VC Classes are Adversarially Robustly Learnable, but Only Improperly

Omar Montasser OMAR@TTIC.EDU
Steve Hanneke STEVE.HANNEKE@GMAIL.COM
Nathan Srebro NATI@TTIC.EDU
Toyota Technological Institute at Chicago, Chicago IL, USA

Editors: Alina Beygelzimer and Daniel Hsu

Abstract

We study the question of learning an adversarially robust predictor. We show that any hypothesis class $\mathcal{H}$ with finite VC dimension is robustly PAC learnable with an *improper* learning rule. The requirement of being improper is necessary as we exhibit examples of hypothesis classes $\mathcal{H}$ with finite VC dimension that are *not* robustly PAC learnable with any *proper* learning rule.

Keywords: adversarial robustness, PAC learning, sample complexity, improper learning.

1. Introduction

Learning predictors that are robust to adversarial perturbations is an important challenge in contemporary machine learning. There has been a lot of interest lately in how predictors learned by deep learning are *not* robust to adversarial examples (Szegedy et al., 2013; Biggio et al., 2013; Goodfellow et al., 2014), and there is an ongoing effort to devise methods for learning predictors that *are* adversarially robust. In this paper, we consider the problem of learning, based on a (non-adversarial) i.i.d. sample, a predictor that is robust to adversarial examples at test time. We emphasize that this is distinct from the learning process itself being robust to an adversarial training set.

Given an instance space $\mathcal{X}$ and label space $\mathcal{Y} = \{+1, -1\}$, we formalize an adversary we would like to protect against as $\mathcal{U} : \mathcal{X} \mapsto 2^{\mathcal{X}}$, where $\mathcal{U}(x) \subseteq \mathcal{X}$ represents the set of perturbations (adversarial examples) that can be chosen by the adversary at test time. For example, $\mathcal{U}$ could be perturbations of distance at most γ w.r.t. some metric ρ, such as the ℓ_∞ metric considered in many applications: $\mathcal{U}(x) = \{z \in \mathcal{X} : \|x - z\|_\infty \leq \gamma\}$. Our only (implicit) restriction on the specification of $\mathcal{U}$ is that $\mathcal{U}(x)$ should be nonempty for every x. For a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, we observe m i.i.d. samples $S \sim \mathcal{D}^m$, and our goal is to learn a predictor $\hat{h} : \mathcal{X} \mapsto \mathcal{Y}$ having small robust risk,

$$\mathrm{R}_{\mathcal{U}}(\hat{h}; \mathcal{D}) := \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\sup_{z\in\mathcal{U}(x)} \mathbb{1}[\hat{h}(z) \neq y]\right].$$

The common approach to adversarially robust learning is to pick a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ (e.g. neural networks) and learn through robust *empirical* risk minimization:

$$\hat{h} \in \mathrm{RERM}_{\mathcal{H}}(S) := \operatorname*{argmin}_{h\in\mathcal{H}} \hat{\mathrm{R}}_{\mathcal{U}}(h; S)$$

where $\hat{\mathrm{R}}_{\mathcal{U}}(h; S) = \frac{1}{m}\sum_{(x,y)\in S} \sup_{z\in\mathcal{U}(x)} \mathbb{1}[h(z) \neq y]$. Most work on the problem has focused on computational approaches to solve this empirical optimization problem, or related problems of

minimizing a robust version of some surrogate loss instead of the 0/1 loss (Madry et al., 2017; Wong and Kolter, 2018; Raghunathan et al., 2018a,b). But of course our true objective is not the empirical robust risk $\hat{\mathrm{R}}_{\mathcal{U}}(h;S)$, but rather the population robust risk $\mathrm{R}_{\mathcal{U}}(h;\mathcal{D})$.

How can we ensure that $\mathrm{R}_{\mathcal{U}}(h;\mathcal{D})$ is small? All prior approaches that we are aware of for ensuring adversarially robust generalization are based on uniform convergence, i.e. showing that w.h.p. for all predictors $h \in \mathcal{H}$, the estimation error $|\mathrm{R}_{\mathcal{U}}(h;\mathcal{D}) - \hat{\mathrm{R}}_{\mathcal{U}}(h;S)|$ is small, perhaps for some surrogate loss (Bubeck et al., 2018; Cullina et al., 2018; Khim and Loh, 2018; Yin et al., 2018). Such approaches justify RERM, and in particular yield M-estimation type *proper* learning rules: we are learning a hypothesis class by choosing a predictor in the class that minimizes some empirical functional. For standard supervised learning we know that proper learning, and specifically ERM, is sufficient for learning, and so it is sensible to limit attention to such methods.

But it has also been observed in practice that the adversarial error does not generalize as well as the standard error, i.e. there can be a large gap between $\mathrm{R}_{\mathcal{U}}(h;\mathcal{D})$ and $\hat{\mathrm{R}}_{\mathcal{U}}(h;S)$ even when their non-robust versions are similar (Schmidt et al., 2018). This suggests that perhaps the robust risk does not concentrate as well as the standard risk, and so RERM in adversarially robust learning might not work as well as ERM in standard supervised learning. Does this mean that such problems are not adversarially robustly learnable? Or is it perhaps that proper learners might not be sufficient?

In this paper we aim to characterize which hypothesis classes are adversarially robustly learnable, and using what learning rules. That is, for a given hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ and adversary $\mathcal{U}$, we ask whether it is possible, based on an i.i.d. sample to learn a predictor h that has population robust risk almost as good as any predictor in $\mathcal{H}$ (see Definition 1 in Section 2). We discover a stark contrast between *proper* learning rules which output predictors in $\mathcal{H}$, and *improper* learning rules which are not constrained to predictors in $\mathcal{H}$. Our main results are:

- We show that there exists an adversary $\mathcal{U}$ and a hypothesis class $\mathcal{H}$ with finite VC dimension that *cannot* be robustly PAC learned with any *proper* learning rule (including RERM).
- We show that for any adversary $\mathcal{U}$ and any hypothesis class $\mathcal{H}$ with finite VC dimension, there exists an *improper* learning rule that can robustly PAC learn $\mathcal{H}$ (although with sample complexity that is sometimes exponential in the VC dimension).

Our results suggest that we should start considering *improper* learning rules to ensure adversarially robust generalization. They also demonstrate that previous approaches to adversarially robust generalization are not always sufficient, as all prior work we are aware of is based on uniform convergence of the robust risk, either directly for the loss of interest (Bubeck et al., 2018; Cullina et al., 2018) or some carefully constructed surrogate loss (Khim and Loh, 2018; Yin et al., 2018), which would still justify the use of M-estimation type proper learning. The approach of Attias et al. (2018) for the case where $|\mathcal{U}(x)| \leq k$ (i.e. finite number of perturbations) is most similar to ours, as it uses an improper learning rule, but their analysis is still based on uniform convergence and so would apply also to RERM (the improperness is introduced only for computational, not statistical, reasons). Also, in this specific case, our approach would give an improved sample complexity that scales only roughly logarithmically with k, as opposed to the roughly linear scaling in Attias et al. (2018)—see discussion at the end of Section 4 for details.

A related negative result was presented by Schmidt et al. (2018), where they showed that there exists a family of distributions (namely, mixtures of two d-dimensional spherical Gaussians) where the sample complexity for standard learning is $O(1)$, but the sample complexity for adversarially

robust learning is at least $\Omega(\frac{\sqrt{d}}{\log d})$. This an interesting instance where there is a large separation in sample complexity between standard learning and robust learning. But distribution-specific learning is known to be less easily characterizable, with the uniform convergence not being necessary for learning, and ERM not always being optimal, even for standard (non-robust) supervised learning. In this paper we focus on "worst case" distribution-free robust learning, as in standard PAC learnability.

A different notion of robust learning was studied by Xu and Mannor (2012). They use empirical robustness as a design technique for learning rules, but their goal, and the guarantees they establish are on the standard non-robust population risk, and so do not inform us about robust learnability.

2. Problem Setup

We are interested in studying the sample complexity of adversarially robust PAC learning in the realizable and agnostic settings. Given a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$, our goal is to design a learning rule $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \mapsto \mathcal{Y}^{\mathcal{X}}$ such that for any distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, the rule $\mathcal{A}$ will find a predictor that competes with the best predictor $h^* \in \mathcal{H}$ in terms of the robust risk using a number of samples that is independent of the distribution $\mathcal{D}$. The following definitions formalize the notion of robust PAC learning in the realizable and agnostic settings:[1]

Definition 1 (Agnostic Robust PAC Learnability) *For any $\varepsilon, \delta \in (0,1)$, the* sample complexity of agnostic robust (ε, δ)−PAC learning of $\mathcal{H}$ with respect to adversary $\mathcal{U}$, *denoted $\mathcal{M}_{\mathrm{AG}}(\varepsilon, \delta; \mathcal{H}, \mathcal{U})$, is defined as the smallest $m \in \mathbb{N} \cup \{0\}$ for which there exists a learning rule $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \mapsto \mathcal{Y}^{\mathcal{X}}$ such that, for every data distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, with probability at least $1 - \delta$ over $S \sim \mathcal{D}^m$,*

$$\mathrm{R}_{\mathcal{U}}(\mathcal{A}(S); \mathcal{D}) \leq \inf_{h \in \mathcal{H}} \mathrm{R}_{\mathcal{U}}(h; \mathcal{D}) + \varepsilon.$$

If no such m exists, define $\mathcal{M}_{\mathrm{AG}}(\varepsilon, \delta; \mathcal{H}, \mathcal{U}) = \infty$. We say that $\mathcal{H}$ is robustly PAC learnable in the agnostic setting with respect to adversary $\mathcal{U}$ if $\forall \varepsilon, \delta \in (0,1)$, $\mathcal{M}_{\mathrm{AG}}(\varepsilon, \delta; \mathcal{H}, \mathcal{U})$ is finite.

Definition 2 (Realizable Robust PAC Learnability) *For any $\varepsilon, \delta \in (0,1)$, the* sample complexity of realizable robust (ε, δ)-PAC learning of $\mathcal{H}$ with respect to adversary $\mathcal{U}$, *denoted $\mathcal{M}_{\mathrm{RE}}(\varepsilon, \delta; \mathcal{H}, \mathcal{U})$, is defined as the smallest $m \in \mathbb{N} \cup \{0\}$ for which there exists a learning rule $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \mapsto \mathcal{Y}^{\mathcal{X}}$ such that, for every data distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ where there exists a predictor $h^* \in \mathcal{H}$ with zero robust risk, $\mathrm{R}_{\mathcal{U}}(h^*; \mathcal{D}) = 0$, with probability at least $1 - \delta$ over $S \sim \mathcal{D}^m$,*

$$\mathrm{R}_{\mathcal{U}}(\mathcal{A}(S); \mathcal{D}) \leq \varepsilon.$$

If no such m exists, define $\mathcal{M}_{\mathrm{RE}}(\varepsilon, \delta; \mathcal{H}, \mathcal{U})) = \infty$. We say that $\mathcal{H}$ is robustly PAC learnable in the realizable setting with respect to adversary $\mathcal{U}$ if $\forall \varepsilon, \delta \in (0,1)$, $\mathcal{M}_{\mathrm{RE}}(\varepsilon, \delta; \mathcal{H}, \mathcal{U})$ is finite.

Definition 3 (Proper Learnability) *We say that $\mathcal{H}$ is* properly *robustly PAC learnable (in the agnostic or realizable setting) if it can be learned as in Definitions 1 or 2 using a learning rule $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \mapsto \mathcal{H}$ that always outputs a predictor in $\mathcal{H}$. We refer to learning using any learning rule $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \mapsto \mathcal{Y}^{\mathcal{X}}$, as in the definitions above, as* improper *learning.*

1. We implicitly suppose that the hypotheses h in $\mathcal{H}$ and their losses $\sup_{z \in \mathcal{U}(x)} \mathbb{1}[h(z) \neq y]$ are measurable, and that standard mild restrictions on $\mathcal{H}$ are imposed to guarantee measurability of empirical processes, so that the standard tools of VC theory apply. See Blumer, Ehrenfeucht, Haussler, and Warmuth (1989); van der Vaart and Wellner (1996) for discussion of such measurability issues, which we will not mention again in the remainder of this article.

We also denote $\mathrm{er}(h;\mathcal{D}) = \mathbb{P}(h(x) \neq y)$, the (non-robust) error rate under the 0-1 loss, and $\hat{\mathrm{er}}(h;S) = \frac{1}{|S|}\sum_{(x,y)\in S}\mathbb{1}[h(x) \neq y]$ the empirical error rate. These agree with the robust variant when $\mathcal{U}(x) = \{x\}$, and so robust learnability agrees with standard supervised learning when $\mathcal{U}(x) = \{x\}$. For more powerful adversaries, robust learnability is a special case of Vapink's "General Learning" (Vapnik, 1982), but can not, in general, be phrased in terms of supervised learning of some modified hypothesis class or loss. We recall the Vapnik-Chervonenkis dimension (VC dimension) is defined as follows,

Definition 4 (VC dimension) *We say that a sequence $\{x_1,\ldots,x_k\} \in \mathcal{X}$ is shattered by $\mathcal{H}$ if $\forall y_1,\ldots,y_k \in \mathcal{Y}, \exists h \in \mathcal{H}$ such that $\forall i \in [k], h(x_i) = y_i$. The VC dimension of $\mathcal{H}$ (denoted $\mathrm{vc}(\mathcal{H})$) is then defined as the largest integer k for which there exists $\{x_1,\ldots,x_k\} \in \mathcal{X}$ that is shattered by $\mathcal{H}$. If no such k exists, then $\mathrm{vc}(\mathcal{H})$ is said to be infinite.*

In the standard PAC learning framework, we know that a hypothesis class $\mathcal{H}$ is PAC learnable if and only if the VC dimension of $\mathcal{H}$ is finite (Vapnik and Chervonenkis, 1971, 1974; Blumer et al., 1989; Ehrenfeucht et al., 1989). In particular, $\mathcal{H}$ is properly PAC learnable with $\mathrm{ERM}_{\mathcal{H}}$ and therefore proper learning is sufficient for supervised learning. A natural question to ask, based on the definition of robust PAC learning, is what is a necessary and sufficient condition on $\mathcal{H}$ that implies that it is robustly PAC learnable with respect to adversary $\mathcal{U}$. We can easily obtain a sufficient condition based on Vapink's "General Learning" (Vapnik, 1982). Denote by $\mathcal{L}^{\mathcal{U}}_{\mathcal{H}}$ the robust loss class of $\mathcal{H}$,

$$\mathcal{L}^{\mathcal{U}}_{\mathcal{H}} = \left\{ (x,y) \mapsto \sup_{z\in\mathcal{U}(x)} \mathbb{1}[h(z) \neq y] : h \in \mathcal{H} \right\}.$$

If the robust loss class $\mathcal{L}^{\mathcal{U}}_{\mathcal{H}}$ has finite VC dimension ($\mathrm{vc}(\mathcal{L}^{\mathcal{U}}_{\mathcal{H}}) < \infty$), then $\mathcal{H}$ is robustly PAC learnable with $\mathrm{RERM}_{\mathcal{H}}$ and sample complexity that scales linearly with $\mathrm{vc}(\mathcal{L}^{\mathcal{U}}_{\mathcal{H}})$. One might then wish to relate the VC dimension of the hypothesis class ($\mathrm{vc}(\mathcal{H})$) to the VC dimension of the robust loss class ($\mathrm{vc}(\mathcal{L}^{\mathcal{U}}_{\mathcal{H}})$). But as we show in Sections 3 and 5, there can be arbitrarily large gaps between them.

As mentioned earlier, for supervised learning finite VC dimension of the loss class (which is equal to the VC dimension of the hypothesis class) is also necessary for learning. For general learning, unlike supervised learning, the loss class having finite VC dimension, and uniform convergence over this class, is not, in general, necessary, and rules other than ERM might be needed for learning (e.g. Vapnik, 1982; Shalev-Shwartz et al., 2009; Daniely et al., 2015). In the following Sections, we show that this is also the case for robust learning. We show that $\mathrm{vc}(\mathcal{L}^{\mathcal{U}}_{\mathcal{H}})$ can be arbitrarily larger, we might not have uniform convergence, RERM might not ensure learning, while the problem is still learnable with a different (improper, in our case) learning rule.

3. Sometimes There are no Proper Robust Learners

We start by showing that even for hypothesis classes with finite VC dimension, indeed even if $\mathrm{vc}(\mathcal{H}) = 1$, robust PAC learning might not be possible using *any* proper learning rule. In particular, even if there is a robust predictor in $\mathcal{H}$, and even with an unbounded number of samples, RERM (or any other M-estimator or other proper learning rules), will not ensure a low robust risk.

Theorem 1 *There exists a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ with $\mathrm{vc}(\mathcal{H}) \leq 1$ and an adversary $\mathcal{U}$ such that $\mathcal{H}$ is not properly robustly PAC learnable with respect to $\mathcal{U}$ in the realizable setting.*

This result implies that finite VC dimension of a hypothesis class $\mathcal{H}$ is not sufficient for robust PAC learning if we want to use *proper* learning rules. For the proofs in this section, we will fix an instance space $\mathcal{X} = \mathbb{R}^d$ equipped with a metric ρ, and an adversary $\mathcal{U} : \mathcal{X} \mapsto 2^{\mathcal{X}}$ such that $\mathcal{U}(x) = \{z \in \mathcal{X} : \rho(x, z) \leq \gamma\}$ for all $x \in \mathcal{X}$ for some $\gamma > 0$. First, we prove a lemma that shows that there exists a hypothesis class $\mathcal{H}$ where there is an arbitrarily large gap between the VC dimension of $\mathcal{H}$ and the VC dimension of the robust loss class of $\mathcal{H}$,

Lemma 2 *Let $m \in \mathbb{N}$. Then, there exists $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ such that $\mathrm{vc}(\mathcal{H}) \leq 1$ but $\mathrm{vc}(\mathcal{L}^{\mathcal{U}}_{\mathcal{H}}) \geq m$.*

Proof Pick m points $x_1, \ldots, x_m$ in $\mathcal{X}$ such that for all $i, j \in [m], \mathcal{U}(x_i) \cap \mathcal{U}(x_j) = \emptyset$. In other words, we want the perturbation sets $\mathcal{U}(x_1), \ldots, \mathcal{U}(x_m)$ to be mutually disjoint.

We will construct a hypothesis class $\mathcal{H}$ in the following iterative manner. Initialize set $\mathcal{Z} = \{x_1, \ldots, x_m\}$. For each bit string $b \in \{0, 1\}^m$, initialize $Z_b = \emptyset$. For each $i \in [m]$, if $b_i = 1$ then pick a point $z \in \mathcal{U}(x_i) \setminus \mathcal{Z}$ and add it to Z_b, i.e. $Z_b = Z_b \cup \{z\}$. Once we finish picking points based on all bits that are set to 1, we add Z_b to $\mathcal{Z}$ (i.e. $\mathcal{Z} = \mathcal{Z} \cup Z_b$). We define $h_b : \mathcal{X} \to \mathcal{Y}$ as:

$$h_b(x) = \begin{cases} +1 & \text{if } x \notin Z_b \\ -1 & \text{if } x \in Z_b \end{cases}$$

Then, let $\mathcal{H} = \{h_b : b \in \{0, 1\}^m\}$. We can think of each mapping h_b as being characterized by a unique signature Z_b that indicates the points that it labels with -1. These points are carefully picked such that, first, they are inside the perturbation sets of $x_1, \ldots, x_m$; and second, no two mappings label the same point with -1, i.e. for any $b, b' \in \{0, 1\}^m$, where $b \neq b'$, $Z_b \cap Z'_b = \emptyset$. Also, we make sure that all mappings in $\mathcal{H}$ label the set $\{x_1, \ldots, x_m\}$ with $+1$.

Next, we proceed with proving two claims about $\mathcal{H}$. First, that $\mathrm{vc}(\mathcal{H}) \leq 1$. Pick any two points $z_1, z_2 \in \mathcal{X}$. Consider the following cases. In case z_1 or z_2 is in $\mathcal{X} \setminus \mathcal{Z}$. Suppose W.L.O.G that $z_2 \in \mathcal{X} \setminus \mathcal{Z}$. Then we know that all mappings label z_2 in the same way with label $+1$, because for all $b \in \{0, 1\}^m, z_2 \notin Z_b$. Therefore, we cannot shatter z_1, z_2 with $\mathcal{H}$. In case z_1 and z_2 are both in $\mathcal{Z}$. Since by our construction, $\mathcal{Z} = \cup_{b \in \{0,1\}^m} Z_b$ and $Z_b \cap Z'_b = \emptyset$ for any $b \neq b'$, we have two sub-cases. Either $z_1, z_2 \in Z_b$ for some $b \in \{0, 1\}^m$, which means that the only labelings we can obtain are $(-1, -1)$ with h_b, and $(+1, +1)$ with h'_b for any $b' \neq b$. Second case is that $z_1 \in Z_b$ and $z_2 \in Z_{b'}$ for $b \neq b', b, b' \in \{0, 1\}^m$. By our construction, we know that we cannot label both points z_1 and z_2 with $(-1, -1)$, because they don't belong to the same set. Therefore, in both subcases, we cannot shatter z_1, z_2 with $\mathcal{H}$. This concludes that $\mathrm{vc}(\mathcal{H}) \leq 1$.

Second, we will show that $\mathrm{vc}(\mathcal{L}^{\mathcal{U}}_{\mathcal{H}}) \geq m$. Consider the set $S = \{(x_1, +), \ldots, (x_m, +)\}$. We will show that $\mathcal{L}^{\mathcal{U}}_{\mathcal{H}}$ shatters S. Pick any labeling $y \in \{0, 1\}^m$. Note that by construction of $\mathcal{H}$, $\exists h_b \in \mathcal{H}$ such that $b = y$. Then, for each $i \in [m]$, $\sup_{z \in \mathcal{U}(x_i)} \mathbb{1}[h_b(z) \neq +1] = b_i = y_i$. This shows that $\mathcal{L}^{\mathcal{U}}_{\mathcal{H}}$ shatters S, and therefore $\mathrm{vc}(\mathcal{L}^{\mathcal{U}}_{\mathcal{H}}) \geq m$. ■

The following lemma (proof provided in Appendix A) establishes that for any sample size $m \in \mathbb{N}$, there exists a hypothesis class $\mathcal{H}$ with $\mathrm{vc}(\mathcal{H}) \leq 1$ such that any *proper* learning rule will fail in learning a robust classifier if it observes at most m samples but not more.

Lemma 3 *Let $m \in \mathbb{N}$. Then, there exists $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ with $\mathrm{vc}(\mathcal{H}) \leq 1$ such that for any proper learning rule $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \mapsto \mathcal{H}$,*

- *$\exists$ a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ and a predictor $h^* \in \mathcal{H}$ where $\mathrm{R}_{\mathcal{U}}(h^*; \mathcal{D}) = 0$.*
- *With probability at least $1/7$ over $S \sim \mathcal{D}^m$, $\mathrm{R}_{\mathcal{U}}(\mathcal{A}(S); \mathcal{D}) > 1/8$.*

We now proceed with the proof of Theorem 1.

Proof [of Theorem 1] Let $(X_m)_{m\in\mathbb{N}}$ be an infinite sequence of sets such that each set X_m contains $3m$ distinct points from $\mathcal{X}$, where for any $x_i, x_j \in \cup_{m=1}^{\infty} X_m$ such that $x_i \neq x_j$ we have $\mathcal{U}(x_i) \cap \mathcal{U}(x_j) = \emptyset$. Foreach $m \in \mathbb{N}$, construct $\mathcal{H}_m$ on X_m as in Lemma 3. We want to ensure that predictors in $\mathcal{H}_m$ are non-robust on the points in $X_{m'}$ for all $m' \neq m$, by doing the following adjustment for each $h_b \in \mathcal{H}_m$ (recall from Lemma 2 that each predictor has its own unique signature Z_b),

$$h_b(x) = \begin{cases} -1 & \text{if } x \in Z_b \text{ or } x \in X_{m'} \text{ for } m' \neq m \\ +1 & \text{otherwise} \end{cases}$$

Let $\mathcal{H} = \cup_{m=1}^{\infty} \mathcal{H}_m$. We will show that $\mathrm{vc}(\mathcal{H}) \leq 1$. Pick any two points $z_1, z_2 \in \mathcal{X}$. There are six cases to consider. In case both z_1 and z_2 are in X_m for some $m \in \mathbb{N}$, then we only obtain the labelings $(+1,+1)$ (by predictors from $\mathcal{H}_m$) and $(-1,-1)$ (by predictors from $\mathcal{H}_{m'}$ with $m' \neq m$). In case both z_1 and z_2 are in $\mathcal{U}(X_m) \setminus X_m$, then they are not shattered by Lemma 2. In case $z_1 \in X_i$ and $z_2 \in X_j$ for $i \neq j$, then we can only obtain the labelings $(+1,-1)$ (by predictors in $\mathcal{H}_i$), $(-1,+1)$ (by predictors in $\mathcal{H}_j$), and $(-1,-1)$ (by predictors in $\mathcal{H}_k$ for $k \neq i,j$). In case $z_1 \in X_i$ and $z_2 \in \mathcal{U}(X_j) \setminus X_j$ for $j \neq i$, then we can't obtain the labeling $(+1,-1)$. In case $z_1 \in \mathcal{U}(X_i) \setminus X_i$ and $z_2 \in \mathcal{U}(X_j) \setminus X_j$ for $i \neq j$, then we can't obtain the labeling $(-1,-1)$. Finally, if either z_1 or z_2 is in $\mathcal{X}$ but not in $\cup_{m=1}^{\infty} X_m$ and not in $\cup_{m=1}^{\infty} \mathcal{U}(X_m)$, then all predictors label z_1 or z_2 with $+1$, and so we can't shatter them. This shows that $\mathrm{vc}(\mathcal{H}) \leq 1$.

By Lemma 3, it follows that for any proper learning rule $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \mapsto \mathcal{H}$ and for any $m \in \mathbb{N}$, we can construct a distribution $\mathcal{D}$ over $X_m \times \mathcal{Y}$ where there exists a predictor $h^* \in \mathcal{H}_m$ with $\mathrm{R}_{\mathcal{U}}(h^*; \mathcal{D}) = 0$, but with probability at least $1/7$ over $S \sim \mathcal{D}^m$, $\mathrm{R}_{\mathcal{U}}(\mathcal{A}(S); \mathcal{D}) > 1/8$. This works because classifiers from classes $\mathcal{H}_{m'}$ where $m' \neq m$ make mistakes on points in X_m and so they are non-robust. Thus, rule $\mathcal{A}$ will do worse if it picks predictors from these classes. This shows that the sample complexity to properly robustly PAC learn $\mathcal{H}$ is infinite. This concludes that $\mathcal{H}$ is not properly robustly PAC learnable. ■

4. Finite VC Dimension is Sufficient for (Improper) Robust Learnability

In the previous section we saw that finite VC dimension is *not* sufficient for *proper* robust learnability. We now show that it *is* sufficient for *improper* robust learnability, thus (1) establishing that if $\mathcal{H}$ is learnable, it is also robustly learnable, albeit possibly with a higher sample complexity; and (2) unlike the standard supervised learning setting, to achieve learnability we might need to escape properness, as improper learning is necessary for some hypothesis classes.

We begin, in Section 4.1 with the realizable case, i.e. where there exists $h^* \in \mathcal{H}$ with zero robust risk. Then in Section 4.2 we turn to the agnostic setting, and observe that a version of a recent reduction by David, Moran, and Yehudayoff (2016) from agnostic to realizable learning applies also for robust learning. We thus establish agnostic robust learnability of finite VC classes by using this reduction and relying on the realizable learning result of Section 4.1.

4.1. Realizable Robust Learnability

We will in fact establish a bound in terms of the *dual VC dimension*. Formally, for each $x \in \mathcal{X}$, define a function $g_x : \mathcal{H} \to \mathcal{Y}$ such that $g_x(h) = h(x)$ for each $h \in \mathcal{H}$. Then the dual VC dimension of $\mathcal{H}$, denoted $\mathrm{vc}^*(\mathcal{H})$, is defined as the VC dimension of the set $\mathcal{G} = \{g_x : x \in \mathcal{X}\}$. This quantity is known to satisfy $\mathrm{vc}^*(\mathcal{H}) < 2^{\mathrm{vc}(\mathcal{H})+1}$ (Assouad, 1983), though for many spaces it satisfies $\mathrm{vc}^*(\mathcal{H}) = O(\mathrm{poly}(\mathrm{vc}(\mathcal{H})))$ or even, as is the case for linear separators, $\mathrm{vc}^*(\mathcal{H}) = O(\mathrm{vc}(\mathcal{H}))$.

Theorem 4 *For any $\mathcal{H}$ and $\mathcal{U}$, $\forall \varepsilon, \delta \in (0, 1/2)$,*

$$\mathcal{M}_{\mathrm{RE}}(\varepsilon, \delta; \mathcal{H}, \mathcal{U}) = O\left(\mathrm{vc}(\mathcal{H})\mathrm{vc}^*(\mathcal{H})\frac{1}{\varepsilon}\log\left(\frac{\mathrm{vc}(\mathcal{H})\mathrm{vc}^*(\mathcal{H})}{\varepsilon}\right) + \frac{1}{\varepsilon}\log\left(\frac{1}{\delta}\right)\right),$$

Since Assouad (1983) has shown $\mathrm{vc}^*(\mathcal{H}) < 2^{\mathrm{vc}(\mathcal{H})+1}$, this implies the following corollary.

Corollary 5 *For any $\mathcal{H}$ and $\mathcal{U}$, $\forall \varepsilon, \delta \in (0, 1/2)$,*

$$\mathcal{M}_{\mathrm{RE}}(\varepsilon, \delta; \mathcal{H}, \mathcal{U}) = 2^{O(\mathrm{vc}(\mathcal{H}))}\frac{1}{\varepsilon}\log\left(\frac{1}{\varepsilon}\right) + O\left(\frac{1}{\varepsilon}\log\left(\frac{1}{\delta}\right)\right).$$

Our approach to this proof is via *sample compression* arguments. Specifically, we make use of a lemma (Lemma 11 in Appendix 4.2), which extends to the robust loss the classic compression-based generalization guarantees from the 0-1 loss. We now proceed with the proof of Theorem 4.

Proof [of Theorem 4] The learning algorithm achieving this bound is a modification of a sample compression scheme recently proposed by Moran and Yehudayoff (2016), or more precisely, a variant of that method explored by Hanneke, Kontorovich, and Sadigurschi (2019). Our modification forces the compression scheme to also have zero empirical *robust* loss. Fix $\varepsilon, \delta \in (0, 1)$ and a sample size $m > 2\mathrm{vc}(\mathcal{H})$, and denote by P any distribution with $\inf_{h\in\mathcal{H}} R_{\mathcal{U}}(h; P) = 0$.

By classic PAC learning guarantees (Vapnik and Chervonenkis, 1974; Blumer et al., 1989), there is a positive integer $n = O(\mathrm{vc}(\mathcal{H}))$ with the property that, for any distribution D over $\mathcal{X} \times \mathcal{Y}$ with $\inf_{h\in\mathcal{H}} \mathrm{er}(h; D) = 0$, for n iid D-distributed samples $S' = \{(x'_1, y'_1), \ldots, (x'_n, y'_n)\}$, with nonzero probability, every $h \in \mathcal{H}$ satisfying $\hat{\mathrm{er}}(h; S') = 0$ also has $\mathrm{er}(h; D) < 1/3$.

Fix a deterministic function $\mathrm{RERM}_{\mathcal{H}}$ mapping any labeled data set to a classifier in $\mathcal{H}$ robustly consistent with the labels in the data set, if a robustly consistent classifier exists (i.e., having zero $\hat{\mathrm{R}}_{\mathcal{U}}$ on the given data set). Suppose we are given training examples $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ as input to the learner. Under the assumption that this is an iid sample from a robustly realizable distribution, we suppose $\hat{R}_{\mathcal{U}}(\mathrm{RERM}_{\mathcal{H}}(S); S) = 0$, which should hold with probability one. Denote by $I(x) = \min\{i \in \{1, \ldots, m\} : x \in \mathcal{U}(x_i)\}$ for every $x \in \bigcup_{i\le m} \mathcal{U}(x_i)$. Before we can apply the compression approach, we first need to *inflate* the data set to a (potentially infinite) larger set, and then *discretize* it to again reduce it back to a finite sample size. Denote by $\hat{\mathcal{H}} = \{\mathrm{RERM}_{\mathcal{H}}(L) : L \subseteq S, |L| = n\}$. Note that $|\hat{\mathcal{H}}| \le |\{L : L \subseteq S, |L| = n\}| = \binom{m}{n} \le \left(\frac{em}{n}\right)^n$. Define an *inflated* data set $S_{\mathcal{U}} = \bigcup_{i\le m}\{(x, y_{I(x)}) : x \in \mathcal{U}(x_i)\}$. As it is difficult to handle this potentially-infinite set in an algorithm, we consider a discretized version of it. Specifically, consider a *dual space* $\mathcal{G}$: a set of functions $g_{(x,y)} : \mathcal{H} \to \{0, 1\}$ defined as $g_{(x,y)}(h) = \mathbb{1}[h(x) \ne y]$, for each $h \in \mathcal{H}$ and each $(x, y) \in S_{\mathcal{U}}$. The VC dimension of $\mathcal{G}$ is at most the *dual VC dimension* of $\mathcal{H}$: $\mathrm{vc}^*(\mathcal{H})$, which is known to satisfy $\mathrm{vc}^*(\mathcal{H}) < 2^{\mathrm{vc}(\mathcal{H})+1}$ (Assouad, 1983). Now denote by $\hat{S}_{\mathcal{U}}$ a subset of $S_{\mathcal{U}}$ which includes exactly one $(x, y) \in S_{\mathcal{U}}$ for each distinct classification $\{g_{(x,y)}(h)\}_{h\in\hat{\mathcal{H}}}$ of $\hat{\mathcal{H}}$ realized by functions $g_{(x,y)} \in \mathcal{G}$. In particular, by Sauer's lemma (Vapnik and Chervonenkis, 1971; Sauer, 1972), $|\hat{S}_{\mathcal{U}}| \le \left(\frac{e|\hat{\mathcal{H}}|}{\mathrm{vc}^*(\mathcal{H})}\right)^{\mathrm{vc}^*(\mathcal{H})}$, which for $m > 2\mathrm{vc}(\mathcal{H})$ is at most $\left(e^2 m/\mathrm{vc}(\mathcal{H})\right)^{\mathrm{vc}(\mathcal{H})\mathrm{vc}^*(\mathcal{H})}$. In particular, note that for any $T \in \mathbb{N}$ and $h_1, \ldots, h_T \in \hat{\mathcal{H}}$, if $\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}[h_t(x) = y] > \frac{1}{2}$ for every $(x, y) \in \hat{S}_{\mathcal{U}}$, then $\frac{1}{T}\sum_{t=1}^{T} \mathbb{1}[h_t(x) = y] > \frac{1}{2}$ for every $(x, y) \in S_{\mathcal{U}}$ as well, which would further imply $\hat{R}_{\mathcal{U}}(\mathrm{Majority}(h_1, \ldots, h_T); S) = 0$. We will next go about finding such a set of h_t functions.

By our choice of n, we know that for any distribution D over $\hat{S}_{\mathcal{U}}$, n iid samples S' sampled from D would have the property that, with nonzero probability, all $h \in \mathcal{H}$ with $\hat{\mathrm{er}}(h; S') = 0$

also have $\mathrm{er}(h; D) < 1/3$. In particular, this implies at least that there *exists* a subset $S' \subseteq \hat{S}_{\mathcal{U}}$ with $|S'| \leq n$ such that every $h \in \mathcal{H}$ with $\hat{\mathrm{er}}(h; S') = 0$ has $\mathrm{er}(h; D) < 1/3$. For such a set S', note that $\{(x_{I(x)}, y) : (x, y) \in S'\} \subseteq S$, and therefore there exists a set L with $|L| = n$ and $\{(x_{I(x)}, y) : (x, y) \in S'\} \subseteq L \subseteq S$. Furthermore, since $x \in \mathcal{U}(x_{I(x)})$ for every $(x, y) \in S'$, we know $\hat{\mathrm{er}}(\mathrm{RERM}_{\mathcal{H}}(L); S') = 0$, and hence $\mathrm{er}(\mathrm{RERM}_{\mathcal{H}}(L); D) < 1/3$. Altogether, we have that, for any distribution D over $\hat{S}_{\mathcal{U}}$, $\exists h_D \in \hat{\mathcal{H}}$ with $\mathrm{er}(h_D; D) < 1/3$.

We will use the above h_D as a *weak hypothesis* in a boosting algorithm. Specifically, we run the α-Boost algorithm (Schapire and Freund, 2012, Section 6.4.2) with $\hat{S}_{\mathcal{U}}$ as its data set, using the above mapping to produce the weak hypotheses for the distributions D_t produced on each round of the algorithm. As proven in (Schapire and Freund, 2012), for an appropriate a-priori choice of α in the α-Boost algorithm, running this algorithm for $T = O(\log(|\hat{S}_{\mathcal{U}}|))$ rounds suffices to produce a sequence of hypotheses $\hat{h}_1, \ldots, \hat{h}_T \in \hat{\mathcal{H}}$ s.t.

$$\forall (x, y) \in \hat{S}_{\mathcal{U}}, \tfrac{1}{T}\textstyle\sum_{i=1}^{T} \mathbb{1}[h_i(x) = y] \geq \tfrac{5}{9}.$$

From this observation, we already have a sample complexity bound, only slightly worse than the claimed result. Specifically, the above implies that $\hat{h} = \mathrm{Majority}(\hat{h}_1, \ldots, \hat{h}_T)$ satisfies $\hat{R}_{\mathcal{U}}(\hat{h}; S) = 0$. Note that each of these classifiers $\hat{h}_t$ is equal $\mathrm{RERM}_{\mathcal{H}}(L_t)$ for some $L_t \subseteq S$ with $|L_t| = n$. Thus, the classifier $\hat{h}$ is representable as the value of an (order-dependent) reconstruction function ϕ with a compression set size

$$nT = O(\mathrm{vc}(\mathcal{H}) \log(|\hat{S}_{\mathcal{U}}|)) = O(\mathrm{vc}(\mathcal{H})^2 \mathrm{vc}^*(\mathcal{H}) \log(m/\mathrm{vc}(\mathcal{H}))). \tag{1}$$

Thus, invoking Lemma 11, if $m > c\mathrm{vc}(\mathcal{H})^2 \mathrm{vc}^*(\mathcal{H}) \log(\mathrm{vc}(\mathcal{H})\mathrm{vc}^*(\mathcal{H}))$ (for a sufficiently large numerical constant c), we have that with probability at least $1 - \delta$,

$$R_{\mathcal{U}}(\hat{h}; P) \leq O\big(\mathrm{vc}(\mathcal{H})^2 \mathrm{vc}^*(\mathcal{H}) \tfrac{1}{m} \log(m/\mathrm{vc}(\mathcal{H})) \log(m) + \tfrac{1}{m} \log(1/\delta)\big),$$

and setting this less than ε and solving for a sufficient size of m to achieve this yields a sample complexity bound, which is slightly larger than that claimed in Theorem 4. We next proceed to further refine this bound via a sparsification step. However, as an aside, we note that the above intermediate step will be useful in a discussion below, where the size of this compression scheme in the second expression in (1) offers an improvement over a result of Attias, Kontorovich, and Mansour (2018).

Via a technique of (Moran and Yehudayoff, 2016) we can further reduce the above bound. Specifically, since all of $\hat{h}_1, \ldots, \hat{h}_T$ are in $\mathcal{H}$, classic uniform convergence results of Vapnik and Chervonenkis (1971) imply that taking $N = O(\mathrm{vc}^*(\mathcal{H}))$ independent random indices $i_1, \ldots, i_N \sim \mathrm{Uniform}(\{1, \ldots, T\})$, we have $\sup_{(x,y) \in \mathcal{X} \times \mathcal{Y}} \left| \frac{1}{N} \sum_{j=1}^{N} \mathbb{1}[h_{i_j}(x) = y] - \frac{1}{T} \sum_{i=1}^{T} \mathbb{1}[h_i(x) = y] \right| < \frac{1}{18}$. In particular, together with the above guarantee from α-Boost, this implies that there exist indices $i_1, \ldots, i_N \in \{1, \ldots, T\}$ (which may be chosen deterministically) satisfying

$$\forall (x, y) \in \hat{S}_{\mathcal{U}}, \tfrac{1}{T}\textstyle\sum_{j=1}^{N} \mathbb{1}[h_{i_j}(x) = y] \geq -\tfrac{1}{18} + \tfrac{1}{T}\textstyle\sum_{i=1}^{T} \mathbb{1}[h_i(x) = y] > -\tfrac{1}{18} + \tfrac{5}{9} = \tfrac{1}{2},$$

so that the majority vote predictor $\hat{h}'(x) = \mathrm{Majority}(\hat{h}_{i_1}, \ldots, \hat{h}_{i_N})$ satisfies $\hat{\mathrm{er}}(\hat{h}'; \hat{S}_{\mathcal{U}}) = 0$, and hence $\hat{R}_{\mathcal{U}}(\hat{h}'; S) = 0$. Since again, each $\hat{h}_{i_j}$ is the result of $\mathrm{RERM}_{\mathcal{H}}(L_{i_j})$ for some $L_{i_j} \subseteq S$ of size n, we have that $\hat{h}'$ can be represented as the value of an (order-dependent) reconstruction function ϕ with a compression set size $nN = O(\mathrm{vc}(\mathcal{H}))\mathrm{vc}^*(\mathcal{H}))$. Thus, Lemma 11 implies that, for $m \geq c\mathrm{vc}(\mathcal{H})\mathrm{vc}^*(\mathcal{H})$ (for an appropriately large numerical constant c), with probability at least $1 - \delta$, $R_{\mathcal{U}}(\hat{h}'; P) \leq O\big(\mathrm{vc}(\mathcal{H})\mathrm{vc}^*(\mathcal{H}) \frac{1}{m} \log(m) + \frac{1}{m} \log(1/\delta)\big)$. Setting this less than ε and solving for a sufficient size of m to achieve this yields the stated bound. ■

4.2. Agnostic Robust Learnability

For the agnostic case, we can establish an upper bound via reduction to the realizable case, following an argument from David, Moran, and Yehudayoff (2016). Specifically, we have the following result.

Theorem 6 *For any $\mathcal{H}$ and $\mathcal{U}$, $\forall \varepsilon, \delta \in (0, 1/2)$,*

$$\mathcal{M}_{\mathrm{AG}}(\varepsilon, \delta; \mathcal{H}, \mathcal{U}) = O\left(\mathrm{vc}(\mathcal{H})\mathrm{vc}^*(\mathcal{H}) \log(\mathrm{vc}(\mathcal{H})\mathrm{vc}^*(\mathcal{H}))\tfrac{1}{\varepsilon^2} \log^2\left(\tfrac{\mathrm{vc}(\mathcal{H})\mathrm{vc}^*(\mathcal{H})}{\varepsilon}\right) + \tfrac{1}{\varepsilon^2}\log\left(\tfrac{1}{\delta}\right)\right).$$

As above, since Assouad (1983) has shown $\mathrm{vc}^*(\mathcal{H}) < 2^{\mathrm{vc}(\mathcal{H})+1}$, this implies the following corollary.

Corollary 7 *For any $\mathcal{H}$ and $\mathcal{U}$, $\forall \varepsilon, \delta \in (0, 1/2)$,*

$$\mathcal{M}_{\mathrm{AG}}(\varepsilon, \delta; \mathcal{H}, \mathcal{U}) = 2^{O(\mathrm{vc}(\mathcal{H}))} \frac{1}{\varepsilon^2} \log^2\left(\frac{1}{\varepsilon}\right) + O\left(\frac{1}{\varepsilon^2} \log\left(\frac{1}{\delta}\right)\right).$$

We establish the theorem via a reduction to the realizable case, following an approach used by David, Moran, and Yehudayoff (2016), except here applied to the robust loss. The reduction is summarized in the following Theorem, whose proof can be found in Appendix C:

Theorem 8 *Denote $\mathcal{M}_{\mathrm{RE}} = \mathcal{M}_{\mathrm{RE}}(1/3, 1/3; \mathcal{H}, \mathcal{U})$. Then*

$$\mathcal{M}_{\mathrm{AG}}(\varepsilon, \delta; \mathcal{H}, \mathcal{U}) = O\left(\frac{\mathcal{M}_{\mathrm{RE}}}{\varepsilon^2} \log^2\left(\frac{\mathcal{M}_{\mathrm{RE}}}{\varepsilon}\right) + \frac{1}{\varepsilon^2} \log\left(\frac{1}{\delta}\right)\right).$$

From this, Theorem 6 follows immediately by combining Theorem 8 with Theorem 4.

Bounded cardinality confusion sets: As noted in the proof of Theorem 4, the compression size (1) further implies an improvement over a theorem of Attias, Kontorovich, and Mansour (2018). Specifically, Attias et al. considered the case $\max_{x \in \mathcal{X}} |\mathcal{U}(x)| \leq k$ for some fixed $k \in \mathbb{N}$, and presented a learning rule establishing the sample complexity gurantee:

$$\mathcal{M}_{\mathrm{AG}}(\varepsilon, \delta; \mathcal{H}, \mathcal{U}) = O\left(\tfrac{\mathrm{vc}(\mathcal{H})k\log(k)}{\varepsilon^2} + \tfrac{1}{\varepsilon^2}\log\left(\tfrac{1}{\delta}\right)\right). \tag{2}$$

Their analysis proceeds by bounding the Rademacher complexity of the robust loss class of the convex hull of $\mathcal{H}$, which implies the sample complexity (2) can also be achieved by $\mathrm{RERM}_{\mathcal{H}}$ (they propose an alternative, improper, learning rule for computational reasons). But when $\max |\mathcal{U}(x)| \leq k$, the second expression in our (1) would be at most $O(\mathrm{vc}(\mathcal{H}) \log(mk))$. Thus, following the compression argument as in the proof of Theorem 4 would yield the following sample complexity for our improper rule:

$$\mathcal{M}_{\mathrm{RE}}(\varepsilon, \delta; \mathcal{H}, \mathcal{U}) = O\left(\tfrac{\mathrm{vc}(\mathcal{H})\log(k)}{\varepsilon} \log\left(\tfrac{\mathrm{vc}(\mathcal{H})\log(k)}{\varepsilon}\right) + \tfrac{\mathrm{vc}(\mathcal{H})}{\varepsilon} \log^2\left(\tfrac{\mathrm{vc}(\mathcal{H})}{\varepsilon}\right) + \tfrac{1}{\varepsilon}\log\left(\tfrac{1}{\delta}\right)\right),$$

and hence by Theorem 8:

$$\mathcal{M}_{\mathrm{AG}}(\varepsilon, \delta; \mathcal{H}, \mathcal{U}) = O\left(\tfrac{\mathrm{vc}(\mathcal{H})\log(k)}{\varepsilon^2} \mathrm{polylog}\left(\tfrac{\mathrm{vc}(\mathcal{H})\log(k)}{\varepsilon}\right) + \tfrac{1}{\varepsilon^2}\log\left(\tfrac{1}{\delta}\right)\right).$$

In particular, our approach reduces the dependence on k from $k\log(k)$ in (2) as obtained by Attias, Kontorovich, and Mansour (2018), to $\log(k)(\log\log(k))^3$. To do so, our approach *does* rely on improper learning, and our arguments are not valid for $\mathrm{RERM}_{\mathcal{H}}$. We do not know whether improperness is *required* to obtain this improvement, or whether in this case a $\mathrm{polylog}k$ dependence is possible even with RERM or some other proper learning rule. It follows from the construction of our negative result for proper learning in Theorem 1, that at least a $\log(k)$ factor is sometimes necessary for proper learning (regardless of the VC dimension), whereas our Corollary 7 implies that improper learning can achieve a sample complexity that is entirely independent of k (albeit with a worse dependence on the VC dimension).

5. Necessary and Sufficient conditions for Robust Learnability

In the previous section, we saw that having finite VC dimension is *sufficient* for robust learnability. But a simple construction shows that it is not *necessary*: consider an infinite domain $\mathcal{X}$, the hypothesis class of all possible predictors $\mathcal{H} = \{-,+\}^{\mathcal{X}}$, and an all-powerful adversary specified by $\mathcal{U}(x) = \mathcal{X}$. In this case, the hypothesis minimizing the population robust risk $\mathrm{R}_{\mathcal{U}}(h;\mathcal{D})$ would always be the all-positive or the all-negative hypothesis, and so these are the only two hypothesis we should compete with. And so, even though $\mathrm{vc}(\mathcal{H}) = \infty$, a single example suffices to inform the learner of whether to produce the all-positive or all-negative function.

Can we then have a tight characterization of robust learnability? Is there a weaker notion that is both necessary and sufficient for learning? A simple complexity measure one might consider is the maximum number of points $x_1, \ldots, x_m$ such that the entire perturbation sets $\mathcal{U}(x_1), \ldots, \mathcal{U}(x_m)$ are shattered by $\mathcal{H}$. That is, such that $\forall y_1, \ldots, y_m \in \{+1, -1\}, \exists h \in \mathcal{H}, \forall i \forall x' \in \mathcal{U}(x_i),\ h(x') = y_i$. We denote this as $\dim_{\mathcal{U}^\times}(\mathcal{H})$. When $\mathcal{U}(x)$ are balls around x, which is the typical case in metric-based robustness, this can be thought of shattering with a margin in input space. Indeed, for linear predictors and when $\mathcal{U}(x) = \{x' | \|x - x'\|_2 \leq \gamma\}$ is a Euclidean ball around x, $\dim_{\mathcal{U}^\times}(\mathcal{H})$ exactly agrees with the fat shattering dimension at scale γ (or the VC_γ dimension).

While it is fairly obvious that $\dim_{\mathcal{U}^\times}(\mathcal{H})$ provides a lower bound on the sample complexity of robust learning, and thus its finiteness is *necessary* for learning, we construct an example in Appendix D showing that it is not *sufficient*. Specifically, there are classes where *no* points can be shattered in this way, and yet the classes are not robustly learnable. Formally,

Proposition 9 *There exist $\mathcal{X}$, $\mathcal{H}$, $\mathcal{U}$ such that $\dim_{\mathcal{U}^\times}(\mathcal{H}) = 0$ but $\mathcal{M}_{\mathrm{RE}}(\varepsilon, \delta; \mathcal{H}, \mathcal{U}) = \infty$.*

We now attempt to refine the above measure, and introduce a weaker notion of robust shattering that that can still be used to lower bound the sample complexity for robust learnability. Given an adversary $\mathcal{U}$ and a hypothesis class $\mathcal{H}$, consider the following notion of $\mathcal{U}$-robust shattering,

Definition 5 (Robust Shattering Dimension) *A sequence $x_1, \ldots, x_m \in \mathcal{X}$ is said to be $\mathcal{U}$-robustly shattered by $\mathcal{H}$ if $\exists z_1^+, z_1^-, \ldots, z_m^+, z_m^- \in \mathcal{X}$ with $x_i \in \mathcal{U}(z_i^+) \cap \mathcal{U}(z_i^-)\ \forall i \in [m]$, and $\forall y_1, \ldots, y_m \in \{-,+\}$, $\exists h \in \mathcal{H}$ with $h(z') = y_i$, $\forall z' \in \mathcal{U}(z_i^{y_i})$, $\forall i \in [m]$. The $\mathcal{U}$-robust shattering dimension $\dim_{\mathcal{U}}(\mathcal{H})$ is defined as the largest m for which there exist m points $\mathcal{U}$-robustly shattered by $\mathcal{H}$.*

We have that $\dim_{\mathcal{U}^\times}(\mathcal{H}) \leq \dim_{\mathcal{U}}(\mathcal{H}) \leq \mathrm{vc}(\mathcal{H})$, where the first inequality follows since disjoint robust shattering is a special case of robust shattering with $z_i^y = x_i$, and so $\dim_{\mathcal{U}}(\mathcal{H})$ is a plausible candidate for a necessary and sufficient dimension of robust learnability. The following theorem (proof provided in appendix D) establishes that the sample complexity of robust learnability is indeed lower bounded by the *$\mathcal{U}$-robust shattering dimension* $\dim_{\mathcal{U}}(\mathcal{H})$,

Theorem 10 *For any $\mathcal{X}$, $\mathcal{H}$, and $\mathcal{U}$,*
$\mathcal{M}_{\mathrm{RE}}(\varepsilon, \delta; \mathcal{H}, \mathcal{U}) = \Omega\Big(\frac{\dim_{\mathcal{U}}(\mathcal{H})}{\varepsilon} + \frac{1}{\varepsilon}\log\big(\frac{1}{\delta}\big)\Big)$ *and* $\mathcal{M}_{\mathrm{AG}}(\varepsilon, \delta; \mathcal{H}, \mathcal{U}) = \Omega\Big(\frac{\dim_{\mathcal{U}}(\mathcal{H})}{\varepsilon^2} + \frac{1}{\varepsilon^2}\log\big(\frac{1}{\delta}\big)\Big)$.

Based on Corollary 5 and Theorem 10, for any adversary $\mathcal{U}$ and any hypothesis class $\mathcal{H}$, we have

$$\Omega\Big(\frac{\dim_{\mathcal{U}}(\mathcal{H})}{\varepsilon} + \frac{1}{\varepsilon}\log\big(\frac{1}{\delta}\big)\Big) \leq \mathcal{M}_{\mathrm{RE}}(\varepsilon, \delta; \mathcal{H}, \mathcal{U}) \leq 2^{O(\mathrm{vc}(\mathcal{H}))}\frac{1}{\varepsilon}\log\big(\frac{1}{\varepsilon}\big) + O\big(\frac{1}{\varepsilon}\log\big(\frac{1}{\delta}\big)\big)\,. \tag{3}$$

That is, the VC dimension is sufficient, and the robust shattering dimension is necessary for robust learnability. As discussed at the beginning of the Section, we know the VC dimension is *not* necessary and there can be an arbitrary large, even infinite, gap in the second inequality. We do not know whether the robust shattering dimension is also sufficient for learning, or whether there can also be a big gap in the first inequality. Establishing a complexity measure that characterizes robust learnability thus remains an open question.

6. Discussion and Future Directions

Perhaps one of the most interesting takeaways from this work is that we should start considering *improper* learning algorithms for adversarially robust learning. Even though our improper learning rule might not be practical, our results suggest to consider departing from robust empirical risk minimization and M-estimation (as in almost all published work), and considering *improper* learning rules such as bagging or other ensemble methods.

Although we settled the question of robust learnability of VC classes, there remains a large gap in the question of what is the optimal sample complexity for robust learning. Can the exponential dependence on $\mathrm{vc}(\mathcal{H})$ in Corollaries 5 and 7 be improved to a linear dependence? Perhaps this is possible with a new analysis of our learning rule or a different *improper* learning rule. Since our learning rule and analysis stem from recent progress on compression schemes for VC classes (Moran and Yehudayoff, 2016), it is certainly possible that further progress on the celebrated open problem regarding the existence of $\mathrm{vc}(\mathcal{H})$ compression schemes (Floyd and Warmuth, 1995; Warmuth, 2003) could also assist in progress on adversarially robust learning.

Our results demonstrate that there *exist* hypothesis classes with large gaps between what can be done with proper vs. improper robust learning. This means that when studying a particular class, such as classes corresponding to neural networks, one should consider the possibility that there *might* be such a gap and that improper learning *might* be necessary. It remains open to establish whether such gaps actually exist for specific interesting neural net classes (e.g., functions representable by a specific architecture, possibly with a bounded weight norm).

Throughout the paper we ignored computational considerations. Our learning rule can be viewed as an algorithm with black-box access to $\mathrm{RERM}_{\mathcal{H}}$, but making order $m^{\mathrm{vc}(\mathcal{H})}$ such calls, and additionally requiring order $m^{\mathrm{vc}(\mathcal{H})\mathrm{vc}^*(\mathcal{H})}$ time and space to represent and update the distributions used by the boosting algorithm. Without significantly increasing the sample complexity, is it possible to robustly learn with an algorithm making only a polynomial (in $\mathrm{vc}(\mathcal{H}), \mathrm{vc}^*(\mathcal{H}), m$) number of calls to $\mathrm{RERM}_{\mathcal{H}}$ or even $\mathrm{ERM}_{\mathcal{H}}$, plus polynomial additional time and space? What about $\mathrm{poly}(\mathrm{vc}(\mathcal{H}), m)$? This question becomes even more interesting if there is such an algorithm that also only requires sample size $m = \mathrm{poly}(\mathrm{vc}(\mathcal{H}), 1/\varepsilon, \log(1/\delta))$, rather than the $m = \mathrm{poly}(\mathrm{vc}(\mathcal{H}), \mathrm{vc}^*(\mathcal{H}), 1/\varepsilon, \log(1/\delta))$ sufficient for our algorithm. Would another type of oracle be useful? For example, can one devise efficient methods that rely on black-box access to ERM on the dual of the hypothesis class (i.e. finding an example that is correct for the largest number of hypotheses in a given finite set of hypotheses)? More ambitiously, one may ask whether efficient PAC learnability implies efficient robust PAC learnability, roughly translating to asking whether access to *any* (non-robust) learning rule is sufficient for efficient robust learning.

As a final remark, we note that our results easily extend to the multiclass setting ($|\mathcal{Y}| > 2$). In that case, by essentially the same algorithms and proofs, Theorems 4 and 6 (and Corollaries 5 and 7) will hold with $\mathrm{vc}(\mathcal{H})$ replaced by the *graph dimension* (Natarajan, 1989; Ben-David et al., 1995; Daniely et al., 2015). The lower bound in Theorem 10 also holds, by the same arguments, but with $\dim_{\mathcal{U}}(\mathcal{H})$ generalized analogous to the *Natarajan dimension* (Natarajan, 1989): that is, in the definition of robust shattering, after "and", we now require $\forall i \exists y_{i,-}, y_{i,+} \in \mathcal{Y}$ s.t. $\forall b_1, \ldots, b_m \in \{-,+\}$, $\exists h \in \mathcal{H}$ with $h(z') = y_{i,b_i}$, $\forall z' \in \mathcal{U}(z_i^{b_i})$, $\forall i$. We leave as an open question whether one can also express an upper bound controlled by this quantity.

Acknowledgments

This work is partially funded by NSF-BSF award 1718970 and NSF award 1764032.

References

M. Anthony and P. L. Bartlett. *Neural Network Learning: Theoretical Foundations*. Cambridge University Press, 1999.

P. Assouad. Densité et dimension. *Annales de l'Institut Fourier (Grenoble)*, 33(3):233–282, 1983.

Idan Attias, Aryeh Kontorovich, and Yishay Mansour. Improved generalization bounds for robust learning. *arXiv preprint arXiv:1810.02180*, 2018.

S. Ben-David, N. Cesa-Bianchi, D. Haussler, and P. Long. Characterizations of learnability for classes of $\{0, \ldots, n\}$-valued functions. *Journal of Computer and System Sciences*, 50:74–86, 1995.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

A. Blumer, A. Ehrenfeucht, D. Haussler, and M. Warmuth. Learnability and the Vapnik-Chervonenkis dimension. *Journal of the Association for Computing Machinery*, 36(4):929–965, 1989.

Sébastien Bubeck, Eric Price, and Ilya Razenshteyn. Adversarial examples from computational constraints. *arXiv preprint arXiv:1805.10204*, 2018.

Daniel Cullina, Arjun Nitin Bhagoji, and Prateek Mittal. PAC-learning in the presence of evasion adversaries. *arXiv preprint arXiv:1806.01471*, 2018.

A. Daniely, S. Sabato, S. Ben-David, and S. Shalev-Shwartz. Multiclass learnability and the ERM principle. *Journal of Machine Learning Research*, 16:2377–2404, 2015.

O. David, S. Moran, and A. Yehudayoff. Supervised learning through the lens of compression. In *Advances in Neural Information Processing Systems 29*, pages 2784–2792, 2016.

A. Ehrenfeucht, D. Haussler, M. Kearns, and L. Valiant. A general lower bound on the number of examples needed for learning. *Information and Computation*, 82(3):247–261, 1989.

S. Floyd and M. Warmuth. Sample compression, learnability, and the Vapnik-Chervonenkis dimension. *Machine Learning*, 21(3):269–304, 1995.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

T. Graepel, R. Herbrich, and J. Shawe-Taylor. PAC-Bayesian compression bounds on the prediction error of learning algorithms for classification. *Machine Learning*, 59(1-2):55–76, 2005.

S. Hanneke, A. Kontorovich, and M. Sadigurschi. Sample compression for real-valued learners. In *Proceedings of the 30th International Conference on Algorithmic Learning Theory*, 2019.

Justin Khim and Po-Ling Loh. Adversarial risk bounds for binary classification via function transformation. *arXiv preprint arXiv:1810.09519*, 2018.

N. Littlestone and M. Warmuth. Relating data compression and learnability. *Unpublished manuscript*, 1986.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

S. Moran and A. Yehudayoff. Sample compression schemes for VC classes. *Journal of the ACM*, 63(3):21:1–21:10, 2016.

B. K. Natarajan. On learning sets and functions. *Machine Learning*, 4:67–97, 1989.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. *arXiv preprint arXiv:1801.09344*, 2018a.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Semidefinite relaxations for certifying robustness to adversarial examples. *arXiv preprint arXiv:1811.01057*, 2018b.

N. Sauer. On the density of families of sets. *Journal of Combinatorial Theory (A)*, 13(1):145–147, 1972.

R. E. Schapire and Y. Freund. *Boosting*. Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2012.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *arXiv preprint arXiv:1804.11285*, 2018.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge university press, 2014.

Shai Shalev-Shwartz, Ohad Shamir, Nathan Srebro, and Karthik Sridharan. Stochastic convex optimization. In *COLT*, 2009.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

A. W. van der Vaart and J. A. Wellner. *Weak Convergence and Empirical Processes*. Springer, 1996.

V. Vapnik. *Estimation of Dependencies Based on Empirical Data*. Springer-Verlag, New York, 1982.

V. Vapnik and A. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. *Theory of Probability and its Applications*, 16(2):264–280, 1971.

V. Vapnik and A. Chervonenkis. *Theory of Pattern Recognition*. Nauka, Moscow, 1974.

M. Warmuth. Compressing to VC dimension many points. In *Proceedings of the* 16^{th} *Conference on Learning Theory*, 2003.

Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International Conference on Machine Learning*, pages 5283–5292, 2018.

Huan Xu and Shie Mannor. Robustness and generalization. *Machine learning*, 86(3):391–423, 2012.

Dong Yin, Kannan Ramchandran, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. *arXiv preprint arXiv:1810.11914*, 2018.

Appendix A. Auxilliary Proofs Related to Proper Robust Learnability

Proof [of Lemma 3] This proof follows standard lower bound techniques that use the probabilistic method (Shalev-Shwartz and Ben-David, 2014, Chapter 5). Let $m \in \mathbb{N}$. Construct $\mathcal{H}_0$ as before, according to Lemma 2, on $3m$ points $x_1, \ldots, x_{3m}$. By construction, we know that $\mathcal{L}^{\mathcal{U}}_{\mathcal{H}_0}$ shatters the set $C = \{(x_1, +1), \ldots, (x_{3m}, +1)\}$. We will only keep a subset $\mathcal{H}$ of $\mathcal{H}_0$ that includes classifiers that are robustly correct only on subsets of size $2m$, i.e. $\mathcal{H} = \{h_b \in \mathcal{H}_0 : \sum_{i=1}^{3m} b_i = m\}$. Let $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \mapsto \mathcal{H}$ be an arbitrary proper learning rule. The main idea here is to construct a family of distributions that are supported only on $2m$ points of C, which would force rule $\mathcal{A}$ to choose which points it can afford to be not correctly robust on. If rule $\mathcal{A}$ observes only m points, it can't do anything better than guessing which of the remaining $2m$ points of C are actually included in the support of the distribution.

Consider a family of distributions $\mathcal{D}_1, \ldots, \mathcal{D}_T$ where $T = \binom{3m}{2m}$, each distribution $\mathcal{D}_i$ is uniform over only $2m$ points in C. For every distribution $\mathcal{D}_i$, by construction of $\mathcal{H}$, there exists a classifier $h^* \in \mathcal{H}$ such that $\mathrm{R}_{\mathcal{U}}(h^*; \mathcal{D}_i) = 0$. This satisfies the first requirement. For the second requirement, we will use the probabilistic method to show that there exists a distribution $\mathcal{D}_i$ such that $\mathbb{E}_{S \sim \mathcal{D}_i^m}\left[\mathrm{R}_{\mathcal{U}}(A(S); \mathcal{D}_i)\right] \geq 1/4$, and finish the proof using a variant of Markov's inequality.

Pick an arbitrary sequence $S \in C^m$. Consider a uniform weighting over the distributions $\mathcal{D}_1, \ldots, \mathcal{D}_T$. Denote by E_S the event that $S \subset \mathrm{supp}(\mathcal{D}_i)$ for a distribution $\mathcal{D}_i$ that is picked uniformly at random. We will lower bound the expected robust loss of the classifier that rule $\mathcal{A}$ outputs, namely $\mathcal{A}(S) \in \mathcal{H}$, given the event E_S,

$$\mathbb{E}_{\mathcal{D}_i}[\mathrm{R}_{\mathcal{U}}(\mathcal{A}(S); \mathcal{D}_i) | E_S] = \mathbb{E}_{\mathcal{D}_i}\left[\mathbb{E}_{(x,y)\sim\mathcal{D}_i}\left[\sup_{z \in \mathcal{U}(x)} \mathbb{1}[\mathcal{A}(S)(z) \neq y]\right] \middle| E_S\right]. \tag{4}$$

We can lower bound the robust loss of the classifier $\mathcal{A}(S)$ by conditioning on the event that $(x, y) \notin S$ denoted $E_{(x,y)\notin S}$,

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_i}\left[\sup_{z \in \mathcal{U}(x)} \mathbb{1}[\mathcal{A}(S)(z) \neq y]\right] \geq \mathbb{P}_{(x,y)\sim\mathcal{D}_i}[E_{(x,y)\notin S}] \mathbb{E}_{(x,y)\sim\mathcal{D}_i}\left[\sup_{z \in \mathcal{U}(x)} \mathbb{1}[\mathcal{A}(S)(z) \neq y] | E_{(x,y)\notin S}\right].$$

Since $|S| = m$, and $\mathcal{D}_i$ is uniform over its support of size $2m$, we have $\mathbb{P}_{(x,y)\sim\mathcal{D}_i}[E_{(x,y)\notin S}] \geq 1/2$. This allows us to get a lower bound on (4),

$$\mathbb{E}_{\mathcal{D}_i}\left[\mathrm{R}_{\mathcal{U}}(\mathcal{A}(S); \mathcal{D}_i) | E_S\right] \geq \frac{1}{2}\mathbb{E}_{\mathcal{D}_i}\left[\mathbb{E}_{(x,y)\sim\mathcal{D}_i}\left[\sup_{z \in \mathcal{U}(x)} \mathbb{1}[\mathcal{A}(S)(z) \neq y] \middle| E_{(x,y)\notin S}\right] \middle| E_S\right]. \tag{5}$$

Since $\mathcal{A}(S) \in \mathcal{H}$, by construction of $\mathcal{H}$, we know that there are at least m points in C where $\mathcal{A}(S)$ is not robustly correct. We can unroll the expectation over $\mathcal{D}_i$ as follows

$$\mathbb{E}_{\mathcal{D}_i}\left[\mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}_i}\left[\sup_{z\in\mathcal{U}(x)} \mathbb{1}[\mathcal{A}(S)(z)\neq y] | E_{(x,y)\notin S}\right] \middle| E_S\right]$$
$$\geq \frac{1}{m}\sum_{(x,y)\notin S} \mathbb{E}_{\mathcal{D}_i}[\mathbb{1}_{(x,y)\in\mathrm{supp}(\mathcal{D}_i)}|E_S] \sup_{z\in\mathcal{U}(x)} \mathbb{1}[\mathcal{A}(S)(z)\neq y] \geq \frac{1}{m}\sum_{(x,y)\notin S}\frac{1}{2}\sup_{z\in\mathcal{U}(x)} \mathbb{1}[\mathcal{A}(S)(z)\neq y] \geq \frac{1}{2}.$$

Thus, it follows by (5) that $\mathbb{E}_{\mathcal{D}_i}\left[\mathrm{R}_{\mathcal{U}}(\mathcal{A}(S);\mathcal{D}_i)|E_S\right] \geq \frac{1}{4}$. Now, by law of total expectation,

$$\mathbb{E}_{\mathcal{D}_i}\left[\mathbb{E}_{S\sim\mathcal{D}_i^m}[\mathrm{R}_{\mathcal{U}}(\mathcal{A}(S);\mathcal{D}_i)]\right] = \mathbb{E}_{S\sim\mathcal{D}_i^m}[\mathbb{E}_{\mathcal{D}_i}[\mathrm{R}_{\mathcal{U}}(\mathcal{A}(S);\mathcal{D}_i)|E_S]] \geq \tfrac{1}{4}.$$

Since the expectation over $\mathcal{D}_1,\ldots,\mathcal{D}_T$ is at least $1/4$, this implies that there exists a distribution $\mathcal{D}_i$ such that $\mathbb{E}_{S\sim\mathcal{D}_i^m}\left[\mathrm{R}_{\mathcal{U}}(\mathcal{A}(S);\mathcal{D}_i)\right] \geq 1/4$. Using a variant of Markov's inequality, for any random variable Z taking values in $[0,1]$, and any $a\in(0,1)$, we have $\mathbb{P}[Z > 1-a] \geq \frac{\mathbb{E}[Z]-(1-a)}{a}$. For $Z = \mathrm{R}_{\mathcal{U}}(\mathcal{A}(S);\mathcal{D}_i)$ and $a = 7/8$, we get $\mathbb{P}_{S\sim\mathcal{D}_i^m}\left[\mathrm{R}_{\mathcal{U}}(\mathcal{A}(S);\mathcal{D}_i) > \frac{1}{8}\right] \geq \frac{1/4-1/8}{7/8} = \frac{1}{7}$. ■

Appendix B. Auxilliary Proofs Related to Realizable Robust Learnability

The following lemma extends the classic compression-based generalization guarantees from the 0-1 loss to also hold for the robust loss. It is used in the proof of Theorem 4. Generally, it is also possible to extend other generalization guarantees for compression schemes to the robust loss, such as improved bounds for permutation-invariant compression schemes, or convergence guarantees for the agnostic case (as discussed in Section 4.2).

Lemma 11 *For any $k\in\mathbb{N}$ and fixed function $\phi:(\mathcal{X}\times\mathcal{Y})^k\to\mathcal{Y}^{\mathcal{X}}$, for any distribution P over $\mathcal{X}\times\mathcal{Y}$ and any $m\in\mathbb{N}$, for $S=\{(x_1,y_1),\ldots,(x_m,y_m)\}$ iid P-distributed random variables, with probability at least $1-\delta$, if $\exists i_1,\ldots,i_k\in\{1,\ldots,m\}$ s.t. $\hat{R}_{\mathcal{U}}(\phi((x_{i_1},y_{i_1}),\ldots,(x_{i_k},y_{i_k}));S)=0$, then*

$$R_{\mathcal{U}}(\phi((x_{i_1},y_{i_1}),\ldots,(x_{i_k},y_{i_k}));P) \leq \frac{1}{m-k}(k\ln(m)+\ln(1/\delta)).$$

Proof For completeness, we include a brief proof, which merely notes that the classic argument of (Littlestone and Warmuth, 1986; Floyd and Warmuth, 1995) establishing generalization guarantees for sample compression schemes under the 0-1 loss remains valid under the robust loss.

For any indices $i_1,\ldots,i_k\in\{1,\ldots,m\}$,

$$\begin{aligned}
&\mathbb{P}\Big(\hat{\mathrm{R}}_{\mathcal{U}}(\phi(\{(x_{i_j},y_{i_j})\}_{j=1}^k);S)=0 \text{ and } \mathrm{R}_{\mathcal{U}}(\phi(\{(x_{i_j},y_{i_j})\}_{j=1}^k);P)>\varepsilon\Big)\\
&\leq \mathbb{E}\Big[\mathbb{P}\Big(\hat{\mathrm{R}}_{\mathcal{U}}(\phi(\{(x_{i_j},y_{i_j})\}_{j=1}^k);S\setminus\{(x_{i_j},y_{i_j})\}_{j=1}^k)=0\Big|\{(x_{i_j},y_{i_j})\}_{j=1}^k\Big)\times\\
&\qquad\qquad \mathbb{1}\Big[\mathrm{R}_{\mathcal{U}}(\phi(\{(x_{i_j},y_{i_j})\}_{j=1}^k);P)>\varepsilon\Big]\Big]\\
&< (1-\varepsilon)^{m-k},
\end{aligned}$$

and a union bound over all m^k possible choices of $i_1,\ldots,i_k$ implies a probability at most $m^k(1-\varepsilon)^{m-k} \leq m^k e^{-\varepsilon(m-k)}$ that there exist $i_1,\ldots,i_k$ with $\mathrm{R}_{\mathcal{U}}(\phi(\{(x_{i_j},y_{i_j})\}_{j=1}^k);\mathcal{P}_{XY}) > \varepsilon$ and yet $\hat{\mathrm{R}}_{\mathcal{U}}(\phi(\{(x_{i_j},y_{i_j})\}_{j=1}^k);S)=0$. This is at most δ for a choice of $\varepsilon=\frac{1}{m-k}(k\ln(m)+\ln(1/\delta))$. ■

Appendix C. Proof of Agnostic Robust Learnability

Proof [of Theorem 8] The argument follows closely a proof of an analogous result by David, Moran, and Yehudayoff (2016) for non-robust learning. Denote by $\mathbb{A}$ the optimal realizable-case learner achieving sample complexity $\mathcal{M}_{\mathrm{RE}}(1/3, 1/3; \mathcal{H}, \mathcal{U})$, and denote $\mathcal{M}_{\mathrm{RE}} = \mathcal{M}_{\mathrm{RE}}(1/3, 1/3; \mathcal{H}, \mathcal{U})$, as above.

Then, in the agnostic case, given a data set $S \sim \mathcal{D}^m$, we first do robust-ERM to find a maximal-size subsequence S' of the data where the robust loss can be zero: that is, $\inf_{h \in \mathcal{H}} \hat{\mathrm{R}}_{\mathcal{U}}(h; S') = 0$. Then for any distribution D over S', there exists a sequence $S_D \in (S')^{\mathcal{M}_{\mathrm{RE}}}$ such that $h_D := \mathbb{A}(S_D)$ has $\mathrm{R}_{\mathcal{U}}(h_D; D) \leq 1/3$; this follows since, by definition of $\mathcal{M}_{\mathrm{RE}}(1/3, 1/3; \mathcal{H}, \mathcal{U})$, there is a $1/3$ chance that $\hat{S}$ a random draw from $D^{\mathcal{M}_{\mathrm{RE}}}$ yields $\mathrm{R}_{\mathcal{U}}(\mathbb{A}(\hat{S}); D) \leq 1/3$, so at least one such S_D exists. We use this to define a weak robust-learner for distributions D over S': i.e., for any D, the weak learner chooses h_D as its weak hypothesis.

Now we run the α-Boost boosting algorithm (Schapire and Freund, 2012, Section 6.4.2) on data set S', but using the robust loss rather than 0-1 loss. That is, we start with D_1 uniform on S'.[2] Then for each round t, we get h_{D_t} as a weak robust classifier with respect to D_t, and for each $(x, y) \in S'$ we define a distribution $\mathcal{D}_{t+1}$ over S' satisfying

$$D_{t+1}(\{(x,y)\}) \propto D_t(\{(x,y)\}) \exp\left\{-2\alpha \mathbb{1}[\forall x' \in \mathcal{U}(x), h_{D_t}(x') = y]\right\},$$

where α is a parameter we can set. Following the argument from Schapire and Freund (2012, Section 6.4.2), after T rounds we are guaranteed

$$\min_{(x,y) \in S'} \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}[\forall x' \in \mathcal{U}(x), h_{D_t}(x') = y] \geq \frac{2}{3} - \frac{2}{3}\alpha - \frac{\ln(|S'|)}{2\alpha T},$$

so we will plan on running until round $T = 1 + 48 \ln(|S'|)$ with value $\alpha = 1/8$ to guarantee

$$\min_{(x,y) \in S'} \frac{1}{T} \sum_{t=1}^{T} \mathbb{1}[\forall x' \in \mathcal{U}(x), h_{D_t}(x') = y] > \frac{1}{2},$$

so that the classifier $\hat{h}(x) := \mathbb{1}\left[\frac{1}{T}\sum_{t=1}^{T} h_{D_t}(x) \geq \frac{1}{2}\right]$ has $\hat{\mathrm{R}}_{\mathcal{U}}(\hat{h}; S') = 0$.

Furthermore, note that, since each h_{D_t} is given by $\mathbb{A}(S_{D_t})$, where S_{D_t} is an $\mathcal{M}_{\mathrm{RE}}$-tuple of points in S', the classifier $\hat{h}$ is specified by an ordered sequence of $\mathcal{M}_{\mathrm{RE}} T$ points from S. Altogether, $\hat{h}$ is a function specified by an ordered sequence of $\mathcal{M}_{\mathrm{RE}} T$ points from S, and which has

$$\hat{\mathrm{R}}_{\mathcal{U}}(\hat{h}; S) \leq \min_{h \in \mathcal{H}} \hat{\mathrm{R}}_{\mathcal{U}}(h; S).$$

Similarly to the realizable case (see the proof of Lemma 11), uniform convergence guarantees for sample compression schemes (see Graepel, Herbrich, and Shawe-Taylor, 2005) remain valid for the robust loss, by essentially the same argument; the essential argument is the same as in the proof of Lemma 11 except using Hoeffding's inequality to get concentration of the empirical robust risks for each fixed index sequence, and then a union bound over the possible index sequnces as before. We

2. We ignore the possibility of repeats; for our purposes we can just remove any repeats from S' before this boosting step.

omit the details for brevity. In particular, denoting $T_m = 1 + 48\ln(m)$, for $m > \mathcal{M}_{\mathrm{RE}} T_m$, with probability at least $1 - \delta/2$,

$$\mathrm{R}_{\mathcal{U}}(\hat{h}; \mathcal{D}) \leq \hat{\mathrm{R}}_{\mathcal{U}}(\hat{h}; S) + \sqrt{\frac{\mathcal{M}_{\mathrm{RE}} T_m \ln(m) + \ln(2/\delta)}{2m - 2\mathcal{M}_{\mathrm{RE}} T_m}}.$$

Let $h^* = \operatorname{argmin}_{h \in \mathcal{H}} \mathrm{R}_{\mathcal{U}}(h; \mathcal{D})$ (supposing the min is realized, for simplicity; else we could take an h^* with very-nearly minimal risk). By Hoeffding's inequality, with probability at least $1 - \delta/2$,

$$\hat{\mathrm{R}}_{\mathcal{U}}(h^*; S) \leq \mathrm{R}_{\mathcal{U}}(h^*; \mathcal{D}) + \sqrt{\frac{\ln(2/\delta)}{2m}}.$$

By the union bound, if $m \geq 2\mathcal{M}_{\mathrm{RE}} T_m$, with probability at least $1 - \delta$,

$$\begin{aligned}
\mathrm{R}_{\mathcal{U}}(\hat{h}; \mathcal{D}) &\leq \min_{h \in \mathcal{H}} \hat{\mathrm{R}}_{\mathcal{U}}(h; S) + \sqrt{\frac{\mathcal{M}_{\mathrm{RE}} T_m \ln(m) + \ln(2/\delta)}{m}} \\
&\leq \hat{\mathrm{R}}_{\mathcal{U}}(h^*; S) + \sqrt{\frac{\mathcal{M}_{\mathrm{RE}} T_m \ln(m) + \ln(2/\delta)}{m}} \\
&\leq \mathrm{R}_{\mathcal{U}}(h^*; \mathcal{D}) + 2\sqrt{\frac{\mathcal{M}_{\mathrm{RE}} T_m \ln(m) + \ln(2/\delta)}{m}}.
\end{aligned}$$

Since $T_m = O(\log(m))$, the above is at most ε for an appropriate choice of sample size $m = O\Big(\frac{\mathcal{M}_{\mathrm{RE}}}{\varepsilon^2} \log^2\Big(\frac{\mathcal{M}_{\mathrm{RE}}}{\varepsilon}\Big) + \frac{1}{\varepsilon^2}\log\big(\frac{1}{\delta}\big)\Big)$. ■

Appendix D. Auxilliary Proofs Related to Necessary Conditions for Robust Learnability

Proof [of Proposition 9] Let $\mathcal{X} = \mathbb{R}^d$ equipped with a metric ρ, and $\mathcal{U} : \mathcal{X} \mapsto 2^{\mathcal{X}}$ such that $\mathcal{U}(x) = \{z \in \mathcal{X} : \rho(x, z) \leq \gamma\}$ for all $x \in \mathcal{X}$ for some $\gamma > 0$. Consider two infinite sequences of points $(x_m)_m \in \mathbb{N}$ and $(z_m)_m \in \mathbb{N}$ such that for any $i \neq j, \mathcal{U}(x_i) \cap \mathcal{U}(x_j) = \emptyset, \mathcal{U}(x_i) \cap \mathcal{U}(z_j) = \emptyset$, $\mathcal{U}(x_j) \cap \mathcal{U}(z_i) = \emptyset$, but $\mathcal{U}(x_i) \cap \mathcal{U}(z_i) = u_i$. In other words, we want the γ-balls of pairs with different indices to be mutually disjoint, and the γ-balls for a pair with the same index to intersect at a single point (this is possible because we are considering closed balls).

Next, we proceed with the construction of $\mathcal{H}$. For each bit string $b \in \{0,1\}^{\mathbb{N}}$, we will define a predictor $h_b : \mathcal{X} \mapsto \mathcal{Y}$ just on the γ-balls of the points $x_1, z_1, x_2, z_2, \ldots$ (it labels the rest of the $\mathcal{X}$ space with $+1$). Foreach $i \in \mathbb{N}$, if $b_i = 0$, set

$$h_b\Big(\mathcal{U}(x_i)\Big) = +1 \quad \text{and} \quad h_b\Big(\mathcal{U}(z_i) \setminus \mathcal{U}(x_i)\Big) = -1$$

and if $b_i = 1$, set

$$h_b\Big(\mathcal{U}(x_i) \setminus \mathcal{U}(z_i)\Big) = -1 \quad \text{and} \quad h_b\Big(\mathcal{U}(z_i)\Big) = +1$$

Let $\mathcal{H} = \{h_b : b \in \{0,1\}^{\mathbb{N}}\}$. Notice that $\dim_{\mathcal{U}\times}(\mathcal{H}) = 0$, because there is no single γ-ball that is labeled in both ways ($+1$ and -1). By construction of $\mathcal{H}$, all classifiers $h_b \in \mathcal{H}$ behave the same way on all points in $\mathcal{X}$, except at points in the intersections $u_1, u_2, \ldots$ which get shattered. However,

the $\mathcal{U}$-robust shattering dimension (see definition 5) is infinite in this construction ($\dim_{\mathcal{U}}(\mathcal{H}) = \infty$), which by Theorem 10 (see below) implies that $\mathcal{M}_{\mathrm{RE}}(\varepsilon, \delta; \mathcal{H}, \mathcal{U}) = \infty$. ■

Proof [Sketch of Theorem 10] We first start with the realizable case. The proof follows a standard argument from (Mohri et al., 2018, Chapter 3). Let $\mathrm{d} = \dim_{\mathcal{U}}(\mathcal{H})$, and fix $x_1, \ldots, x_{\mathrm{d}}$ a sequence $\mathcal{U}$-robustly shattered by $\mathcal{H}$, and let $z_1^+, z_1^-, \ldots, z_{\mathrm{d}}^+, z_{\mathrm{d}}^- \in \mathcal{X}$ be as in definition 5; in particular, note that any y, y' and any $i \neq j$ necessarily have $z_i^y \neq z_j^{y'}$. For each $\mathbf{y} = (y_1, \ldots, y_{\mathrm{d}}) \in \{+1, -1\}^{\mathrm{d}}$, let $h^{\mathbf{y}} \in \mathcal{H}$ be such that $\forall i \in [m]$, $\forall z' \in \mathcal{U}(z_i^{y_i})$, $h^{\mathbf{y}}(z') = y_i$. Let $\mathcal{D}$ be a distribution over $\{1, 2, \ldots, \mathrm{d}\}$ such that $\mathbb{P}_{i\sim\mathcal{D}}[i = 1] = 1 - 8\varepsilon$ and $\mathbb{P}_{i\sim\mathcal{D}}[i = 1] = 8\varepsilon/(\mathrm{d} - 1)$ for $2 \leq i \leq d$. Now choose $\mathbf{y} \sim \mathrm{Uniform}(\{+1, -1\}^{\mathrm{d}})$, and let $\mathcal{D}_{\mathbf{y}}$ be the induced distribution over $\mathcal{X} \times \mathcal{Y}$ such that

$$\mathbb{P}_{(x,y)\sim\mathcal{D}_{\mathbf{y}}}\left[(x, y) = (z_1^{y_1}, y_1)\right] = 1 - 8\varepsilon \text{ and } \mathbb{P}_{(x,y)\sim\mathcal{D}_{\mathbf{y}}}\left[(x, y) = (z_i^{y_i}, y_i)\right] = 8\varepsilon/(\mathrm{d} - 1)$$

for $2 \leq i \leq \mathrm{d}$.

Note that by construction we have $\mathrm{R}_{\mathcal{U}}(h^{\mathbf{y}}; \mathcal{D}) = 0$. Now, consider an arbitrary learning rule $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \mapsto \mathcal{Y}^{\mathcal{X}}$. We will assume that $\mathcal{A}$ always gets the prediction of $z_1^{y_1}$ correct. Let $I = \{2, \ldots, d\}$ and let $\mathcal{S}$ be the set of all sequences of size m containing at most $(\mathrm{d} - 1)/2$ elements from I. Fix an arbitrary sequence $S \in \mathcal{S}$. Denote by $S_{\mathbf{y}} = ((z_i^{y_i}, y_i) : i \in \mathcal{S})$ the sequence of examples induced by the indices sequence S. Then,

$$\begin{aligned}
\mathbb{E}_{\mathbf{y}}\left[\mathrm{R}_{\mathcal{U}}(\mathcal{A}(S_{\mathbf{y}}); \mathcal{D}_{\mathbf{y}})\right] &\geq \mathbb{E}_{\mathbf{y}}\left[\sum_{i \notin S} \mathbb{P}_{\mathcal{D}_{\mathbf{y}}}(z_i^{y_i}) \sup_{z' \in \mathcal{U}(z_i^{y_i})} \mathbb{1}[\mathcal{A}(S_{\mathbf{y}})(z') \neq y_i]\right] \\
&\geq \frac{\mathrm{d} - 1}{2} \times \frac{8\varepsilon}{\mathrm{d} - 1} \times \mathbb{E}_{\mathbf{y}}\left[\sup_{z' \in \mathcal{U}(z)} \mathbb{1}[\mathcal{A}(S_{\mathbf{y}})(z') \neq y]\right] \\
&= \frac{\mathrm{d} - 1}{2} \times \frac{8\varepsilon}{\mathrm{d} - 1} \times \frac{1}{2} \\
&= 2\varepsilon
\end{aligned}$$

Since the inequality above holds for any sequence $S \in \mathcal{S}$, it follows that

$$\mathbb{E}_{S\sim\mathcal{D}^m}[\mathbb{E}_{\mathbf{y}}\left[\mathrm{R}_{\mathcal{U}}(\mathcal{A}(S_{\mathbf{y}}); \mathcal{D}_{\mathbf{y}})\mathbb{1}_{S\in\mathcal{S}}\right]] = \mathbb{E}_{\mathbf{y}}[\mathbb{E}_{S\sim\mathcal{D}^m}\left[\mathrm{R}_{\mathcal{U}}(\mathcal{A}(S_{\mathbf{y}}); \mathcal{D}_{\mathbf{y}})|E_{S\in\mathcal{S}}\right]] \geq 2\varepsilon$$

Which implies that there exists $\mathbf{y}_0$ such that $\mathbb{E}_{S\sim\mathcal{D}^m}\left[\mathrm{R}_{\mathcal{U}}(\mathcal{A}(S_{\mathbf{y}_0}); \mathcal{D}_{\mathbf{y}_0})|E_{S\in\mathcal{S}}\right] \geq 2\varepsilon$. Since $\mathbb{P}_{\mathcal{D}}[i \in I] \leq 8\varepsilon$, the robust risk $\mathrm{R}_{\mathcal{U}}(\mathcal{A}(S_{\mathbf{y}_0})); \mathcal{D}_{\mathbf{y}}) \leq 8\varepsilon$. Then, by law of total expectation, we have

$$\begin{aligned}
2\varepsilon &\leq \mathbb{E}_{S\sim\mathcal{D}^m}\left[\mathrm{R}_{\mathcal{U}}(\mathcal{A}(S_{\mathbf{y}_0}); \mathcal{D}_{\mathbf{y}_0})|E_{S\in\mathcal{S}}\right] \\
&\leq 8\varepsilon\mathbb{P}_{S\sim\mathcal{D}^m}\left[\mathrm{R}_{\mathcal{U}}(\mathcal{A}(S_{\mathbf{y}_0}); \mathcal{D}_{\mathbf{y}}) \geq \varepsilon|E_{S\in\mathcal{S}}\right] + \varepsilon(1 - \mathbb{P}_{S\sim\mathcal{D}^m}\left[\mathrm{R}_{\mathcal{U}}(\mathcal{A}(S_{\mathbf{y}_0}); \mathcal{D}_{\mathbf{y}_0}) \geq \varepsilon|\, E_{S\in\mathcal{S}}\right])
\end{aligned}$$

By collecting terms, we obtain that $\mathbb{P}_{S\sim\mathcal{D}^m}\left[\mathrm{R}_{\mathcal{U}}(\mathcal{A}(S_{\mathbf{y}_0}); \mathcal{P}_{XY}) \geq \varepsilon|\, E_{S\in\mathcal{S}}\right] \geq 1/7$. Then, by law of total probability, the probability over all sequences (not necessarily in $\mathcal{S}$) can be lower bounded,

$$\mathbb{P}_{S\sim\mathcal{D}^m}\left[\mathrm{R}_{\mathcal{U}}(\mathcal{A}(S_{\mathbf{y}_0}); \mathcal{D}_{\mathbf{y}_0}) \geq \varepsilon\right] \geq \mathbb{P}[E_{S\in\mathcal{S}}]\mathbb{P}_{S\sim\mathcal{D}^m}\left[\mathrm{R}_{\mathcal{U}}(\mathcal{A}(S_{\mathbf{y}_0}); \mathcal{D}_{\mathbf{y}_0}) \geq \varepsilon|E_{S\in\mathcal{S}}\right] \geq \frac{1}{7}\mathbb{P}[E_{S\in\mathcal{S}}]$$

By a standard application of Chernoff bounds, for $\varepsilon = \frac{\mathrm{d}-1}{32m}$ and $\delta \leq 1/100$, we get that $\mathbb{P}[E_{S\in\mathcal{S}}] \geq 7\delta$ and by the above this concludes that $\mathbb{P}_{S\sim\mathcal{D}^m}\left[\mathrm{R}_{\mathcal{U}}(\mathcal{A}(S_{\mathbf{y}_0});\mathcal{D}_{\mathbf{y}_0}) \geq \varepsilon\right] \geq \delta$. This establishes that

$$\mathcal{M}_{\mathrm{RE}}(\varepsilon,\delta;\mathcal{H},\mathcal{U}) \geq \Omega\left(\frac{\mathrm{d}}{\varepsilon}\right)$$

To finish the proof, we need to show that

$$\mathcal{M}_{\mathrm{RE}}(\varepsilon,\delta;\mathcal{H},\mathcal{U}) \geq \Omega\left(\frac{1}{\varepsilon}\log\left(\frac{1}{\delta}\right)\right)$$

For this just consider a distribution P_1 with mass $1-\varepsilon$ on $(z_1^+,+1)$ and mass ε on $(z_2^+,+1)$, and another distribution P_2 with mass $1-\varepsilon$ on $(z_1^+,+1)$ and mass ε on $(z_2^-,-1)$. If $m \leq (1/2\varepsilon)\ln(1/\delta)$, with probability at least δ, we will only observe m samples of $(z_1^+,+1)$, and thus learning rule $\mathcal{A}$ will make a mistake on x_2 (which is in $\mathcal{U}(z_2^+) \cap \mathcal{U}(z_2^-)$) with probability at least $1/2$, therefore having error at least $\varepsilon/2$. By combining both parts, we arrive at the theorem statement.

For the agnostic case, we briefly describe the construction. The remainder of the proof more or less follows a standard argument, for instance see Anthony and Bartlett (1999, Chapter 5). Let $\mathrm{d} = \dim_{\mathcal{U}}(\mathcal{H})$, and fix $x_1,\ldots,x_{\mathrm{d}}$ a sequence $\mathcal{U}$-robustly shattered by $\mathcal{H}$, and let $z_1^+, z_1^-, \ldots, z_{\mathrm{d}}^+, z_{\mathrm{d}}^- \in \mathcal{X}$ be as in definition 5; in particular, note that any y, y' and any $i \neq j$ necessarily have $z_i^y \neq z_j^{y'}$. For $b \in \{0,1\}^{\mathrm{d}}$, define distribution $\mathcal{D}_b$ as follows, for $i \in [\mathrm{d}]$:

- If $b_i = 0$, then set $\mathbb{P}_{\mathcal{D}_b}((z_i^+,+1)) = (1-\alpha)/(2\mathrm{d})$ and $\mathbb{P}_{\mathcal{D}_b}((z_i^-,-1)) = (1+\alpha)/(2\mathrm{d})$.
- If $b_i = 1$, then set $\mathbb{P}_{\mathcal{D}_b}((z_i^+,+1)) = (1+\alpha)/(2\mathrm{d})$ and $\mathbb{P}_{\mathcal{D}_b}((z_i^-,-1)) = (1-\alpha)/(2\mathrm{d})$.

where $) < \alpha < 1$ is appropriately chosen based on ε and δ. ■