A Characterization of Semi-Supervised Adversarially Robust PAC Learnability

Idan Attias
Department of Computer Science
Ben-Gurion University of the Negev
idanatti@post.bgu.ac.il

Steve Hanneke
Department of Computer Science
Purdue University
steve.hanneke@gmail.com

Yishay Mansour
Blavatnik School of Computer Science
Tel Aviv University and Google Research
mansour.yishay@gmail.com

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., 18, 58, 51, 46, 1, 55, 36]. 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

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

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 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 Montasser et al. [40] (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, for any hypothesis class $\mathcal{H}$, it holds that $\mathrm{VC}_{\mathcal{U}}(\mathcal{H}) \leq \mathrm{VC}(\mathcal{H})$, and the gap can be arbitrarily large. That is, there exist $\mathcal{H}_0$ such that $\mathrm{VC}_{\mathcal{U}}(\mathcal{H}_0) = 0$ and $\mathrm{VC}(\mathcal{H}_0) = \infty$ (see Proposition 3.2).

For an improved lower bound on the sample complexity, Montasser et al. [40, 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 $\tilde{\mathcal{O}}\left(\frac{\mathrm{VC}\cdot\mathrm{VC}^*}{\epsilon} + \frac{\log\frac{1}{\delta}}{\epsilon}\right)$[1] 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. [40] 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

[1] $\tilde{\mathcal{O}}(\cdot)$ stands for omitting poly-logarithmic factors of $\mathrm{VC}, \mathrm{VC}^*, \mathrm{VC}_{\mathcal{U}}, \mathrm{RS}_{\mathcal{U}}, 1/\epsilon, 1/\delta$.

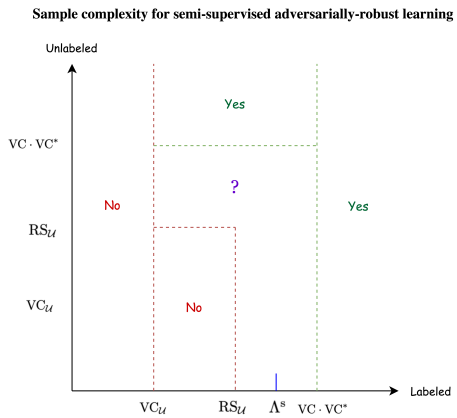

Figure 1: Summary of the sample complexity regimes for semi-supervised robust learning, for the realizable model and the agnostic model with error $3\eta + \epsilon$, where η is the minimal agnostic error in the hypothesis class.
Obtaining an error of $\eta+\epsilon$ requires at least $\mathrm{RS}_{\mathcal{U}}$ 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 $\mathrm{RS}_{\mathcal{U}}$.

the minimal agnostic error. If we allow an error of $3\eta + \epsilon$, 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 $\gamma > 0$. 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 work of Montasser et al. [40] studied the setting of fully-supervised robust PAC learning. In this paper, we propose a semi-supervised method with a significant improvement on the labeled sample size. We show that the labeled and unlabeled sample complexities are controlled by different complexity measures. Adversarially robust learning has been extensively studied in several supervised learning models [e.g., 25, 49, 34, 57, 20, 7, 35, 6, 10, 44, 41–43, 3, 11, 21, 9, 15, 56, 4]. For semi-supervised robust learning, Ashtiani et al. [3] showed that under some assumptions, robust PAC learning is possible with $\mathcal{O}(\mathrm{VC}(\mathcal{H}))$ labeled examples and additional unlabeled samples. Carmon et al. [18] studied a robust semi-supervised setting where the distribution is a mixture of Gaussians and the hypothesis class is linear separators.

Semi-supervised (non-robust) learning. There is substantial interest in semi-supervised (non-robust) learning, and many contemporary practical problems significantly benefit from it [e.g., 16, 19, 59]. This was formalized in theoretical frameworks. Urner et al. [52] suggested a semi-supervised learning (non-robust) framework, with an algorithmic idea that is similar to our method. Their framework consists of two steps; using labeled data to learn a classifier with small error (not necessarily a member of the target class $\mathcal{H}$), and then labeling an unlabeled input sample in order to use a fully-supervised proper learner. They investigate scenarios where saving of labeled examples occurs. In our paper, we are interested in the robust loss function. We use labeled data in order to learn a classifier (with the 0-1 loss function) from a class with a potentially smaller complexity measure, then we label an unlabeled input sample, and use a fully-supervised method using the robust loss function. The sample

complexity of learning the robust loss class is controlled by a larger complexity measure. Fortunately, this affects our unlabeled sample size and not the labeled sample size as in the fully-supervised setting. Göpfert et al. [27] studied circumstances where the learning rate can be improved given unlabeled data. Darnstädt et al. [23] showed that the label complexity gap between the semi-supervised and the fully supervised setting can become arbitrarily large for concept classes of infinite VC-dimension, and that this gap is bounded when a function class contains the constant zero and the constant one functions. Balcan and Blum [13, 12] introduced an augmented version of the PAC model designed for semi-supervised learning and analyzed when unlabeled data can help. The main idea is to augment the notion of learning a concept class, with a notion of compatibility between a function and the data distribution that we hope the target function will satisfy.

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 $\operatorname{supp}(\mathcal{D}) = \{(x, y) \in \mathcal{X} \times \mathcal{Y} : \mathcal{D}(x, y) > 0\}$. Denote the marginal distribution $\mathcal{D}_{\mathcal{X}}$ on $\mathcal{X}$ and its support by $\operatorname{supp}(\mathcal{D}_{\mathcal{X}}) = \{x \in \mathcal{X} : \mathcal{D}(x, y) > 0\}$. 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,

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

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 $S \in (\mathcal{X} \times \mathcal{Y})^*$,

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

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

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

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 $\mathrm{R}(h; \mathcal{D}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell_{0\text{-}1}(h; x, y)] = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\mathbb{I}[h(x) \neq y]]$, 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 $\widehat{\mathrm{R}}(h; S) = \frac{1}{|S|} \sum_{(x,y)\in S} \ell_{0\text{-}1}(h; x, y) = \frac{1}{|S|} \sum_{(x,y)\in S} [\mathbb{I}[h(x) \neq y]]$. 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 $\mathrm{ERM}_{\mathcal{H}}(S) \in \operatorname{argmin}_{h \in \mathcal{H}} \widehat{\mathrm{R}}(h; S)$.

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 $\mathrm{R}_{\mathcal{U}}(\mathcal{H}; D) = 0$, for a random sample $S \sim \mathcal{D}^m$, it holds that

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

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 $\mathrm{R}_{\mathcal{U}}(\mathcal{H}; D) = 0$, 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 $\mathrm{R}_{\mathcal{U}}(\mathcal{H}; \mathcal{D}) > 0$, 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 $\eta = \mathrm{R}_{\mathcal{U}}(\mathcal{H}; \mathcal{D})$. We say that a function f is (α, ϵ)-optimal if $\mathrm{R}_{\mathcal{U}}(f; D) \leq \alpha\eta + \epsilon$.

Definition 2.3 (Agnostic robust PAC **learnability)** For any $\epsilon, \delta \in (0, 1)$, the sample complexity of agnostic robust $(\alpha, \epsilon, \delta)$-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

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

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

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

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, $\mathrm{supp}(\mathcal{D}_{\mathcal{X}}) = \{x \in \mathcal{X} : \exists y \in \mathcal{Y},\ \text{s.t. } \mathcal{D}(x,y) > 0\}$, 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

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

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 $\mathcal{H}_{\mathcal{U}\text{-cons}}$. [2] 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. [40, Theorem 10]. For completeness, we prove the following in Appendix B.

Proposition 3.2 ([40], 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 prove 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.

[2] See Mohri et al. [39, 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 [30] 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., 8, 22, 31, 29, 37, 26, 17, 14].

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 by Alon et al. [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 Montasser et al. [40].) 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 any $h \in \mathcal{H}$, we define a corresponding partial concept $h^\star : \mathcal{X} \to \{0,1,\star\}$, and 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 on h, regardless of the prediction of h. 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 $\frac{\epsilon}{3}, \frac{\delta}{2}$. 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. (On points where h_1 predicts $\star$, we can arbitrarily choose a label of 0 or 1.)
4. Execute the agnostic adversarially robust supervised PAC learner $\mathcal{B}$ on S^u with parameters $\frac{\epsilon}{3}, \frac{\delta}{2}$. 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 guarantee 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 $\epsilon/3$ 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 at most $\frac{\epsilon}{3}$, i.e., $\mathrm{R}_{\mathcal{U}}(\mathcal{H}; \tilde{\mathcal{D}}) = \frac{\epsilon}{3}$. Indeed, the function with zero robust error on $\mathcal{D}$, $h_{\text{opt}} \in \operatorname{argmin}_{h \in \mathcal{H}} \mathrm{R}_{\mathcal{U}}(h; \mathcal{D})$ has a robust error of at most $\frac{\epsilon}{3}$ 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 $\frac{\epsilon}{3}$. Moreover, the total variation distance between $\mathcal{D}$ and $\tilde{\mathcal{D}}$ is at most $\frac{\epsilon}{3}$. We are guaranteed that the resulting hypothesis h_2 has a robust error of at most $\frac{\epsilon}{3} + \frac{\epsilon}{3} + \frac{\epsilon}{3} = \epsilon$ on $\mathcal{D}$. We conclude that a size of $|S^u_{\mathcal{X}}| = m_u = \Lambda_{\mathrm{AG}}\left(1, \frac{\epsilon}{3}, \frac{\delta}{2}, \mathcal{H}, \mathcal{U}, \eta = \frac{\epsilon}{3}\right)$ unlabeled samples suffices, this completes the proof for Theorem 4.2. For a specific instantiation of such algorithm ([40]), we deduce the sample complexity in Theorem 4.4. 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*

$$m_l = \mathcal{O}\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon} \log^2 \frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon} + \frac{\log \frac{1}{\delta}}{\epsilon}\right) \ , \ m_u = \Lambda_{\mathrm{AG}}\left(1, \frac{\epsilon}{3}, \frac{\delta}{2}, \mathcal{H}, \mathcal{U}, \eta = \frac{\epsilon}{3}\right),$$

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., $\eta = \mathrm{R}_{\mathcal{U}}(\mathcal{H}; \mathcal{D})$.

Remark 4.3 Note that if we simply invoke a PAC learner (for total concept classes) on $\mathcal{H}$, with the 0-1 loss, instead of steps 1 and 2 in the algorithm, we would get a labeled sample complexity of roughly $\mathcal{O}(\mathrm{VC}(\mathcal{H}))$. This is already an exponential improvement upon previous results that require roughly $\mathcal{O}\left(2^{\mathrm{VC}(\mathcal{H})}\right)$ labeled samples. The purpose of using partial concept classes is to further reduce the labeled sample complexity to $\mathcal{O}(\mathrm{VC}_{\mathcal{U}}(\mathcal{H}))$.

The following result follows by using the agnostic supervised robust learner suggested by Montasser et al. [40]. 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 4.4 *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. [40, Theorem 10].

Corollary 4.5 *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 [40, 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 4.6 *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 $\eta+\epsilon/3$ (η is the optimal robust error of a hypothesis in $\mathcal{H}$ and not 0). This leads eventually to error of $3\eta+\epsilon$ 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 $(3,\epsilon,\delta)$- PAC *learns $\mathcal{H}$ with respect to the robust loss function, in the agnostic robust case, with samples of size*

$$m_l=\mathcal{O}\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon^2}\log^2\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon^2}+\frac{\log\frac{1}{\delta}}{\epsilon^2}\right),\quad m_u=\Lambda_{\mathrm{AG}}\left(1,\frac{\epsilon}{3},\frac{\delta}{2},\mathcal{H},\mathcal{U},2\eta+\frac{\epsilon}{3}\right),$$

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 Montasser et al. [40], we have the following upper bound on the unlabeled sample size, $m_u=\tilde{\mathcal{O}}\left(\frac{\mathrm{VC}(\mathcal{H})\,\mathrm{VC}^*(\mathcal{H})}{\epsilon^2}+\frac{\log\frac{1}{\delta}}{\epsilon^2}\right)$.

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*

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

Open question. What is the optimal error rate in the agnostic setting when using only $\mathcal{O}(\mathrm{VC}_{\mathcal{U}})$ labeled examples?

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 $S \sim \mathcal{D}^m$.*

Acknowledgments

We are grateful to Omar Montasser for his helpful input, particularly inspiring steps 3 and 4 of the GRASS learning algorithm. We would like to thank Vinod Raman for his enlightening comments regarding the correctness of our algorithm. Finally, we thank the anonymous reviewers for their thoughtful comments, which helped us improve the presentation of our paper.

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation program (grant agreement No. 882396), by the Israel Science Foundation (grants 993/17, 1602/19), Tel Aviv University Center for AI and Data Science (TAD), and the Yandex Initiative for Machine Learning at Tel Aviv University. I.A. is supported by the Vatat Scholarship from the Israeli Council for Higher Education and by Kreitman school of Advanced Graduate Studies.

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] H. Ashtiani, V. Pathak, and R. Urner. Adversarially robust learning with tolerance. *arXiv preprint arXiv:2203.00849*, 2022.

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

[6] I. Attias and S. Hanneke. Adversarially robust learning of real-valued functions. *arXiv preprint arXiv:2206.12977*, 2022.

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

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

[9] 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.

[10] 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.

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

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

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

[14] 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.

[15] 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.

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

[17] 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.

[18] 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.

[19] 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.

[20] 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.

[21] 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.

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

[23] M. Darnstädt, H. U. Simon, and B. Szörényi. Unlabeled data does provably help. 2013.

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

[25] 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.

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

[27] 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.

[28] 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.

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

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

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

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

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

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

[35] A. Kontorovich and I. Attias. Fat-shattering dimension of k-fold maxima. *arXiv preprint arXiv:2110.04763*, 2021.

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

[37] 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.

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

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

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

[41] 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.

[42] 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.

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

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

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

[46] 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.

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

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

[49] 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.

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

[51] 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.

[52] R. Urner, S. Shalev-Shwartz, and S. Ben-David. Access to unlabeled data can speed up prediction time. In *ICML*, 2011.

[53] 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.

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

[55] 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.

[56] 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.

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

[58] 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.

[59] 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]

A Additional preliminaries for Section 2

Complexity measures. The capacity measures, $\mathrm{VC}_{\mathcal{U}}$, $\mathrm{RS}_{\mathcal{U}}$ and VC, play an important role in our results. See Definitions 1.1 and 1.2 for the $\mathrm{VC}_{\mathcal{U}}$ and $\mathrm{RS}_{\mathcal{U}}$ dimensions. It holds that $\mathrm{VC}_{\mathcal{U}}(\mathcal{H}) \leq \mathrm{RS}_{\mathcal{U}}(\mathcal{H}) \leq \mathrm{VC}(\mathcal{H})$, in Proposition 3.2 we demonstrate an arbitrary gap between $\mathrm{VC}_{\mathcal{U}}$ and $\mathrm{RS}_{\mathcal{U}}$, the key parameters controlling the sample complexity of robust learnability.

Denote the projection of a hypothesis class $\mathcal{H}$ on set $S = \{x_1, \ldots, x_k\}$ by $\mathcal{H}|_S = \{(h(x_1), \ldots, h(x_k)) : h \in \mathcal{H}\}$. We say that a set $S \subseteq \mathcal{X}$ is shattered by $\mathcal{H}$ if $\{0,1\}^S = \mathcal{H}|_S$, the VC-dimension [53] of $\mathcal{H}$ is defined as the maximal size of a shattered set S. The dual hypothesis class $\mathcal{H}^* \subseteq \{0,1\}^{\mathcal{H}}$ is defined as the set of all functions $f_x : \mathcal{H} \to \{0,1\}$ where $f_x(h) = h(x)$. We denote the VC-dimension of the dual class by $\mathrm{VC}^*(\mathcal{H})$. It is known that $\mathrm{VC}^*(\mathcal{H}) < 2^{\mathrm{VC}(\mathcal{H})+1}$ [5].

Definition A.1 (Sample compression scheme) A pair of functions (κ, ρ) is a sample compression scheme of size ℓ for class $\mathcal{H}$ if for any $n \in \mathbb{N}$, $h \in \mathcal{H}$ and sample $S = \{(x_i, h(x_i))\}_{i=1}^n$, it holds for the compression function that $\kappa(S) \subseteq S$ and $|\kappa(S)| \leq \ell$, and the reconstruction function $\rho(\kappa(S)) = \hat{h}$ satisfies $\hat{h}(x_i) = h(x_i)$ for any $i \in [n]$.

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. However, the uniform convergence argument does not hold, and the ERM principle does not ensure learning. The proof hinges on a combination of sample compression scheme and a variant of the *one-Inclusion-Graph* algorithm [33]. In Section 4 we elaborate on the sample complexity of partial concept classes, and in Appendix F we elaborate on the learning algorithms. The definitions of realizability and agnostic learning in the partial concepts sense generalizes the classic definitions for total concept classes. See [2, Section 2 and Appendix C] for more details.

B Proofs for Section 3

Proof of Proposition 3.2 We overview the construction by Montasser et al. [40], which exemplifies an arbitrarily large gap between $\mathrm{VC}_{\mathcal{U}}$ and $\mathrm{RS}_{\mathcal{U}}$. In this example $\mathrm{VC}_{\mathcal{U}}(\mathcal{H}) = 0$, $\mathrm{RS}_{\mathcal{U}}(\mathcal{H}) = \infty$, and $\mathrm{VC}(\mathcal{H}) = \infty$.

Define the Euclidean ball of radius r perturbation function $\mathcal{U}(x) = B_r(x)$. Consider infinite sequences $(x_n)_{n\in\mathbb{N}}$ and $(z_n)_{n\in\mathbb{N}}$ of points such that $\forall i \neq j,\ \mathcal{U}(x_i) \cap \mathcal{U}(x_j) = U(x_i) \cap \mathcal{U}(z_j) = \mathcal{U}(x_j) \cap \mathcal{U}(z_i) = \emptyset$, and $\forall i,\ \left|\mathcal{U}(x_i) \cap \mathcal{U}(z_i)\right| = 1$.

For a bit string $b \in \{0,1\}^{\mathbb{N}}$, define a hypothesis $h_b : \{\mathcal{U}(x_i) \cup \mathcal{U}(z_i)\}_{i\in\mathbb{N}} \to \{0,1\}$ as follows.

$$h_b = \begin{cases} h_b\Big(\mathcal{U}(x_i)\Big) = 1 \ \wedge\ h_b\Big(\mathcal{U}(z_i) \setminus \mathcal{U}(x_i)\Big) = -1, & b_i = 0 \\ h_b\Big(\mathcal{U}(z_i)\Big) = 1 \ \wedge\ h_b\Big(\mathcal{U}(x_i) \setminus \mathcal{U}(z_i)\Big) = -1, & b_i = 1. \end{cases}$$

Define the hypothesis class $\mathcal{H} = \Big\{h_b : b \in \{0,1\}^{\mathbb{N}}\Big\}$. It holds that $\mathrm{VC}_{\mathcal{U}}(\mathcal{H}) = 0$ and $\mathrm{RS}_{\mathcal{U}} = \infty$. ■

C Proofs for Section 4

Before proceeding to the proof, we present the following result on learning partial concept classes. Recall the definition of VC is in the context of partial concepts (see Appendix A).

Theorem C.1 ([2], Theorem 34) *Any partial concept class $\mathcal{H}$ with $\mathrm{VC}(\mathcal{H}) < \infty$ is* PAC *learnable in the realizable setting with sample complexity,*

- $\Lambda_{\mathrm{RE}}(\epsilon, \delta, \mathcal{H}) = \mathcal{O}\left(\min\left\{\frac{\mathrm{VC}(\mathcal{H})}{\epsilon}\log\frac{1}{\delta}, \frac{\mathrm{VC}(\mathcal{H})}{\epsilon}\log^2\left(\frac{\mathrm{VC}(\mathcal{H})}{\epsilon}\right) + \frac{1}{\epsilon}\log\frac{1}{\delta}\right\}\right)$
- $\Lambda_{\mathrm{RE}}(\epsilon, \delta, \mathcal{H}) = \Omega\left(\frac{\mathrm{VC}(\mathcal{H})}{\epsilon} + \frac{1}{\epsilon}\log\frac{1}{\delta}\right)$.

Proof of Theorem 4.2 At first, we convert the hypothesis class $\mathcal{H}$ to $\mathcal{H}^\star_{\mathcal{U}}$ as described in Definition 4.1. Then, we employ the learning algorithm $\mathcal{A}$ for partial concepts on the partial concept class $\mathcal{H}^\star_{\mathcal{U}}$ and S^l, denote the resulting hypothesis by h_1. Note that we reduced the complexity of the class, since $\mathrm{VC}(\mathcal{H}^\star_{\mathcal{U}}) = \mathrm{VC}_{\mathcal{U}}(\mathcal{H})$. Theorem C.1 implies that whenever $m_l = |S^l| \geq \tilde{\mathcal{O}}\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon} + \frac{1}{\epsilon}\log\frac{1}{\delta}\right)$, the hypothesis h_1 has a <u>non-robust</u> error at most $\frac{\epsilon}{3}$ with probability $1 - \frac{\delta}{2}$, with respect to the <u>0-1 loss</u>. Note that there exists $h \in \mathcal{H}$ that classifies correctly any point in $\mathcal{D}$ with respect to the robust loss function. So when we convert $\mathcal{H}$ to $\mathcal{H}^\star_{\mathcal{U}}$, the "partial version" of h still classifies correctly any point in S^l, and does not return any $\star$, which always counts as a mistake. Algorithm $\mathcal{A}$ guarantees to return a hypothesis that is ϵ-optimal with respect to the 0-1 loss, with high probability. Observe that after these two steps, we obtain the following intermediate result. Whenever a distribution $\mathcal{D}$ is robustly realizable by a hypothesis class $\mathcal{H}$, i.e., $\mathrm{R}_{\mathcal{U}}(\mathcal{H}; \mathcal{D}) = 0$, we have an algorithm that learns this class with respect to the <u>0-1 loss</u>, with sample complexity of

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

The sample complexity of this model is defined formally in Definition E.1. in Section 6 present more results for this model.

In the third step, we label an independent unlabeled sample $S^u_{\mathcal{X}} \sim \mathcal{D}^{m_u}_{\mathcal{X}}$ with h_1, denote this 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 at most $\frac{\epsilon}{3}$, i.e., $\mathrm{R}_{\mathcal{U}}(\mathcal{H}; \tilde{\mathcal{D}}) \leq \frac{\epsilon}{3}$. Indeed, we show that $h_{\text{opt}} \in \operatorname{argmin}_{h \in \mathcal{H}} \mathrm{R}_{\mathcal{U}}(h; \mathcal{D})$ has a robust error of at most $\frac{\epsilon}{3}$ on $\tilde{\mathcal{D}}$. Note that,

$$\mathrm{R}_{\mathcal{U}}(\mathcal{H}; \tilde{\mathcal{D}}) \leq \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell_{\mathcal{U}}(h_{\text{opt}}; x, h_1(x))\right] = \mathbb{E}_{(x,y)\sim\tilde{\mathcal{D}}}\left[\ell_{\mathcal{U}}(h_{\text{opt}}; x, y)\right]. \tag{2}$$

Observe that the following holds for any (x, y),

$$\ell_{\mathcal{U}}(h_{\text{opt}}; x, h_1(x)) \leq \ell_{\mathcal{U}}(h_{\text{opt}}; x, y) + \ell_{\text{0-1}}(h_1; x, y). \tag{3}$$

Indeed, the right hand side is 0, whenever h_1 classifies (x, y) correctly, and h_{opt} robustly classifies (x, y) correctly, which implies that the left hand side is 0 as well.

By taking the expectation on Eq. (3) we have,

$$\mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell_{\mathcal{U}}(h_{\text{opt}}; x, h_1(x))] \leq \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell_{\mathcal{U}}(h_{\text{opt}}; x, y)] + \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell_{\text{0-1}}(h_1; x, y)]. \tag{4}$$

Combining it together, we obtain

$$\begin{aligned}
\mathrm{R}_{\mathcal{U}}(\mathcal{H}; \tilde{\mathcal{D}}) &\leq \mathbb{E}_{(x,y)\sim\tilde{\mathcal{D}}}\left[\ell_{\mathcal{U}}(h_{\text{opt}}; x, y)\right] \\
&\overset{\text{(i)}}{=} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell_{\mathcal{U}}(h_{\text{opt}}; x, h_1(x))] \\
&\overset{\text{(ii)}}{\leq} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell_{\mathcal{U}}(h_{\text{opt}}; x, y)] + \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell_{\text{0-1}}(h_1; x, y)] \\
&\leq \frac{\epsilon}{3}
\end{aligned}$$

where (i) follows from Eq. (2) and (ii) follows from Eq. (4).

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 $\frac{\epsilon}{3}$. We are guaranteed that the resulting hypothesis h_2 has a robust error of at most $\frac{\epsilon}{3} + \frac{\epsilon}{3} = \frac{2\epsilon}{3}$ on $\tilde{\mathcal{D}}$, with probability $1 - \frac{\delta}{2}$. We observe that the total variation distance between $\mathcal{D}$ and $\tilde{\mathcal{D}}$ is at most $\frac{\epsilon}{3}$, and as a result, h_2 has a robust error of at most $\frac{2\epsilon}{3} + \frac{\epsilon}{3} = \epsilon$ on $\mathcal{D}$, with probability $1 - \delta$.

We conclude that a size of $|S^u_{\mathcal{X}}| = m_u = \Lambda_{\mathrm{AG}}\left(1, \frac{\epsilon}{3}, \frac{\delta}{2}, \mathcal{H}, \mathcal{U}, \eta = \frac{\epsilon}{3}\right)$ unlabeled samples suffices, in addition to $m_l = \tilde{\mathcal{O}}\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon} + \frac{1}{\epsilon}\log\frac{1}{\delta}\right)$ labeled samples which are required in the first 2 steps. ■

We now prove Theorem 4.4. The following data-dependent compression based generalization bound is a variation of the classic bound by Graepel et al. [28]. It follows the same arguments while using the empirical Bernstein bound instead of Hoeffding's inequality. A variation of this bound, with respect to the 0-1 loss, appears in [2, Lemma 42], and [38, Section 5]. The exact same arguments follows for the robust loss as well.

This bound includes the empirical error factor, and as soon as we call the compression based learner on a sample that is "nearly" realizable (Step 4 in the algorithm), we can improve the sample complexity of the agnostic robust supervised learner, such that the dependence on ϵ^2 is reduced to ϵ, for the unlabeled sample size.

Lemma C.2 (Agnostic sample compression generalization bound) *For any sample compression scheme (κ, ρ), for any $m \in \mathbb{N}$ and $\delta \in (0,1)$, for any distribution $\mathcal{D}$ over $\mathcal{X} \times \{0,1\}$, for $S \sim \mathcal{D}^m$, with probability $1-\delta$,*

$$\left|\mathrm{R}_{\mathcal{U}}\left(\rho(\kappa(S)); \mathcal{D}\right) - \widehat{\mathrm{R}}_{\mathcal{U}}\left(\rho(\kappa(S)); S\right)\right| \leq \mathcal{O}\left(\sqrt{\widehat{\mathrm{R}}_{\mathcal{U}}\left(\rho(\kappa(S)); S\right)\frac{\left(|\kappa(S)|\log(m) + \log\frac{1}{\delta}\right)}{m}} + \frac{|\kappa(S)|\log(m) + \log\frac{1}{\delta}}{m}\right).$$

Proof of Theorem 4.4 Montasser et al. [40, Theorem 6] introduced an agnostic robust supervised learner that requires the following labeled sample size,

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

Their argument for generalization is based on classic compression generalization bound by Graepel et al. [28], adapted to the robust loss. See Montasser et al. [40, Lemma 11].

We show that in our use case we can deduce a stronger bound. We employ the agnostic learner on a distribution which is "close" to realizable, the error of the optimal $h \in \mathcal{H}$ is at most $\eta = \frac{\epsilon}{3}$, and so we need $\Lambda_{\mathrm{AG}}\left(1, \frac{\epsilon}{3}, \frac{\delta}{2}, \mathcal{H}, \mathcal{U}, \eta = \frac{\epsilon}{3}\right)$ unlabeled examples. As a result, we obtain an improved bound by using a data-dependant generalization bound described in Lemma C.2.

This improves the unlabeled sample size (denoted by m_u), and reduces its dependence on ϵ^2 to ϵ. Overall we obtain a sample complexity of

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

■

Proof of Theorem 4.6 This proof is identical to [40, Lemma 3], We overview the idea of the proof. If the proof is true for a labeled sample, it remains true when some of the labels are missing.

Define the following hypothesis class $\mathcal{H}_m \subseteq [0,1]^{\mathcal{X}}$. Define the instance space $\mathcal{X} = \{x_1, \ldots, x_m\} \subseteq \mathbb{R}$ and a perturbation function $\mathcal{U} : \mathcal{X} \to 2^{\mathcal{X}}$, such that the perturbation sets of the instances do not intersect, that is, $\forall i, j \in [m] : \mathcal{U}(x_i) \cap \mathcal{U}(x_j)$. We can simply take the perturbations sets to be ℓ_2 unit balls, $\mathcal{U}(x) = \{z \in \mathbb{R} : \|z - x\|_2 \leq 1\}$ such that $\forall i, j \in [m] : \|x_i - x_j\|_2 > 2$. Now, each

$h_b \in \mathcal{H}_m$ is represented by a bit string $b = \{0,1\}^m$, such that if $b_i = 1$, then there exist an adversarial example in $\mathcal{U}(x_i)$ that is unique for each h_b, and otherwise, the function is consistent on $\mathcal{U}(x_i)$.

Formally, for each $i \in [m]$ define a bijection $\psi_i : x_i \times \mathcal{H}_m \to \mathcal{U}(x_i) \setminus \{x_i\}$. Define $\mathcal{H}_m = \{h_b : b \in \{0,1\}^m\}$, such that for any $x_i \in \mathcal{X}$, h_b is defined by

$$h_b(x_i) = \begin{cases} h_b\Big(\mathcal{U}(x_i) \setminus \psi_i(x_i, h_b)\Big) = 0 \ \wedge \ h_b\Big(\psi_i(x_i, h_b)\Big) = 1, & b_i = 1, \\ h_b\Big(\mathcal{U}(x_i)\Big) = 0, & b_i = 0. \end{cases}$$

Note that since ψ_i is a bijection, and different functions with $b_i = 1$ have a different perturbation for x_i that causes a misclassification.

For a function class $\mathcal{H}$, define the robust loss class $\mathcal{L}^{\mathcal{U}}_{\mathcal{H}} = \Big\{(x,y) \mapsto \sup_{z \in \mathcal{U}(x)} \mathbb{I}\{h(z) \neq y\} : h \in \mathcal{H}\Big\}$. It holds that $\mathrm{VC}(\mathcal{H}_m) \leq 1$ and $\mathrm{VC}\left(\mathcal{L}^{\mathcal{U}}_{\mathcal{H}_m}\right) = m$ (see [40, Lemma 2]).

We define a function class $\tilde{\mathcal{H}}_{3m} = \Big\{h_b \in \mathcal{H}_{3m} : \sum_{i=1}^{3m} b_i = m\Big\}$. In words, we are keeping only functions in $\mathcal{H}_{3m}$ that are robustly correct on exactly $2m$ points. Note that the function $h_{\vec{0}}$ (bit string of all zeros) which is robustly correct on all $3m$ points, is not the class.

The idea is that we can construct a family of $\binom{3m}{2m}$ distributions, such that each distribution is supported on $2m$ points from $\mathcal{X} = \{x_1, \ldots, x_{3m}\}$. Now, if we have a proper learning rule, observing only m points, the algorithm has no information which are the remaining m points in the support (out of $2m$ possible points in $\mathcal{X}$). For each such a distribution there exists $h \in \tilde{\mathcal{H}}_{3m}$, with zero robust error. We can follow a standard proof of the no-free-lunch theorem [e.g., 50, Section 5], showing via the probabilistic method, that there exists a distribution on which the algorithm has constant error, although there is an optimal function in $\tilde{\mathcal{H}}_{3m}$. See [40, Lemma 3] for the full proof. ■

D Proofs for Section 5

Before proceeding to the proof, we present the following result on agnostically learning partial concept classes. Recall the definition of VC is in the context of partial concepts (see Appendix A).

Theorem D.1 ([2], Theorem 41) *Any partial concept class $\mathcal{H}$ with $\mathrm{VC}(\mathcal{H}) < \infty$ is agnostically* PAC *learnable with sample complexity,*

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

Proof of Theorem 5.1 We follow the same steps as in the proof of the realizable case, with the following difference. In the first two steps of the algorithm we learn with respect to the 0-1 loss, with error of η (the optimal robust error of a hypothesis in $\mathcal{H}$) and not 0, which leads eventually to approximation of 3η for learning with the robust loss.

At first, we convert the class $\mathcal{H}$ into $\mathcal{H}^{\star}_{\mathcal{U}}$, on which we employ the learning algorithm $\mathcal{A}$ for partial concepts with with the sample S^l. Theorem D.1 implies that whenever $m_l = |S^l| \geq \tilde{\mathcal{O}}\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon^2} + \frac{1}{\epsilon^2}\log\frac{1}{\delta}\right)$, the resulting hypothesis h_1 returned by algorithm $\mathcal{A}$ has a <u>non-robust</u> error at most $\eta + \frac{\epsilon}{3}$ with probability $1 - \frac{\delta}{2}$, with respect to the <u>0-1 loss</u>, where $\eta = \mathrm{R}_{\mathcal{U}}(\mathcal{H}; \mathcal{D})$. Note that there exists $h \in \mathcal{H}$ with robust error of η on $\mathcal{D}$. The "partial version" of h has an error of η on $\mathcal{D}$ with respect to the 0-1 loss. As a result, algorithm $\mathcal{A}$ guarantees to return a hypothesis that is ϵ-optimal with respect to the 0-1 loss, with high probability.

We label an independent unlabeled sample $S^u_{\mathcal{X}} \sim \mathcal{D}^{m_u}_{\mathcal{X}}$ with h_1, denote this labeled sample by S^u. Similarly to the realizable case, 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 at most $2\eta + \frac{\epsilon}{3}$, i.e., $\mathrm{R}_{\mathcal{U}}(\mathcal{H}; \tilde{\mathcal{D}}) = 2\eta + \frac{\epsilon}{3}$, by showing that $h_{\text{opt}} = \operatorname{argmin}_{h \in \mathcal{H}} \mathrm{R}_{\mathcal{U}}(h; \mathcal{D})$ has a robust error of at most $2\eta + \frac{\epsilon}{3}$ on $\tilde{\mathcal{D}}$.

Eqs. (2) to (4) still hold as in the realizable case proof. Combining it together, we have

$$\begin{aligned}
\mathrm{R}_{\mathcal{U}}(\mathcal{H}; \tilde{\mathcal{D}}) &\leq \mathbb{E}_{(x,y)\sim\tilde{\mathcal{D}}}\left[\ell_{\mathcal{U}}(h_{\text{opt}}; x, y)\right] \\
&\overset{\text{(i)}}{=} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell_{\mathcal{U}}(h_{\text{opt}}; x, h_1(x))] \\
&\overset{\text{(ii)}}{\leq} \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell_{\mathcal{U}}(h_{\text{opt}}; x, y)] + \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell_{\text{0-1}}(h_1; x, y)] \\
&\leq \eta + \eta + \frac{\epsilon}{3} \\
&= 2\eta + \frac{\epsilon}{3},
\end{aligned}$$

where (i) follows from Eq. (2) and (ii) follows from Eq. (4).

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 $2\eta + \frac{\epsilon}{3}$. We are guaranteed that the resulting hypothesis h_2 has a robust error of at most $2\eta + \frac{\epsilon}{3} + \frac{\epsilon}{3} = 2\eta + \frac{2\epsilon}{3}$ on $\tilde{\mathcal{D}}$, with probability $1 - \frac{\delta}{2}$. We observe that the total variation distance between $\mathcal{D}$ and $\tilde{\mathcal{D}}$ is at most $\eta + \frac{\epsilon}{3}$, and as a result, h_2 has a robust error of at most $2\eta + \frac{2\epsilon}{3} + \eta + \frac{\epsilon}{3} = 3\eta + \epsilon$ on $\mathcal{D}$, with probability $1 - \delta$.

We conclude that a size of $|S^u_{\mathcal{X}}| = m_u = \Lambda_{\text{AG}}\left(1, \frac{\epsilon}{3}, \frac{\delta}{2}, \mathcal{H}, \mathcal{U}, 2\eta + \frac{\epsilon}{3}\right)$ unlabeled sample suffices, in addition to the $m_l = \mathcal{O}\left(\frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon^2} \log^2 \frac{\mathrm{VC}_{\mathcal{U}}(\mathcal{H})}{\epsilon^2} + \frac{\log \frac{1}{\delta}}{\epsilon^2}\right)$ labeled samples which are required in the first 2 steps. We remark that the best known value of $\Lambda_{\text{AG}}(1, \epsilon, \delta, \mathcal{H}, \mathcal{U}, \eta)$ is $\tilde{\mathcal{O}}\left(\frac{\mathrm{VC}(\mathcal{H})\,\mathrm{VC}^*(\mathcal{H})}{\epsilon^2} + \frac{\log \frac{1}{\delta}}{\epsilon^2}\right)$.

■

Proof of Theorem 5.2 We give a proof sketch, this is similar to [40, Theorem 10], knowing the marginal distribution $\mathcal{D}_{\mathcal{X}}$ does not give more power to the learner. The argument is based on the standard lower bound for VC classes (for example [39, Section 3]). Let $S = \{x_1, \ldots, x_k\}$ be a maximal set that is $\mathcal{U}$-robustly shattered by $\mathcal{H}$.

Let $z_1^+, z_1^-, \ldots, z_k^+, z_k^-$ be as in Definition 1.2, and note that for $i \neq j$, $z_i^+ \neq z_j^+$ and $z_i^- \neq z_j^-$. Define a distribution $\mathcal{D}_{\boldsymbol{\sigma}}$ for any possible labeling $\boldsymbol{\sigma} = (\sigma_1, \ldots, \sigma_k) \in \{0,1\}^k$ of S.

$$\forall j \in [k] : \begin{cases} \mathcal{D}_{\boldsymbol{\sigma}}(z_j^+, 1) = \frac{1-\alpha}{2k} \;\wedge\; \mathcal{D}_{\boldsymbol{\sigma}}(z_j^-, 0) = \frac{1+\alpha}{2k} & \sigma_j = 0, \\ \mathcal{D}_{\boldsymbol{\sigma}}(z_j^+, 1) = \frac{1+\alpha}{2k} \;\wedge\; \mathcal{D}_{\boldsymbol{\sigma}}(z_j^-, 0) = \frac{1-\alpha}{2k} & \sigma_j = 1. \end{cases}$$

We can now choose α as a function of ϵ, δ in order to get a lower bound on the sample complexity $|S| \gtrsim \frac{\mathrm{RS}_{\mathcal{U}}}{\epsilon^2}$.

■

Proof of Theorem 5.3 We take the construction in Proposition 3.2, where there is an arbitrary gap between $\mathrm{VC}_{\mathcal{U}}$ and $\mathrm{RS}_{\mathcal{U}}$.

Recall that on every pair (x, z) in Proposition 3.2 the optimal error is $\eta = 1/2$. On such unlabeled pairs, the learner can only randomly choose a prediction, and the error is $3/4$. We have $\mathrm{VC}_{\mathcal{U}} = 0$, and the labeled sample size is $\frac{1}{\epsilon^2}\log\frac{1}{\delta}$. As $(\mathrm{RS}_{\mathcal{U}} - \frac{1}{\epsilon^2}\log\frac{1}{\delta})$ grows, the gap between the learner and the optimal classifier is approaching $3/2$, which means that for any $\gamma > 0$ we can pick $\mathrm{RS}_{\mathcal{U}}$ such that error of $(\frac{2}{3} - \gamma)\eta$ is not possible.

In order to prove the case of any $0 < \eta \leq 1/2$, we can just add points such that their perturbation set does not intersect with any other perturbation set, and follow the same argument. ■

E Auxiliary definitions and proofs for Section 6

Definition of the model.

Definition E.1 ((non-robust) PAC learnability for robustly realizable distributions) For any $\epsilon, \delta \in (0, 1)$, the sample complexity of (ϵ, δ)-PAC learning for a class $\mathcal{H}$, denoted by $\Upsilon(\epsilon, \delta, \mathcal{H}, \mathcal{U})$, 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 $\mathcal{H}$ with respect to a perturbation function $\mathcal{U} : \mathcal{X} \to 2^{\mathcal{X}}$, namely $\mathrm{R}_{\mathcal{U}}(\mathcal{H}; D) = 0$, for a random sample $S \sim \mathcal{D}^m$, it holds that

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

If no such m exists, define $\Upsilon(\epsilon, \delta, \mathcal{H}, \mathcal{U}) = \infty$, and $\mathcal{H}$ is not (ϵ, δ)-PAC for distributions that are robustly realizable by $\mathcal{H}$ with respect to $\mathcal{U}$.

Proof of Proposition 6.1 Define the uniform distribution $\mathcal{D}$ over the support $\{(x_1, 1), \ldots, (x_{2m}, 1)\}$, such that $\bigcap_{i=1}^{2m} \mathcal{U}(x_i) \neq \emptyset$. Define $\mathcal{H} : \mathcal{X} \to 2^{\mathcal{X}}$ to be all binary functions over $\mathcal{X}$. Note that the $\mathcal{D}$ is robustly realizable by $\mathcal{H}$, the constant function that return always 1 has no error. Moreover we have $\mathrm{VC}_{\mathcal{U}} = 1$, and $\mathrm{VC} = 2m$, for any $m \in \mathbb{N}$. ■

Proof of Theorem 6.2 We follow only the first two steps of the generic Algorithm 1. Namely, take a labeled sample S and a hypothesis class $\mathcal{H}$ and create the partial hypothesis class $\mathcal{H}^\star_{\mathcal{U}}$. Assuming that the distribution is robustly realizable by $\mathcal{H}$, we end up in a realizable setting of learning a partial concept class $\mathcal{H}^\star_{\mathcal{U}}$.

In the second step of the algorithm we call a learning algorithm for partial concept classes (Appendix F) in order to do so. The sample complexity is the same as Theorem C.1, $\Upsilon(\epsilon, \delta, \mathcal{H}, \mathcal{U}) = \mathcal{O}\left(\Lambda_{\mathrm{RE}}(\epsilon, \delta, \mathcal{H})\right)$. Has we have shown in the proof of Theorem 4.2, Eq. (1), this implies the Theorem. ■

Proposition E.2 *Consider the distribution $\mathcal{D}$ and the hypothesis class $\mathcal{H}$ in Proposition 6.1. There exists a robust ERM algorithm returning a hypothesis $h_{ERM} \in \mathcal{H}$, such that $\mathrm{R}\left(h_{ERM}; \mathcal{D}\right) \geq \frac{1}{4}$ with probability 1 over $S \sim \mathcal{D}^m$.*

Proof Consider the following robust ERM. For any sample of size m, return 1 on the sample points and randomly choose a label for out of sample points. The error rate of such a robust ERM is at least $1/4$ with probability 1. ■

Proof of Theorem 6.3 This follows from a similar no-free-lunch argument for VC classes [e.g., 50, Section 5]. We briefly explain the proof idea.

Take the distribution $\mathcal{D}$, and the class $\mathcal{H}$ from Proposition E.2 with $\mathrm{VC}_{\mathcal{U}}(\mathcal{H}) = 1$ and $\mathrm{VC}(\mathcal{H}) = 3m$. Keep functions that are robustly self consistent only on $2m$ points. Construct all of distributions

on $2m$ points from the support of $\mathcal{D}$. We have $\binom{3m}{2m}$ such distributions, and on each one of them is robustly realizable by different $h \in \mathcal{H}$. The idea is the that a proper leaner observing only m points should guess which are the remaining m points the support of the distribution. There rest of the proof follows from the no-free-lunch proof. It can be shown formally via the probabilistic method, that for every proper rule, there exist a distribution on which the error is constant with fixed probability. ■

F Learning algorithms for partial concept classes

Here we overview the algorithmic techniques from Alon et al. [2, Theorem 34 and 41], for learning partial concepts in the realizable and agnostic settings. We use these algorithms in step 2 of our Algorithm 1.

One-inclusion graph algorithm for partial concept classes. We briefly discuss the algorithm, for the full picture, see [54, 33]. The one-inclusion algorithm for a class $\mathcal{F} \subseteq \{0, 1, \star\}^{\mathcal{X}}$ gets an input of unlabeled examples $S = (x_1, \ldots, x_m)$ and labels $(y_1, \ldots, y_{i-1}, y_{i+1}, \ldots, y_m)$ that are consistent with some $f \in \mathcal{F}$, that is, $f(x_k) = y_k$ for all $k \neq i$. It guarantees an (ϵ, δ)- PAC learner in the realizable setting, with sample complexity of $\Lambda_{\mathrm{RE}}(\epsilon, \delta, \mathcal{H}) = \mathcal{O}\left(\frac{\mathrm{VC}(\mathcal{H})}{\epsilon} \log \frac{1}{\delta}\right)$ as mentioned in Theorem C.1.

Here is a description of the algorithm. At first, construct the one-inclusion graph. For any $j \in [m]$ and $f \in \mathcal{F}|_S$ define $E_{j,f} = \{f' \in \mathcal{F}|_S : f'(x_k) = f(x_k), \forall k \neq j\}$, that is, all functions in $\mathcal{F}|_S$ that are consistent with f on S, except the point x_j. Define the set of edges $E = \{E_{j,f} : j \in [m], f \in \mathcal{F}|_S\}$, and the set vertices $V = \mathcal{F}|_S$ of the one-inclusion graph $G = (V, E)$. An orientation function $\psi : E \to V$ for an undirected graph G is an assignment of a direction to each edge, turning G into a directed graph. Find an orientation ψ that minimizes the out-degree of G. For prediction of x_i, pick $f \in V$ such that $f(x_k) = y_k$ for all $k \neq i$, and output $\psi(E_{i,f})(x_i)$.

Note that this algorithm is transductive, in a sense that in order to predict the label of a test point, it uses the entire training sample to computes its prediction.

Boosting and compression schemes. Recall the well known boosting algorithm, α-Boost [48, pages 162-163], which is a simplified version of AdaBoost, where the returned function is a simple majority over weak learners, instead of a weighted majority. For a hypothesis class $\mathcal{H}$ and a sample of size m, the algorithm yields a compression scheme of size $\mathcal{O}(\mathrm{VC}(\mathcal{H}) \log(m))$. Recall the following generalization bound based on sample compression scheme.

Lemma F.1 ([28]) *Let a sample compression scheme (κ, ρ), and a loss function $\ell : \mathbb{R} \times \mathbb{R} \to [0, 1]$. In the agnostic case, for any $\kappa(S) \lesssim m$, any $\delta \in (0, 1)$, and any distribution $\mathcal{D}$ over $\mathcal{X} \times \{0, 1\}$, for $S \sim \mathcal{D}^m$, with probability $1 - \delta$,*

$$\left|\mathrm{R}\left(\rho(\kappa(S)); \mathcal{D}\right) - \widehat{\mathrm{R}}\left(\rho(\kappa(S)); S\right)\right| \leq \mathcal{O}\left(\sqrt{\frac{\left(|\kappa(S)| \log(m) + \log \frac{1}{\delta}\right)}{m}}\right).$$

The learning algorithm for the realizable setting is α-Boost, where the weak learners are taken from the one-inclusion graph algorithm. As mentioned in Theorem C.1, this obtains an upper bound of $\Lambda_{\mathrm{RE}}(\epsilon, \delta, \mathcal{H}) = \mathcal{O}\left(\frac{\mathrm{VC}(\mathcal{H})}{\epsilon} \log^2\left(\frac{\mathrm{VC}(\mathcal{H})}{\epsilon}\right) + \frac{1}{\epsilon} \log \frac{1}{\delta}\right)$.

For the agnostic setting, follow a reduction to the realizable case suggested by David et al. [24]. The reduction requires a construction of a compression scheme based on Boosting algorithm. Roughly speaking, the reductions works as follows. Denote $\Lambda_{\mathrm{RE}} = \Lambda_{\mathrm{RE}}(1/3, 1/3, \mathcal{H})$, the sample complexity of $(1/3, 1/3)$-PAC learn $\mathcal{H}$, in the realizable case. Now, Λ_{RE} samples suffice for weak learning for any distribution $\mathcal{D}$ on a given sample S.

Find the maximal subset $S' \subseteq S$ such that $\inf_{h \in \mathcal{H}} \widehat{\mathrm{R}}(h; S') = 0$. Now, Λ_{RE} samples suffice for weak robust learning for any distribution $\mathcal{D}$ on S'. Execute the α-boost algorithm on S', with parameters

$\alpha = \frac{1}{3}$ and number of boosting rounds $T = \mathcal{O}\left(\log\left(|S'|\right)\right)$, where each weak learner is trained on Λ_{RE} samples. The returned hypothesis $\bar{h} = \mathrm{Majority}\left(\hat{h}_1, \ldots, \hat{h}_T\right)$ satisfies that $\widehat{\mathrm{R}}\left(\bar{h}; S'\right) = 0$, and each hypothesis $\hat{h}_t \in \left\{\hat{h}_1, \ldots, \hat{h}_T\right\}$ is representable as set of size $\mathcal{O}(\Lambda_{\mathrm{RE}})$. This defines a compression scheme of size $\Lambda_{\mathrm{RE}} T$, and $\bar{h}$ can be reconstructed from a compression set of points from S of size $\Lambda_{\mathrm{RE}} T$.

Recall that $S' \subseteq S$ is a maximal subset such that $\inf_{h \in \mathcal{H}} \widehat{\mathrm{R}}\left(h; S'\right) = 0$ which implies that $\widehat{\mathrm{R}}\left(\bar{h}; S\right) \leq \inf_{h \in \mathcal{H}} \widehat{\mathrm{R}}\left(h; S\right)$. Plugging it into a data-dependent compression generalization bound (Lemma C.2), we obtain a sample complexity of $\Lambda_{\mathrm{AG}}\left(\epsilon, \delta, \mathcal{H}\right) = \mathcal{O}\left(\frac{\mathrm{VC}(\mathcal{H})}{\epsilon^2} \log^2\left(\frac{\mathrm{VC}(\mathcal{H})}{\epsilon^2}\right) + \frac{1}{\epsilon^2} \log \frac{1}{\delta}\right)$, as mentioned in Theorem D.1.

G Supervised robust learning algorithms

We overview the algorithms of Montasser et al. [40, proofs of Theorems 4 and 8]. Their construction is based on sample compression methods explored in [32, 45].

Let $\mathcal{H} \subseteq \{0,1\}^{\mathcal{X}}$, fix a distribution $\mathcal{D}$ over the input space $\mathcal{X} \times \mathcal{Y}$. Let $S = \{(x_1, y_1), \ldots, (x_m, y_m)\}$ be an i.i.d. training sample from a robustly realizable distribution $\mathcal{D}$ by $\mathcal{H}$, namely $\inf_{h \in \mathcal{H}} \mathrm{Risk}_{\mathcal{U}}\left(h; \mathcal{D}\right) = 0$. Denote $d = \mathrm{VC}(\mathcal{H})$, $d^* = \mathrm{VC}^*(\mathcal{H})$ is the *dual VC-dimension*. Fix $\epsilon, \delta \in (0,1)$.

1. Define the inflated training data set

$$S_{\mathcal{U}} = \bigcup_{i \in [n]} \left\{(z, y_{I(z)}) : z \in \mathcal{U}(x_i)\right\},$$

   where $I(z) = \min\left\{i \in [n] : z \in \mathcal{U}(x_i)\right\}$. The goal is to construct a compression scheme the is consistent with $S_{\mathcal{U}}$.

2. Discretize $S_{\mathcal{U}}$ to a finite set $\bar{S}_{\mathcal{U}}$. Define the class of hypotheses with zero robust error on every d points in S,

$$\hat{\mathcal{H}} = \left\{\mathrm{RERM}_{\mathcal{H}}(S') : S' \subseteq S, |S'| = d\right\},$$

   where $\mathrm{RERM}_{\mathcal{H}}$ maps any labeled set to a hypothesis in $\mathcal{H}$ with zero robust loss on this set. The cardinality of this class is bounded as following

$$|\hat{\mathcal{H}}| = \binom{n}{d} \leq \left(\frac{en}{d}\right)^d.$$

   Discretize $S_{\mathcal{U}}$ to a finite set using the finite class $\hat{\mathcal{H}}$. Define the *dual class* $\mathcal{H}^* \subseteq \{0,1\}^{\mathcal{H}}$ of $\mathcal{H}$ as the set of all functions $f_{(x,y)} : \mathcal{H} \to \{0,1\}$ defined by $f_{(x,y)}(h) = \mathbb{I}\left[h(x) \neq y\right]$, for any $h \in \mathcal{H}$ and $(x,y) \in S_{\mathcal{U}}$. If we think of a binary matrix where the rows consist of the distinct hypotheses and the columns are points, then the dual class corresponds to the transposed matrix where the distinct rows are points and the columns are hypotheses. A discretization $\bar{S}_{\mathcal{U}}$ will be defined by the dual class of $\hat{\mathcal{H}}$. Formally, $\bar{S}_{\mathcal{U}} \subseteq S_{\mathcal{U}}$ consists of exactly one $(x,y) \in S_{\mathcal{U}}$ for each distinct classification $\left\{f_{(x,y)}(h)\right\}_{h \in \hat{\mathcal{H}}}$. In other words, $\hat{\mathcal{H}}$ induces a finite partition of $S_{\mathcal{U}}$ into regions where every $\hat{h} \in \hat{\mathcal{H}}$ suffers a constant loss $\mathbb{I}\left[\hat{h}(x) \neq y\right]$ in each region, and the discretization $\bar{S}_{\mathcal{U}}$ takes one point per region. By Sauer's lemma [53, 47], for $n > 2d$,

$$|\bar{S}_{\mathcal{U}}| \leq \left(\frac{e|\hat{\mathcal{H}}|}{d^*}\right)^{d^*} \leq \left(\frac{e^2 n}{dd^*}\right)^{dd^*},$$

3. Execute the following modified version of the algorithm α-boost [48, pages 162-163] on the discretized set $\bar{S}_{\mathcal{U}}$, with parameters $\alpha = \frac{1}{3}$ and number of boosting rounds $T = \mathcal{O}\left(\log\left(|\bar{S}_{\mathcal{U}}|\right)\right) = \mathcal{O}\left(dd^* \log(n)\right)$.

Algorithm 2 Modified α-boost

Input: $\mathcal{H}, S, \bar{S}_{\mathcal{U}}, d, \mathrm{RERM}_{\mathcal{H}}$.
Parameters: α, T.
Initialize P_1 = Uniform($\bar{S}_{\mathcal{U}}$).
For $t = 1, \ldots, T$:

(a) Find $\mathcal{O}(d)$ points $S_t \subseteq \bar{S}_{\mathcal{U}}$ such that every $h \in \mathcal{H}$ with $\widehat{\mathrm{R}}(h; S_t) = 0$ has $\mathrm{R}(h; P_t) \leq 1/3$.
(b) Let S'_t be the original $\mathcal{O}(d)$ points in S with $S_t \subseteq \bigcup_{(x,y)\in S'_t} \bigcup \{(z, y) : z \in \mathcal{U}(x)\}$.
(c) Let $\hat{h}_t = \mathrm{RERM}_{\mathcal{H}}(S'_t)$.
(d) For each $(x, y) \in \bar{S}_{\mathcal{U}}$:

$$P_{t+1}(x, y) \propto P_t(x, y) e^{-\alpha \mathbb{I}\left\{\hat{h}_t(x)=y\right\}}$$

Output: classifiers $\hat{h}_1, \ldots, \hat{h}_T$ and sets $S'_1, \ldots, S'_T$.

4. Output the majority vote $\bar{h} = \mathrm{Majority}\left(\hat{h}_1, \ldots, \hat{h}_T\right)$.

We are guaranteed that $\widehat{\mathrm{R}}_{\mathcal{U}}\left(\bar{h}; S\right) = 0$, and each hypothesis $\hat{h}_t \in \left\{\hat{h}_1, \ldots, \hat{h}_T\right\}$ is representable as set S'_t of size $\mathcal{O}(d)$. This defines a compression function $\kappa(S) = \bigcup_{t\in[T]} S'_t$. Thus, $\bar{h}$ can be reconstructed from a compression set of size

$$dT = \mathcal{O}\left(d^2 d^* \log(n)\right).$$

This compression size can be further reduced to $\mathcal{O}\left(dd^*\right)$, using a sparsification technique introduced by Moran and Yehudayoff [45], Hanneke et al. [32], by randomly choosing $\mathcal{O}(d^*)$ hypotheses from $\left\{\hat{h}_1, \ldots, \hat{h}_T\right\}$. The proof follows via standard uniform convergence argument. Plugging it into a compression generalization bound, we have a sample complexity of $\tilde{\mathcal{O}}\left(\frac{dd^*}{\epsilon} + \frac{\log\frac{1}{\delta}}{\epsilon}\right)$, in the realizable robust case.

Agnostic case. The construction follows a reduction to the realizable case suggested by David et al. [24]. Denote $\Lambda_{\mathrm{RE}} = \Lambda_{\mathrm{RE}}(1/3, 1/3, \mathcal{H}, \mathcal{U})$, the sample complexity of $(1/3, 1/3)$-PAC learn $\mathcal{H}$ with respect to a perturbation function $\mathcal{U}$, in the realizable robust case.

Using a robust ERM, find the maximal subset $S' \subseteq S$ such that $\inf_{h\in\mathcal{H}} \widehat{\mathrm{R}}_{\mathcal{U}}\left(h; S'\right) = 0$. Now, Λ_{RE} samples suffice for weak robust learning for any distribution $\mathcal{D}$ on S'.

Execute the α-boost algorithm [48, pages 162-163] on S' for the robust loss function, with parameters $\alpha = \frac{1}{3}$ and number of boosting rounds $T = \mathcal{O}\left(\log\left(|S'|\right)\right)$, where each weak learner is trained on Λ_{RE} samples. The returned hypothesis $\bar{h} = \mathrm{Majority}\left(\hat{h}_1, \ldots, \hat{h}_T\right)$ satisfies that $\widehat{\mathrm{R}}_{\mathcal{U}}\left(\bar{h}; S'\right) = 0$, and each hypothesis $\hat{h}_t \in \left\{\hat{h}_1, \ldots, \hat{h}_T\right\}$ is representable as set of size $\mathcal{O}(\Lambda_{\mathrm{RE}})$. This defines a compression scheme of size $\Lambda_{\mathrm{RE}}T$, and $\bar{h}$ can be reconstructed from a compression set of points from S of size $\Lambda_{\mathrm{RE}}T$.

Recall that $S' \subseteq S$ is a maximal subset such that $\inf_{h\in\mathcal{H}} \widehat{\mathrm{R}}_{\mathcal{U}}\left(h; S'\right) = 0$ which implies that $\widehat{\mathrm{R}}_{\mathcal{U}}\left(\bar{h}; S\right) \leq \inf_{h\in\mathcal{H}} \widehat{\mathrm{R}}_{\mathcal{U}}\left(h; S\right)$. Plugging it into a compression generalization bound (Lemma F.1 holds for the robust loss function as well), we have a sample complexity of $\tilde{\mathcal{O}}\left(\frac{\Lambda_{\mathrm{RE}}}{\epsilon^2} + \frac{\log\frac{1}{\delta}}{\epsilon^2}\right)$, which translates into $\tilde{\mathcal{O}}\left(\frac{dd^*}{\epsilon^2} + \frac{\log\frac{1}{\delta}}{\epsilon^2}\right)$, in the agnostic robust case.