Probabilistically Robust PAC Learning

Vinod Raman
Department of Statistics
University of Michigan
Ann Arbor, MI 48104
vkraman@umich.edu

Unique Subedi
Department of Statistics
University of Michigan
Ann Arbor, MI 48104
subedi@umich.edu

Ambuj Tewari
Department of Statistics
University of Michigan
Ann Arbor, MI 48018
tewaria@umich.edu

Abstract

Recently, Robey et al. propose a notion of probabilistic robustness, which, at a high-level, requires a classifier to be robust to most but not all perturbations. They show that for certain hypothesis classes where proper learning under worst-case robustness is *not* possible, proper learning under probabilistic robustness *is* possible with sample complexity exponentially smaller than in the worst-case robustness setting. This motivates the question of whether proper learning under probabilistic robustness is always possible. In this paper, we show that this is *not* the case. We exhibit examples of hypothesis classes $\mathcal{H}$ with finite VC dimension that are *not* probabilistically robustly PAC learnable with *any* proper learning rule.

1 Introduction

As deep neural networks become increasingly ubiquitous, their susceptibility to test-time adversarial attacks has become more and more apparent. Designing learning algorithms that are robust to these test-time adversarial perturbations has garnered increasing attention by machine learning researchers and practitioners alike. Prior work on adversarially robust learning has mainly focused on learnability under the *worst-case* robust risk, defined as

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

where $\mathcal{U}(x)\subset\mathcal{X}$ is an adversarially chosen perturbation set (for example L_p balls). In practice, however, classifiers trained to achieve worst-case adversarial robustness often exhibit degraded nominal performance [DHHR20, RXY+19, SZC+18, TSE+18, YRZ+20, ZYJ+19, RCPH22]. As a result, several works have considered relaxing the worst-case nature of $R_{\mathcal{U}}(h;\mathcal{D})$ [RCPH22, LBSS20, LBSS21, LF19, RBZK21]. In this work, we consider the *probabilistic* relaxation of worst-case adversarial robustness, as introduced by [RCPH22]. In particular, we are interested in understanding the PAC learnability (i.e. sample complexity) of general hypothesis classes under the *probabilistic* robust risk, defined as,

$$R_{\mathcal{U}}(h;\mathcal{D},\rho) := \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\mathbb{1}\{\mathbb{P}_{z\sim\mu_{\mathcal{U}(x)}}(h(z)\neq y)>\rho\}\right],$$

where again $\mathcal{U}(x)\subset\mathcal{X}$ is an adversarially chosen perturbation set, and $\mu_{\mathcal{U}(x)}$ is an adversarially chosen probability measure over $\mathcal{U}(x)$. Roughly speaking, learning under probabilistic robustness asks to find a hypothesis $h\in\mathcal{H}$ that is robust to most, but not all, perturbations for each example in the support of the data distribution $\mathcal{D}$. As highlighted in [RCPH22], this notion of robustness is desirable as it nicely interpolates between worst and average case robustness via an interpretable parameter ρ, while being more computationally tractable compared to existing relaxations. We note that probabilistic robustness is a *strict* relaxation of worst-case robustness. While $R_{\mathcal{U}}(h;\mathcal{D})\leq\epsilon$ implies $R_{\mathcal{U}}(h;\mathcal{D},\rho)\leq\epsilon$ for every $\rho\in[0,1]$ and every hypothesis h, the converse is not true *even for* $\rho=0$. Indeed, when $\rho=0$, there exists problems were a classifier that is non-robust to a countably infinite number of perturbations for every $x\in\mathcal{X}$ still achieves $R_{\mathcal{U}}(h;\mathcal{D},\rho)=0$.

ML Safety Workshop, 36th Conference on Neural Information Processing Systems (NeurIPS 2022).

2 Notation, Preliminaries, and Problem Setup

Throughout the paper we will let $[k]$ denote the set of integers $\{1, ..., k\}$. Let $\mathcal{X}$ denote an instance space, $\mathcal{Y} = \{0, 1\}$ denote our label space, and $\mathcal{D}$ be any distribution over $\mathcal{X} \times \mathcal{Y}$. Let $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ denote a hypothesis class mapping examples in $\mathcal{X}$ to labels in $\mathcal{Y}$. In probabilistic robust learning, an adversary picks for each point $x \in \mathcal{X}$, a perturbation set $\mathcal{U}(x)$ and a perturbation measure $\mu_{\mathcal{U}(x)}$ fully supported on $\mathcal{U}(x)$. At test time, the adversary receives a test example $(x, y) \sim D$, samples a perturbation $\tilde{x} \sim \mu_{\mathcal{U}(x)}$, and passes $(\tilde{x}, y)$ to the learner. We make no assumptions on $\mathcal{U}(x)$ or $\mu_{\mathcal{U}(x)}$ other than the fact that $\mathcal{U}(x)$ must be non-empty for each $x \in \mathcal{X}$. From a learning perspective, given a hypothesis class $\mathcal{H}$, our goal is to design a learning rule $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{Y}^{\mathcal{X}}$ such that for any distribution D over $\mathcal{X} \times \mathcal{Y}$, the learning rule $\mathcal{A}$ finds a predictor that competes with the best predictor $h^* \in \mathcal{H}$ in terms of the probabilistic robust risk using a number of samples that is independent from $\mathcal{D}$. If $\mathcal{A}$ always outputs a hypothesis in $\mathcal{H}$, then we call $\mathcal{A}$ a *proper* learner. The main result in this paper is showing that sometimes proper learning is not possible, even under the probabilistic robustness.

We now recall the Vapnik-Chervonekis dimension (VC dimension) which plays an important role in characterizing PAC learnability under the standard 0-1 risk.

Definition 1. *A set $\{x_1, ..., x_n\} \in \mathcal{X}$ is shattered by $\mathcal{H}$, if $\forall y_1, ..., y_n \in \mathcal{Y}$, $\exists h \in \mathcal{H}$, s.t. $\forall i \in [n]$, $h(x_i) = y_i$. The VC dimension of $\mathcal{H}$, is defined as the largest natural number $n \in \mathbb{N}$ such that there exists a set $\{x_1, ..., x_n\} \in \mathcal{X}$ that is shattered by $\mathcal{H}$.*

In traditional PAC learning framework, a hypothesis class $\mathcal{H}$ is PAC learnable if and only if its VC dimension is finite [VC71]. In fact, when the VC dimension is finite, $\mathcal{H}$ is *properly* learnable via an Empirical Risk Minimization (ERM) oracle. As in the worst-case robust setting, we are interested in understanding what are necessary and sufficient condition on $\mathcal{H}$ that enable probabilistic robust PAC learnability against an arbitrary adversary. A sufficient condition, based on Vapnik's "General Learning" [Vap06], is the finiteness of VC dimension of the probabilistic robust loss class $\mathcal{L}_{\mathcal{H}}^{\mathcal{U},\rho}$:

$$\mathcal{L}_{\mathcal{H}}^{\mathcal{U},\rho} = \{(x, y) \mapsto \mathbb{1}\{\mathbb{P}_{z \sim \mu_{\mathcal{U}(x)}}(h(z) \neq y) > \rho\} : h \in \mathcal{H}\}.$$

In particular, if the VC dimension of the probabilistic robust loss class $\mathcal{L}_{\mathcal{H}}^{\mathcal{U},\rho}$ is finite, then $\mathcal{H}$ is probabilistically robustly PAC learnable via oracle access to a Probabilistic Robust Empirical Risk Minimizer (PRERM) with sample complexity that scales linearly with VC($\mathcal{L}_{\mathcal{H}}^{\mathcal{U},\rho}$). In this sense, if one can upper bound VC($\mathcal{L}_{\mathcal{H}}^{\mathcal{U},\rho}$) in terms of VC($\mathcal{H}$), then finite VC dimension is sufficient for proper learnability. However, as we will show in the next section, there can be an arbitrarily large gap between these two quantities, and proper learning overall might not be possible.

3 Proper Learning Is Not Always Possible

In this section, we show that even for hypothesis classes with finite VC dimension, probabilistic robust PAC learning might not be possible using *any* proper learning rule. In particular, this implies that even if there is a probabilistic robust hypothesis in $\mathcal{H}$, and even with arbitrarily large number of samples, the PRERM may not guarantee low risk. For the proofs in this section we fix $\mathcal{X} = \mathbb{R}^d$ equipped with some metric τ, an adversary $\mathcal{U} : \mathcal{X} \to 2^{\mathcal{X}}$ such that $\mathcal{U}(x) = \{z \in \mathcal{X} : \tau(x, z) \leq \gamma\}$ for all $x \in \mathcal{X}$ for some $\gamma > 0$, and uniform perturbation measures $\mu_{\mathcal{U}(x)}$ for all $x \in \mathcal{X}$.

We start by showing that for every $\rho \in [0, 1)$, there can be an arbitrary gap between the VC dimension of the loss class and the VC dimension of the hypothesis class.

Lemma 1. *For every $\rho \in [0, 1)$ and $m \in \mathbb{N}$, there exists a hypothesis class $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ s.t. VC($\mathcal{H}$) ≤ 1 but VC($\mathcal{L}_{\mathcal{H}}^{\mathcal{U},\rho}$) $\geq m$.*

Proof. Fix $\rho \in [0, 1)$. Let $m \in \mathbb{N}$. Pick m center points $c_1, ..., c_m$ in $\mathcal{X}$ such that for all $i, j \in [m]$, $\mathcal{U}(c_i) \cap \mathcal{U}(c_j) = \emptyset$. For each $c \in \mathcal{X}$, let $\mu_{\mathcal{U}(c)}$ denote a probability measure fully supported on the perturbation set $\mathcal{U}(c)$. In particular, let $\mu_{\mathcal{U}(c)}$ be uniform over $\mathcal{U}(c)$. For each center c_i, consider $2^{m-1} + 1$ disjoint subsets of its perturbation set $\mathcal{U}(c_i)$ which **do not** contain c_i. Label 2^{m-1} of these subsets with a unique bitstring $b \in \{0, 1\}^m$ fixing $b_i = 1$. Let $\mathcal{B}_i^b$ denote the subset labelled by bitstring b and let $\mathcal{B}_i$ denote the single remaining subset that was not labelled. Let $\mu_{\mathcal{U}(c_i)}(\mathcal{B}_i) = \rho$

and $0 < \mu_{\mathcal{U}(c_i)}(\mathcal{B}_i^b) < \frac{1-\rho}{2^{m-1}}$ for every $b \in \{\{0,1\}^m | b_i = 1\}$. If $\rho = 0$, let $\mathcal{B}_i = \emptyset$ for all $i \in [m]$. Observe that indeed $\mu_{\mathcal{U}(c_i)}\left(\mathcal{B}_i \cup \left(\bigcup_b \mathcal{B}_i^b\right)\right) < 1$. For bitstring $b \in \{0,1\}^m$, define the hypothesis h_b as

$$h_b(z) = \begin{cases} 0 & \text{if } z \in \bigcup_{i=1}^m \mathcal{B}_i^b \cup \mathcal{B}_i \\ 1 & \text{otherwise} \end{cases}$$

and consider the hypothesis class $\mathcal{H} = \{h_b | b \in \{0,1\}^m\}$ which consists of all 2^m hypothesis, one for each bitstring. Finally, define $\mathcal{B} = \bigcup_{i=1}^m \bigcup_{b \in \{\{0,1\}^m | b_i = 1\}} \mathcal{B}_i^b \cup \mathcal{B}_i$ as the union of all the subsets. We first show that $\mathcal{H}$ has VC dimension at most 1. Consider two points $x_1, x_2 \in \mathcal{X}$. We will show case by case that every possible pair of points cannot be shattered by $\mathcal{H}$. First, consider the case where, wlog, $x_1 \notin \mathcal{B}$. Then, $\forall h \in \mathcal{H}$, $h(x_1) = 1$, and thus shattering is not possible. Now, consider the case where both $x_1 \in \mathcal{B}$ and $x_2 \in \mathcal{B}$. If either x_1 or x_2 is in $\bigcup_{i=1}^m \mathcal{B}_i$, then every hypothesis $h \in \mathcal{H}$ will label it as 0, and thus these two points cannot be shattered. If $x_1 \in \mathcal{B}_i^b$ and $x_2 \in \mathcal{B}_j^b$ for $i \neq j$, then $h_b(x_1) = h_b(x_2) = 0$, but $\forall h \in \mathcal{H}$ s.t. $h \neq h_b, h(x_1) = h(x_2) = 1$. If $x_1 \in \mathcal{B}_i^{b_1}$ and $x_2 \in \mathcal{B}_j^{b_2}$ for $b_1 \neq b_2$, then there exists no hypothesis in $\mathcal{H}$ that can label (x_1, x_2) as $(0,0)$. Thus, overall, no two points $x_1, x_2 \in \mathcal{X}$ can be shattered by $\mathcal{H}$ implying that $\text{VC}(\mathcal{H}) \leq 1$.

Now we are ready to show that the VC dimension of the loss class is at least m. Specifically, given the sample of labelled points $S = \{(c_1, 1), ..., (c_m, 1)\}$, we will show that the loss behavior corresponding to hypothesis h_b on the sample S is exactly b. Since $\mathcal{H}$ contains all the hypotheses corresponding to every single bitstring $b \in \{0,1\}^m$, the loss class of $\mathcal{H}$ will shatter S. In order to prove that the loss behavior of h_b on the sample S is exactly b, it suffices to show that the probabilistic loss of h_b on example $(c_i, 1)$ is b_i, where b_i denotes the ith bit of b. By definition,

$$\begin{aligned} \ell(h_b; \rho) &= \mathbb{1}\{\mathbb{P}_{z \sim \mu_{\mathcal{U}(c_i)}}(h_b(z) \neq 1) > \rho\} \\ &= \mathbb{1}\{\mathbb{P}_{z \sim \mu_{\mathcal{U}(c_i)}}(h_b(z) = 0) > \rho\} \\ &= \mathbb{1}\{\mathbb{P}_{z \sim \mu_{\mathcal{U}(c_i)}}\left(z \in \mathcal{B}_i^b \cup \mathcal{B}_i\right) > \rho\} \\ &= \mathbb{1}\{\mu_{\mathcal{U}(c_i)}(\mathcal{B}_i^b \cup \mathcal{B}_i) > \rho\} \\ &= b_i. \end{aligned}$$

Thus, the loss behavior of h_b on S is b, and the total number of distinct loss behaviors over each hypothesis in $\mathcal{H}$ on S is 2^m, implying that the VC dimension of the loss class is at least m. This completes the construction and proof of the claim. □

Similar to [MHS19], the hypothesis class construction in Lemma 1 can be used to show the existence of a hypothesis class that cannot be learned properly. Specifically, the lemma below follows exactly from Lemma 3 in [MHS19]. We include the full proof of Lemma 2 in the Appendix.

Lemma 2. *Let $m \in \mathbb{N}$. For every $\rho \in [0,1)$ there exists $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ with $\text{VC}(\mathcal{H}) \leq 1$ such that for any proper learner $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{H}$: (1) there is a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ and a hypothesis $h^* \in \mathcal{H}$ where $R_{\mathcal{U}}(h^*; \mathcal{D}, \rho) = 0$ and (2) with probability at least $1/7$ over $S \sim \mathcal{D}^m$, $R_{\mathcal{U}}(\mathcal{A}(S); \mathcal{D}, \rho) > 1/8$.*

We now state our main theorem indicating that proper learning is **not** possible for any $\rho \in [0,1)$, even for ρ arbitrarily close to 1. Again, the proof of Theorem 3 closely follows that in [MHS19], however, since our hypothesis class construction in Lemma 1 is different, we include a complete proof below.

Theorem 3. *Fix $\rho \in [0,1)$. There exists a hypothesis class $\mathcal{H} \subset \mathcal{Y}^{\mathcal{X}}$ with $\text{VC}(\mathcal{H}) \leq 1$ and an adversary $(\mathcal{U}, \mu)$, such that $\mathcal{H}$ is not properly probabilistically robustly PAC learnable with respect to $(\mathcal{U}, \mu)$.*

Proof. Fix $\rho \in [0,1)$. Let $(C_m)_{m \in \mathbb{N}}$ be an infinite sequence of disjoint sets such that each set C_m contains $3m$ distinct center points from $\mathcal{X}$, where for any $c_i, c_j \in \bigcup_{m=1}^{\infty} C_m$ such that $c_i \neq c_j$, we have $\mathcal{U}(c_i) \cap \mathcal{U}(c_j) = \emptyset$. For every $m \in \mathbb{N}$, construct $\mathcal{H}_m$ on C_m as in Lemma 1. In addition, a key part of this proof is to ensure that the hypothesis in $\mathcal{H}_m$ are non-robust to points in $C_{m'}$ for all $m' \neq m$. To do so, we will need to adjust each hypothesis $h_b \in \mathcal{H}_m$ carefully. By definition, for every $m \in \mathbb{N}$, $\mathcal{H}_m$ consists of 2^{3m} hypothesis of the form

$$h_b(z) = \begin{cases} 0 & \text{if } z \in \bigcup_{i=1}^{3m} \mathcal{B}_i^b \cup \mathcal{B}_i \\ 1 & \text{otherwise} \end{cases}$$

for each bitstring $b \in \{0,1\}^{3m}$. Note that the same set $\bigcup_{i=1}^{3m} \mathcal{B}_i$ is shared across every hypothesis $h_b \in \mathcal{H}_m$. For each $m \in \mathbb{N}$, let $\mathcal{B}^m = \bigcup_{i=1}^{3m} \mathcal{B}_i$ be exactly the union of these $3m$ sets. Next, from the construction in Lemma 1, for every center $c_i \in C_m$, $\mu_{\mathcal{U}(c_i)}\left(\mathcal{B}_i \cup \left(\bigcup_b \mathcal{B}_i^b\right)\right) < 1$. Thus, there exists a set $\tilde{\mathcal{B}}_i \subset \mathcal{U}(c_i)$ s.t. $\mu_{\mathcal{U}(c_i)}(\tilde{\mathcal{B}}_i) > 0$ and $\tilde{\mathcal{B}}_i \cap \left(\mathcal{B}_i \cup \left(\bigcup_b \mathcal{B}_i^b\right)\right) = \emptyset$. Consider one such subset $\tilde{\mathcal{B}}_i$ from each of the $3m$ centers in C_m and let $\tilde{\mathcal{B}}^m = \bigcup_{i=1}^{3m} \tilde{\mathcal{B}}_i$. Finally, make the following adjustment to each $h_b \in \mathcal{H}_m$,

$$h_b(z) = \begin{cases} 0 & \text{if } z \in \bigcup_{i=1}^{3m} \mathcal{B}_i^b \cup \mathcal{B}_i \text{ or } z \in \mathcal{B}^{m'} \cup \tilde{\mathcal{B}}^{m'} \text{ for } m' \neq m \\ 1 & \text{otherwise} \end{cases}$$

One can verify that every hypothesis in $\mathcal{H}_m$ has a non-robust region (i.e. $\mathcal{B}^{m'} \cup \tilde{\mathcal{B}}^{m'}$ for $m' \neq m$) with mass strictly bigger than ρ in every center in $C_{m'}$ for every $m' \neq m$. Thus, the hypotheses in $\mathcal{H}_m$ are non-robust to points in $C_{m'}$ for all $m' \neq m$. Finally, as we did in Lemma 2, for each m, we only keep the subset of hypothesis $\mathcal{H}'_m = \{h_b \in \mathcal{H}_m : \sum_{i=1}^{3m} b_i = m\} \subset \mathcal{H}$. Note that for each $m \in \mathbb{N}$, the hypothesis class $\mathcal{H}'_m$ behaves exactly like the hypothesis class from Lemma 2 on C_m.

Let $\mathcal{H} = \bigcup_{m=1}^{\infty} \mathcal{H}'_m$ and $\mathcal{U}(C_m) = \bigcup_{i=1}^{3m} \mathcal{U}(c_i)$. Since we have modified the hypothesis class, we need to reprove that its VC dimension is still at most 1.

Consider two points $x_1, x_2 \in \mathcal{X}$. If either x_1 or x_2 is not in $\bigcup_{m=1}^{\infty} \mathcal{U}(C_m)$ and not in $\bigcup_{m=1}^{\infty} \mathcal{B}^m \cup \tilde{\mathcal{B}}^m$, then all hypothesis predict x_1 or x_2 as 1. If both x_1 and x_2 are in $\mathcal{B}^m \cup \tilde{\mathcal{B}}^m$ for some $m \in \mathbb{N}$, then:

- if either x_1 or x_2 are in $\mathcal{B}^m$, every hypothesis in $\mathcal{H}$ labels either x_1 or x_2 as 0.
- if both x_1 and x_2 are in $\tilde{\mathcal{B}}^m$, we can only get the labeling $(1,1)$ from hypotheses in $\mathcal{H}_m$ and the labelling $(0,0)$ from the hypotheses in $\mathcal{H}_{m'}$ for $m' \neq m$.

In the case both x_1 and x_2 are in $\mathcal{U}(C_m) \setminus (\mathcal{B}^m \cup \tilde{\mathcal{B}}^m)$, then, they cannot be shattered by Lemma 1. In the case $x_1 \in \mathcal{B}^m \cup \tilde{\mathcal{B}}^m$ and $x_2 \in \mathcal{U}(C_m) \setminus (\mathcal{B}^m \cup \tilde{\mathcal{B}}^m)$:

- if x_1 is in $\mathcal{B}^m$, every hypothesis in $\mathcal{H}$ labels x_1 as 0.
- if x_1 is in $\tilde{\mathcal{B}}^m$ then, we can never get the labelling $(0,0)$.

If $x_1 \in \mathcal{B}^i \cup \tilde{\mathcal{B}}^i$ and $x_2 \in \mathcal{B}^j \cup \tilde{\mathcal{B}}^j$ for $i \neq j$, then:

- if either x_1 or x_2 are in $\mathcal{B}^i$ or $\mathcal{B}^j$ respectively, every hypothesis in $\mathcal{H}$ labels either x_1 or x_2 as 0.
- if both x_1 and x_2 are in $\tilde{\mathcal{B}}^i$ and $\tilde{\mathcal{B}}^j$ respectively, we can never get the labelling $(1,1)$.

In the case $x_1 \in \mathcal{B}^i \cup \tilde{\mathcal{B}}^i$ and $x_2 \in \mathcal{U}(C_j) \setminus (\mathcal{B}^j \cup \tilde{\mathcal{B}}^j)$ for $j \neq i$, then we cannot obtain the labelling $(1,0)$. If $x_1 \in \mathcal{U}(C_i) \setminus (\mathcal{B}^i \cup \tilde{\mathcal{B}}^i)$ and $x_2 \in \mathcal{U}(C_j) \setminus (\mathcal{B}^j \cup \tilde{\mathcal{B}}^j)$ for $i \neq j$, then we cannot obtain the labelling $(0,0)$. Since we shown that for all possible x_1 and x_2, $\mathcal{H}$ cannot shatter them, $\text{VC}(\mathcal{H}) \leq 1$.

We now use the same reasoning in [MHS19], to show that no proper learning rule works. By Lemma 2, for any proper learning rule $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{H}$ and for any $m \in \mathbb{N}$, we can construct a distribution $\mathcal{D}$ over C_m (which has $3m$ points from $\mathcal{X}$) where there exists a hypothesis $h^* \in \mathcal{H}'_m$ that achieves $R_{\mathcal{U}}(h^*; \mathcal{D}, \rho) = 0$, but with probability at least $1/7$ over $S \sim \mathcal{D}^m$, $R_{\mathcal{U}}(\mathcal{A}(S); \mathcal{D}, \rho) > 1/8$. Note that it suffices to only consider hypothesis in $\mathcal{H}'_m$ because, by construction, all hypothesis in $\mathcal{H}'_{m'}$ for $m' \neq m$ are not probabilistically robust on C_m, and thus always achieve loss 1 on all points in C_m. Thus, rule $\mathcal{A}$ will do worse if it picks hypotheses from these classes. This shows that the sample complexity of properly probabilistically robustly PAC learning $\mathcal{H}$ is arbitrarily large, allowing us to conclude that $\mathcal{H}$ is not properly learnable. □

4 Discussion

The ability to achieve test-time robustness via *proper* learning rules is important from a practical standpoint. It aligns better with the current approaches used in practice and proper learning algorithms

are often more simpler to implement than improper ones. Indeed, for worst-case adversarial robustness, the improper learning algorithm proposed by [MHS19] is complicated and computationally intractable. This motivates understanding when proper learning is possible under our weaker notion of probabilistic robustness. In particular, is proper learning under $R_{\mathcal{U}}(h; \mathcal{D}, \rho)$ possible if we assume a *stronger* realizability assumption, namely $\min_{h \in \mathcal{H}} R_{\mathcal{U}}(h; \mathcal{D}, \rho^*) = 0$ for $\rho^* < \rho$? Crucially, we note that our construction in Lemma 2 fails if this is the case. Answering whether this is sufficient for proper learning is an interesting future question.

References

[DHHR20] Edgar Dobriban, Hamed Hassani, David Hong, and Alexander Robey. Provable tradeoffs in adversarially robust classification. *arXiv preprint arXiv:2006.05161*, 2020.

[LBSS20] Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization. *arXiv preprint arXiv:2007.01162*, 2020.

[LBSS21] Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. On tilted losses in machine learning: Theory and applications. *arXiv preprint arXiv:2109.06141*, 2021.

[LF19] Cassidy Laidlaw and Soheil Feizi. Functional adversarial attacks. *Advances in neural information processing systems*, 32, 2019.

[MHS19] Omar Montasser, Steve Hanneke, and Nathan Srebro. Vc classes are adversarially robustly learnable, but only improperly. In *Conference on Learning Theory*, pages 2512–2530. PMLR, 2019.

[RBZK21] Leslie Rice, Anna Bair, Huan Zhang, and J Zico Kolter. Robustness between the worst and average case. *Advances in Neural Information Processing Systems*, 34:27840–27851, 2021.

[RCPH22] Alexander Robey, Luiz FO Chamon, George J Pappas, and Hamed Hassani. Probabilistically robust learning: Balancing average-and worst-case performance. *arXiv preprint arXiv:2202.01136*, 2022.

[RXY+19] Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C Duchi, and Percy Liang. Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032*, 2019.

[SZC+18] Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. Is robustness the cost of accuracy?–a comprehensive study on the robustness of 18 deep image classification models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018.

[TSE+18] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

[Vap06] Vladimir Vapnik. *Estimation of dependences based on empirical data*. Springer Science & Business Media, 2006.

[VC71] Vladimir Naumovich Vapnik and Aleksei Yakovlevich Chervonenkis. On uniform convergence of the frequencies of events to their probabilities. *Teoriya Veroyatnostei i ee Primeneniya*, 16(2):264–279, 1971.

[YRZ+20] Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Russ R Salakhutdinov, and Kamalika Chaudhuri. A closer look at accuracy vs. robustness. *Advances in neural information processing systems*, 33:8588–8601, 2020.

[ZYJ+19] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [N/A]
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [N/A]
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [N/A]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [N/A]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]

A Auxiliary Proofs Related to Proper Probabilistic Robust Learnability

Proof. [of Lemma 2] This proof closely follows Lemma 3 from [MHS19]. In fact, the only difference is in the construction of the hypothesis class, which we will describe below.

Fix $\rho \in [0, 1)$. Let $m \in \mathbb{N}$. Construct a hypothesis class $\mathcal{H}_0$ as in Lemma 1 on $3m$ centers $c_1, ..., c_{3m}$ based on ρ. By the construction in Lemma 1, we know that $\mathcal{L}_{\mathcal{H}}^{\mathcal{U},\rho}$ shatters the sample $C = \{(c_1, 1), ..., (c_{3m}, 1)\}$. Instead of keeping all of $\mathcal{H}_0$, we will only keep a subset $\mathcal{H}$ of $\mathcal{H}_0$, namely those classifiers that are probabilistically robustly correct on subsets of size $2m$ of C. More specifically, recall from the construction in Lemma 1, that each hypothesis $h_b \in \mathcal{H}_0$ is parameterized by a bitstring $b \in \{0, 1\}^{3m}$ where if $b_i = 1$, then h_b is not robust to example $(c_i, 1)$. Therefore, $\mathcal{H} = \{h_b \in \mathcal{H}_0 : \sum_{i=1}^{3m} b_i = m\}$. Now, let $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{H}$ be an arbitrary proper learning rule. Consider a set of distributions $\mathcal{D}_1, ..., \mathcal{D}_L$ where $L = \binom{3m}{2m}$. Each distribution $\mathcal{D}_i$ is uniform over exactly $2m$ centers in C. Critically, note that by our construction of $\mathcal{H}$, every distribution $\mathcal{D}_i$ is probabilistically robustly realizable by a hypothesis in $\mathcal{H}$. That is, for all $\mathcal{D}_i$, there exists a hypothesis

$h^* \in \mathcal{H}$ s.t. $R_{\mathcal{U}}(h^*; \mathcal{D}, \rho) = 0$. Observe that this satisfies the first condition in Lemma 2. For the second condition, at a high-level, the idea is to use the probabilistic method to show that there exists a distribution $\mathcal{D}_i$ where $\mathbb{E}_{S\sim\mathcal{D}_i^m}\left[R_{\mathcal{U}}(\mathcal{A}(S); \mathcal{D}, \rho)\right] \geq \frac{1}{4}$ and then use a variant of Markov's inequality to show that with probability at least $1/7$ over $S \sim \mathcal{D}^m$, $R_{\mathcal{U}}(\mathcal{A}(S); \mathcal{D}, \rho) > 1/8$.

Let $S \in C^m$ be an arbitrary set of m points. Let $\mathcal{C}$ be a uniform distribution over C. Let $\mathcal{P}$ be a uniform distribution over $\mathcal{D}_1, ..., \mathcal{D}_T$. Let E_S denote the event that $S \subset \text{supp}(\mathcal{D}_i)$ for $\mathcal{D}_i \sim \mathcal{P}$. Given the event E_S, we will lower bound the expected probabilistic robust loss of the hypothesis the proper learning rule $\mathcal{A}$ outputs,

$$\mathbb{E}_{\mathcal{D}_i\sim\mathcal{P}}\left[R_{\mathcal{U}}(\mathcal{A}(S); \mathcal{D}_i, \rho)|E_S\right] = \mathbb{E}_{\mathcal{D}_i\sim\mathcal{P}}\left[\mathbb{E}_{(x,y)\sim\mathcal{D}_i}\left[\mathbb{1}\{\mathbb{P}_{z\sim\mu_{\mathcal{U}(x)}}\left(\mathcal{A}(S)(z) \neq y\right) > \rho\}\right]|E_S\right].$$

Conditioning on the event that $(x, y) \notin S$, denoted, $E_{(x,y)\notin S}$,

$$\begin{aligned}\mathbb{E}_{(x,y)\sim\mathcal{D}_i}\left[\mathbb{1}\{\mathbb{P}_{z\sim\mu_{\mathcal{U}(x)}}\left(\mathcal{A}(S)(z) \neq y\right) > \rho\}\right] &\geq \mathbb{P}_{(x,y)\sim\mathcal{D}_i}\left[E_{(x,y)\notin S}\right] \\ &\quad\times \mathbb{E}_{(x,y)\sim\mathcal{D}_i}\left[\mathbb{1}\{\mathbb{P}_{z\sim\mu_{\mathcal{U}(x)}}\left(\mathcal{A}(S)(z) \neq y\right) > \rho\}|E_{(x,y)\notin S}\right]\end{aligned}$$

Since $\mathcal{D}_i$ is supported over $2m$ points and $|S| = m$, $\mathbb{P}_{(x,y)\sim\mathcal{D}_i}\left[E_{(x,y)\notin S}\right] \geq \frac{1}{2}$ since in the worst-case $S \subset \text{supp}(\mathcal{D}_i)$. Thus, we obtain the lower bound,

$$\mathbb{E}_{\mathcal{D}_i\sim\mathcal{P}}\left[R_{\mathcal{U}}(\mathcal{A}(S); \mathcal{D}_i, \rho)|E_S\right] \geq \frac{1}{2}\mathbb{E}_{\mathcal{D}_i\sim\mathcal{P}}\left[\mathbb{E}_{(x,y)\sim\mathcal{D}_i}\left[\mathbb{1}\{\mathbb{P}_{z\sim\mu_{\mathcal{U}(x)}}\left(\mathcal{A}(S)(z) \neq y\right) > \rho\}|E_{(x,y)\notin S}\right]|E_S\right].$$

Unravelling the expectation over the draw from $\mathcal{D}_i$, we have,

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_i}\left[\mathbb{1}\{\mathbb{P}_{z\sim\mu_{\mathcal{U}(x)}}\left(\mathcal{A}(S)(z) \neq y\right) > \rho\}|E_{(x,y)\notin S}\right] \geq \frac{1}{m}\sum_{(x,y)\in\text{supp}(\mathcal{D}_i)\setminus S}\mathbb{1}\{\mathbb{P}_{z\sim\mu_{\mathcal{U}(x)}}\left(\mathcal{A}(S)(z) \neq y\right) > \rho\}$$

Observing that $\mathbb{E}_{\mathcal{D}_i\sim\mathcal{P}}\left[\mathbb{1}\{(x, y) \in \text{supp}(\mathcal{D}_i)\}|E_S\right] \geq \frac{1}{2}$ yields,

$$\mathbb{E}_{\mathcal{D}_i\sim\mathcal{P}}\left[\mathbb{E}_{(x,y)\sim\mathcal{D}_i}\left[\mathbb{1}\{\mathbb{P}_{z\sim\mu_{\mathcal{U}(x)}}\left(\mathcal{A}(S)(z) \neq y\right) > \rho\}|E_{(x,y)\notin S}\right]|E_S\right] \geq \frac{1}{2m}\sum_{(x,y)\notin S}\mathbb{1}\{\mathbb{P}_{z\sim\mu_{\mathcal{U}(x)}}\left(\mathcal{A}(S)(z) \neq y\right) > \rho\}.$$

Since $\mathcal{A}(S) \in \mathcal{H}$, by construction of $\mathcal{H}$, there are at least m points in C where $\mathcal{A}$ is not probabilistically robustly correct. Therefore,

$$\frac{1}{2m}\sum_{(x,y)\notin S}\mathbb{1}\{\mathbb{P}_{z\sim\mu_{\mathcal{U}(x)}}\left(\mathcal{A}(S)(z) \neq y\right) > \rho\} \geq \frac{1}{2},$$

from which we have that, $\mathbb{E}_{\mathcal{D}_i\sim\mathcal{P}}\left[R_{\mathcal{U}}(\mathcal{A}(S); \mathcal{D}_i, \rho)|E_S\right] \geq \frac{1}{4}$. By the law of total expectation, we have that

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_i\sim\mathcal{P}}\left[\mathbb{E}_{S\sim\mathcal{D}_i^m}\left[R_{\mathcal{U}}(\mathcal{A}(S); \mathcal{D}_i, \rho)\right]\right] &= \mathbb{E}_{S\sim\mathcal{C}}\left[\mathbb{E}_{\mathcal{D}_i\sim\mathcal{P}|E_S}\left[R_{\mathcal{U}}(\mathcal{A}(S); \mathcal{D}_i, \rho)\right]\right] \\ &= \mathbb{E}_{S\sim\mathcal{C}}\left[\mathbb{E}_{\mathcal{D}_i\sim\mathcal{P}}\left[R_{\mathcal{U}}(\mathcal{A}(S); \mathcal{D}_i, \rho)|E_S\right]\right] \\ &\geq 1/4\end{aligned}$$

Since the expectation over $\mathcal{D}_1, ..., \mathcal{D}_T$ is at least $1/4$, there must exist a distribution $\mathcal{D}_i$ where $\mathbb{E}_{S\sim\mathcal{D}_i^m}\left[R_{\mathcal{U}}(\mathcal{A}(S); \mathcal{D}_i, \rho)\right] \geq 1/4$. Using a variant of Markov's inequality, we have that

$$\mathbb{P}_{S\sim\mathcal{D}_i^m}\left[R_{\mathcal{U}}(\mathcal{A}(S); \mathcal{D}_i, \rho) > 1/8\right] \geq 1/7$$

which completes the proof. □