained by first controlling packing numbers and then passing to covering numbers, which necessarily incurs a loss in the scale. Our approach departs from this line of work by bounding covering numbers directly, without passing through packing arguments. This allows us to obtain sharp bounds at the optimal scale. Our second main result establishes the optimal scale at which the metric entropy becomes sublinear, and refines the known bounds on its growth across different scales.

Theorem 2. *Let $\mathcal{F} \subset [-R, R]^{\mathcal{X}}$ be a bounded function class, and let*

$$\gamma^\star = \inf\{\gamma > 0 : \mathrm{fat}_\gamma(\mathcal{F}) < \infty\}.$$

Then[a],

$$H(\mathcal{F}, \gamma, n) = \begin{cases} \Theta(n) & \gamma \in (0, \frac{\gamma^\star}{2}), \\ O(\log^2 n) & \gamma \in (\frac{\gamma^\star}{2}, 2\gamma^\star], \\ O(\log n) & \gamma \in (2\gamma^\star, R). \end{cases}$$

Moreover, the second regime is tight for $\gamma \in \left(\frac{\gamma^\star}{2}, \gamma^\star\right)$: there exist $\mathcal{F}$ s.t. for all $\gamma < \gamma^\star$,

$$H(\mathcal{F}, \gamma, n) = \Omega\left(\log^{2-o(1)} n\right).$$

The third regime is tight: there exist $\mathcal{F}$ such that $H(\mathcal{F}, \gamma, n) = \Omega(\log n)$ for $\gamma \in (2\gamma^\star, R)$.

[a]Throughout, the asymptotic notation treats n as the parameter tending to infinity; the implicit constants may depend on γ and on the fat-shattering dimension at scale γ.

The proof appears in Section A.3.

In light of the $O(\log^{1+\varepsilon} n)$ bounds of Rudelson and Vershynin [RV06] mentioned above, which left open whether the exponent can be improved to 1, our result resolves this question. It also answers the question of Alon et al. [ABDCBH97] on whether the $\log^2 n$ bound can be improved to $\log n$, by showing that such an improvement is possible at scales above $2\gamma^\star$, but impossible (up to lower-order terms) below $\gamma^\star$.

The first and third regimes are tight and well understood from prior work: below $\gamma^\star/2$, linear entropy growth holds for every class, while above $2\gamma^\star$ there exist classes with entropy $\Omega(\log n)$. Thus, on the lower bound side our main contribution is that the $O(\log^2 n)$ bound is sometimes tight (up to lower-order terms). We remark that the case $\gamma = \gamma^\star/2$ is not covered by the theorem, and there are simple examples exhibiting either behavior; see Proposition 4.

We leave open the problem of characterizing the precise behavior of the entropy in the intermediate range $\gamma^\star < \gamma \leq 2\gamma^\star$, as well as obtaining sharp sample complexity bounds for learning and uniform convergence at the optimal scale. We also leave open the problem of obtaining optimal relationships between the fat-shattering dimension and covering numbers under other ℓ_p metrics. Such covering numbers have been studied extensively, e.g. [BKP97, AB99, Men02, MV03].

In Section 1.2, we illustrate a simple application of our results by deriving closure properties and composition theorems for uniform convergence and learnability. We next turn to a different - and perhaps more surprising - application, where we use the sharp scale-sensitive results to resolve an open question of [AMMS26] concerning the evaluability and estimability of generative models. Remarkably, the *quantitative* improvement from a factor of 2 to 1 in the scale unlocks a *qualitative* characterization: a sharp dichotomy between estimability and evaluability. We elaborate on this connection in the following section.

1.1 Evaluability and Estimability of Generative Models

How can one test whether a generative model truly generalizes? Consider a model trained to generate images or short musical tunes from a dataset. When it produces new samples, are these genuinely novel, or merely recombinations - or even copies - of the training data? Given only sample access to the true distribution, can we reliably determine whether the generator generalizes well? Can indistinguishability between generated and real samples be tested from finite data?

In practice, such questions are addressed via evaluation metrics that quantify the discrepancy between the generated and target distributions. A recent work by [AMMS26] introduced a general framework for studying such metrics as tools for generalization testing. A central class of metrics in this framework is given by integral probability metrics (IPMs):

$$d_{\mathcal{F}}(q, q^\star) := \sup_{f \in \mathcal{F}} \left| \mathbb{E}_q[f] - \mathbb{E}_{q^\star}[f] \right|,$$

where $\mathcal{F}$ is a class of real-valued functions $f : \mathcal{X} \to \mathbb{R}$ defined on the data domain $\mathcal{X}$. Intuitively, $d_{\mathcal{F}}$ measures how well one can distinguish q from $q^\star$ using tests from $\mathcal{F}$.

Classical examples of IPMs include total variation distance, obtained when $\mathcal{F}$ is the set of all measurable functions bounded in $[0, 1]$, and the Wasserstein-1 distance, which arises when $\mathcal{F}$ is the class of 1-Lipschitz functions with respect to a given metric on $\mathcal{X}$.

We next formalize the setup and introduce evaluability and estimability, following the framework of [AMMS26].

Let $\mathcal{X}$ denote the data domain. A *model* is a probability distribution over $\mathcal{X}$, representing a generative model of the data. We let $\mathcal{M} = \Delta(\mathcal{X})$ denote the set of all probability distributions over $\mathcal{X}$. We assume there exists an unknown ground-truth model $q^\star \in \mathcal{M}$ from which data is sampled.

An evaluation algorithm $\mathcal{A}$ takes as input two candidate models $q_1, q_2 \in \mathcal{M}$ and an evaluation sample S_{eval} drawn from $q^\star$, and outputs one of the two models, interpreted as the model it judges to be better.

A common class of evaluation algorithms operates by assigning each model a score based on the evaluation data and selecting the model with the smaller score.

Evaluability. For $c \geq 1$, we say that $d_{\mathcal{F}}$ is *c-weakly evaluable* if there exists a sample complexity function $m_{\text{evl}} : (0,1)^2 \to \mathbb{N}$ and an evaluation algorithm $\mathcal{A}$ such that for every pair of models $q_1, q_2 \in \mathcal{M}$, every $\varepsilon, \delta \in (0,1)$ and every ground-truth model $q^\star$, given IID evaluation data $S_{\text{eval}} = \{x_1, \ldots, x_m\}$ of size $m \geq m_{\text{evl}}(\varepsilon, \delta)$ with $x_i \sim q^\star$, with probability at least $1 - \delta$,

$$\mathcal{A}(\{q_1, q_2\}, S_{\text{eval}}) = q_1 \implies d_{\mathcal{F}}(q_1, q^\star) \leq c \cdot d_{\mathcal{F}}(q_2, q^\star) + \varepsilon\,,$$

and

$$\mathcal{A}(\{q_1, q_2\}, S_{\text{eval}}) = q_2 \implies d_{\mathcal{F}}(q_2, q^\star) \leq c \cdot d_{\mathcal{F}}(q_1, q^\star) + \varepsilon\,.$$

Thus, evaluability asks whether two candidate models can be reliably compared using only samples from the ground-truth distribution. The evaluation algorithm need not estimate the true distances themselves; it only needs to select a model whose distance to $q^\star$ is no worse than that of the alternative, up to a multiplicative factor c and an additive error ε.

We say that $d_{\mathcal{F}}$ is *weakly evaluable with optimal scale* c if it is c-weakly evaluable and there does not exist any $c' < c$ such that it is c'-weakly evaluable. When $c = 1$, we say that $d_{\mathcal{F}}$ is *strongly evaluable*.

Estimability. We say that $d_{\mathcal{F}}$ is *estimable* if there exists a sample complexity function $m_{\text{est}} : (0,1)^2 \to \mathbb{N}$ and a score function s such that for any $\varepsilon, \delta \in (0,1)$, any ground-truth model $q^\star$, and any model $q \in \mathcal{M}$, given IID evaluation data $S_{\text{eval}} = \{x_1, \ldots, x_m\}$ of size $m \geq m_{\text{est}}(\varepsilon, \delta)$ with $x_i \sim q^\star$, with probability at least $1 - \delta$,

$$\left| s(q, S_{\text{eval}}) - d_{\mathcal{F}}(q, q^\star) \right| \leq \varepsilon.$$

In contrast to evaluability, which concerns relative comparison, estimability requires accurately approximating the *value* of the metric itself. In particular, estimability implies strong evaluability.

A central result of [AMMS26] establishes a dichotomy for IPMs induced by $\{0,1\}$-valued test functions: such metrics are either estimable (and hence strongly evaluable), or weakly evaluable with optimal scale $c = 3$. Their proof, however, is tailored to the $\{0,1\}$-valued setting, relying on tools such as VC-dimension and uniform convergence for binary-valued classes.

Extending this result to general real-valued test classes is significantly more challenging. In particular, the binary setting allows one to leverage VC-dimension arguments directly, whereas in the real-valued case the appropriate complexity measure is the fat-shattering dimension, and its connection to uniform convergence is more delicate. A key obstacle is that the classical results of [BL98] only guarantee uniform convergence at scale 2γ, which is insufficient for the evaluability guarantees required here.

Our approach overcomes this difficulty by establishing uniform convergence at the optimal scale γ. Crucially, eliminating this factor-2 loss is essential: without it, the argument breaks down and yields no non-trivial weak evaluability guarantee for real-valued IPMs.

Theorem 3. *Every bounded IPM is either:*

1. *estimable, and hence strongly evaluable; or*
2. *weakly evaluable with optimal scale $c = 3$.*

We further characterize this dichotomy in terms of the fat-shattering dimension of the test class.

Proposition 1. *Let $\mathcal{F}$ be a class of bounded real-valued functions.*

1. *If $\mathrm{fat}_\gamma(\mathcal{F}) < \infty$ for all $\gamma > 0$, then $d_{\mathcal{F}}$ is estimable (and hence strongly evaluable).*
2. *If there exists $\gamma \in (0, 1/2)$ such that $\mathrm{fat}_\gamma(\mathcal{F}) = \infty$, then $d_{\mathcal{F}}$ is weakly evaluable with optimal scale $c = 3$.*

The proof is in Section B.2. The above results are qualitative in nature and are most meaningful over infinite domains. Indeed, over a finite domain $\mathcal{X}$ of size N, every bounded IPM is estimable with $O(N/\varepsilon^2)$ samples, and hence strongly evaluable.

This raises the question of whether quantitative analogues of the above dichotomy hold in the finite-domain setting. For example, as a function of the domain size N, how does the sample complexity of weak evaluability behave as the approximation factor improves from $c = 3$ to $c = 1$? Moreover, is the sample complexity required for strong evaluability comparable to that required for estimability?

Understanding these questions would provide a more refined picture of the trade-offs between sample complexity and evaluation in practical settings.

1.2 Composition and Closure Properties

We now return to the setting of real-valued function learning and uniform convergence, and describe a more direct and standard application of Theorem 1, in contrast to the more involved application to evaluability discussed above.

Characterizations of learnability using combinatorial dimensions are ubiquitous in learning theory, and a natural litmus test for their usefulness is whether they can be used to derive new results that are otherwise harder to prove. While the usefulness of the fat-shattering dimension is well established, we use it here to illustrate another application of Theorem 1.

Specifically, we show that the properties of being γ-learnable and of satisfying γ-uniform convergence are invariant under natural operations. By Theorem 1, this reduces to showing that the property of having finite γ-fat-shattering dimension is invariant under these operations. The latter is essentially well known and relies on arguments from prior work, which we include here for completeness. The novel ingredient is therefore Theorem 1, which establishes the equivalence between finite fat-shattering dimension and learnability and uniform convergence at the optimal scale, thereby yielding closure properties for these notions.

As a first example, consider the dual class. For a function class $\mathcal{F} \subset [0,1]^{\mathcal{X}}$, its dual class $\mathcal{F}^\star$ is a class of functions over $\mathcal{F}$, defined by

$$\mathcal{F}^\star = \{\delta_x : x \in \mathcal{X}\},$$

where each $\delta_x : \mathcal{F} \to [0,1]$ is given by $\delta_x(f) = f(x)$. Assouad's bound [Ass83] relates the VC dimension of a class to that of its dual. This result extends naturally to the setting of γ-shattering when a common threshold $r \in (0,1)$ is used across all points. A simple application of the pigeonhole principle then shows that $\mathrm{fat}_\gamma(\mathcal{F})$ is finite then $\mathrm{fat}_{\gamma+\varepsilon}(\mathcal{F}^\star)$ is finite for every $\varepsilon > 0$. As a corollary, we obtain the following result, which shows that γ-uniform convergence is preserved under taking duals.

Proposition 2. *Let $\gamma > 0$, and let $\mathcal{F}$ be a bounded function class that satisfies γ-uniform convergence. Then the dual class $\mathcal{F}^\star$ also satisfies γ-uniform convergence.*

The proof is in Section A.4.

Another closure property is stability under aggregation. Let $G : [0,1]^k \to [0,1]$ be a function, and let $f_1, \ldots, f_k : \mathcal{X} \to [0,1]$. Their aggregation with respect to G is defined by

$$G(f_1, \ldots, f_k)(x) = G\big(f_1(x), \ldots, f_k(x)\big).$$

Given a sequence of function classes $\mathcal{F}_1, \ldots, \mathcal{F}_k$, we define the aggregated class

$$G(\mathcal{F}_1, \ldots, \mathcal{F}_k) = \{\, G(f_1, \ldots, f_k) : f_i \in \mathcal{F}_i \text{ for all } i \,\}.$$

The fat-shattering dimension of such aggregated classes was studied in [AK24]. Our next result is similar in spirit to Theorem 2 of [AK24], but avoids the blow-up in the fat-shattering scale. As a consequence, we obtain that uniform convergence is closed under aggregation.

Proposition 3. *Let $\mathcal{F}_1, \ldots, \mathcal{F}_k \subset [0,1]^{\mathcal{X}}$ be bounded function classes, each satisfying γ-uniform convergence, and let $G : [0,1]^k \to [0,1]$ be a 1-Lipschitz function with respect to the ℓ_∞ metric. Then the aggregated class $G(\mathcal{F}_1, \ldots, \mathcal{F}_k)$ also satisfies γ-uniform convergence.*

The proof is in Section A.4, and follows as a consequence of Theorem 1, by showing that an aggregation of covers is a cover for the aggregated class.

2 Related Work

The relationship between fat-shattering dimension, uniform convergence, and covering numbers is well established in statistical learning theory. The characterization of uniform convergence via covering numbers is due to Dudley [Dud78, Dud84, DGZ91]. The fat-shattering dimension was introduced by Kearns and Schapire [KS94] as a complexity measure capturing the sample complexity of approximate learnability. In a seminal work, Alon et al. [ABDCBH97] showed that finite γ-fat-shattering dimension implies a $O(\log^2 n)$ bound on the metric entropy at scale 2γ, thereby establishing a connection between fat-shattering and uniform convergence. They also conjectured that the $\log^2 n$ dependence could be improved to $\log n$.

Building on these ideas, Bartlett and Long [BL98] made significant progress toward understanding the optimal dependence on the scale. In particular, they improved the result of [ABDCBH97] by showing that finite γ-fat-shattering dimension implies a $O(\log^2 n)$ bound on the metric entropy at scale $\gamma + \varepsilon$. They also established complementary lower bounds which, combined with classical results of Dudley and Pollard [DGZ91, Pol84], yield an equivalence between γ-fat-shattering and uniform convergence up to a factor of 2 in the scale. In subsequent work, Long [Lon01] gave improvements on the scale at which one can get sub-liner covering bounds and gave a tight equivalence between finite fat-shattering dimension and learnability. A comprehensive overview of those techniques appears in [AB99].

A complementary line of work focuses on improving covering number bounds themselves, rather than the scale at which they are obtained. These works often consider ℓ_1 or ℓ_2 metrics instead of ℓ_∞, leading to improved quantitative bounds on sample complexity. Early contributions include [BKP97, CBH98], followed by a sequence of works by Mendelson and Vershynin [MV02, MV03, Men02], which developed refined entropy bounds. This line of research culminated in the work of Rudelson and Vershynin [RV06], who obtained strong covering number bounds under various norms. Notably, they made progress toward the conjecture of [ABDCBH97] by improving the $\log^2 n$ dependence to $\log^{1+\varepsilon} n$, at the cost of a multiplicative factor of $1/\varepsilon$ in the scale. They further conjectured that this dependence could be improved to $O(\log n)$.

There are several additional directions related to these questions. Colomboni et al. [CEP25] used refined covering bounds to derive sharp sample complexity guarantees for uniform convergence. Hanneke et al. [HKS19] and Attias et al. [AHKS24] established connections between fat-shattering dimension and approximate sample compression schemes. Finally, recent works such as [AK24, KS23] study structural properties of fat-shattering dimension, including closure under aggregation and duality.

The framework of evaluability was recently introduced by [AMMS26], which provides a theoretical formulation of the problem of statistically evaluating the performance of generative models from finite samples. This line of work is motivated in part by limitations of commonly used empirical scores, such as perplexity and its variants [FWL+25, HHT+24, PSZ+21, MMM20]. Commonly used metrics include KL divergence, total variation distance, and other divergences.

One general class of metrics is integral probability metrics (IPMs), introduced by [Mü97], which subsume important examples such as total variation distance and the Wasserstein-1 distance. IPMs have been extensively studied in the context of density estimation, where, given sample access to a ground-truth distribution and a class of candidate distributions, the goal is to output the closest distribution in the class. In this setting, [Yat85] established an upper bound with multiplicative factor 3, and [BKM19] showed that this factor is optimal. [AMMS26] extended this factor of 3 from density estimation to evaluability for binary $\mathcal{F}$. In this work, we further generalize these results to real-valued function classes.

3 Proof Overviews

3.1 Proof Overview of Theorem 1

The fact that a finite fat-shattering dimension is a necessary condition for uniform convergence and learnability is well established, and follows by considering the uniform distribution over large shattered sets; see, Theorem 22 and Theorem 26 in [BL98]. The main challenge is to show the converse direction, namely that finite fat-shattering dimension implies uniform convergence and learnability.

The standard approach for deriving uniform convergence bounds, as in [BL98, ABDCBH97, CEP25], proceeds via symmetrization combined with classical probabilistic arguments, reducing the problem to bounding covering numbers. Concretely, one needs to show that for all $\varepsilon > 0$,

$$\lim_{n\to\infty} \frac{H(\mathcal{F}, \frac{\gamma}{2} + \varepsilon, n)}{n} = 0,$$

where we recall $H(\mathcal{F}, \gamma, n) = \log N(\mathcal{F}, \gamma, n)$ denotes the empirical metric entropy of $\mathcal{F}$ with respect to the ℓ_∞ metric.

In prior works, the covering number is typically upper bounded via the packing number, which is easier to control. This reduction allows one to assume the existence of a well-separated set and bound its size using Sauer-type lemmas or volume arguments. However, this approach has an inherent limitation: a finite γ-fat-shattering dimension only guarantees a bound on the γ-packing number. This, in turn, yields a bound on the γ-covering number, leading to 2γ-uniform convergence and γ-learnability. To improve upon the result of [BL98], one must therefore bound the covering number directly, without relying on packing arguments.

This introduces the main technical difficulty: *bounding the covering number directly requires constructing an explicit cover, which may not consist of functions from the class itself.* To address this, we take a similar approach to [Lon01], and employ the theory of partial concept classes [Lon01, AHHM22].

Let $R > 0$, and let $\mathcal{F} \subset [-\frac{R}{2}, \frac{R}{2}]^n$ be a bounded function class[3] with finite γ-fat-shattering dimension $\mathrm{fat}_\gamma(\mathcal{F}) = d$. We associate to each $f \in \mathcal{F}$ a partial binary function h_f defined by

$$h_f(x) = \begin{cases} 1 & f(x) \geq \frac{\gamma}{2}, \\ -1 & f(x) \leq -\frac{\gamma}{2}, \\ \text{undefined} & \text{otherwise.} \end{cases}$$

The VC dimension of the induced partial function class will be bounded by d. Hence, by [AHHM22], it admits a disambiguation of sub-exponential size: that is, there exists a set $H \subset \{-1, 1\}^n$ with $|H| = n^{O(d \log n)}$ such that for every $f \in \mathcal{F}$, there exists $h \in H$ satisfying $h(x) = h_f(x)$ whenever $h_f(x)$ is defined.

This implies that if $h(x) = 1$, then $f(x) \in [-\frac{\gamma}{2}, \frac{R}{2}]$, and if $h(x) = -1$, then $f(x) \in [-\frac{R}{2}, \frac{\gamma}{2}]$. Consequently, by associating, with each $h \in H$, a function ϕ_h, such that $\phi_h(x) = \frac{R-\gamma}{4}$ when $h(x) = 1$ and $\phi_h(x) = \frac{-R+\gamma}{4}$ when $h(x) = -1$, we obtain a cover of $\mathcal{F}$ by $n^{O(d \log n)}$ boxes of radius $\frac{\gamma+R}{4}$.

To obtain finer scales, we iterate this construction. Starting from a coarse cover and recursively refining each box, the radius decreases geometrically and converges to $\frac{\gamma}{2}$, while at each step we incur a multiplicative increase of $n^{O(d \log n)}$ in the covering number size. Thus after k iterations we have a bound of $n^{O(dk \log n)}$ on the covering number size at scale $\frac{\gamma}{2} + \frac{1}{2^k}$, yielding an effective bound on the covering number at all scales strictly larger than $\frac{\gamma}{2}$

$$H(\mathcal{F}, \tfrac{\gamma}{2} + \varepsilon, n) = O(d \log \frac{1}{\varepsilon} \log^2 n).$$

The $\frac{\gamma}{2}$-learnability result follows from a covering argument based on the algorithm of Theorem 21 in [BL98], combined with our improved covering bound to obtain a sharp rate. Given a sample $(x_i, y_i)_{i=1}^{2n}$ and a test point x, the algorithm constructs a $(\frac{\gamma}{2} + \varepsilon)$-cover of the restriction of $\mathcal{F}$ to these $2n + 1$ points. It then outputs $f(x)$, where $f \in \mathcal{F}$ minimizes the empirical loss over the first n points. Standard concentration inequalities ensure that the excess loss is controlled by the covering scale, provided the size of the cover is appropriately bounded.

3.2 Proof Overview of Theorem 2

We begin with the first regime. Suppose a set of size n is γ-shattered. Then, for any $\gamma' < \gamma/2$, every γ'-cover of its projection must contain at least 2^n functions. It follows that $H(\mathcal{F}, \gamma', n) = \Omega(n)$, as

[3] It suffices to assume that $\mathcal{F}$ is inside an ℓ_∞ ball of radius $\frac{R}{2}$. The assumption it is centered at 0 is for notational convenience

observed in [ABDCBH97, BL98]. On the other hand, standard volume estimates yield the existence of a γ-cover of size at most $(1/\gamma)^n$. Hence, for $\gamma \in (0, \gamma^\star/2)$,

$$H(\mathcal{F}, \gamma, n) = \Theta(n).$$

The $O(\log^2 n)$ upper bound in the intermediate regime follows from the same covering argument used in the proof of Theorem 1.

To establish tightness for $\gamma \in (\gamma^\star/2, \gamma^\star)$, we appeal to a result of [AHHM22], which shows that the disambiguation bound is tight. Specifically, they construct a partial concept class $\mathcal{F}$ with finite VC dimension for which any disambiguation must have size at least $n^{(\log n)^{(1-o(1))}}$. We embed this class into the real-valued setting by mapping 1-labels to $\gamma^\star$, -1-labels to $-\gamma^\star$, and undefined points to 0.

In this construction, any γ-cover with $\gamma \in (\gamma^\star/2, \gamma^\star)$ induces a disambiguation as follows: given a γ-cover H, define $\Phi_h(x) = \operatorname{sign}(h(x))$. If $|f(x) - h(x)| < \gamma^\star$ for all x, then f and h agree in sign on all points where $f(x) \neq 0$. Thus Φ_h disambiguates the original partial class, yielding the desired lower bound.

Finally, we consider the regime $\gamma \in (2\gamma^\star, 1)$, where we rely on sample compression schemes. We begin by briefly recalling the notion of sample compression and how it leads to covering bounds.

Formally, we say that $\mathcal{F}$ admits a γ-sample compression scheme of size k if there exists a reconstruction function

$$\rho : (\mathcal{X} \times \mathbb{R})^k \to \mathbb{R}^{\mathcal{X}}$$

such that for every finite set $X \subseteq \mathcal{X}$ and every function $f \in \mathcal{F}$, there exists a sequence

$$T = \{(x_i, f(x_i))\}_{i=1}^k, \quad x_i \in X,$$

for which the reconstructed function $\rho(T)$ satisfies

$$|\rho(T)(x) - f(x)| \leq \gamma \quad \text{for all } x \in X.$$

Such a compression scheme yields a cover of polynomial size when $k = O(1)$ (i.e., independent of the size of X), since discretizing the function values at scale ε and enumerating all possible reconstructions from k labeled points yields at most $(n/\varepsilon)^k$ candidate functions. This gives a cover of size polynomial in n, and hence a logarithmic bound on the metric entropy.

It therefore remains to construct such a compression scheme. We follow the approach of [MY16], who studied sample compression in the binary-labeled setting, and adapt their argument to the real-valued case. Fix an empirical risk minimization (ERM) rule that maps any realizable dataset $T = \{(x_i, y_i)\}$ (that is, a dataset for which there exists $f \in \mathcal{F}$ with $y_i = f(x_i)$ for all i) to a hypothesis $h_T \in \mathcal{F}$ that is consistent with the dataset.

By uniform convergence, if we take T to be of size $m = m(\varepsilon, 1/3)$, where $m(\varepsilon, \delta)$ is the uniform convergence sample complexity (and $1/3$ can be replaced by any constant $\delta < 1$), then for any finite domain $X = \{x_1, \ldots, x_n\}$, any distribution p over X, and any function $f \in \mathcal{F}$, there exists a realizable dataset $T \subseteq X$ (with labels given by f) such that the corresponding hypothesis h_T satisfies

$$\mathbb{E}_{x \sim p}[|f(x) - h_T(x)|] \leq \gamma + \varepsilon.$$

This follows from the uniform convergence guarantee applied to a random draw of T from p.

A minimax argument now implies that there exists a distribution over such datasets T such that, when T is drawn from this distribution, the function $h_T(x)$ is close to $f(x)$ in expectation for every point $x \in X$.

We then use uniform convergence for the dual class to sparsify this distribution. Specifically, we approximate it by a uniform distribution over a finite multiset of datasets $T_1, \ldots, T_{m^\star}$, where $m^\star = m^\star(\varepsilon, \frac{1}{3})$ is the uniform convergence sample complexity of the dual class at scale γ. This sparsification incurs an additional error of $\gamma + \varepsilon$.

Averaging the corresponding hypotheses h_{T_i} then yields a function that is $(2\gamma + O(\varepsilon))$-close to f on all of X, giving the desired compression scheme.

Crucially, this argument relies on the fact that the dual class satisfies γ-uniform convergence. This is not automatic, and follows from our Theorem 1 (see Proposition 2). Without this property, the sparsification step would not be possible.

3.3 Proof Overview of Theorem 3

Theorem 3 follows as a direct corollary of Proposition 1. Indeed, the proposition provides a complete characterization of the behavior of $d_{\mathcal{F}}$ in terms of the fat-shattering dimension: if $\mathrm{fat}_\gamma(\mathcal{F})$ is finite for all $\gamma > 0$ then $d_{\mathcal{F}}$ is estimable, while if $\mathrm{fat}_\gamma(\mathcal{F}) = \infty$ for some $\gamma \in (0, 1/2)$ then $d_{\mathcal{F}}$ is weakly evaluable with optimal scale 3. Thus, we focus on the proof of Proposition 1.

The first item and the first part of the second item (that $d_{\mathcal{F}}$ is 3-weakly evaluable) follow from standard arguments presented in [AMMS26]. The main challenge is to show that the factor 3 is optimal. The proof builds on the argument of [BKM19], which applies to total variation distance - an example of an IPM induced by a class with infinite fat-shattering dimension. While their proof is technically involved, most of the required adaptations are straightforward. We therefore focus on the main conceptual step: reducing a general IPM with infinite fat-shattering dimension to the total variation setting.

The result of [BKM19] shows that on a domain of size n, any evaluation algorithm with approximation factor strictly smaller than 3 requires at least $\sqrt{n}$ samples. Letting $n \to \infty$, this rules out c-evaluability for any $c < 3$ on infinite domains. Thus, it suffices to embed arbitrarily large total variation instances into $d_{\mathcal{F}}$.

Let $\gamma^\star = \inf\{\gamma : \mathrm{fat}_\gamma(\mathcal{F}) < \infty\}$, and fix $\gamma^- < \gamma^\star < \gamma^+$ arbitrarily close to $\gamma^\star$. Let $d = \mathrm{fat}_{\gamma^+}(\mathcal{F})$. Since $\mathrm{fat}_{\gamma^-}(\mathcal{F}) = \infty$, we can find arbitrarily large γ^--shattered sets. Let $X = \{x_1, \ldots, x_N\}$ be such a set, and let P be the uniform distribution over X.

We sample n independent batches $X_1, \ldots, X_n$, where each batch X_i consists of $m = \widetilde{O}(d/\varepsilon^2)$ i.i.d. samples from P. With high probability, the following properties hold:

1. The batches $X_1, \ldots, X_n$ are pairwise disjoint (by taking N large enough to avoid collisions).
2. For every pair $i \neq j$, the distance between the corresponding empirical distributions satisfies
$$d_{\mathcal{F}}(X_i, X_j) \leq \gamma^+ + \varepsilon.$$
This follows from a uniform convergence argument at scale γ^+

Let P_i denote the uniform distribution over X_i, and let Δ be the set of all convex combinations of $\{P_1, \ldots, P_n\}$. For any mixtures $P_\alpha = \sum_i \alpha_i P_i$ and $P_\beta = \sum_i \beta_i P_i$, we show that
$$d_{\mathcal{F}}(P_\alpha, P_\beta) \approx \gamma^\star \cdot \mathrm{TV}(\alpha, \beta).$$

The lower bound uses the γ^--shattering property. Since the batches are disjoint subsets of the shattered set, for the subset of indices $S = \{i : \alpha_i > \beta_i\}$, there exists a function $f \in \mathcal{F}$ that separates $\bigcup_{i \in S} X_i$ from its complement at scale γ^-. This function witnesses a discrepancy of
$$\gamma^- \cdot \mathrm{TV}(\alpha, \beta).$$

The upper bound uses the fact that, by uniform convergence, any two batch distributions P_i and P_j are close under $d_{\mathcal{F}}$: namely, $d_{\mathcal{F}}(P_i, P_j) \leq \gamma^+ + \varepsilon$. A triangle-inequality argument then shows that moving mass between the P_i's can increase $d_{\mathcal{F}}$ by at most roughly $\gamma^+ + \varepsilon$ times the total amount of mass moved, giving
$$d_{\mathcal{F}}(P_\alpha, P_\beta) \leq (\gamma^+ + \varepsilon)\mathrm{TV}(\alpha, \beta).$$
Since γ^- and γ^+ can be chosen arbitrarily close to $\gamma^\star$, and since ε can be taken arbitrarily small, these two bounds match up to a factor arbitrarily close to 1.

This reduction embeds a total variation problem of dimension n into $d_{\mathcal{F}}$, allowing us to invoke the lower bound argument of [BKM19].

A crucial feature of the construction is that γ^- and γ^+ can be taken arbitrarily close to $\gamma^\star$. This ensures that, for mixtures in Δ, the comparison between $d_{\mathcal{F}}$ and total variation has distortion arbitrarily close to 1:
$$\gamma^- \cdot \mathrm{TV}(\alpha, \beta) \lesssim d_{\mathcal{F}}(P_\alpha, P_\beta) \lesssim \gamma^+ \cdot \mathrm{TV}(\alpha, \beta).$$

The critical point is that the γ^+, γ^- must be taken arbitrarily close to each other, which require uniform convergence at the optimal scale.

Acknowledgments and Disclosure of Funding

Yishay Mansour received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 882396), by the Israel Science Foundation (1357/24), the Blavatnik foundation at Tel Aviv University and a grant from the Tel Aviv University Center for AI and Data Science (TAD).

Shay Moran acknowledges support by Israel by ISF grant 2362/26, by the Technion Center for Machine Learning and Intelligent Systems (MLIS), and by the European Union (ERC, GENERALIZATION, 101039692). Views and opinions expressed are, however, those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

Han Shao acknowledges support from an Adobe Research gift.

References

[AB99] Martin Anthony and Peter L Bartlett. *Neural network learning: Theoretical foundations*. cambridge university press, 1999.

[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.

[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.

[AHKS24] Idan Attias, Steve Hanneke, Aryeh Kontorovich, and Menachem Sadigurschi. Agnostic sample compression schemes for regression. In *Forty-first International Conference on Machine Learning*, 2024.

[AK24] Idan Attias and Aryeh Kontorovich. Fat-shattering dimension of k-fold aggregations. *Journal of Machine Learning Research*, 25(144):1–29, 2024.

[AMMS26] Shashaank Aiyer, Yishay Mansour, Shay Moran, and Han Shao. A theoretical framework for statistical evaluability of generative models. *arXiv preprint arXiv:2604.05324*, 2026.

[Ass83] Patrick Assouad. Densité et dimension. In *Annales de l'institut Fourier*, volume 33, pages 233–282, 1983.

[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.

[BKM19] Olivier Bousquet, Daniel Kane, and Shay Moran. The optimal approximation factor in density estimation. In *Conference on Learning Theory*, pages 318–341. PMLR, 2019.

[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.

[CBH98] Nicolò Cesa-Bianchi and David Haussler. A graph-theoretic generalization of the sauer-shelah lemma. *Discrete Applied Mathematics*, 86(1):27–35, 1998. Vapnik-Chervonenkis dimension.

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

[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.

[Dud84] Richard M. Dudley. A course on empirical processes. 1984.

[FWL+25] Lizhe Fang, Yifei Wang, Zhaoyang Liu, Chenheng Zhang, Stefanie Jegelka, Jinyang Gao, Bolin Ding, and Yisen Wang. What is wrong with perplexity for long-context language modeling?, 2025.

[HHT+24] Yutong Hu, Quzhe Huang, Mingxu Tao, Chen Zhang, and Yansong Feng. Can perplexity reflect large language model's ability in long text understanding?, 2024.

[HKS19] Steve Hanneke, Aryeh Kontorovich, and Menachem Sadigurschi. Sample compression for real-valued learners. In *Algorithmic Learning Theory*, pages 466–488. PMLR, 2019.

[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.

[KS23] Pieter Kleer and Hans Simon. Primal and dual combinatorial dimensions. *Discrete Applied Mathematics*, 327:185–196, 2023.

[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.

[Lor86] G. G. Lorentz. *Approximation of functions*. Chelsea Pub. Co., New York, N.Y, 2nd ed. edition, 1986.

[M+89] Colin McDiarmid et al. On the method of bounded differences. *Surveys in combinatorics*, 141(1):148–188, 1989.

[Men02] Shahar Mendelson. Rademacher averages and phase transitions in Glivenko–Cantelli classes. *IEEE Transactions on Information Theory*, 48(1):251–263, 2002.

[MMM20] Pedro Henrique Martins, Zita Marinho, and André F. T. Martins. Sparse text generation, 2020.

[MV02] Shahar Mendelson and Roman Vershynin. Entropy, combinatorial dimensions and random averages. In *International Conference on Computational Learning Theory*, pages 14–28. Springer, 2002.

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

[MY16] Shay Moran and Amir Yehudayoff. Sample compression schemes for vc classes. *Journal of the ACM (JACM)*, 63(3):1–10, 2016.

[Mü97] Alfred Müller. Integral probability metrics and their generating classes of functions. *Advances in Applied Probability*, 29(2):429–443, 1997.

[Pol84] David Pollard. Convergence of stochastic processes. 1984.

[PSZ+21] Krishna Pillutla, Swabha Swayamdipta, Rowan Zellers, John Thickstun, Sean Welleck, Yejin Choi, and Zaid Harchaoui. Mauve: Measuring the gap between neural text and human text using divergence frontiers, 2021.

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

[Sau72] Norbert Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13(1):145–147, 1972.

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

[Tik60] V. M. Tikhomirov. Diameters of sets in function spaces and the theory of best approximations. *Russian Mathematical Surveys*, 15(3):75–111, 1960.

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

[Yat85] Yannis G. Yatracos. Rates of convergence of minimum distance estimators and kolmogorov's entropy. *The Annals of Statistics*, 13(2):768–774, 1985.

A Proofs of Scale-Sensitive Learnability Results

A.1 Covering Lemmas

Lemma 1. *Let $\mathcal{F} \subset [\frac{-r}{2}, \frac{r}{2}]^{\mathcal{X}}$ be a bounded set of functions, let $\gamma \in (0, r)$, and denote $\mathrm{fat}_\gamma(\mathcal{F}) = d > 0$. Then for any $n > 1$ we have*

$$N(\mathcal{F}, \frac{r+\gamma}{4}, n) \leq n^{7d \log n}$$

Proof. By restricting to an arbitrary set of size n we may assume that $|\mathcal{X}| = n$, and we wish to find a cover for the function class $\mathcal{F}$. For each $f \in \mathcal{F}$ define the partial function h_f by

$$h_f(x) = \begin{cases} 1 & \text{if } f(x) \geq \frac{\gamma}{2}, \\ -1 & \text{if } f(x) \leq -\frac{\gamma}{2}, \\ \text{undefined} & \text{otherwise.} \end{cases}$$

By the definition of the fat-shattering dimension, the VC dimension of the induced partial concept class

$$\{h_f : f \in F\}$$

is at most d. Hence, by Theorem 13 of [AHHM22], there exists a set of functions $H \subset \{-1, 1\}^{\mathcal{X}}$ of size

$$|H| \leq (n+1)^{(d+1)\log_2(n)+2} \leq n^{7d \log n},$$

that disambiguates this partial concept class; that is, for every $f \in F$ there exists some $h \in H$ such that $h(x) = h_f(x)$ whenever $h_f(x)$ is defined.

For each $h \in H$, define $\Phi_h : \mathcal{X} \to [-r, r]$ by

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

We claim that $\{\Phi_h \ : \ h \in H\}$ is the desired $\frac{r+\gamma}{4}$-cover. Indeed for any $f \in F$, if h disambiguates h_f, then for any $x \in X$ with $h(x) = 1$ we have $f(x) \in (\frac{-\gamma}{2}, \frac{r}{2}]$, and for any $x \in X$ with $h(x) = -1$ we have $f(x) \in [\frac{-r}{2}, \frac{\gamma}{2})$. Thus we indeed get that for all $x \in \mathcal{X}$ we have

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

This gives the desired result. □

Lemma 2 (Covering Number Bound)**.** *Let $\mathcal{F} \subset [-R, R]^{\mathcal{X}}$ be a bounded function class, let $\gamma > 0$, and denote $\mathrm{fat}_\gamma(\mathcal{F}) = d > 0$. Then for any $\varepsilon > 0$, and $n > 1$ we have*

$$N(\mathcal{F}, \tfrac{\gamma}{2} + \varepsilon, n) \leq n^{14d \log \frac{2R}{\varepsilon} \log n},$$

Proof. By restricting to an arbitrary set of size n we may assume that $|\mathcal{X}| = n$, and we wish to find a cover for the function class $\mathcal{F}$. We prove by induction on k that for every integer $k > 0$,

$$N\left(\mathcal{F}, \tfrac{\gamma}{2} + \frac{2R}{2^{k+1}}\right) \leq n^{7dk \log n}.$$

Choosing k such that $\frac{2R}{2^{k+1}} \leq \varepsilon \leq \frac{2R}{2^k}$ then yields the desired bound.

For the base case $k = 0$, observe that $\mathcal{F}$ can be covered by a single ball of radius R in the ℓ_∞ metric.

For the inductive step, assume that $\mathcal{F}$ can be covered by $n^{7d(k-1)\log n}$ balls of radius $\frac{\gamma}{2} + \frac{2R}{2^k}$. For each such ball F with center f we may apply Lemma 1 to the translated class $F - f = \{g - f \ : \ g \in \mathcal{F}\}$, and find a covering of F of size $n^{7d \log n}$ with balls of radius $\frac{\gamma + (\gamma + \frac{4R}{2^k})}{4} = \frac{\gamma}{2} + \frac{2R}{2^{k+1}}$. Combining the bounds give a $(\frac{\gamma}{2} + \frac{2R}{2^{k+1}})$-covering of size

$$n^{7(k-1)d \log n} n^{7d \log n} = n^{7kd \log n}.$$

From which we deduce the desired result. □

A.2 Proof for Theorem 1

The following is a standard refinement of an argument by [DGZ91] relating uniform convergence and covering number. Similar results can be found in [AB99, BL98, Pol84]

Lemma 3 (Uniform Convergence Rate vs Covering Number)**.** *Let $\mathcal{F} \subseteq [-R,R]^{\mathcal{X}}$ be a bounded function class, and let $\gamma > 0$. Then for any distribution q over $\mathcal{X}$ and any $\delta, \varepsilon > 0$, if*

$$n > \frac{128R^2}{\varepsilon^2}\left(H\left(\mathcal{F}, \frac{\gamma}{2} + \frac{\varepsilon}{8}, n\right) + \log\frac{1}{\delta}\right),$$

then

$$\Pr_{S\sim q^n}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{x\in S} f(x) - \mathop{\mathbb{E}}_{x\sim q}[f(x)]\right| > \gamma + \varepsilon\right] \leq \delta.$$

In particular, if $H(\mathcal{F}, \frac{\gamma}{2} + \varepsilon, n) = o(n)$ for all $\varepsilon > 0$, then $\mathcal{F}$ satisfies γ-uniform convergence.

Proof. Let $\varepsilon > 0$, let q be some distribution over $\mathcal{X}$ and for every dataset $S = \{x_i\}_{i=1}^n$ define the random variable Z_S by

$$Z_S = \sup_{f\in\mathcal{F}} |\mathop{\mathbb{E}}_{x\sim q}[f(x)] - \frac{1}{n}\sum_{i=1}^n f(x_i)|.$$

We need to show that $\Pr[Z_S > \gamma + \varepsilon] < \delta$. We will do that by showing

$$\mathop{\mathbb{E}}_{S\sim q^n}[Z_S] \leq \gamma + \frac{\varepsilon}{2},$$
$$Pr[Z_S > \mathbb{E}Z_S + \frac{\varepsilon}{2}] \leq \delta.$$

The second claim follows from McDiarmid's inequality [M+89], which tells us that if Z_S is such that $|Z_S - Z_{S'}| \leq c_j$ for any $S = \{x_i\}_{i=1}^n$, $S' = \{x_i'\}_{i=1}^n$ satisfying $x_i = x_i'$ for all $i \neq j$, then we have

$$\Pr[Z_S > \mathbb{E}Z_S + \eta] \leq \exp\left(\frac{-2\eta^2}{\sum_{i=1}^n c_i^2}\right).$$

Note that in our case this holds with $c_j = \frac{2R}{n}$ for all j, hence we have

$$Pr[Z_S > \mathbb{E}Z_S + \frac{\varepsilon}{2}] \leq e^{\frac{-\varepsilon^2 n}{8R^2}} \leq \delta.$$

So we just need to bound the expectation. For that we use symmetrization

$$\mathop{\mathbb{E}}_{S\sim q^n}\left[\sup_{f\in\mathcal{F}}\left|\mathop{\mathbb{E}}_{x\sim q}[f(x)] - \frac{1}{n}\sum f(x_i)\right|\right] \tag{1}$$
$$= \mathop{\mathbb{E}}_{S\sim q^n}\left[\sup_{f\in\mathcal{F}}\left|\mathop{\mathbb{E}}_{S'\sim q^n}\left[\frac{1}{n}\sum_{i=1}^n f(x_i')\right] - \frac{1}{n}\sum f(x_i)\right|\right] \tag{2}$$
$$\leq \mathop{\mathbb{E}}_{S,S'\sim q^n}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(x_i') - \frac{1}{n}\sum f(x_i)\right|\right]. \tag{3}$$

We generate the samples $S = \{x_i\}_{i=1}^n$ and $S' = \{x_i'\}_{i=1}^n$ via the following symmetrization procedure. First draw a sample

$$T = \{t_i\}_{i=1}^n \cup \{t_i'\}_{i=1}^n$$

of size $2n$. Then, for each $1 \leq i \leq n$, independently set $(x_i, x_i') = (t_i, t_i')$ with probability $\frac{1}{2}$, and $(x_i, x_i') = (t_i', t_i)$ with probability $\frac{1}{2}$. This construction yields

$$\mathop{\mathbb{E}}_{S,S'\sim q^n}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n f(x_i') - \frac{1}{n}\sum f(x_i)\right|\right] = \mathop{\mathbb{E}}_{T,\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^n \sigma_i\big(f(t_i) - f(t_i')\big)\right|\right]$$

where $\sigma = (\sigma_1, \ldots, \sigma_n) \in \{\pm 1\}^n$ are independent Rademacher random variables, with $\sigma_i = 1$ if $(x_i, x_i') = (t_i, t_i')$ and $\sigma_i = -1$ otherwise. By the triangle inequality and sub-additivity of the supremum we have

$$\mathop{\mathbb{E}}_{T,\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\big(f(t_i)-f(t_i')\big)\right|\right] \le 2\sup_{x\in\mathcal{X}^n}\mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(x_i)\right|\right].$$

The quantity above is known as the Rademacher complexity of $\mathcal{F}$. To bound fix some $x \in \mathcal{X}^n$ and let H be a cover of $\mathcal{F}|_x$ of size $N = N(\mathcal{F}, \frac{\gamma}{2} + \frac{\varepsilon}{8}, n)$. Then we may use Massart Lemma (see lemma 26.8 in [SSBD14]) for bounding the Rademacher complexity of finite classes to get

$$2\mathop{\mathbb{E}}_{\sigma}\left[\sup_{f\in\mathcal{F}}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i f(x_i)\right|\right] \le \gamma + \frac{\varepsilon}{4} + 2\mathop{\mathbb{E}}_{\sigma}\left[\sup_{h\in H}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i h(x_i)\right|\right] \le \gamma + \frac{\varepsilon}{4} + \sqrt{\frac{8R^2\log 2|H|}{n}}.$$

This yields the bound

$$\mathop{\mathbb{E}}_{S\sim q^n}[Z_S] \le \gamma + \frac{\varepsilon}{4} + \sqrt{\frac{8R^2\cdot H(\mathcal{F}, \frac{\gamma}{2}+\frac{\varepsilon}{8}, n) + 8R^2}{n}} \le \gamma + \frac{\varepsilon}{2}.$$

Giving the desired result. □

The following result follows from analyzing the proof of Theorem 21 in [BL98], replacing their specific covering bound with a generic one. The proof is exactly the same and is repeated here for completeness.

Lemma 4 (Learning Via Covering)**.** *Let $\gamma > 0$, and let $\mathcal{F}$ be a bounded function class. Assume that for all $\varepsilon > 0$ we have a bound*

$$H(\mathcal{F}, \gamma + \varepsilon, n) = o(n)$$

on the empirical entropy of $\mathcal{F}$. Then $\mathcal{F}$ is γ-learnable.

Proof. By normalization and translation assume $\mathcal{F} \subset [0,1]^{\mathcal{X}}$. The learning algorithm is as follows: Given a sample $S = \{(x_i, y_i)\}_{i=1}^{2n}$, and a test point x set $X = \{x_i\}_{i=1}^{2n} \cup \{x\}$, and let H be a $(\gamma+\varepsilon)$-cover of $\mathcal{F}|_X$ of size at most $N(\mathcal{F}, \gamma+\varepsilon, 2n+1)$. Split S, as $S_1 = \{x_i\}_{i=1}^{n}$, $S_2 = \{x_i\}_{i=n+1}^{2n}$, and let $h \in H$ be such that

$$L_{S_1}(h) := \frac{1}{n}\sum_{i=1}^{n}|h(x_i) - y_i|,$$

is minimized over all functions in H. Our learner will return the value of h on the test point

$$\mathcal{A}_S(x) = h(x).$$

To see this gives a valid learner let μ be a distribution on $\mathcal{X}\times[-R,R]$, and let $\varepsilon, \delta > 0$. Let $f^\star \in \mathcal{F}$ be a function for which

$$L_\mu(f^\star) := \mathop{\mathbb{E}}_{(x,y)\sim q}[|f^\star(x) - y|] \le \inf_{f\in\mathcal{F}}\mathop{\mathbb{E}}_{(x,y)\sim q}[|f(x)-y|] + \varepsilon.$$

Now by Hoeffding inequality we have

$$\Pr_{S\sim\mu^{2n}}[L_S(f^\star) \le L_\mu(f^\star) + \varepsilon] > 1 - 2e^{-4\varepsilon^2 n}.$$

On this event, since H is a cover of $\mathcal{F}|_X$ there is some $h' \in H$ such that

$$L_S(h') \le L_S(f^\star) + \gamma + \varepsilon \le L_\mu(f^\star) + \gamma + 2\varepsilon \le \inf_{f\in\mathcal{F}} L_\mu(f) + \gamma + 3\varepsilon.$$

By symmetrization

$$\begin{aligned}
&\Pr_{S_1,S_2\sim\mu^n}[\sup_{h\in H}|L_{S_1}(h) - L_{S_2}(h)| > \varepsilon] \\
&\le \sup_{S\in(\mathcal{X}\times[0,1])^{2n}}\Pr_{\sigma}\left[\sup_{h\in H}\left|\frac{1}{n}\sum_{i=1}^{n}\sigma_i\Big(|h(x_i)-y_i| - |h(x_{n+i})-y_{n+i}|\Big)\right| > \varepsilon\right] \\
&\le 2|H|e^{\frac{-\varepsilon^2 n}{2}} \le 2N(\mathcal{F}, \gamma+\varepsilon, 2n+1)e^{\frac{-\varepsilon^2 n}{2}}.
\end{aligned}$$

Now since L_S is the average of L_{S_1} and L_{S_2}, and h is the minimizer of L_{S_1} we deduce that on the above event

$$L_{S_2}(h) \leq L_{S_1}(h) + \varepsilon \leq L_{S_1}(h') + \varepsilon \leq L_S(h') + 2\varepsilon \leq \inf_{f \in \mathcal{F}} L_\mu(f) + \gamma + 5\varepsilon.$$

Averaging over permutation of the last n points and the test point, which does not affect h, yields that the above gives the same bound for $L_\mu(h)$. This establish an in expectation bound. Standard repeated sampling technique gives the high probability result

$$\Pr\left(L_\mu(h) \geq \inf_{f \in \mathcal{F}} L_\mu(f) + \gamma + 5\varepsilon\right) \leq 4N(\mathcal{F}, \gamma + \varepsilon, 2n+1)e^{\frac{-\varepsilon^2 n}{2}}.$$

From which the desired result follows. □

Theorem 1. *Let $\mathcal{F} \subset [-R, R]^{\mathcal{X}}$ be a bounded function class and let $\gamma > 0$. Then, the following are equivalent:*

1. *$\mathcal{F}$ is $\frac{\gamma}{2}$-learnable;*
2. *$\mathcal{F}$ satisfies γ-uniform convergence;*
3. *for every $\gamma' > \gamma$, the γ'-fat-shattering dimension of $\mathcal{F}$ is finite.*

Proof. (1) $\implies$ (3) Follow from Theorem 26 in [BL98] and (2) $\implies$ (3) follow from Theorem 22 in the same work. The implication (3) $\implies$ (1) follow from Lemma 2 and Lemma 4. Thus we only need to prove (3) $\implies$ (2). Let $\gamma > 0$ be such that $\mathrm{fat}_{\gamma'}(\mathcal{F})$ is finite for all $\gamma' > \gamma$, and we will show that $\mathcal{F}$ satisfies γ-uniform convergence.

Let $\varepsilon, \delta > 0$, and denote $d = \mathrm{fat}_{\gamma + \frac{\varepsilon}{16}}$, which is finite by assumption. By Lemma 2 we have

$$H\left(\mathcal{F}, \frac{\gamma}{2} + \frac{\varepsilon}{8}, n\right) \leq 14d \log \frac{32R}{\varepsilon} \log^2 n.$$

Thus, by Lemma 3 if n is large enough such that

$$\frac{n}{\log^2 n} > \frac{128R^2}{\varepsilon^2}\left(14d \log \frac{32R}{\varepsilon} + \log \frac{1}{\delta}\right)$$

then we have

$$\Pr_{S \sim q^n}\left[\sup_{f \in \mathcal{F}} \left|\frac{1}{n}\sum_{x \in S} f(x) - \mathbb{E}_{x \sim q}[f(x)]\right| > \gamma + \varepsilon\right] \leq \delta.$$

From which the desired result follow. We note that the induced sample size bound will be of order

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

□

A.3 Proof for Theorem 2

The following Lemma concerns approximate sample compression schemes, which were defined in [HKS19] as a natural generalization of sample compression schemes for discrete classes. We say that $\mathcal{F}$ has a γ-sample compression scheme of size k, if there exists a reconstructor

$$\rho : (\mathcal{X} \times \mathbb{R})^k \to \mathbb{R}^{\mathcal{X}}$$

such that for any finite set X and function $f \in \mathcal{F}$ there exists a sequence of length k

$$T = \{(x_i, f(x_i))\}_{i=1}^k, \quad x_i \in X \text{ for all } 1 \leq i \leq k,$$

satisfying

$$|\rho(T)(x) - f(x)| \leq \gamma \quad \text{for all } x \in X.$$

Lemma 5 (Sample compression)**.** *Let $\mathcal{F} \subseteq [-R, R]^{\mathcal{X}}$ be a bounded function class, and assume that $\mathrm{fat}_\gamma(\mathcal{F}) < \infty$. Then for every $\varepsilon > 0$, $\mathcal{F}$ has a $(2\gamma + \varepsilon)$-compression scheme of constant size.*

Proof. The proof is a real-valued analogue of [MY16]. Fix $\varepsilon > 0$. We will choose $m, m^\star$ below.

For every realizable sample

$$T = \{(x_i, f(x_i))\}_{i=1}^m,$$

where $x_i \in \mathcal{X}$ and $f \in \mathcal{F}$, choose some $h_T \in \mathcal{F}$ such that

$$h_T(x_i) = f(x_i) \qquad \text{for all } i \in [m].$$

We will show that for every finite set $X = \{x_1, \ldots, x_n\} \subseteq \mathcal{X}$ and every $f \in \mathcal{F}$, there exist subsets $J_1, \ldots, J_{m^\star} \subseteq X$ of size m such that, writing

$$T_i = \{(x, f(x)) : x \in J_i\},$$

we have, for every $x \in X$,

$$\left| \frac{1}{m^\star} \sum_{i=1}^{m^\star} h_{T_i}(x) - f(x) \right| \le 2\gamma + \varepsilon.$$

The compression scheme stores the samples $T_1, \ldots, T_{m^\star}$ and reconstructs by

$$\rho(T_1, \ldots, T_{m^\star}) = \frac{1}{m^\star} \sum_{i=1}^{m^\star} h_{T_i}.$$

Fix X and f, and define

$$\mathcal{H} = \{h_T : \ T = \{(x_i, f(x_i))\}_{i=1}^m, \ x_i \in X\}.$$

We first claim that, for every distribution q over X, there exists $h \in \mathcal{H}$ such that

$$\mathop{\mathbb{E}}_{x \sim q} [|h(x) - f(x)|] \le \gamma + \frac{\varepsilon}{2}.$$

For any function $h : X \to \mathbb{R}$, define

$$\phi_h(x) = |h(x) - f(x)|,$$

and let

$$L_{\mathcal{H},f} = \{\phi_h : h \in \mathcal{H}\}.$$

For any functions g, h and any $x \in X$,

$$|\phi_g(x) - \phi_h(x)| = \big||g(x) - f(x)| - |h(x) - f(x)|\big| \le |g(x) - h(x)|.$$

Hence, for every finite sample size n and every scale $\alpha > 0$,

$$N(L_{\mathcal{H},f}, \alpha, n) \le N(\mathcal{H}, \alpha, n) \le N(\mathcal{F}, \alpha, n).$$

Since $\mathrm{fat}_\gamma(\mathcal{F}) < \infty$, the covering bound Lemma 2 and the uniform convergence bound Lemma 3 imply that the classes $L_{\mathcal{H},f}$ satisfy γ-uniform convergence with a sample size depending only on γ, ε, R and $\mathrm{fat}_\gamma(\mathcal{F})$, but not on X, f, or q. Choose m large enough so that, with probability at least $2/3$ over $x_1, \ldots, x_m \sim q$, we have simultaneously for all $h \in \mathcal{H}$,

$$\left| \mathop{\mathbb{E}}_{x \sim q} [\phi_h(x)] - \frac{1}{m} \sum_{i=1}^m \phi_h(x_i) \right| \le \gamma + \frac{\varepsilon}{2}.$$

For the sampled set

$$T = \{(x_i, f(x_i))\}_{i=1}^m,$$

the corresponding function $h_T \in \mathcal{H}$ satisfies $h_T(x_i) = f(x_i)$ for all i. Therefore

$$\frac{1}{m} \sum_{i=1}^m \phi_{h_T}(x_i) = 0,$$

and so

$$\mathop{\mathbb{E}}_{x\sim q}\left[|h_T(x)-f(x)|\right] = \mathop{\mathbb{E}}_{x\sim q}\left[\phi_{h_T}(x)\right] \le \gamma + \frac{\varepsilon}{2}.$$

This proves the claim.

Now consider the zero-sum game in which Player 1 chooses $h \in \mathcal{H}$, Player 2 chooses $x \in X$, and the payoff is

$$|h(x) - f(x)|.$$

The claim says that for every mixed strategy q of Player 2, there is $h \in \mathcal{H}$ whose expected payoff is at most $\gamma + \frac{\varepsilon}{2}$. By minimax, there exists a distribution p over $\mathcal{H}$ such that, for every distribution q over X,

$$\mathop{\mathbb{E}}_{h\sim p,\ x\sim q}\left[|h(x)-f(x)|\right] \le \gamma + \frac{\varepsilon}{2}.$$

In particular, for every $x \in X$,

$$\left|\mathop{\mathbb{E}}_{h\sim p}\left[h(x)\right] - f(x)\right| \le \mathop{\mathbb{E}}_{h\sim p}\left[|h(x)-f(x)|\right] \le \gamma + \frac{\varepsilon}{2}.$$

Finally, by Proposition 2, the dual class $\mathcal{H}^\star$ satisfies γ-uniform convergence. Choose $m^\star$ large enough so that, with probability at least $2/3$ over $h_1, \dots, h_{m^\star} \sim p$,

$$\sup_{x\in X}\left|\mathop{\mathbb{E}}_{h\sim p}\left[h(x)\right] - \frac{1}{m^\star}\sum_{i=1}^{m^\star} h_i(x)\right| \le \gamma + \frac{\varepsilon}{2}.$$

Therefore there exist $h_1, \dots, h_{m^\star} \in \mathcal{H}$ satisfying this bound. Writing each h_i as h_{T_i}, we obtain, for every $x \in X$,

$$\left|\frac{1}{m^\star}\sum_{i=1}^{m^\star} h_{T_i}(x) - f(x)\right| \le \gamma + \frac{\varepsilon}{2} + \gamma + \frac{\varepsilon}{2} = 2\gamma + \varepsilon.$$

Since m and $m^\star$ depend only on γ, ε, R and the relevant fat-shattering bounds, and not on X, f, or n, the resulting compression scheme has constant size $mm^\star$. We remark that our uniform convergence bound imply the compression size is of order

$$mm^\star = O\left(\frac{\mathrm{fat}_\gamma(\mathcal{F})\mathrm{fat}_\gamma(\mathcal{F}^\star)}{\varepsilon^4}\log^6\frac{\mathrm{fat}_\gamma(\mathcal{F})\mathrm{fat}_\gamma(\mathcal{F}^\star)}{\varepsilon}\right).$$

□

Theorem 2. *Let $\mathcal{F} \subset [-R,R]^{\mathcal{X}}$ be a bounded function class, and let*

$$\gamma^\star = \inf\{\gamma > 0 : \mathrm{fat}_\gamma(\mathcal{F}) < \infty\}.$$

Then[4]*,*

$$H(\mathcal{F},\gamma,n) = \begin{cases} \Theta(n) & \gamma \in (0, \frac{\gamma^\star}{2}), \\ O(\log^2 n) & \gamma \in (\frac{\gamma^\star}{2}, 2\gamma^\star], \\ O(\log n) & \gamma \in (2\gamma^\star, R). \end{cases}$$

Moreover, the second regime is tight for $\gamma \in \left(\frac{\gamma^\star}{2}, \gamma^\star\right)$: there exist $\mathcal{F}$ s.t. for all $\gamma < \gamma^\star$,

$$H(\mathcal{F},\gamma,n) = \Omega\left(\log^{2-o(1)} n\right).$$

The third regime is tight: there exist $\mathcal{F}$ such that $H(\mathcal{F},\gamma,n) = \Omega(\log n)$ for $\gamma \in (2\gamma^\star, R)$.

[4]Throughout, the asymptotic notation treats n as the parameter tending to infinity; the implicit constants may depend on γ and on the fat-shattering dimension at scale γ.

Proof. The first part is well established [ABDCBH97, BL98]. If S is a γ'-shattered set then the γ-covering number of $\mathcal{F}|_S$ is at least 2^n for any $\gamma < \frac{\gamma'}{2}$. For the second part , let $\gamma > \frac{\gamma^\star}{2}$ and pick some γ' such that $\gamma > \gamma' > \frac{\gamma^\star}{2}$. By definition $\mathrm{fat}_{2\gamma'}(\mathcal{F})$ is finite hence by Lemma 2 we have

$$N(\mathcal{F}, \gamma, n) \leq n^{14\mathrm{fat}_{2\gamma'}(\mathcal{F}) \log \frac{1}{\gamma - \gamma'} \log n}.$$

Taking $\log$ gives the desired bound. To show this bound can be tight in the $\gamma \in (\frac{\gamma^\star}{2}, \gamma^\star)$ regime we use Theorem 11 of [AHHM22], which gives, for any $n > 1$, the existence of a partial concept class $\mathcal{F}_n \subset \{-1, 1, \star\}^n$, with VC dimension 1, such that any disambiguation of $\mathcal{F}_n$ is of size at least $n^{(\log n)^{1-o(1)}}$.

For each $f \in \mathcal{F}_n$ define $\widehat{f} : \mathcal{X} \to [-\gamma^\star, \gamma^\star]$ by

$$\widehat{f}(x) = \begin{cases} \gamma^\star & f(x) = 1, \\ -\gamma^\star & f(x) = -1, \\ 0 & f(x) = \star, \end{cases}$$

And set $\widehat{\mathcal{F}}_n = \{\widehat{f} \ : \ f \in \mathcal{F}_n\}$. Then $\mathrm{fat}_\gamma(\widehat{\mathcal{F}}_n)$ is finite (And equal to 1) if $\gamma > \gamma^\star$, other properties of this class will imply this is an if and only if but this is not necessary for our bound. Now let H be some γ-cover of $\widehat{\mathcal{F}}_n$, for $\gamma < \gamma^\star$. Define $\{\Phi_h \ : \ h \in H\}$ a disambiguation of $\mathcal{F}_n$ by

$$\Phi_h(x) = \begin{cases} 1 & h(x) \geq 0, \\ -1 & h(x) < 0. \end{cases}$$

For any $f \in \mathcal{F}_n$ there is $h \in H$ such that $\left\| \widehat{f} - h \right\|_\infty \leq \gamma$, hence we have $h(x) \geq \gamma^\star - \gamma > 0$, whenever $f(x) = 1$ and $h(x) \leq \gamma - \gamma^\star < 0$ whenever $f(x) = -1$. Hence ϕ_h disambiguates f and $\{\phi_h \ : \ h \in H\}$ is a disambiguation of $\mathcal{F}_n$, implying that its size is at least $n^{(\log n)^{1-o(1)}}$. This gives the result for a fixed n, to get $\mathcal{F}$ for which the bound will hold for all n simply take a disjoint union of all the $\mathcal{F}_n$. each defined over its own domain and extend to be 0 on all other domains.

For the $\log n$ bound let $\gamma > 2\gamma^\star$ and define $\varepsilon = \frac{\gamma - 2\gamma^\star}{8}$ so $\gamma = 2\gamma^\star + 8\varepsilon$. First discretize the space at scale ε, replacing each function $f \in \mathcal{F}$ with f_ε defined by

$$f_\varepsilon(x) = \left\lfloor \frac{f(x)}{\varepsilon} \right\rfloor \varepsilon,$$

and Set $\mathcal{F}_\varepsilon = \{f_\varepsilon \ : \ f \in \mathcal{F}\}$. Note that $|f(x) - f_\varepsilon(x)| \leq \varepsilon$ for all $x \in \mathcal{X}$, hence we have

$$\mathrm{fat}_{\frac{\gamma}{2} - \varepsilon}(\mathcal{F}_\varepsilon) \leq \mathrm{fat}_{\frac{\gamma}{2} - 3\varepsilon}(\mathcal{F}) = \mathrm{fat}_{\gamma^\star + \varepsilon}(\mathcal{F}) < \infty.$$

Thus, by Theorem 1 we have that $\mathcal{F}_\varepsilon$ satisfies $(\frac{\gamma}{2} - \varepsilon)$-uniform convergence. By Proposition 2 we also have that its dual class $\mathcal{F}_\varepsilon^\star$ satisfies $(\frac{\gamma}{2} - \varepsilon)$-uniform convergence. Hence by Lemma 5 $\mathcal{F}_\varepsilon$ admits a $(\gamma - \varepsilon)$-sample compression of constant size k, i.e. there is a reconstructor ρ, such that for any dataset $S = \{(x_i, f(x_i))\}_{i=1}^n$ there is $T = \{(x_{i_j}, f(x_{i_j}))\}_{i=1}^k$ such that the reconstructed function $\rho(T)$ is at most $(\gamma - \varepsilon)$-away from f on any point in S. We use this compression scheme to construct a covering as follow : given a dataset X of size n define the covering H by

$$H = \{\rho(T) \ : \ T = \{(x_i, f_\varepsilon(x_i)\}_{i=1}^k, f_\varepsilon \in \mathcal{F}_\varepsilon, \ x_i \in X\}.$$

So H is all possible outputs of reconstructor on realizable subset of X of size k. Note that each function in H is defined by k points from X and the values of f_ε on them. By discretization f_ε has no more than $\frac{4R}{\varepsilon}$ possible outputs at every point, yielding a bound of

$$|H| \leq (\frac{4Rn}{\varepsilon})^k.$$

So it just remains to show that H is an γ-cover for $\mathcal{F}$. By definition of compression scheme H forms a $(\gamma - \varepsilon)$-cover of $\mathcal{F}_\varepsilon$, and since $|f(x) - f_\varepsilon(x)| \leq \varepsilon$ for all x we indeed have that it is also a γ cover of $\mathcal{F}$. Taking $\log$, and noting that k, ε are independent on n gives the desired result.

To see this $\log n$ bound can be tight let $\mathcal{F} \subset \{-1, 1\}^{\mathbb{N}}$ be the class of all singleton functions, so $\mathcal{F} = \{f_n \ : \ n \in \mathbb{N}\}$, where $f_n(k)$ is 1 if $n = k$, and -1 otherwise. It is easy to see that $\mathrm{fat}_\gamma(\mathcal{F}) = 1$, and $N(\mathcal{F}, \gamma, n) = n$ for all $\gamma \in (0, 1)$. $\square$

Proposition 4. *Let $\gamma^\star > 0$. There exists bounded function classes $\mathcal{F}_1, \mathcal{F}_2$ such that*

$$\inf\{\gamma \ : \ \mathrm{fat}_\gamma(\mathcal{F}_1) < \infty\} = \gamma^\star = \inf\{\gamma \ : \ \mathrm{fat}_\gamma(\mathcal{F}_2) < \infty\}.$$

But for all $n > 0$, we have

$$H(\mathcal{F}_1, \frac{\gamma^\star}{2}, n) = 0,$$
$$H(\mathcal{F}_2, \frac{\gamma^\star}{2}, n) = \Omega(\frac{n}{\log n}).$$

Proof. Let $\mathcal{F}_1 = [-\frac{\gamma^\star}{2}, \frac{\gamma^\star}{2}]^{\mathbb{N}}$. Clearly we have $\mathrm{fat}_\gamma(\mathcal{F}_1) = 0$ for $\gamma > \gamma^\star$ and $\mathrm{fat}_\gamma(\mathcal{F}_1) = \infty$ for $\gamma \leq \gamma^\star$ so $\gamma^\star$ is indeed the critical threshold for finite fat shattering. It is also clear that $N(\mathcal{F}_1, \frac{\gamma^\star}{2}, n) = 1$ for all n since the zero function is distance $\frac{\gamma^\star}{2}$ from any function in $\mathcal{F}_1$,

Now let $\mathcal{F}_2 \subset [0, 1 + \gamma^\star]^{\mathbb{N}}$ be the class of all function f from $\mathbb{N}$ to $[0, 1 + \gamma^\star]$ satisfying

$$f(n) \leq \gamma^\star + \frac{1}{\log n}.$$

It is simple to see that $\mathrm{fat}_\gamma(\mathcal{F}_2) > n$ if and only if $\gamma \leq \gamma^\star + \frac{1}{\log n}$, thus $\gamma^\star$ is indeed the critical threshold for finite fat shattering. Note that the restriction of $\mathcal{F}_2$ to $1, 2 \ldots n$, contains $[0, \gamma^\star + \frac{1}{\log n}]^n$ as a subset. It has a volume of $(\gamma^\star + \frac{1}{\log n})^n$, hence if we want to cover it with balls of radius $\frac{\gamma^\star}{2}$, each of volume $(\gamma^\star)^n$ the number of functions we must use is at least

$$\left(\frac{\gamma^\star + \frac{1}{\log n}}{\gamma^\star}\right)^n = \left(1 + \frac{1}{\gamma^\star \log n}\right)^n \sim e^{\frac{n}{\gamma^\star \log n}}.$$

This shows that $H(\mathcal{F}_2, \frac{\gamma^\star}{2}, n) = \Omega(\frac{n}{\log n})$. Note that by replacing $\frac{1}{\log n}$ with a slower decaying sequence we may get arbitrarily close to a $\log n$ lower bound. □

A.4 Composition and Closure

Lemma 6. *Let $\mathcal{F}$ be a bounded function class. Let $\gamma^\star = \inf\{\gamma \ : \ \mathrm{fat}_\gamma(\mathcal{F}) < \infty\}$. Then for any $\gamma^- < \gamma^\star$, and any $n \in \mathbb{N}$, there exists $r \in \mathbb{R}$ such that there are at least n points that can be γ^--shattered at the same threshold r by $\mathcal{F}$.*

Proof. By rescaling and translating $\mathcal{F}$, we assume $\mathcal{F} \subset [0, 1]^{\mathcal{X}}$. Fix $n \in \mathbb{N}$ and let $M \in \mathbb{N}$ be chosen later. Fix a scale γ' such that $\gamma^- < \gamma' < \gamma^\star$. By definition, there exists a subset $\{x_1, ..., x_M\} \subseteq \mathcal{X}$ that is $\gamma'-$shattered by $\mathcal{F}$. with thresholds $r_1, ..., r_M$. Observe that since $\mathcal{F} \subseteq [0, 1]^{\mathcal{X}}$, any threshold $r_i \in [\gamma'/2, 1 - \gamma'/2]$. Now, partition the interval $[\gamma'/2, 1 - \gamma'/2]$ into subintervals of length $\alpha \leq \frac{\gamma' - \gamma^-}{2}$. Note that there can be at most $S = \lceil \frac{1-\gamma'}{\alpha} \rceil$ subintervals and that we can identify any subinterval with an index set $I \subseteq [M]$. It follows by the pigeonhole principle that if we take $M \geq Sn$, there must exist an index set $I^\star \subseteq [M]$ with $|I^\star| \geq n$ for which there exists $r \in \mathbb{R}$ such that

$$r_i \in [r, r + \alpha] \quad \forall i \in I^\star.$$

Now, we claim that $\{x_i : i \in I^\star\}$ is γ^--shattered by $\mathcal{F}$ with common threshold r. To see this, fix any labeling $y \in \{+1, -1\}^M$ (taking the subset of labelings corresponding to $I^\star$ gives an arbitrary labeling on $I^\star$). By definition, there exists $f \in \mathcal{F}$ such that for all $i \in I^\star$:

If $y_i = +1$, then

$$f(x_i) \geq r_i + \frac{\gamma'}{2} \geq r + \frac{\gamma'}{2} \geq r + \frac{\gamma^-}{2}$$

If $y_i = -1$, then

$$f(x_i) \leq r_i - \frac{\gamma'}{2} \leq r + \alpha - \frac{\gamma'}{2} \leq r - \frac{\gamma^-}{2},$$

where the last inequality uses the fact that $\alpha \leq \frac{\gamma' - \gamma^-}{2}$. The result follows. □

Proposition 2. *Let $\gamma > 0$, and let $\mathcal{F}$ be a bounded function class that satisfies γ-uniform convergence. Then the dual class $\mathcal{F}^\star$ also satisfies γ-uniform convergence.*

Proof. By Theorem 1 it is enough to show that for any $\gamma' > \gamma$ we have that $\mathrm{fat}_{\gamma'}(\mathcal{F}^\star) < \infty$. Indeed, let $\gamma' > \gamma$ and choose some γ'' such that $\gamma < \gamma'' < \gamma'$. Assume toward contradiction that $\mathrm{fat}_{\gamma'}(\mathcal{F}^\star) = \infty$, and denote $n = 2^{\mathrm{fat}_{\gamma''}(\mathcal{F})+2}$, which is finite by Theorem 1. By Lemma 6 there is some $r \in \mathbb{R}$ such that there are n points which are γ''-shattered by $\mathcal{F}^\star$ at this same threshold r. Now that the points are shattered at the same threshold, standard Assouad type bounds, such as Theorem 3.7 in [KS23], will imply that $\mathcal{F}$ γ''-shatters at least $\log(n) - 1 = \mathrm{fat}_{\gamma''}(\mathcal{F}) + 1$ points. Giving the desired contradiction. □

Proposition 3. *Let $\mathcal{F}_1, \ldots, \mathcal{F}_k \subset [0,1]^{\mathcal{X}}$ be bounded function classes, each satisfying γ-uniform convergence, and let $G : [0,1]^k \to [0,1]$ be a 1-Lipschitz function with respect to the ℓ_∞ metric. Then the aggregated class $G(\mathcal{F}_1, \ldots, \mathcal{F}_k)$ also satisfies γ-uniform convergence.*

Proof. Let $\varepsilon > 0$. By Theorem 1 we know that $\mathcal{F}_i$ has finite $(\gamma + \varepsilon)$-fat-shattering dimension for each $1 \le i \le k$. Denote $d_i = \mathrm{fat}_{\gamma+\varepsilon}(\mathcal{F}_i)$, then by Lemma 2 we have

$$N(\mathcal{F}_i, \frac{\gamma}{2} + \varepsilon) \le n^{14 d_i \log \frac{1}{\varepsilon} \log n}$$

For each $1 \le i \le k$ let H_i be such $(\frac{\gamma}{2} + \varepsilon)$-cover for $\mathcal{F}_i$ and define

$$G(H) = \{G(h_1, h_2 \ldots h_k) \ : \ \forall i, \ h_i \in H_i\}.$$

We claim that $G(H)$ is a $(\frac{\gamma}{2} + \varepsilon)$-cover for $G(\mathcal{F})$. Indeed, for any $f = (f_1, f_2, \ldots f_k) \in \mathcal{F}$ there is $h = (h_1, h_2, \ldots h_k)$, such that $h_i \in H_i$, and $\|f_i - h_i\|_\infty \le \frac{\gamma}{2} + \varepsilon$ for all i. Then since G is 1-Lipschitz with respect to the ℓ_∞ norm we have that

$$\|G(f) - G(h)\|_\infty \le \|f - h\|_\infty = \sup_{1 \le i \le k} \sup_{x \in \mathcal{X}} |f_i(x) - h_i(x)| \le \frac{\gamma}{2} + \varepsilon.$$

Thus we get that

$$N(G(\mathcal{F}), \frac{\gamma}{2} + \varepsilon, n) \le \prod_{i=1}^{k} N(\mathcal{F}_i, \frac{\gamma}{2} + \varepsilon, n) \le \exp\left(14 \log^2 n \log \frac{1}{\varepsilon} \sum_{i=1}^{k} d_i\right).$$

The above with Lemma 3 implies the desired result. □

B Proofs of Evaluability Results

B.1 Lower bound construction of [BKM19]

Lemma 7 (Lemma 22 of [BKM19])**.** *Let $\mathcal{D}_1$ and $\mathcal{D}_2$ be two families of probability distributions, $\mathcal{D}_i^{\oplus m}$ denotes the distribution obtained by sampling $p \sim \mathcal{D}_i$ (assuming some given fixed distribution over $\mathcal{D}_i$) and then drawing m independent samples from p. Consider an algorithm (which can be randomized) that determines, given m i.i.d. examples from some $p \in \mathcal{D}_1 \cup \mathcal{D}_2$, whether $p \in \mathcal{D}_1$ or $p \in \mathcal{D}_2$. Then such an algorithm will have a probability of making a mistake lower bounded by*

$$\frac{1}{2}\left(1 - \mathrm{TV}(\mathcal{D}_1^{\oplus m}, \mathcal{D}_2^{\oplus m})\right).$$

Proof. We first assume that the algorithm is deterministic. Any deterministic algorithm deciding whether p comes from $\mathcal{D}_1$ or $\mathcal{D}_2$ is associated with a set $A \subseteq \mathcal{X}^m$ (the set such that if the sample falls in it, it decides $i = 1$, and $i = 2$ otherwise). The worst-case probability of the algorithm to err is given by

$$\max\left(\max_{p \in \mathcal{D}_2} p^m(A), \max_{p \in \mathcal{D}_1} p^m(\overline{A})\right)$$

which can be lower bounded by the expectation under first choosing between $i = 1$ and $i = 2$ with probability $1/2$ and then picking $p \sim \mathcal{D}_i$:

$$\frac{1}{2}\left(\mathbb{E}_{p\sim\mathcal{D}_1}p^m(A) + \mathbb{E}_{p\sim\mathcal{D}_2}p^m(\overline{A})\right) = \frac{1}{2}\left(1 + \mathcal{D}_1^{\oplus m}(A) - \mathcal{D}_2^{\oplus m}(A)\right)$$
$$\geq \frac{1}{2}\left(1 - \mathrm{TV}(\mathcal{D}_1^{\oplus m}, \mathcal{D}_2^{\oplus m})\right).$$

If the algorithm is randomized then it may pick A randomly, so there is an additional expectation with respect to the distribution over sets A which also leads to the same lower bound. □

Lemma 8 (Lemma 23 of [BKM19])**.** *Given two probability distributions P, Q on a domain $\mathcal{X}$ and an event $E \subseteq \mathcal{X}$, denoting by $P_{|E}$ and $Q_{|E}$ the corresponding conditional distributions (i.e. $P_{|E}(A) := P(A \mid E)$), we have*

$$\mathrm{TV}(P, Q) \leq \mathrm{TV}(P_{|E}, Q_{|E}) + 2P(\overline{E}) + 2Q(\overline{E}).$$

Proof.

$$\begin{aligned}
\mathrm{TV}(P,Q) &= \sup_A |P(A) - Q(A)| \\
&\leq \sup_A |P(A\cap E) - Q(A\cap E)| + \sup_A |P(A\cap \overline{E}) - Q(A\cap \overline{E})| \\
&\leq \sup_A \left|P(E)\big(P(A \mid E) - Q(A \mid E)\big) + Q(A \mid E)\big(P(E) - Q(E)\big)\right| \\
&\quad + P(\overline{E}) + Q(\overline{E}) \\
&\leq P(E)\sup_A |P(A \mid E) - Q(A \mid E)| + |P(E) - Q(E)| \\
&\quad + P(\overline{E}) + Q(\overline{E}) \\
&\leq \mathrm{TV}(P_{|E}, Q_{|E}) + |P(E) - Q(E)| + P(\overline{E}) + Q(\overline{E}) \\
&= \mathrm{TV}(P_{|E}, Q_{|E}) + |P(\overline{E}) - Q(\overline{E})| + P(\overline{E}) + Q(\overline{E}) \\
&\leq \mathrm{TV}(P_{|E}, Q_{|E}) + 2P(\overline{E}) + 2Q(\overline{E}).
\end{aligned}$$

□

Lower bound construction. We now present the construction of [BKM19]. Fix $N \in \mathbb{N}, \beta \in (0, 1)$ such that $\frac{1+\beta}{\beta} \in \mathbb{N}$ and let $k = \frac{1+\beta}{\beta}$. Given a collection of $2kN$ non-overlapping distributions over $\mathcal{X}$, which we denote by $\boldsymbol{\mu} = \{\mu_{i,j} | i \in [2N], j \in [k]\}$, we construct mixtures over $\boldsymbol{\mu}$. We refer to distributions in $\boldsymbol{\mu}$ as elements to avoid ambiguity.

We define hypothesis class $\mathcal{H} = \{q_1, q_2\}$:

- q_1 assigns probability $\frac{1-\beta}{2kN}$ to each $\mu_{i,j}$ and $\frac{1+\beta}{2kN}$ to each $\mu_{i+N,j}$ for all $i \in [N]$ and $j \in [k]$.
- q_2 assigns probability $\frac{1+\beta}{2kN}$ to each $\mu_{i,j}$ and $\frac{1-\beta}{2kN}$ to each $\mu_{i+N,j}$ for all $i \in [N]$ and $j \in [k]$.

For each $i \in [N]$, we pick one entry by an index function $l : [2N] \mapsto [k]$ and then define distributions that concentrate all mass on that picked entries:

- $p_{1,l}$: Starting from q_1, for all $i \in [N]$, we increase the probability of $\mu_{i,l(i)}$ from $\frac{1-\beta}{2kN}$ to $\frac{1}{kN}$ and set the probability of $\mu_{i+N,l(i+N)}$ to 0. All other elements retain their probabilities from q_1.
- $p_{2,l}$: Starting from q_2, for all $i \in [N]$, we increase the probability of $\mu_{i+N,l(i+N)}$ from $\frac{1-\beta}{2kN}$ to $\frac{1}{kN}$ and set the probability of $\mu_{i,l(i)}$ to 0. All other elements retain their probabilities from q_2.

We define two distributions $\mathcal{D}_1$ and $\mathcal{D}_2$: $\mathcal{D}_1$ uniformly chooses from $\{p_{1,l} | l \in [k]^{[2N]}\}$ and $\mathcal{D}_2$ uniformly chooses from $\{p_{2,l} | l \in [k]^{[2N]}\}$. The following claim comes directly from the information-theoretic part of Theorem 19 in [BKM19]. We reproduce the argument here for clarity.

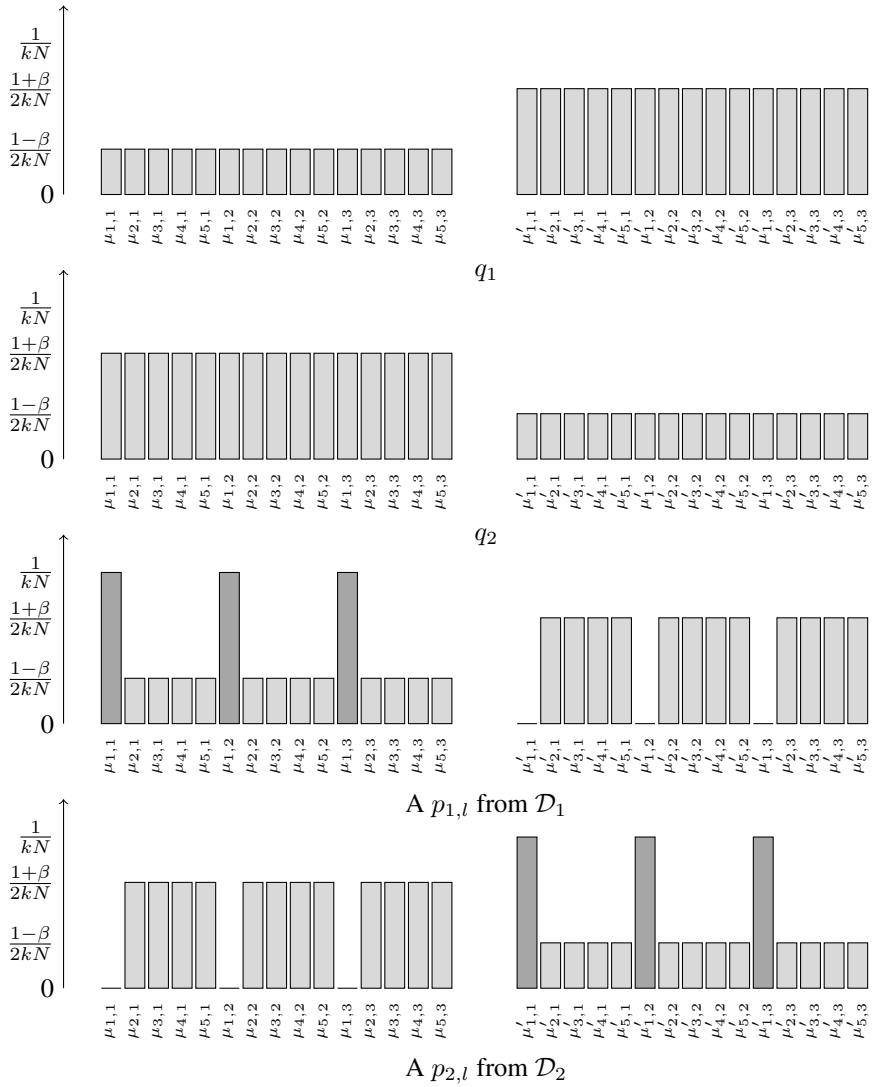


Figure 1: Illustration of the distribution construction for $k = 5$ and $N = 3$.

Lemma 9. *Let $\mathcal{F}$ be a bounded function class. Fix $N \in \mathbb{N}, \beta \in \mathbb{N}$ such that $\frac{1+\beta}{\beta} \in \mathbb{N}$, and let $k = \frac{1+\beta}{\beta}$. Then, construct $q_1, q_2, \mathcal{D}_1$ and $\mathcal{D}_2$ as above. For any evaluation algorithm $\mathcal{A}$, there exists a $q^\star \in \mathcal{D}_1 \cup \mathcal{D}_2$ such that, given a sample S_{eval} of size at most $\sqrt{2kN\beta}$ from $q^\star$, with probability at least $1/3$, $\widehat{q} = \mathcal{A}(\{q_1, q_2\}, S_{eval})$ will be the wrong hypothesis, i.e., $d_{\mathcal{F}}(\widehat{q}, q^\star) = \max\{d_{\mathcal{F}}(q_1, q^\star), d_{\mathcal{F}}(q_2, q^\star)\}$.*

Proof. We can view $\mathcal{A}$ as a distinguisher between $\mathcal{D}_1$ and $\mathcal{D}_2$. In particular, if the underlying distribution belongs to $\mathcal{D}_1$, then $\mathcal{A}$ errs exactly when, given S_{eval}, $\mathcal{A}(\{q_1, q_2\}, S_{\text{eval}}) = q_2$; if the underlying distribution belongs to $\mathcal{D}_2$, then $\mathcal{A}$ errs exactly when $\mathcal{A}(\{q_1, q_2\}, S_{\text{eval}}) = q_1$. Therefore, it suffices to lower bound the error probability of distinguishing $\mathcal{D}_1$ from $\mathcal{D}_2$.

By Lemma 7,

$$\Pr[\mathcal{A} \text{ errs}] \geq \frac{1}{2}\left(1 - TV(\mathcal{D}_1^{\oplus m}, \mathcal{D}_2^{\oplus m})\right),$$

where $\mathcal{D}_i^{\oplus m}$ denotes the distribution obtained by first drawing $p \sim \mathcal{D}_i$ and then drawing m independent samples from p. Hence it remains to show that

$$TV(\mathcal{D}_1^{\oplus m}, \mathcal{D}_2^{\oplus m}) \leq \frac{1}{3}.$$

For each $i \in [2N]$, let

$$B_i = \bigcup_{j=1}^{k} \mathrm{supp}(\mu_{i,j})$$

be the i-th big block. By construction, every distribution $p \in \mathcal{D}_1 \cup \mathcal{D}_2$ assigns mass exactly

$$p(B_i) = \frac{1}{2N}$$

to each big block B_i. Let E be the event that the m samples fall in m distinct big blocks. Since the big-block marginal is uniform over $[2N]$ under every $p \in \mathcal{D}_1 \cup \mathcal{D}_2$, we have, under both $\mathcal{D}_1^{\oplus m}$ and $\mathcal{D}_2^{\oplus m}$,

$$\Pr(E^c) \leq \binom{m}{2} \frac{1}{2N} \leq \frac{m^2}{4N}.$$

We next show that

$$(\mathcal{D}_1^{\oplus m})|E = (\mathcal{D}_2^{\oplus m})|E.$$

Conditioned on E, each occupied big block contains exactly one sample. The occupied big blocks have the same law under $\mathcal{D}_1^{\oplus m}$ and $\mathcal{D}_2^{\oplus m}$, because every distribution in $\mathcal{D}_1 \cup \mathcal{D}_2$ gives mass $1/(2N)$ to each big block.

It remains to check the conditional distribution inside an occupied big block. Fix a block B_i, and write

$$\overline{\mu}_i = \frac{1}{k} \sum_{j=1}^{k} \mu_{i,j}.$$

We claim that, after averaging over the random selector $\ell(i)$, the conditional distribution of the unique sample inside B_i is $\overline{\mu}_i$, regardless of whether the underlying family is $\mathcal{D}_1$ or $\mathcal{D}_2$.

Indeed, consider a left block $i \leq N$. For $p_{1,\ell}$, conditional on landing in B_i, the selected component $\mu_{i,\ell(i)}$ has conditional weight $2/k$, while each unselected component has conditional weight $(1-\beta)/k$. Since $\ell(i)$ is uniform on $[k]$, the averaged conditional weight of any fixed component $\mu_{i,j}$ is

$$\frac{1}{k} \cdot \frac{2}{k} + \frac{k-1}{k} \cdot \frac{1-\beta}{k} = \frac{2 + (k-1)(1-\beta)}{k^2} = \frac{1}{k},$$

where we used $k = (1+\beta)/\beta$. Thus the averaged conditional distribution is $\overline{\mu}_i$.

For $p_{2,\ell}$ on the same left block, the selected component has conditional weight 0, while each unselected component has conditional weight $(1+\beta)/k$. Hence the averaged conditional weight of any fixed component is

$$\frac{1}{k} \cdot 0 + \frac{k-1}{k} \cdot \frac{1+\beta}{k} = \frac{(k-1)(1+\beta)}{k^2} = \frac{1}{k}.$$

Thus the averaged conditional distribution is again $\overline{\mu}_i$. The calculation for right blocks $i > N$ is identical, with the roles of $\mathcal{D}_1$ and $\mathcal{D}_2$ reversed.

Because E ensures that no big block contributes more than one sample, there is no dependence coming from reusing the same hidden selector $\ell(i)$ twice within a block. Across different big blocks, the selectors $\ell(i)$ are independent. Therefore the full conditional joint distributions agree:

$$(\mathcal{D}_1^{\oplus m})|E = (\mathcal{D}_2^{\oplus m})|E.$$

Applying Lemma 8 with this event E, in the form

$$TV(P, Q) \leq TV(P|E, Q|E) + 2P(E^c) + 2Q(E^c),$$

gives

$$TV(\mathcal{D}_1^{\oplus m}, \mathcal{D}_2^{\oplus m}) \leq 0 + 2\mathcal{D}_1^{\oplus m}(E^c) + 2\mathcal{D}_2^{\oplus m}(E^c) \leq 4 \cdot \frac{m^2}{4N} = \frac{m^2}{N}.$$

Choosing $N \geq 3m^2$, we obtain

$$TV(\mathcal{D}_1^{\oplus m}, \mathcal{D}_2^{\oplus m}) \leq \frac{1}{3}.$$

Therefore Lemma 7 implies

$$\Pr[\mathcal{A} \text{ errs}] \geq \frac{1}{2}\left(1 - \frac{1}{3}\right) = \frac{1}{3}.$$

Equivalently, there exists some

$$q^\star \in \{p_{1,\ell} : \ell \in [k]^{[2N]}\} \cup \{p_{2,\ell} : \ell \in [k]^{[2N]}\}$$

such that, with probability at least $1/3$ over $S_{\text{eval}} \sim (q^\star)^m$, the evaluation algorithm selects the hypothesis associated with the wrong family. □

B.2 Proof for Theorem 3

Lemma 10. *Let $\mathcal{F} \subset [0,1]^{\mathcal{X}}$ be a bounded function class, and denote*

$$\gamma^\star = \inf\{\gamma \ : \ \mathrm{fat}_\gamma(\mathcal{F}) < \infty\}.$$

Assume that $\gamma^\star > 0$. Then, for any γ^-, γ^+, such that $\gamma^+ > \gamma^\star > \gamma^-$, and any $n \in \mathbb{N}$, there exists $r \in (0,1)$, a set X and n subsets $X_1, \ldots, X_n \subset X$ such that:

1. *The X_i's are pairwise disjoint.*
2. *The union of the sets $\bigcup_{i=1}^n X_i$ is γ^--shattered at the same threshold r by $\mathcal{F}$.*
3. *Let P_i be the uniform distribution over X_i. Then for every i, j and every $\phi \in \mathcal{F}$*

$$|\mathop{\mathbb{E}}_{P_i}[\phi] - \mathop{\mathbb{E}}_{P_j}[\phi]\,| \leq \gamma^+.$$

Proof of Lemma 10. Choose some $\overline{\gamma}$ such that $\gamma^+ > \overline{\gamma} > \gamma^\star$. It follows that $\mathrm{fat}_{\overline{\gamma}}(\mathcal{F}) < \infty$. Let $\eta = 2(\gamma^+ - \overline{\gamma})$, and let $m, K > 0$ to be chosen later. By Lemma 6, there is a set X of size K that is γ^--shattered by $\mathcal{F}$ at the same threshold. We generate the sets X_i by sampling m points from q, the uniform distribution over X. Note that by definition condition $(2.)$ will hold. For condition $(3.)$ we use the same symmetrization argument as in Lemma 3 (see also Lemma 16 in [BL98]) to deduce that the probability it will fail for any fixed pair (i,j) is bounded by

$$\Pr_{X_i, X_j \sim q^m}\left[\sup_{f \in \mathcal{F}}\left|\frac{1}{m}\sum_{x \in X_i} f(x) - \frac{1}{m}\sum_{x \in X_j} f(x)\right| > \overline{\gamma} + \frac{\eta}{2}\right] \leq 2N\left(\mathcal{F}, \overline{\gamma}/2 + \eta/8, 2m\right)\exp\left(\frac{-\eta^2 m}{32}\right).$$

So by the union bound, item $(3.)$ will fail with probability at most

$$\binom{n}{2} 2N\left(\mathcal{F}, \overline{\gamma}/2 + \eta/8, 2m\right)\exp\left(\frac{-\eta^2 m}{32}\right),$$

Applying Lemma 2 we see that the above probability goes to 0 as m goes to infinity, hence we may pick some m for which it is at most $\frac{1}{4}$.

Now for the first item, note that the probability of two sampled points colliding is at most

$$\binom{2nm}{2}\frac{1}{|X|}.$$

Hence if we take K the size of X to be larger than $4\binom{2nm}{2}$, the above will happen with probability at most $\frac{1}{4}$. Thus by union bound all the conditions will hold with positive probability, implying the existence of such sets. □

Theorem 3. *Every bounded IPM is either:*

1. *estimable, and hence strongly evaluable; or*
2. *weakly evaluable with optimal scale $c = 3$.*

Proof. The proof follows immediately from Proposition 1, with $\mathrm{fat}_\gamma(\mathcal{F})$ underpinning the dichotomy. □

Proposition 1. *Let $\mathcal{F}$ be a class of bounded real-valued functions.*

1. *If $\mathrm{fat}_\gamma(\mathcal{F}) < \infty$ for all $\gamma > 0$, then $d_\mathcal{F}$ is estimable (and hence strongly evaluable).*
2. *If there exists $\gamma \in (0, 1/2)$ such that $\mathrm{fat}_\gamma(\mathcal{F}) = \infty$, then $d_\mathcal{F}$ is weakly evaluable with optimal scale $c = 3$.*

Proof. When $\mathrm{fat}_\gamma(\mathcal{F}) < \infty$ for all $\gamma > 0$, from Proposition 3.4 of [AMMS26], $d_\mathcal{F}$ is estimable (and hence strongly evaluable) with sample complexity

$$m(\varepsilon, \delta) = O\left(\frac{1}{\varepsilon^2} d \log^2 \frac{4d}{\varepsilon^2} + \frac{1}{\varepsilon} \log \frac{1}{\delta}\right),$$

where $d = \mathrm{fat}_{\varepsilon/24}(\mathcal{F})$. This proves the first item.

Now, consider the case where there exists $\gamma \in (0, 1/2)$ such that $\mathrm{fat}_\gamma(\mathcal{F}) = \infty$. We know again from Proposition 3.4 of [AMMS26] that $d_\mathcal{F}$ is 3-weakly evaluable. It remains to show that for any $c' < 3$, $d_\mathcal{F}$ is not c'-weakly evaluable.

For any $\beta \in (0, 1)$ and a collection of elements $\boldsymbol{\mu}$, following the construction and guarantee by Lemma 9, we can obtain a hypothesis class $\mathcal{H} = \{q_1, q_2\}$ and two distribution families $\mathcal{D}_1 = \{p_{1,l} | l : [2N] \mapsto [k]\}$ and $\mathcal{D}_2 = \{p_{2,l} | l : [2N] \mapsto [k]\}$ such that no evaluation algorithm can evaluate $\mathcal{H}$. Fix any $\tau > 0$. The only thing that remains is to find a collection of elements $\boldsymbol{\mu}$ such that $d_\mathcal{F}(q_2, p_{1,l}) \geq (3 - \tau) d_\mathcal{F}(q_1, p_{1,l})$ for an index l.

Let $k = \frac{1+\beta}{\beta}$. Let $\gamma^\star = \inf\{\gamma > 0 : \mathrm{fat}_\gamma(\mathcal{F}) < \infty\}$. For a small $\alpha > 0$, let $\gamma^+ = \gamma^\star + \alpha$ and $\gamma^- = \gamma^\star - \alpha$. Now, choose β so that $\frac{3-\beta}{1+\beta} \cdot \frac{\gamma^-}{\gamma^+} \geq 3 - \tau$.

By Lemma 10, we can find a set X and a sequence of $2kN$ disjoint subsets $\{X_{i,j} | i \in [2N], j \in [k]\}$ such that $\bigcup X_{i,j}$ is γ^--shattered by $\mathcal{F}$ at the same threshold and $| \mathop{\mathbb{E}}_{\mu_{i,j}} [\phi] - \mathop{\mathbb{E}}_{\mu_{i',j'}} [\phi] | \leq \gamma^+$ for every pair $(i, j), (i', j') \in [2N] \times [k]$ and every $\phi \in \mathcal{F}$, where $\mu_{i,j}$ is the uniform distribution over $X_{i,j}$.

For any index function l, we have

$$d_\mathcal{F}(q_1, p_{1,l}) = \frac{1+\beta}{2kN} \sup_{\phi \in \mathcal{F}} \sum_{i=1}^{N} \left(\mathop{\mathbb{E}}_{\mu_{i,l(i)}} [\phi] - \mathop{\mathbb{E}}_{\mu_{i+N,l(i+N)}} [\phi] \right).$$

Note the function within the supremum is a weighted sum of $\mathop{\mathbb{E}}_{\mu_{i,l(i)}} [\phi]$'s and $\mathop{\mathbb{E}}_{\mu_{i+N,l(i+N)}} [\phi]$'s. We define the normalized weighted sum $a(\boldsymbol{\mu}, l)$ as

$$a(\boldsymbol{\mu}, l) := \frac{2k}{1+\beta} \cdot d_\mathcal{F}(q_1, p_{1,l}) = \sup_{\phi \in \mathcal{F}} \frac{1}{N} \sum_{i=1}^{N} \left(\mathop{\mathbb{E}}_{\mu_{i,l(i)}} [\phi] - \mathop{\mathbb{E}}_{\mu_{i+N,l(i+N)}} [\phi] \right).$$

We also have

$$d_\mathcal{F}(q_2, p_{1,l}) = \sup_{\phi \in \mathcal{F}} \frac{1}{2kN} \sum_{i=1}^{N} \left((1-\beta) \left(\mathop{\mathbb{E}}_{\mu_{i,l(i)}} [\phi] \right) - \mathop{\mathbb{E}}_{\mu_{i+N,l(i+N)}} [\phi] \right) + 2\beta \sum_{j \neq l(i)} \left(\mathop{\mathbb{E}}_{\mu_{i+N,j}} [\phi] - \mathop{\mathbb{E}}_{\mu_{i,j}} [\phi] \right) \right).$$

Given $\boldsymbol{\mu}$ and l, we define the normalized weighted sum $b(\boldsymbol{\mu}, l)$ as

$$\begin{aligned} b(\boldsymbol{\mu}, l) :=& \frac{2k}{3-\beta} \cdot d_\mathcal{F}(q_2, p_{1,l}) \\ =& \sup_{\phi \in \mathcal{F}} \frac{1}{(3-\beta)N} \sum_{i=1}^{N} \left((1-\beta) \left(\mathop{\mathbb{E}}_{\mu_{i,l(i)}} [\phi] \right) - \mathop{\mathbb{E}}_{\mu_{i+N,l(i+N)}} [\phi] \right) + 2\beta \sum_{j \neq l(i)} \left(\mathop{\mathbb{E}}_{\mu_{i+N,j}} [\phi] - \mathop{\mathbb{E}}_{\mu_{i,j}} [\phi] \right) \right). \end{aligned}$$

We can show that $a(\boldsymbol{\mu}, l) \leq \gamma^+$ and $b(\boldsymbol{\mu}, l) \geq \gamma^-$. By Lemma 10, we have that $a(\boldsymbol{\mu}, l) \leq \gamma^+$ for all $l \in [k]^{[2N]}$. Since $\bigcup X_{i,j}$ can be γ^--shattered by $\mathcal{F}$ at the same threshold, we also have

$$b(\boldsymbol{\mu}, l) \geq \gamma^- .$$

Combining the two bounds, we obtain

$$\frac{d_{\mathcal{F}}(q_2, p_{1,l})}{d_{\mathcal{F}}(q_1, p_{1,l})} \geq \frac{3-\beta}{1+\beta} \cdot \frac{\gamma_-}{\gamma_+} \geq 3 - \tau.$$

Thus

$$d_{\mathcal{F}}(q_2, p_{1,l}) \geq (3-\tau)\, d_{\mathcal{F}}(q_1, p_{1,l}).$$

So conditional on the event that the evaluation algorithm returns the wrong hypothesis, we have

$$d_{\mathcal{F}}(q^\star, \widehat{q}) \geq (3-\tau) \min\left\{d_{\mathcal{F}}(q^\star, q_1),\, d_{\mathcal{F}}(q^\star, q_2)\right\}.$$

Now, for any $c' < 3$, if we take τ such that $3 - \tau > c'$, by Lemma 9, we have that $d_{\mathcal{F}}$ is not c'-weakly evaluable, proving that the scale $c = 3$ is optimal. □

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction provide a clear description of our main results.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: This is a theory paper, so we present any limitation as an open question. Refer to the discussion after Theorem 2.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

Justification: We provide full proofs in the appendix and have detailed proof overviews in the main paper.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A]

   Justification: This is a theoretical work, and the paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A]

Justification: This is a theoretical work, and the paper does not include experiments.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (https://neurips.cc/public/guides/CodeSubmissionPolicy) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A]

   Justification: This is a theoretical work, and the paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A]

   Justification: This is a theoretical work, and the paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A]

   Justification: This is a theoretical work, and the paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our work does not involve any human participants, datasets, or potentially harmful societal impact.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A]

    Justification: There is no major societal impact of our work, as this is a theory paper.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: This is a theoretical work, and we have no experiments involving data or models.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A]

    Justification: The paper does not use existing assets.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A]

    Justification: This is a theoretical work and does not release new assets.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: This work does not involve LLMs in any important, original, or non-standard components.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.