A Characterization of Semi-Supervised Adversarially Robust PAC Learnability

Anonymous Author(s)
Affiliation
Address
email

Abstract

We study the problem of learning an adversarially robust predictor to test time attacks in the *semi-supervised* PAC model. We address the question of how many *labeled* and *unlabeled* examples are required to ensure learning. We show that having enough unlabeled data (the size of a labeled sample that a fully-supervised method would require), the labeled sample complexity can be arbitrarily smaller compared to previous works, and is sharply characterized by a *different* complexity measure. We prove nearly matching upper and lower bounds on this sample complexity. This shows that there is a significant benefit in semi-supervised robust learning even in the worst-case distribution-free model, and establishes a gap between supervised and semi-supervised label complexities which is known not to hold in standard non-robust PAC learning.

1 Introduction

The problem of learning predictors that are immune to adversarial corruptions at inference time is central in modern machine learning. The phenomenon of fooling learning models by adding imperceptible perturbations to their input illustrates a basic vulnerability of learning based models, and is named *adversarial examples*. We study the model of adversarially-robust PAC learning, in a *semi-supervised* setting.

Adversarial robustness has been shown to significantly benefit from semi-supervised learning, mostly empirically, but also theoretically in some specific cases of distributions (e.g., [16, 56, 50, 45, 1, 53, 34]). In this paper we ask the following natural question. To what extent can we benefit from *unlabeled* data in the learning process of robust models in the general case? More specifically, what is the sample complexity in a distribution free model?

Our semi-supervised model is formalized as follows. Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ be a hypothesis class. We formalize the adversarial attack by a perturbation function $\mathcal{U} : \mathcal{X} \to 2^{\mathcal{X}}$, where $\mathcal{U}(x)$ is the set of possible perturbations (attacks) on x. In practice, we usually consider $\mathcal{U}(x)$ to be the ℓ_p ball centered at x. In this paper, we have no restriction on $\mathcal{U}$, besides $x \in \mathcal{U}(x)$. The robust error of hypothesis h on a pair (x, y) is $\sup_{z \in \mathcal{U}(x)} \mathbb{I}[h(z) \neq y]$. The learner has access to both *labeled* and *unlabeled* examples drawn i.i.d. from unknown distribution $\mathcal{D}$, and the goal is to find $h \in \mathcal{H}$ with low robust error on a random point from $\mathcal{D}$. The sample complexity in semi-supervised learning has two parameters, the number of labeled examples and the number of unlabeled examples which suffice to ensure learning. The learner would like to restrict the amount of labeled data, which is significantly more expensive to obtain than unlabeled data.

In this paper, we show a gap between supervised and semi-supervised label complexities of adversarially robust learning in a distribution free model. The label complexity in semi-supervised may be arbitrarily smaller compared to the supervised case, and is characterized by a different complexity

Submitted to 36th Conference on Neural Information Processing Systems (NeurIPS 2022). Do not distribute.

measure. Importantly, we are not using more data, just less labeled data. The unlabeled sample size is the same as how much labeled data a fully-supervised method would require, and so this is a strict improvement. This kind of gap is known not to hold in standard (non-robust) PAC learning, this is a unique property of robust learning.

Background. The following complexity measure $\mathrm{VC}_{\mathcal{U}}$ was introduced by [39] (and denoted there by $\dim_{\mathcal{U}\times}$) as a candidate for determining the sample complexity of supervised robust learning. It was shown that indeed its finiteness is necessary, but not sufficient. This parameter is our primary object in this work, as we will show that it characterizes the labeled sample complexity of *semi-supervised* robust PAC-learning.

Definition 1.1 ($\mathrm{VC}_{\mathcal{U}}$**-dimension)** A sequence of points $\{x_1, \ldots, x_k\}$ is $\mathcal{U}$*-shattered* by $\mathcal{H}$ if $\forall y_1, \ldots, y_k \in \{0,1\}$, $\exists h \in \mathcal{H}$ such that $\forall i \in [k], \forall z \in \mathcal{U}(x_i)$, $h(z) = y_i$. The $\mathrm{VC}_{\mathcal{U}}(\mathcal{H})$ is largest integer k for which there exists a sequence $\{x_1, \ldots, x_k\}$ $\mathcal{U}$-shattered by $\mathcal{H}$.

Intuitively, this dimension relates to shattering of the entire perturbation sets, instead of one point in the standard VC-dimension. When $\mathcal{U}(x) = \{x\}$, this parameter coincides with the standard VC. Moreover, it holds that $\mathrm{VC}_{\mathcal{U}} \leq \mathrm{VC}$ and the gap can be arbitrarily large, that is, there exist $\mathcal{H}$ such that $\mathrm{VC}_{\mathcal{U}}(\mathcal{H}) = 0$ and $\mathrm{VC}(\mathcal{H}) = \infty$ (see Proposition 3.2).

For an improved lower bound on the sample complexity, Montasser et al. [39, Theorem 10] introduced the Robust Shattering dimension, denoted by $\mathrm{RS}_{\mathcal{U}}$ (and denoted there by $\dim_{\mathcal{U}}$).

Definition 1.2 ($\mathrm{RS}_{\mathcal{U}}$**-dimension)** A sequence $x_1, \ldots, x_k$ is said to be $\mathcal{U}$-robustly shattered by $\mathcal{F}$ if $\exists z_1^+, z_1^-, \ldots, z_k^+, z_k^-$ such that $x_i \in \mathcal{U}\left(z_i^+\right) \cap \mathcal{U}\left(z_i^-\right) \forall i \in [k]$ and $\forall y_1, \ldots, y_k \in \{+,-\}, \exists f \in \mathcal{F}$ with $f(\zeta) = y_i, \forall \zeta \in \mathcal{U}\left(z_i^{y_i}\right), \forall i \in [k]$. The $\mathcal{U}$-robust shattering dimension $\mathrm{RS}_{\mathcal{U}}(\mathcal{H})$ is defined as the maximum size of a set that is $\mathcal{U}$-robustly shattered by $\mathcal{H}$.

Specifically, the lower bound on the sample complexity is $\Omega\left(\frac{\mathrm{RS}_{\mathcal{U}}}{\epsilon} + \frac{1}{\epsilon}\log\frac{1}{\delta}\right)$ for realizable robust learning, and $\Omega\left(\frac{\mathrm{RS}_{\mathcal{U}}}{\epsilon^2} + \frac{1}{\epsilon^2}\log\frac{1}{\delta}\right)$ for agnostic robust learning. They also showed upper bounds of $\tilde{\mathcal{O}}\left(\frac{\mathrm{VC}\cdot\mathrm{VC}^*}{\epsilon} + \frac{\log\frac{1}{\delta}}{\epsilon}\right)$ in the realizable case and $\tilde{\mathcal{O}}\left(\frac{\mathrm{VC}\cdot\mathrm{VC}^*}{\epsilon^2} + \frac{\log\frac{1}{\delta}}{\epsilon^2}\right)$ in the agnostic case, where VC^* is the dual VC dimension (definitions are in Appendix A). Montasser et al. [39] showed that for any $\mathcal{H}$, $\mathrm{VC}_{\mathcal{U}}(\mathcal{H}) \leq \mathrm{RS}_{\mathcal{U}}(\mathcal{H}) \leq \mathrm{VC}(\mathcal{H})$, and there can be an arbitrary gap between them. Specifically, there exists $\mathcal{H}_0$ with $\mathrm{VC}_{\mathcal{U}}(\mathcal{H}_0) = 0$ and $\mathrm{RS}_{\mathcal{U}}(\mathcal{H}_0) = \infty$, and there exists $\mathcal{H}_1$ with $\mathrm{RS}_{\mathcal{U}}(\mathcal{H}_1) = 0$ and $\mathrm{VC}(\mathcal{H}_1) = \infty$.

Main contributions.

- In Section 3, we first analyze the simple case where the support of the marginal distribution on the inputs is fully known to the learner. In this case, we show a tight bound of $\Theta\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon} + \frac{\log\frac{1}{\delta}}{\epsilon}\right)$ on the labeled complexity for learning $\mathcal{H}$.
- In Section 4, we present a generic algorithm that can be applied both for the realizable and agnostic settings. We prove an upper bound and nearly matching lower bounds on the sample complexity in the realizable case.
  For semi-supervised robust learning, we prove a labeled sample complexity bound Λ^{ss} and compare to the sample complexity of supervised robust learning Λ^{s}. Our algorithm uses $\Lambda^{\mathrm{ss}} = \tilde{\mathcal{O}}\left(\frac{\mathrm{VC}_{\mathcal{U}}}{\epsilon} + \frac{1}{\epsilon}\log\frac{1}{\delta}\right)$ *labeled* examples and $\mathcal{O}(\Lambda^{\mathrm{s}})$ *unlabeled* examples. Recall that $\Lambda^{\mathrm{s}} = \Omega(\mathrm{RS}_{\mathcal{U}})$, and since $\mathrm{RS}_{\mathcal{U}}$ can be arbitrarily larger than $\mathrm{VC}_{\mathcal{U}}$, this means our labeled sample complexity represents a strong improvement over the sample complexity of supervised learning.
- In Section 5, we prove upper and lower bounds on the sample complexity in the agnostic setting. We reveal an interesting structure, which is inherently different than the realizable case. Let η be the minimal agnostic error. If we allow an error of $2\eta + \epsilon$, it is sufficient for our algorithm to have $\Lambda^{\mathrm{ss}} = \tilde{\mathcal{O}}\left(\frac{\mathrm{VC}_{\mathcal{U}}}{\epsilon^2} + \frac{\log\frac{1}{\delta}}{\epsilon^2}\right)$ *labeled* examples and $\mathcal{O}(\Lambda^{\mathrm{s}})$ *unlabeled* examples (as in the realizable case). If we insist on having error $\eta + \epsilon$, then there is a lower bound of $\Lambda^{\mathrm{ss}} = \Omega\left(\frac{\mathrm{RS}_{\mathcal{U}}}{\epsilon^2} + \frac{1}{\epsilon^2}\log\frac{1}{\delta}\right)$ labeled examples. Furthermore, an error of $(\frac{3}{2} - \gamma)\eta + \epsilon$ is unavoidable if the learner is restricted to $\mathcal{O}(\mathrm{VC}_{\mathcal{U}})$ labeled examples, for any

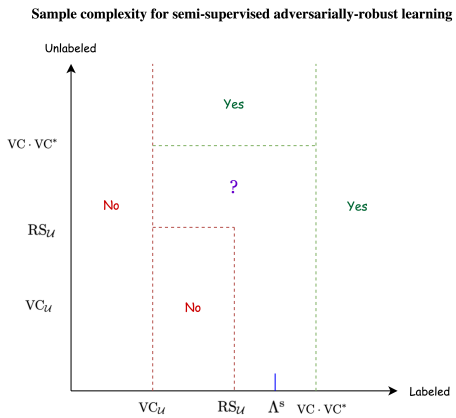

Figure 1: Summary of the sample complexity regimes for semi-supervised robust learning, for the realizable model and the agnostic model with error $2\eta + \epsilon$, where η is the minimal agnostic error in the hypothesis class.
Obtaining an error of $\eta+\epsilon$ requires at least $\mathrm{RS}_{\mathcal{U}}$ labeled examples, as in the supervised case.
Λ^{s} denotes the sample complexity of supervised robust learning. It is an open question whether Λ^{s} equals $\mathrm{RS}_{\mathcal{U}}$.

$\gamma > 0$. We also show that *improper* learning is necessary, similar to the supervised case. We summarize the results in Fig. 1 showing for which labeled and unlabeled samples we have a robust learner.

- The above results show that there is a significant benefit in semi-supervised robust learning. For example, take $\mathcal{H}_0$ with $\mathrm{VC}_{\mathcal{U}}(\mathcal{H}_0) = 0$ and $\mathrm{RS}_{\mathcal{U}}(\mathcal{H}_0) = n$. The labeled sample size for learning $\mathcal{H}_0$ in supervised learning is $\Omega(n)$. In contrast, in semi-supervised learning our algorithms requires only $\mathcal{O}(1)$ *labeled* examples and $\mathcal{O}(n)$ *unlabeled* examples. We are not using more data, just less labeled data. Note that n can be arbitrarily large.
- A byproduct of our result is that if we assume that the distribution is robustly realizable by a hypothesis class (i.e., there exist a hypothesis with zero robust error) then, with respect to the <u>non-robust</u> loss (i.e., the standard 0-1 loss) we can learn with only $\tilde{\mathcal{O}}\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon} + \frac{\log\frac{1}{\delta}}{\epsilon}\right)$ labeled examples, even if the VC is infinite. Recall that there exists $\mathcal{H}_0$ with $\mathrm{VC}_{\mathcal{U}}(\mathcal{H}_0) = 0$, $\mathrm{RS}_{\mathcal{U}}(\mathcal{H}_0) = \infty$ and $\mathrm{VC}(\mathcal{H}_0) = \infty$. Learning linear functions with margin is a special case of this data-dependent assumption. Moreover, we show that this is obtained only by *improper* learning. (See Section 6.)

Related work. *Adversarially robust learning.* The most related work of [39] studied the setting of supervised robust PAC learning. Adversarially robust learning has been extensively studied in several supervised learning models [e.g., 22, 48, 33, 55, 18, 5, 8, 43, 40–42, 3, 9, 19, 7, 13, 54]

Semi-supervised learning. There is a substantial interest in semi-supervised (non-robust) learning, and many contemporary practical problems significantly benefit from it [14, 17, 57]. This was formalized in theoretical frameworks, extending the standard PAC model [11, 10]. Moreover, [25] studied under which circumstances the learning rate can be improved given unlabeled data. In this paper we discuss semi-supervised learning in the standard PAC model, but with respect to the robust loss function. [16] studies a robust semi-supervised setting where the distribution is mixture of Gaussians and the hypothesis class is linear separators.

2 Preliminaries

Let $\mathcal{X}$ be the instance space, $\mathcal{Y}$ a label space, and $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ a hypothesis class. A perturbation function $\mathcal{U} : \mathcal{X} \to 2^{\mathcal{X}}$ maps an input to a set $\mathcal{U}(x) \subseteq \mathcal{X}$. Denote the 0-1 loss of hypothesis h on (x, y) by $\ell_{0\text{-}1}(h; x, y) = \mathbb{I}[h(x) \neq y]$, and the robust loss with respect to $\mathcal{U}$ by $\ell_{\mathcal{U}}(h; x, y) = \sup_{z\in\mathcal{U}(x)} \mathbb{I}[h(z) \neq y]$. Denote the support of a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ by $\mathrm{supp}(\mathcal{D}) = \{(x, y) \in \mathcal{X} \times \mathcal{Y} : \mathcal{D}(x, y) > 0\}$. Denote the marginal distribution $\mathcal{D}_{\mathcal{X}}$ on $\mathcal{X}$ and its support by $\mathrm{supp}(\mathcal{D}_{\mathcal{X}}) = \{x \in \mathcal{X} : \mathcal{D}(x, y) > 0\}$. Define the *robust risk* of a hypothesis $h \in \mathcal{H}$ with respect to distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$,

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

The approximation error of $\mathcal{H}$ on $\mathcal{D}$, namely, the optimal robust error achievable by a hypothesis in $\mathcal{H}$ on $\mathcal{D}$ is denoted by,

$$\mathrm{R}_{\mathcal{U}}(\mathcal{H};\mathcal{D}) = \inf_{h\in\mathcal{H}} \mathrm{R}_{\mathcal{U}}\left(h;\mathcal{D}\right).$$

We say that a distribution $\mathcal{D}$ is *robustly realizable* by a class $\mathcal{H}$ if $\mathrm{R}_{\mathcal{U}}(\mathcal{H};\mathcal{D}) = 0$.

Define the *empirical robust risk* of a hypothesis $h \in \mathcal{H}$ with respect to a sequence $S \in (\mathcal{X}\times\mathcal{Y})^*$,

$$\widehat{\mathrm{R}}_{\mathcal{U}}\left(h;S\right) = \frac{1}{|S|}\sum_{(x,y)\in S} \ell_{\mathcal{U}}(h;x,y) = \frac{1}{|S|}\sum_{(x,y)\in S}\left[\sup_{z\in\mathcal{U}(x)} \mathbb{I}\left[h(z)\neq y\right]\right].$$

The *robust empirical risk minimizer* learning algorithm $\mathrm{RERM} : (\mathcal{X}\times\mathcal{Y})^* \to \mathcal{H}$ for a class $\mathcal{H}$ on a sequence S is defined by

$$\mathrm{RERM}_{\mathcal{H}}(S) \in \operatorname*{argmin}_{h\in\mathcal{H}} \widehat{\mathrm{R}}_{\mathcal{U}}\left(h;S\right).$$

When the perturbation function is the identity, $\mathcal{U}(x) = \{x\}$, we recover the standard notions. The *risk* of a hypothesis $h \in \mathcal{H}$ with respect to distribution $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$ is defined by $\mathrm{R}\left(h;\mathcal{D}\right) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell_{\text{0-1}}(h;x,y)\right] = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\mathbb{I}\left[h(x)\neq y\right]\right]$, and the *empirical risk* of a hypothesis $h \in \mathcal{H}$ with respect to a sequence $S \in (\mathcal{X}\times\mathcal{Y})^*$ is defined by $\widehat{\mathrm{R}}\left(h;S\right) = \frac{1}{|S|}\sum_{(x,y)\in S}\ell_{\text{0-1}}(h;x,y) = \frac{1}{|S|}\sum_{(x,y)\in S}\left[\mathbb{I}\left[h(x)\neq y\right]\right]$. The *empirical risk minimizer* learning algorithm $\mathrm{ERM} : (\mathcal{X}\times\mathcal{Y})^* \to \mathcal{H}$ for a class $\mathcal{H}$ on a sequence S is defined by $\mathrm{ERM}_{\mathcal{H}}(S) \in \operatorname{argmin}_{h\in\mathcal{H}} \widehat{\mathrm{R}}\left(h;S\right)$.

A learning algorithm $\mathcal{A} : (\mathcal{X}\times\mathcal{Y})^* \to \mathcal{Y}^{\mathcal{X}}$ for a class $\mathcal{H}$ is called *proper* if it always outputs a hypothesis in $\mathcal{H}$, otherwise it is called *improper*.

Realizable robust PAC **learning.** We define the supervised and semi-supervised settings.

Definition 2.1 (Realizable robust PAC **learnability)** For any $\epsilon,\delta\in(0,1)$, the sample complexity of realizable robust (ϵ,δ)-PAC learning for a class $\mathcal{H}$, with respect to perturbation function $\mathcal{U}$, denoted by $\Lambda_{\mathrm{RE}}(\epsilon,\delta,\mathcal{H},\mathcal{U})$, is the smallest integer m for which there exists a learning algorithm $\mathcal{A} : (\mathcal{X}\times\mathcal{Y})^* \to \mathcal{Y}^{\mathcal{X}}$, such that for every distribution $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$ robustly realizable by $\mathcal{H}$, namely $\mathrm{R}_{\mathcal{U}}\left(\mathcal{H};D\right) = 0$, for a random sample $S\sim\mathcal{D}^m$, it holds that

$$\mathbb{P}\left(\mathrm{R}_{\mathcal{U}}\left(\mathcal{A}(S);D\right)\leq\epsilon\right) > 1-\delta.$$

If no such m exists, define $\Lambda_{\mathrm{RE}}(\epsilon,\delta,\mathcal{H},\mathcal{U}) = \infty$, and $\mathcal{H}$ is not robustly (ϵ,δ)-PAC learnable with respect to $\mathcal{U}$.

For the standard (non-robust) learning with the 0-1 loss function, we omit the dependence on $\mathcal{U}$ and denote the sample complexity of class $\mathcal{H}$ by $\Lambda_{\mathrm{RE}}(\epsilon,\delta,\mathcal{H})$.

Definition 2.2 (Realizable semi-supervised robust PAC **learnability)** A hypothesis class $\mathcal{H}$ is semi-supervised realizable robust (ϵ,δ)-PAC learnable, with respect to perturbation function $\mathcal{U}$, if for any $\epsilon,\delta\in(0,1)$, there exists $m_u, m_l \in \mathbb{N}\cup\{0\}$, and a learning algorithm $\mathcal{A} : (\mathcal{X}\times\mathcal{Y})^* \cup (\mathcal{X})^* \to \mathcal{Y}^{\mathcal{X}}$, such that for every distribution $\mathcal{D}$ over $\mathcal{X}\times\mathcal{Y}$ robustly realizable by $\mathcal{H}$, namely $\mathrm{R}_{\mathcal{U}}\left(\mathcal{H};D\right) = 0$, for random samples $S^l\sim\mathcal{D}^{m_l}$ and $S^u_{\mathcal{X}}\sim\mathcal{D}^{m_u}_{\mathcal{X}}$, it holds that

$$\mathbb{P}\left(\mathrm{R}_{\mathcal{U}}\left(\mathcal{A}(S^l,S^u_{\mathcal{X}});D\right)\leq\epsilon\right) > 1-\delta.$$

The sample complexity $\mathcal{M}_{\mathrm{RE}}(\epsilon,\delta,\mathcal{H},\mathcal{U})$ includes all such pairs (m_u,m_l). If no such (m_u,m_l) exist, then $\mathcal{M}_{\mathrm{RE}}(\epsilon,\delta,\mathcal{H},\mathcal{U}) = \emptyset$.

Agnostic robust PAC **learning.** In this case we have $\mathrm{R}_{\mathcal{U}}\left(\mathcal{H};\mathcal{D}\right) > 0$, and we would like to compete with the optimal $h\in\mathcal{H}$. We add a parameter to the sample complexity, denoted by η, which is the optimal robust error of a hypothesis in $\mathcal{H}$, namely $\eta = \mathrm{R}_{\mathcal{U}}\left(\mathcal{H};\mathcal{D}\right)$. We say that a function f is (α,ϵ)-optimal if $\mathrm{R}_{\mathcal{U}}\left(f;D\right)\leq\alpha\eta+\epsilon$.

Definition 2.3 (Agnostic robust PAC **learnability)** For any $\epsilon,\delta\in(0,1)$, the sample complexity of agnostic robust (α,ϵ,δ)-PAC learning for a class $\mathcal{H}$, with respect to perturbation function $\mathcal{U}$, denoted by $\Lambda_{\mathrm{AG}}(\alpha,\epsilon,\delta,\mathcal{H},\mathcal{U},\eta)$, is the smallest integer m, for which there exists a learning algorithm

$\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{Y}^{\mathcal{X}}$, such that for every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, for a random sample $S \sim \mathcal{D}^m$, it holds that

$$\mathbb{P}\left(\mathrm{R}_{\mathcal{U}}\left(\mathcal{A}(S); D\right) \leq \alpha \inf_{h \in \mathcal{H}} \mathrm{R}_{\mathcal{U}}\left(h; \mathcal{D}\right) + \epsilon\right) > 1 - \delta.$$

If no such m exists, define $\Lambda_{\mathrm{AG}}(\alpha, \epsilon, \delta, \mathcal{H}, \mathcal{U}, \eta) = \infty$, and $\mathcal{H}$ is not robustly $(\alpha, \epsilon, \delta)$-PAC learnable in the agnostic setting with respect to $\mathcal{U}$. Note that for $\alpha = 1$ we recover the standard agnostic definition, our notation allows for a more relaxed approximation.

Analogously, we define the semi-supervised case.

Definition 2.4 (Agnostic semi-supervised robust PAC **learnability)** A hypothesis class $\mathcal{H}$ is semi-supervised agnostically robust $(\alpha, \epsilon, \delta)$-PAC learnable, with respect to perturbation function $\mathcal{U}$, if for any $\epsilon, \delta \in (0, 1)$, there exists $m_u, m_l \in \mathbb{N} \cup \{0\}$, and a learning algorithm $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \cup (\mathcal{X})^* \to \mathcal{Y}^{\mathcal{X}}$, such that for every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, for random samples $S^l \sim \mathcal{D}^{m_l}$ and $S^u_{\mathcal{X}} \sim \mathcal{D}^{m_u}_{\mathcal{X}}$, it holds that

$$\mathbb{P}\left(\mathrm{R}_{\mathcal{U}}\left(\mathcal{A}(S^l, S^u_{\mathcal{X}}); \mathcal{D}\right) \leq \alpha \inf_{h \in \mathcal{H}} \mathrm{R}_{\mathcal{U}}\left(h; \mathcal{D}\right) + \epsilon\right) > 1 - \delta.$$

The sample complexity $\mathcal{M}_{\mathrm{AG}}(\alpha, \epsilon, \delta, \mathcal{H}, \mathcal{U}, \eta)$ includes all such pairs (m_u, m_l). If no such (m_u, m_l) exist, then $\mathcal{M}_{\mathrm{AG}}(\alpha, \epsilon, \delta, \mathcal{H}, \mathcal{U}, \eta) = \emptyset$.

Partial concept classes [2]. Let a partial concept class $\mathcal{H} \subseteq \{0, 1, \star\}^{\mathcal{X}}$. For $h \in \mathcal{H}$ and input x such that $h(x) = \star$, we say that h is undefined on x. The support of a partial hypothesis $h : \mathcal{X} \to \{0, 1, \star\}$ is the preimage of $\{0, 1\}$, formally, $h^{-1}(\{0, 1\}) = \{x \in \mathcal{X} : h(x) \neq \star\}$. The main motivation of introducing partial concepts classes, is that data-dependent assumptions can be modeled in a natural way that extends the classic theory of total concepts. The VC dimension of a partial class $\mathcal{H}$ is defined as the maximum size of a shattered set $S \subseteq \mathcal{X}$, where S is shattered by $\mathcal{H}$ if the projection of $\mathcal{H}$ on S contains all possible binary patterns, $\{0, 1\}^S \subseteq \mathcal{H}|_S$. The VC-dimension also characterizes verbatim the PAC learnability of partial concept classes, even though uniform convergence does not hold in this setting,

We use the notation $\tilde{\mathcal{O}}(\cdot)$ for omitting poly-logarithmic factors (of $\mathrm{VC}, \mathrm{VC}^*, \mathrm{VC}_{\mathcal{U}}, \mathrm{RS}_{\mathcal{U}}, 1/\epsilon, 1/\delta$). See Appendix A for additional preliminaries on complexity measures, sample compression schemes, and partial concept classes.

3 Warm-up: knowing the support of the marginal distribution

In this section, we provide a tight bound on the labeled sample complexity when the support of marginal distribution is fully known to the learner, under the robust realizable assumption. Studying this setting gives an intuition for the general semi-supervised model. The main idea is that as long as we know the support of the marginal distribution, $\mathrm{supp}(\mathcal{D}_{\mathcal{X}}) = \{x \in \mathcal{X} : \mathcal{D}(x, y) > 0\}$, we can restrict our search to a subspace of functions that are robustly self-consistent, $\mathcal{H}_{\mathcal{U}\text{-cons}} \subseteq \mathcal{H}$, where

$$\mathcal{H}_{\mathcal{U}\text{-cons}} = \left\{h \in \mathcal{H} : \forall x \in \mathrm{supp}(\mathcal{D}_{\mathcal{X}}), \forall z, z' \in \mathcal{U}(x), h(z) = h(z')\right\}.$$

As long as the distribution is robustly realizable, i.e., $\mathrm{R}_{\mathcal{U}}(\mathcal{H}; \mathcal{D}) = 0$, we are guaranteed that the target hypothesis belongs to $\mathcal{H}_{\mathcal{U}\text{-cons}}$. As a result, it suffices to learn the class $\mathcal{H}_{\mathcal{U}\text{-cons}}$ with the 0-1 loss function, in order to robustly learn the original class $\mathcal{H}$. We observe that,

$$\mathrm{VC}(\mathcal{H}_{\mathcal{U}\text{-cons}}) = \mathrm{VC}_{\mathcal{U}}(\mathcal{H}) \leq \mathrm{VC}(\mathcal{H}).$$

Moreover, there exits $\mathcal{H}_0$ with $\mathrm{VC}_{\mathcal{U}}(\mathcal{H}_0) = 0$ and $\mathrm{VC}(\mathcal{H}_0) = \infty$ (see Proposition 3.2). Fortunately, moving from $\mathrm{VC}(\mathcal{H})$ to $\mathrm{VC}_{\mathcal{U}}(\mathcal{H})$ implies a significant sample complexity improvement. Since $\mathrm{supp}(\mathcal{D}_{\mathcal{X}})$ is known, we can now employ any algorithm for learning the hypothesis class $\mathcal{H}_{\mathcal{U}\text{-cons}}$. [1]

[1] See Mohri et al. [38, Chapter 3] for standard upper and lower bounds. In order to remove the superfluous $\log \frac{1}{\epsilon}$ factor of the standard uniform convergence based upper bound, $\mathcal{O}\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon} \log \frac{1}{\epsilon} + \frac{\log \frac{1}{\delta}}{\epsilon}\right)$, we can use the learning algorithm and its analysis from Hanneke [28] that applies for any $\mathcal{H}$ and $\mathcal{D}$, or some other algorithms that are doing so while restricting the hypothesis class or the data distribution [e.g., 6, 20, 29, 27, 36, 24, 15, 12]).

This leads eventually to robustly learn $\mathcal{H}$ with labeled sample complexity that scales linearly with $\mathrm{VC}_{\mathcal{U}}$ (instead of the VC). Formally,

Theorem 3.1 *For hypothesis class $\mathcal{H}$ and adversary $\mathcal{U}$, when the support of the marginal distribution $\mathcal{D}_{\mathcal{X}}$ is known, the labeled sample complexity is* $\Theta\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon} + \frac{\log\frac{1}{\delta}}{\epsilon}\right)$.

The following Proposition demonstrates that semi-supervised robust learning requires much less labeled samples compared to the supervised counterpart. Recall the lower bound on the sample complexity of supervised robust learning, $\Lambda_{\mathrm{RE}}(\epsilon, \delta, \mathcal{H}, \mathcal{U}) = \Omega\left(\frac{\mathrm{RS}_{\mathcal{U}}(\mathcal{H})}{\epsilon} + \frac{1}{\epsilon}\log\frac{1}{\delta}\right)$ given by Montasser et al. [39, Theorem 10]. For completeness, we prove the following in Appendix B.

Proposition 3.2 (Montasser et al. [39], Proposition 9) *There exists a hypothesis class $\mathcal{H}_0$ such that $\mathrm{VC}_{\mathcal{U}}(\mathcal{H}_0) = 0$, $\mathrm{RS}_{\mathcal{U}}(\mathcal{H}_0) = \infty$, and $\mathrm{VC}(\mathcal{H}_0) = \infty$.*

We can now conclude the following separation result on supervised and semi-supervised label complexities.

Corollary 3.3 *The hypothesis class in Proposition 3.2 is not learnable in supervised robust learning (i.e., we need to see the entire data distribution). However, when $\mathrm{supp}(\mathcal{D}_{\mathcal{X}})$ is known, this class can be learned with $\mathcal{O}(\frac{1}{\epsilon}\log\frac{1}{\delta})$ labeled examples.*

In the next section we conclude a stronger separation in the general semi-supervised setting. The size of the labeled data required in the supervised case is lower bounded by $\mathrm{RS}_{\mathcal{U}}$, whereas in the semi-supervised case the *labeled* sample complexity depends only on $\mathrm{VC}_{\mathcal{U}}$ and the *unlabeled* data is lower bounded by $\mathrm{RS}_{\mathcal{U}}$. Moreover, note that in Theorem 3.1, when $\mathrm{supp}(\mathcal{D}_{\mathcal{X}})$ is known, we can use any proper learner. In Section 4 we show that in the general semi-supervised model this is not the case, and sometimes improper learning is necessary, similarly to supervised robust learning.

4 Near-optimal semi-supervised sample complexity

In this section we present our algorithm and its guarantees for the realizable setting. We also prove nearly matching lower bounds on the sample complexity. Finally, we show that improper learning is necessary in semi-supervised robust learning, similar to the supervised case.

We present a generic semi-supervised robust learner, that can be applied on both realizable and agnostic settings. The algorithm uses the following two subroutines. The first one is any algorithm for learning partial concept classes, which controls our *labeled* sample size. (In Appendix F we discuss in detail the algorithm suggested in [2].) The second subroutine, is any algorithm for the agnostic adversarially robust supervised learning, which controls our *unlabeled* sample size. (In Appendix G we discuss in detail the algorithm suggested by [39].) Any progress on one of these problems improves directly the guarantees of our algorithm. We use the following definition that explains how to convert a total concept class into a partial one, in a way that preserves the idea of the robust loss function.

Definition 4.1 Let a hypothesis class $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$ and a perturbation function $\mathcal{U} : \mathcal{X} \to 2^{\mathcal{X}}$. For $h \in \mathcal{H}$ define the corresponding partial concept $h^\star : \mathcal{X} \to \{0, 1, \star\}$, denote this mapping by $\varphi(h) = h^\star$. For $x \in \mathcal{X}$, whenever h is not consistent on the entire set $\mathcal{U}(x)$, i.e., $\exists z, z' \in \mathcal{U}(x), h(z) \neq h(z')$, define $h^\star(x) = \star$. Otherwise, h is robustly self-consistent on x, i.e., $\forall z, z' \in \mathcal{U}(x), h(z) = h(z')$ and h remains unchanged, $h^\star(x) = h(x)$. The corresponding partial concept class is defined by $\mathcal{H}^\star_{\mathcal{U}} = \{h^\star : \varphi(h) = h^\star,\ \forall h \in \mathcal{H}\}$.

The main motivation for the above definition is the following. Fix a hypothesis h. For any point x, as defined above, the adversary can force a mistake of h regardless of h prediction. We would like to mark such points as *mistake*. We do this by defining a partial concept $h^\star$ and setting $h^\star(x) = \star$, which, for partial concepts, implies a mistake. The benefit of this preprocessing is that we reduce the complexity of the hypothesis class from VC to $\mathrm{VC}_{\mathcal{U}}$, which potentially can reduce the labeled sample complexity.

We are now ready to describe the algorithm.

Algorithm 1 Generic Adversarially-Robust Semi-Supervised (GRASS) learner

Input: Labeled data set $S^l \sim \mathcal{D}^{m_l}$, unlabeled data set $S^u_{\mathcal{X}} \sim \mathcal{D}^{m_u}_{\mathcal{X}}$, hypothesis class $\mathcal{H}$, perturbation function $\mathcal{U}$, parameters ϵ, δ.
Algorithms used: PAC learner $\mathcal{A}$ for partial concept classes, agnostic adversarially robust supervised PAC learner $\mathcal{B}$.

1. Given the class $\mathcal{H}$, construct the hypothesis class $\mathcal{H}^\star_{\mathcal{U}}$ using Definition 4.1.
2. Execute the learning algorithm for partial concepts $\mathcal{A}$ on $\mathcal{H}^\star_{\mathcal{U}}$ and sample S^l, with the 0-1 loss and parameters $\frac{\epsilon}{2}, \frac{\delta}{2}$. Denote the resulting hypothesis h_1.
3. Label the unlabeled data set $S^u_{\mathcal{X}}$ with h_1, denote the labeled sample by S^u.
4. Execute the agnostic adversarially robust supervised PAC learner $\mathcal{B}$ on S^u with parameters $\frac{\epsilon}{2}, \frac{\delta}{2}$. Denote the resulting hypothesis h_2.

Output: h_2.

Algorithm Motivation. The main idea behind the algorithm is the following. Given the class $\mathcal{H}^\star_{\mathcal{U}}$, we would like to find a hypothesis $h_1 \in \mathcal{H}^\star_{\mathcal{U}}$ which has a small error, whose existence follows from our realizability assumption. The required sample size scales with $\mathrm{VC}_{\mathcal{U}}$, which is the complexity of $\mathcal{H}^\star_{\mathcal{U}}$, rather than VC. This is where we make a significant gain in the labeled sample complexity. Note that h_1 does not guarantee a small robust error, although it does guarantees a small non-robust error. We utilize an additional unlabeled sample for this task, which we label using h_1. If we would simply minimize the non-robust error on this sample we would simply get back h_1. The main insight is that we would like to minimize the robust error over this sample, which will result in hypothesis h_2. We now need to bound the robust error of h_2. The optimal function h_{opt} has only a slightly increased robust error on this sample, namely, at most on the sample points where it disagrees with h_1. Note that h_1 might have a large robust error due to the perturbation $\mathcal{U}$. However, a robust supervised PAC learner would return a hypothesis h_2 which has robust error similar to h_{opt}, which is at most ϵ.

Algorithm outline and guarantees. In the first step, we convert $\mathcal{H}$ to $\mathcal{H}^\star_{\mathcal{U}}$. Then we employ a learning algorithm $\mathcal{A}$ for partial concepts on $\mathcal{H}^\star_{\mathcal{U}}$ with a labeled sample $S^l \sim \mathcal{D}^{m_l}$. The output of the algorithm is a function h_1 with $\epsilon/2$ on the 0-1 error. Crucially, we needed for this step $|S^l| = \tilde{\mathcal{O}}(\mathrm{VC}_{\mathcal{U}}(\mathcal{H})/\epsilon)$ labeled examples for learning the partial concept $\mathcal{H}^\star_{\mathcal{U}}$, since $\mathrm{VC}(\mathcal{H}^\star_{\mathcal{U}}) = \mathrm{VC}_{\mathcal{U}}(\mathcal{H})$. So our labeled sample size is controlled by the sample complexity for learning partial concepts with the 0-1 loss. In step 3, we label an independent unlabeled sample $S^u_{\mathcal{X}} \sim \mathcal{D}^{m_u}_{\mathcal{X}}$ with h_1, denote his labeled sample by S^u. Define a distribution $\tilde{\mathcal{D}}$ over $\mathcal{X} \times \mathcal{Y}$ by $\tilde{\mathcal{D}}(x, h_1(x)) = \mathcal{D}_{\mathcal{X}}(x)$, and so S^u is an i.i.d. sample from $\tilde{\mathcal{D}}$. We argue that the robust error of $\mathcal{H}$ with respect to $\tilde{\mathcal{D}}$ is $\frac{\epsilon}{2}$, i.e., $\mathrm{R}_{\mathcal{U}}(\mathcal{H}; \tilde{\mathcal{D}}) = \frac{\epsilon}{2}$. Indeed, the function with zero robust error on $\mathcal{D}$, $h_{\mathrm{opt}} \in \mathrm{argmin}_{h \in \mathcal{H}} \mathrm{R}_{\mathcal{U}}(h; \tilde{\mathcal{D}})$ has a robust error of at most $\frac{\epsilon}{2}$ on $\tilde{\mathcal{D}}$. Finally, we employ an agnostic adversarially robust supervised PAC learner $\mathcal{B}$ for the class $\mathcal{H}$ on $S^u \sim \tilde{\mathcal{D}}^{m_u}$, that should be of size of the sample complexity of agnostically robust learn $\mathcal{H}$ with respect to $\mathcal{U}$, when the optimal robust error of hypothesis from $\mathcal{H}$ on $\tilde{\mathcal{D}}$ is at most $\frac{\epsilon}{2}$. We are guaranteed that the resulting hypothesis h_2 has a robust error of at most $\frac{\epsilon}{2} + \frac{\epsilon}{2} = \epsilon$. We conclude that a size of $|S^u_{\mathcal{X}}| = m_u = \Lambda_{\mathrm{AG}}\left(1, \frac{\epsilon}{2}, \frac{\delta}{2}, \mathcal{H}, \mathcal{U}, \eta = \frac{\epsilon}{2}\right)$ unlabeled samples suffices, this completes the proof for Theorem 4.2. For a specific instantiation of such algorithm ([39]), we deduce the sample complexity in Theorem 4.3. A simple analysis of the latter yields a dependence of ϵ^2 for the unlabeled sample size. However, by applying a suitable data-dependent generalization bound, we reduce this dependence to ϵ. (Full proofs appear in Appendix C).

We now formally present the sample complexity of the generic semi-supervised learner for the robust realizable setting. First, in the case of using a generic agnostic robust supervised learner as a subroutine (step 4 in the algorithm). Then we deduce the sample complexity of a specific instantiation of such algorithm.

Theorem 4.2 *For any hypothesis class $\mathcal{H}$ and adversary $\mathcal{U}$, algorithm* GRASS (ϵ, δ)-PAC *learns $\mathcal{H}$ with respect to the robust loss function, in the realizable robust case, with samples of size*

$$m_l = \mathcal{O}\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon} \log^2 \frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon} + \frac{\log \frac{1}{\delta}}{\epsilon}\right) \ , \ m_u = \Lambda_{\mathrm{AG}}\left(1, \frac{\epsilon}{2}, \frac{\delta}{2}, \mathcal{H}, \mathcal{U}, \eta = \frac{\epsilon}{2}\right),$$

where $\Lambda_{\mathrm{AG}}(\alpha, \epsilon, \delta, \mathcal{H}, \mathcal{U}, \eta)$ *is the sample complexity of adversarially-robust agnostic supervised* $(\alpha, \epsilon, \delta)$-PAC *learning, such that* η *is the error of the optimal hypothesis in* $\mathcal{H}$*, i.e.,* $\eta = \mathrm{R}_{\mathcal{U}}(\mathcal{H}; \mathcal{D})$.

The following result follows by using the agnostic supervised robust learner suggested by [39]. A simple analysis of the latter yields a dependence of ϵ^2 for the unlabeled sample size. However, by applying a suitable data-dependent generalization bound, we reduce this dependence to ϵ.

Theorem 4.3 *For any hypothesis class* $\mathcal{H}$ *and adversary* $\mathcal{U}$*, Algorithm* GRASS (ϵ, δ)-PAC *learns* $\mathcal{H}$ *with respect to the robust loss function, in the realizable robust case, with samples of size*

$$m_l = \mathcal{O}\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon}\log^2\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon} + \frac{\log\frac{1}{\delta}}{\epsilon}\right), \; m_u = \tilde{\mathcal{O}}\left(\frac{\mathrm{VC}(\mathcal{H})\,\mathrm{VC}^*(\mathcal{H})}{\epsilon} + \frac{\log\frac{1}{\delta}}{\epsilon}\right).$$

We present nearly matching lower bounds for the realizable setting. The following Corollary stems from Theorem 3.1 and Montasser et al. [39, Theorem 10].

Corollary 4.4 *For any* $\epsilon, \delta \in (0, 1)$*, the sample complexity of realizable robust* (ϵ, δ)-PAC *learning for a class* $\mathcal{H}$*, with respect to perturbation function* $\mathcal{U}$ *is*

$$m_l = \Omega\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon} + \frac{\log\frac{1}{\delta}}{\epsilon}\right), \; m_u = \infty, \quad \textit{or} \quad m_l + m_u = \Omega\left(\frac{\mathrm{RS}_{\mathcal{U}}(\mathcal{H})}{\epsilon} + \frac{\log\frac{1}{\delta}}{\epsilon}\right).$$

Proper vs. improper. In Section 3, we have seen that when the support of the marginal distribution $\mathcal{D}_{\mathcal{X}}$ is known, the labeled sample complexity is $\Theta\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon} + \frac{\log\frac{1}{\delta}}{\epsilon}\right)$. This was obtained by a proper learner: keep the robustly self-consistent hypotheses, $\mathcal{H}_{\mathcal{U}\text{-cons}} \subseteq \mathcal{H}$, and then use ERM on this class. The case when $\mathcal{D}_{\mathcal{X}}$ is unknown is different. We know that there exists a perturbation function $\mathcal{U}$ and a hypothesis class $\mathcal{H}$ with finite VC-dimension that cannot be robustly PAC learned with any proper learning rule [39, Lemma 3]. The same proof holds in the semi-supervised case. Note that both algorithms $\mathcal{A}$ and $\mathcal{B}$ used in Algorithm 1 are improper. (The proof appears in Appendix C).

Theorem 4.5 *There exists* $\mathcal{H}$ *with* $\mathrm{VC}(\mathcal{H}) = 0$ *such that for any proper learning rule* $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \cup (\mathcal{X})^* \to \mathcal{H}$*, there exists a distribution* $\mathcal{D}$ *over* $\mathcal{X} \times \mathcal{Y}$ *that is robustly realizable by* $\mathcal{H}$*, i.e.,* $\mathrm{R}_{\mathcal{U}}(\mathcal{H}; \mathcal{D}) = 0$*. It holds that* $\mathrm{R}_{\mathcal{U}}\left(\mathcal{A}(S^l, S^u_{\mathcal{X}}); D\right) > \frac{1}{8}$ *with probability at least* $\frac{1}{7}$ *over* $S^l \sim \mathcal{D}^{m_l}$ *and* $S^u_{\mathcal{X}} \sim \mathcal{D}^{m_u}$*, where* $m_l, m_u \in \mathbb{N} \cup \{0\}$ *is the size of the labeled and unlabeled samples respectively. Moreover, when the marginal distribution* $\mathcal{D}_{\mathcal{X}}$ *is known, there exists a proper learning rule for any* $\mathcal{H}$*.*

5 Agnostic robust learning

In this section, we prove the guarantees of Algorithm 1 in the more challenging agnostic robust setting. We then prove lower bounds on the sample complexity which exhibit that it is inherently different from the realizable case.

We follow the same steps as in the proof of the realizable case, with the following important difference. In the first two steps of the algorithm, we learn a partial concept class with respect to the 0-1 loss, and obtain a hypothesis with error of $\eta + \epsilon/2$ (the optimal robust error of a hypothesis in $\mathcal{H}$) and not 0. This leads eventually to error of $2\eta + \epsilon$ for learning with respect to the robust loss.

We then present two negative results. In Theorem 5.2 we show that for obtaining error $\eta + \epsilon$ there is a lower bound of $\Omega(\mathrm{RS}_{\mathcal{U}})$ labeled examples, this result coincides with the lower bound of supervised robust learning. In Theorem 5.3, we show that for any $\gamma > 0$ there exist a hypothesis class, such that having access only to $\mathcal{O}(\mathrm{VC}_{\mathcal{U}})$ labeled examples, leads to an error $(\frac{3}{2} - \gamma)\eta + \epsilon$. (All proofs for this section are in Appendix D).

We start with the upper bounds. First, we analyze the case of using a generic agnostic robust learner, then we deduce the sample complexity of a specific instantiation of such algorithm.

Theorem 5.1 *For any hypothesis class* $\mathcal{H}$ *and adversary* $\mathcal{U}$*, Algorithm* GRASS $(2, \epsilon, \delta)$- PAC *learns* $\mathcal{H}$ *with respect to the robust loss function, in the agnostic robust case, with samples of size*

$$m_l = \mathcal{O}\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon^2}\log^2\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon^2} + \frac{\log\frac{1}{\delta}}{\epsilon^2}\right), \quad m_u = \Lambda_{\mathrm{AG}}\left(2, \frac{\epsilon}{2}, \frac{\delta}{2}, \mathcal{H}, \mathcal{U}, 2\eta + \frac{\epsilon}{2}\right),$$

where $\Lambda_{\mathrm{AG}}(\alpha, \epsilon, \delta, \mathcal{H}, \mathcal{U}, \eta)$ is the sample complexity of adversarially-robust agnostic supervised learning, such that η is error of the optimal hypothesis in $\mathcal{H}$, namely $\eta = \mathrm{R}_{\mathcal{U}}(\mathcal{H}; \mathcal{D})$.

By using the agnostic supervised robust learner suggested by [39], we have the following upper bound on the unlabeled sample size, $\Lambda_{\mathrm{AG}}(\alpha, \epsilon, \delta, \mathcal{H}, \mathcal{U}, \eta) = \tilde{\mathcal{O}}\left(\frac{\mathrm{VC}(\mathcal{H})\,\mathrm{VC}^*(\mathcal{H})}{\epsilon^2} + \frac{\log\frac{1}{\delta}}{\epsilon^2}\right)$.

We now present two negative results.

Theorem 5.2 *For any $\epsilon, \delta \in (0,1)$, the sample complexity of agnostic robust $(1, \epsilon, \delta)$-PAC learning for a class $\mathcal{H}$, with respect to perturbation function $\mathcal{U}$ is (even if $\mathcal{D}_{\mathcal{X}}$ is known),*

$$m_l = \Omega\left(\frac{\mathrm{RS}_{\mathcal{U}}(\mathcal{H})}{\epsilon^2} + \frac{1}{\epsilon^2}\log\frac{1}{\delta}\right)\ ,\ m_u = \infty.$$

Theorem 5.3 *For any $\gamma > 0$, there exists a hypothesis class $\mathcal{H}$ and adversary $\mathcal{U}$, such that the sample complexity for $(\frac{3}{2} - \gamma, \epsilon, \delta)$-PAC learn $\mathcal{H}$ is*

$$m_l = \Omega\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon^2} + \frac{1}{\epsilon^2}\log\frac{1}{\delta}\right)\ ,\ m_u = \infty.$$

6 Learning with the 0-1 loss assuming robust realizability

In this section we learn with respect to the 0-1 loss, under robust realizability assumption. A Distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ is robustly realizable by $\mathcal{H}$ given a perturbation function $\mathcal{U}$, if there is $h \in \mathcal{H}$ such that not only h classifies all points in $\mathcal{D}$ correctly, it also does so with respect to the robust loss function, that is, $\mathrm{R}_{\mathcal{U}}(\mathcal{H}; \mathcal{D}) = 0$. Note that our guarantees, only in this section, are with respect to the non-robust risk. The formal definition is in Appendix E. A simple example for this model is the following. Let $\mathcal{H}$ be linear separators on $\mathcal{X}$ the unit ball in $\mathbb{R}^d$, and $\mathcal{U}$ as ℓ_2 balls of radius γ, the robustly realizable distributions are separable with margin γ, where $\mathrm{VC}_{\mathcal{U}}(\mathcal{H}) = \frac{1}{\gamma^2}$ but $\mathrm{VC}(\mathcal{H}) = d + 1$ can be arbitrarily larger. Moreover, we have the following example. (All proofs are in appendix Appendix E.)

Proposition 6.1 *For any $m \in \mathbb{N}$, there exist a hypothesis class $\mathcal{H}_m$ and distribution $\mathcal{D}$, such that $\mathcal{D}$ is robustly realizable by $\mathcal{H}_m$, $\mathrm{VC}_{\mathcal{U}}(\mathcal{H}_m) = 1$, and $\mathrm{VC}(\mathcal{H}_m) = 2m$.*

Standard VC theory does not ensure learning in this case. In this section we explain how we can learn in such a scenario with a small sample complexity (scales linearly in $\mathrm{VC}_{\mathcal{U}}$). Moreover, we show that it cannot be achieved via proper learners.

Theorem 6.2 *The sample complexity for learning a hypothesis class $\mathcal{H}$ with respect to the 0-1 loss, for any distribution $\mathcal{D}$ that is robustly realizable by $\mathcal{H}$, namely $\mathrm{R}_{\mathcal{U}}(\mathcal{H}; D) = 0$,*

$$\mathcal{O}\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon}\log^2\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon} + \frac{\log\frac{1}{\delta}}{\epsilon}\right), \Omega\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon} + \frac{\log\frac{1}{\delta}}{\epsilon}\right).$$

This Theorem was an intermediate step in the proof of Theorem 4.2, and the sample complexity is the same as Theorem C.1, $\mathcal{O}\left(\Lambda_{\mathrm{RE}}(\epsilon, \delta, \mathcal{H})\right)$. We show that there exists a robust ERM that fails in this setting (Proposition E.2 in Appendix E). Then, we claim that every proper learner fails.

Theorem 6.3 *There exists $\mathcal{H}$ with $\mathrm{VC}_{\mathcal{U}}(\mathcal{H}) = 1$, such that for any proper learning rule $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{H}$, there exists a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ that is robustly realizable by $\mathcal{H}$, i.e., $\mathrm{R}_{\mathcal{U}}(\mathcal{H}; \mathcal{D}) = 0$, and it holds that $\mathrm{R}(\mathcal{A}(S); D) > \frac{1}{8}$ with probability at least $\frac{1}{7}$ over $S \sim \mathcal{D}^m$.*

References

[1] J.-B. Alayrac, J. Uesato, P.-S. Huang, A. Fawzi, R. Stanforth, and P. Kohli. Are labels required for improving adversarial robustness? *Advances in Neural Information Processing Systems*, 32, 2019.

[2] N. Alon, S. Hanneke, R. Holzman, and S. Moran. A theory of pac learnability of partial concept classes. *arXiv preprint arXiv:2107.08444*, 2021.

[3] H. Ashtiani, V. Pathak, and R. Urner. Black-box certification and learning under adversarial perturbations. In *International Conference on Machine Learning*, pages 388–398. PMLR, 2020.

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

[5] I. Attias, A. Kontorovich, and Y. Mansour. Improved generalization bounds for robust learning. In *Algorithmic Learning Theory*, pages 162–183. PMLR, 2019.

[6] P. Auer and R. Ortner. A new pac bound for intersection-closed concept classes. *Machine Learning*, 66(2):151–163, 2007.

[7] P. Awasthi, N. Frank, and M. Mohri. Adversarial learning guarantees for linear hypotheses and neural networks. In *International Conference on Machine Learning*, pages 431–441. PMLR, 2020.

[8] P. Awasthi, N. Frank, A. Mao, M. Mohri, and Y. Zhong. Calibration and consistency of adversarial surrogate losses. *Advances in Neural Information Processing Systems*, 34, 2021.

[9] P. Awasthi, N. Frank, and M. Mohri. On the existence of the adversarial bayes classifier. *Advances in Neural Information Processing Systems*, 34, 2021.

[10] M.-F. Balcan and A. Blum. 21 an augmented pac model for semi-supervised learning. 2006.

[11] M.-F. Balcan and A. Blum. A discriminative model for semi-supervised learning. *Journal of the ACM (JACM)*, 57(3):1–46, 2010.

[12] M.-F. Balcan and P. Long. Active and passive learning of linear separators under log-concave distributions. In *Conference on Learning Theory*, pages 288–316. PMLR, 2013.

[13] R. Bhattacharjee, S. Jha, and K. Chaudhuri. Sample complexity of robust linear classification on separated data. In *International Conference on Machine Learning*, pages 884–893. PMLR, 2021.

[14] A. Blum. Semi-supervised learning. In *Encyclopedia of Algorithms*, pages 1936–1941. 2016.

[15] N. H. Bshouty, Y. Li, and P. M. Long. Using the doubling dimension to analyze the generalization of learning algorithms. *Journal of Computer and System Sciences*, 75(6):323–335, 2009.

[16] Y. Carmon, A. Raghunathan, L. Schmidt, J. C. Duchi, and P. S. Liang. Unlabeled data improves adversarial robustness. *Advances in Neural Information Processing Systems*, 32, 2019.

[17] O. Chapelle, B. Scholkopf, and A. Zien. Semi-supervised learning (chapelle, o. et al., eds.; 2006)[book reviews]. *IEEE Transactions on Neural Networks*, 20(3):542–542, 2009.

[18] D. Cullina, A. N. Bhagoji, and P. Mittal. Pac-learning in the presence of adversaries. In *Advances in Neural Information Processing Systems*, pages 230–241, 2018.

[19] C. Dan, Y. Wei, and P. Ravikumar. Sharp statistical guaratees for adversarially robust gaussian classification. In *International Conference on Machine Learning*, pages 2345–2355. PMLR, 2020.

[20] M. Darnstädt. The optimal pac bound for intersection-closed concept classes. *Information Processing Letters*, 115(4):458–461, 2015.

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

[22] U. Feige, Y. Mansour, and R. Schapire. Learning and inference in the presence of corrupted inputs. In *Conference on Learning Theory*, pages 637–657, 2015.

[23] S. Floyd and M. Warmuth. Sample compression, learnability, and the vapnik-chervonenkis dimension. *Machine learning*, 21(3):269–304, 1995.

[24] E. Giné and V. Koltchinskii. Concentration inequalities and asymptotic results for ratio type empirical processes. *The Annals of Probability*, 34(3):1143–1216, 2006.

[25] C. Göpfert, S. Ben-David, O. Bousquet, S. Gelly, I. Tolstikhin, and R. Urner. When can unlabeled data improve the learning rate? In *Conference on Learning Theory*, pages 1500–1518. PMLR, 2019.

[26] 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.

[27] S. Hanneke. *Theoretical foundations of active learning*. Carnegie Mellon University, 2009.

[28] S. Hanneke. The optimal sample complexity of pac learning. *The Journal of Machine Learning Research*, 17(1):1319–1333, 2016.

[29] S. Hanneke. Refined error bounds for several learning algorithms. *The Journal of Machine Learning Research*, 17(1):4667–4721, 2016.

[30] S. Hanneke and A. Kontorovich. Stable sample compression schemes: New applications and an optimal svm margin bound. In *Algorithmic Learning Theory*, pages 697–721. PMLR, 2021.

[31] S. Hanneke, A. Kontorovich, and M. Sadigurschi. Sample compression for real-valued learners. In *Algorithmic Learning Theory*, pages 466–488, 2019.

[32] D. Haussler, N. Littlestone, and M. K. Warmuth. Predicting $\{0, 1\}$-functions on randomly drawn points. *Information and Computation*, 115(2):248–292, 1994.

[33] J. Khim and P.-L. Loh. Adversarial risk bounds via function transformation. *arXiv preprint arXiv:1810.09519*, 2018.

[34] M. Levi, I. Attias, and A. Kontorovich. Domain invariant adversarial learning. *arXiv preprint arXiv:2104.00322*, 2021.

[35] N. Littlestone and M. Warmuth. Relating data compression and learnability. 1986.

[36] P. M. Long. An upper bound on the sample complexity of pac-learning halfspaces with respect to the uniform distribution. *Information Processing Letters*, 87(5):229–234, 2003.

[37] A. Maurer and M. Pontil. Empirical bernstein bounds and sample variance penalization. *arXiv preprint arXiv:0907.3740*, 2009.

[38] M. Mohri, A. Rostamizadeh, and A. Talwalkar. *Foundations of machine learning*. MIT press, 2018.

[39] O. Montasser, S. Hanneke, and N. Srebro. Vc classes are adversarially robustly learnable, but only improperly. *arXiv preprint arXiv:1902.04217*, 2019.

[40] O. Montasser, S. Goel, I. Diakonikolas, and N. Srebro. Efficiently learning adversarially robust halfspaces with noise. In *International Conference on Machine Learning*, pages 7010–7021. PMLR, 2020.

[41] O. Montasser, S. Hanneke, and N. Srebro. Reducing adversarially robust learning to non-robust pac learning. *Advances in Neural Information Processing Systems*, 33:14626–14637, 2020.

[42] O. Montasser, S. Hanneke, and N. Srebro. Adversarially robust learning with unknown perturbation sets. In *Conference on Learning Theory*, pages 3452–3482. PMLR, 2021.

[43] O. Montasser, S. Hanneke, and N. Srebro. Transductive robust learning guarantees. *arXiv preprint arXiv:2110.10602*, 2021.

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

[45] A. Najafi, S.-i. Maeda, M. Koyama, and T. Miyato. Robustness to adversarial perturbations in learning from incomplete data. *Advances in Neural Information Processing Systems*, 32, 2019.

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

[47] R. E. Schapire and Y. Freund. Boosting: Foundations and algorithms. *Kybernetes*, 2013.

[48] L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, and A. Madry. Adversarially robust generalization requires more data. *Advances in neural information processing systems*, 31, 2018.

[49] S. Shalev-Shwartz and S. Ben-David. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

[50] J. Uesato, J.-B. Alayrac, P.-S. Huang, R. Stanforth, A. Fawzi, and P. Kohli. Are labels required for improving adversarial robustness? *arXiv preprint arXiv:1905.13725*, 2019.

[51] V. N. Vapnik and A. Y. Chervonenkis. On the uniform convergence of relative frequencies of events to their probabilities. In *Measures of complexity*, pages 11–30. Springer, 2015.

[52] M. K. Warmuth. The optimal pac algorithm. In *International Conference on Computational Learning Theory*, pages 641–642. Springer, 2004.

[53] C. Wei, K. Shen, Y. Chen, and T. Ma. Theoretical analysis of self-training with deep networks on unlabeled data. *arXiv preprint arXiv:2010.03622*, 2020.

[54] Y. Xing, R. Zhang, and G. Cheng. Adversarially robust estimate and risk analysis in linear regression. In *International Conference on Artificial Intelligence and Statistics*, pages 514–522. PMLR, 2021.

[55] D. Yin, R. Kannan, and P. Bartlett. Rademacher complexity for adversarially robust generalization. In *International Conference on Machine Learning*, pages 7085–7094. PMLR, 2019.

[56] R. Zhai, T. Cai, D. He, C. Dan, K. He, J. Hopcroft, and L. Wang. Adversarially robust generalization just requires more unlabeled data. *arXiv preprint arXiv:1906.00555*, 2019.

[57] X. Zhu and A. B. Goldberg. Introduction to semi-supervised learning. *Synthesis lectures on artificial intelligence and machine learning*, 3(1):1–130, 2009.

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]