Black-box Certification and Learning under Adversarial Perturbations

Hassan Ashtiani [*1] **Vinayak Pathak** [*2] **Ruth Urner** [*3]

Abstract

We formally study the problem of classification under adversarial perturbations from a learner's perspective as well as a third-party who aims at certifying the robustness of a given black-box classifier. We analyze a PAC-type framework of semi-supervised learning and identify possibility and impossibility results for proper learning of VC-classes in this setting. We further introduce a new setting of black-box certification under limited query budget, and analyze this for various classes of predictors and perturbation. We also consider the viewpoint of a black-box adversary that aims at finding adversarial examples, showing that the existence of an adversary with polynomial query complexity can imply the existence of a sample efficient robust learner.

1. Introduction

We formally study the problem of classification under *adversarial perturbations*. An adversarial perturbation is an imperceptible alteration of a classifier's input which changes its prediction. The existence of adversarial perturbations for real-world input instances and typical classifiers (Szegedy et al., 2014) has contributed to a lack of trust in predictive tools derived from automated learning. Recent years have thus seen a surge of studies proposing various heuristics to enhance robustness to adversarial attacks (Chakraborty et al., 2018). Existing solutions often either (i) modify the learning procedure to increase the adversarial robustness, e.g. by modifying the training data or the loss function used for training (Sinha et al., 2018; Cohen et al., 2019; Salman et al., 2019), or (ii) post-process an existing classifier to enhance its robustness (Cohen et al., 2019).

A user of a predictive tool, however, may not oftentimes be involved in the training of the classifier nor have the technical access or capabilities to modify its input/output behavior. Instead, the predictor may have been provided by a third party and the user may have merely a *black-box access* to the predictor. That is, the predictor h presents itself as an oracle that takes input *query* x and responds with the label $h(x)$. The provider of the predictive tool, while not necessarily assumed to have malicious intent, is still naturally considered untrusted, and the user thus has an interest in verifying the predictor's performance (including adversarial robustness) on its own application domain. While the standard notion of classification accuracy can be easily estimated from an i.i.d. sample generated from user's data generating distribution, estimating the expected robust loss is not that easy: Given a labeled instance (x, y), the user can immediately verify whether the instance is misclassified ($h(x) \neq y$) using a single query to h, but understanding whether x is vulnerable under adversarial perturbations may require many more queries to the oracle.

We introduce and analyze a formal model for *black-box certification* under query access, and provide examples of hypothesis classes and perturbation types[1] for which such a certifier exists. We further introduce the notion of *witness sets* for certification, and identify more general classes of problems and perturbation types that admit black-box certification with finite queries. On the contrary, we demonstrate cases of simple classes where the query complexity of certification is unbounded.

We further look at the problem from the viewpoint of the adversary, connecting the query complexity of an adversary (for finding adversarial examples) and that of the certifier. An intriguing question that we explore is whether the sample complexity of learning a robust classifier with respect to a hypothesis class is related to the query complexity of an optimal adversary (or certifier) for that class. We uncover such a connection, showing that the existence of a successful adversary with polynomial query complexity for a properly compressible class implies sample efficient robust learnability of that class. For this, we adapt a compression-based argument, demonstrating a sample complexity upper bound

[*] Equal contribution [1] Department of Computing and Software, McMaster University, Hamilton, ON, Canada [2] Scotiabank, Toronto, ON, Canada [3] Lassonde School of Engineering, EECS Department, York University, Toronto, ON, Canada. Correspondence to: Hassan Ashtiani <zokaeiam@mcmaster.ca>, Vinayak Pathak <vpathak@uwaterloo.ca>, Ruth Urner <ruth@eecs.yorku.ca>.

Proceedings of the 37^{th} International Conference on Machine Learning, Online, PMLR 119, 2020. Copyright 2020 by the author(s).

[1] A perturbation type captures the set of admissible perturbations the adversary is allowed to make at each point (See Section 2).

for robust learning that is smaller than what was previously known (Montasser et al., 2019) (assuming that a linear adversary exists and the class is properly compressible).

We start our investigations with the problem of robustly (PAC-)learning classes of finite VC-dimension. It has been shown recently that, while the VC-dimension characterizes the proper learnability of a hypothesis class under the binary (misclassification) loss, there are classes of small VC-dimension that are not properly learnable under the robust loss (Montasser et al., 2019). We define the notion of the margin class (associated with a hypothesis class and a perturbation type) and show that, if both the class and the margin class are simple (measured by their VC-dimension), then proper learning under robust loss is not significantly more difficult than learning with respect to the binary loss.

The corresponding complexity of the margin class, however, can be potentially large for specific choices of perturbation types and hypothesis classes. We thus investigate and provide scenarios where a form of semi-supervised learning can overcome the impossibility of proper robust learning. We believe our investigations of robust learnability in these scenarios may help shed some light on where the difficulty of general robust classification stems from.

1.1. Related work

Recent years have produced a surge of work on adversarial attack (and defense) mechanisms (Madry et al., 2018; Chakraborty et al., 2018; Chen et al., 2017; Dong et al., 2018; Narodytska & Kasiviswanathan, 2017; Papernot et al., 2017; Akhtar & Mian, 2018; Su et al., 2019), as well as the development of reference implementations of these (Goodfellow et al., 2018). Here, we briefly review some earlier work on theoretical understanding of the problem.

Several recent studies have suggested and analyzed approaches of training under data augmentation (Sinha et al., 2018; Salman et al., 2019). The general idea is to add adversarial perturbations to data points already at training time to promote smoothness around the support of the data generating distribution. These studies then provide statistical guarantees for the robustness of the learned classifier. Similarly, statistical guarantees have been presented for robust training that modifies the loss function rather than the training data (Wong & Kolter, 2018). However, the notion of robustness certification used in these is different from what we propose. While they focus on designing learning methods that are certifiably robust, we aim at certifying an *arbitrary* classifier and for a potentially new distribution.

The robust learnability of finite VC-classes has been studied only recently, often with pessimistic conclusions. An early result demonstrated that there exist distributions where robust learning requires provably more data than its non-robust counterpart (Schmidt et al., 2018). Recent works have studied adversarially robust classification in the PAC-learning framework of computational learning theory (Cullina et al., 2018; Awasthi et al., 2019; Montasser et al., 2020) and presented hardness results for binary distribution and hypothesis classes in this framework (Diochnos et al., 2018; Gourdeau et al., 2019; Diochnos et al., 2019). On the other hand, robust learning has been shown to be possible, for example when the hypothesis class is finite and the adversary has a finite number of options for corrupting the input (Feige et al., 2015). This result has also been extended to the more general case of classes with finite VC-dimension (Attias et al., 2019). It has also been shown that robust learning is possible (by Robust Empirical Risk Minimization (RERM)) under a feasibility assumption on the distribution and bounded covering numbers of the hypothesis class (Bubeck et al., 2019). However, more recent work has presented classes of VC-dimension 1, where the robust loss class has arbitrarily large VC-dimension (Cullina et al., 2018) and, moreover, where proper learning (such as RERM) is impossible in a distribution-free finite sample regime (Montasser et al., 2019). Remarkably, the latter work also presents an improper learning scheme for any VC-class and any adversary type. The sample complexity of this approach, however, depends on the dual VC-dimension which can be exponential in the VC-dimension of the class.

We note that two additional aspects of our work have appeared in the the literature before: considering robust learnability by imposing computational constraints on an adversary has been explored recently (Bubeck et al., 2019; Gourdeau et al., 2019; Garg et al., 2019). Earlier work has also hypothesized that unlabeled data may facilitate adversarially robust learning, and demonstrated a scenario where access to unlabeled data yields a better bound on the sample complexity under a specific data generative model (Carmon et al., 2019; Alayrac et al., 2019).

Less closely related to our work, the theory of adversarially robust learnability has been studied for non-parametric learners. A first study in that framework showed that a nearest neighbor classifier's robust loss converges to that of the Bayes optimal (Wang et al., 2018). A follow-up work then derived a characterization of the best classifier with respect to the robust loss (analogous to the notion of the Bayes optimal), and suggested a training data pruning approach for non-parametric robust classification (Yang et al., 2019).

1.2. Outline and summary of contributions

Problem setup and the adversarial loss formulation. In Section 2, we provide the formal setup for the problem of adversarial learning. We also decompose the adversarial loss, and define the notion of the margin class associated with a hypothesis class and a perturbation type (Def. 4).

Using unlabeled data for adversarial learning of VC-classes. In Section 3, we study the sample complexity of *proper* robust learning. While this sample complexity can be infinite for general VC-classes (Cullina et al., 2018; Montasser et al., 2019; Yin et al., 2019), we show that VC-classes are properly robustly learnable if the margin class also has finite VC-dim (Thm. 7). We formalize an idealized notion of semi-supervised learning where the learner has additional oracle access to probability weights of the margin sets. We show that, perhaps counter intuitively, oracle access to both (exact) margin weights and (exact) binary losses, does not suffice for identifying the minimizer of the adversarial loss in a class $\mathcal{H}$ (Thm. 9), even in the $0/1$-realizable case (Thm. 12). However, under the additional assumption of robust realizability, proper learning becomes feasible with access to the marginal or sufficient unlabled data (Thms. 10 and 11).

Black-box certification with query access. We formally define the problem of *black-box certification through query access* (Def. 15), and demonstrate examples where certification is possible (Obs. 16) or impossible (Obs. 17). Motivated by this impossibility result, we also introduce a *tolerant* notion of certification (Def. 19). We show that while more classes are certifiable with this definition (Obs. 20), some simple classes remain impossible to certify (Obs. 21). We identify a sufficient condition for certifiability of a hypothesis class w.r.t. a perturbation type through the notion of *witness sets* (Def. 22 and Thm. 23). We then consider the query complexity of the adversary (as opposed to that of the certifier) for finding adversarial instances (Def. 24, 25, and 26) and—for the case of a non-adaptive adversary—relate it to the existence of a witness set (Obs. 27).

Connecting adversarial query complexity and PAC-learnability. The culminating result connects the two themes of our work: robust (PAC-)learnability, and query complexity of an adversary. With Theorem 28, we show that existence of a perfect adversary with small query complexity implies sample-efficient robust learning for properly compressible classes.

We include the proof sketches in the paper, and refer the reader to the supplementary material for detailed proofs.

2. Setup and Definitions

We let X denote the domain (often $X \subseteq \mathbb{R}^d$) and Y (mostly $Y = \{0, 1\}$) a (binary) label space. We assume that data is generated by some distribution P over $X \times Y$ and let P_X denote the marginal of P over X. A *hypothesis* is a function $h : X \to Y$, and can naturally be identified with a subset of $X \times Y$, namely $h = \{(x, y) \in X \times Y \mid x \in X,\ y = f(x)\}$. Since we are working with binary labels, we also sometimes identify a hypothesis h with the pre-image of 1 under h, that is the domain subset $\{x \in X \mid h(x) = 1\}$. We let $\mathcal{F}$ denote the set of all Borel functions[2] from X to Y (or all functions in case of a countable domain). A *hypothesis class* is a subset of $\mathcal{F}$, often denoted by $\mathcal{H} \subseteq \mathcal{F}$.

The quality of prediction of a hypothesis on a labeled example (x, y) is measured by a *loss function* $\ell : (\mathcal{F} \times X \times Y) \to \mathbb{R}$. For classification problems, the quality of prediction is typically measured with the *binary* loss

$$\ell^{0/1}(h, x, y) = \mathbb{1}\left[h(x) \neq y\right]$$

, where $\mathbb{1}\left[\alpha\right]$ denotes the indicator function for predicate α. For (adversarially) robust classification, we let $\mathcal{U} : X \to 2^X$, the *perturbation type*, be a function that maps each instance to the set of admissible perturbations at point x. We assume that the perturbation type satisfies $x \in \mathcal{U}(x)$ for all $x \in X$. If X is equipped with a metric dist, then a natural choice for the set of perturbations at x is a ball $\mathcal{B}_r(x) = \{z \in X \mid \mathrm{dist}(x, z) \leq r\}$ of radius r around x. For an $x \in X$ and $h \in \mathcal{H}$, we say that $x' \in \mathcal{U}(x)$ is an *adversarial* point of x with respect to h if $h(x) \neq h(x')$. We use the following definition of the adversarially robust loss with respect to perturbation type $\mathcal{U}$

$$\ell^{\mathcal{U}}(h, x, y) = \mathbb{1}\left[\exists z \in \mathcal{U}(x) \ :\ h(z) \neq y\right].$$

If $\mathcal{U}(x)$ is always a ball of radius r around x, we will also use the notation $\ell^r(h, x, y) = \ell^{\mathcal{B}_r}(h, x, y)$. We assume that the perturbation type is so that $\ell^{\mathcal{U}}(f, \cdot, \cdot)$ is a measurable function for all $f \in \mathcal{F}$. A sufficient condition for this is that the set $\mathcal{U}(x)$ are open sets (where X is assumed to be equipped with some topology) and the pertubation type further satisfies $z \in \mathcal{U}(x)$ if and only if $x \in \mathcal{U}(z)$ for all $x, z \in X$ (see Appendix B for a proof and an example of a simple perturbation type that renders the the corresponding loss function of a threshold predictor non-measurable).

We denote the *expected loss* (or *true loss*) of a hypothesis h with respect to the distribution P and loss function ℓ by $\mathcal{L}_P(h) = \mathbb{E}_{(x,y)\sim P}[\ell(h, x, y)]$. In particular, we will denote the true binary loss by $\mathcal{L}^{0/1}_P(h)$ and the true robust loss by $\mathcal{L}^{\mathcal{U}}_P(h)$. Further, we denote the *approximation error* of class $\mathcal{H}$ with respect to distribution P and loss function ℓ by $\mathcal{L}_P(\mathcal{H}) = \inf_{h\in\mathcal{H}} \mathcal{L}_P(h)$.

The *empirical loss* of a hypothesis h with respect to loss function ℓ and a sample $S = ((x_1, y_1), \ldots, (x_n, y_n))$ is defined as $\mathcal{L}_S(h) = \frac{1}{n}\sum_{i=1}^n \ell(h, x_i, y_i)$.

A *learner* $\mathcal{A}$ is a function that takes in a finite sequence of labeled instances $S = ((x_1, y_1), \ldots, (x_n, y_n))$ and outputs a hypothesis $h = \mathcal{A}(S)$. The following is a standard notion of (PAC-)learnability from finite samples of a hypothesis

[2] For an uncountable domain, we only consider Borel-measurable hypotheses to avoid dealing with measurability issues.

class (Vapnik & Chervonenkis, 1971; Valiant, 1984; Blumer et al., 1989; Shalev-Shwartz & Ben-David, 2014).

Definition 1 ((Agnostic) Learnability)**.** *A hypothesis class $\mathcal{H}$ is agnostic learnable with respect to set of distributions $\mathcal{P}$ and loss function ℓ, if there exists a learner $\mathcal{A}$ such that for all $\epsilon, \delta \in (0,1)$, there is a sample size $m(\epsilon, \delta)$ such that, for any distribution $P \in \mathcal{P}$, if the input to $\mathcal{A}$ is an iid sample S from P of size $m \geq m(\epsilon, \delta)$, then, with probability at least $(1-\delta)$ over the samples, the learner outputs a hypothesis $h = \mathcal{A}(S)$ with $\mathcal{L}_P(h) \leq \mathcal{L}_P(\mathcal{H}) + \epsilon$.*

$\mathcal{H}$ is said to be learnable in the realizable case *with respect to loss function ℓ, if the above holds under the condition that $\mathcal{L}_P(\mathcal{H}) = 0$. We say that $\mathcal{H}$ is* distribution-free learnable *(or simply* learnable*) if it is learnable when $\mathcal{P}$ is the set of all probability measures over $X \times Y$.*

Definition 2 (VC-dimension)**.** *We say that a collection of subsets $\mathcal{G} \subseteq 2^X$ of some domain X* shatters *a subset $B \subseteq X$ if for every $F \subseteq B$ there exists $G \in \mathcal{G}$ such that $G \cap B = F$. The* VC-dimension *of $\mathcal{G}$, denoted by $\mathrm{VC}(\mathcal{G})$, is defined to be the supremum of the size of the sets that are shattered by $\mathcal{G}$.*

It is easy to see that the VC-dimension of a binary hypothesis class $\mathcal{H}$ is independent of whether we view $\mathcal{H}$ as a subset of $X \times Y$ or pre-images of 1 (thus, subsets of X). It is well known that, for the binary loss, a hypothesis class is (distribution-free) learnable if and only if it has finite VC-dimension (Blumer et al., 1989). Furthermore, any learnable binary hypothesis class can be learned with a *proper learner*.

Definition 3 (Proper Learnability)**.** *We call a learner $\mathcal{A}$ a* proper learner *for the class $\mathcal{H}$ if, for all input samples S, we have $\mathcal{A}(S) \in \mathcal{H}$. A class $\mathcal{H}$ is* properly learnable *if the conditions in Definition 1 hold with a proper learner $\mathcal{A}$.*

It has recently been shown that there are classes of finite VC-dimension that are not properly learnable with respect to the adversarially robust loss (Montasser et al., 2019).

2.1. Decomposing the robust loss

In this work, we adapt the most commonly used notion of a adversarially robust loss (Montasser et al., 2019; Yang et al., 2019). Note that, we have $\ell^{\mathcal{U}}(h, x, y) = 1$ if and only if at least one of the following conditions holds:

- h makes a mistake on x with respect to label y, or
- there is a close-by instance $z \in \mathcal{U}(x)$ that h labels different than x, that is, x is close to h's decision boundary.

The first condition holds when (x, y) falls into the *error region*, $\mathrm{err}_h = (X \times Y) \setminus h$. The notion of error region then naturally captures the (non-adversarial) loss:

$$\mathcal{L}_P^{0/1}(h) = \mathbb{P}_{(x,y)\sim P}[(x,y) \in \mathrm{err}_h] = P(\mathrm{err}_h).$$

The second condition holds when x lies in the *margin area* of h. The following definition makes this notion explicit.

Let $h \in \mathcal{F}$ be some hypothesis. We define the *margin area* of h with respect to perturbation type $\mathcal{U}$, as the subset $\mathrm{mar}_h^{\mathcal{U}} \subseteq X \times Y$ defined by

$$\mathrm{mar}_h^{\mathcal{U}} = \{(x,y) \in X \times Y \mid \exists z \in \mathcal{U}(x) : h(x) \neq h(z)\}$$

Based on these definitions, the adversarially robust loss with respect to $\mathcal{U}$ is 1 if and only if the sample (x, y) falls into the error region err_h and/or the margin area $\mathrm{mar}_h^{\mathcal{U}}$ of h:

$$\mathcal{L}_P^{\mathcal{U}}(h) = P(\mathrm{err}_h \cup \mathrm{mar}_h^{\mathcal{U}}).$$

Definition 4. *For class $\mathcal{H}$, we refer to the collection $\mathcal{H}_{\mathrm{mar}}^{\mathcal{U}} = \{\mathrm{mar}_h^{\mathcal{U}} \mid h \in \mathcal{H}\}$ as the* margin class *of $\mathcal{H}$.*

While we defined that margin areas $\mathrm{mar}_h^{\mathcal{U}}$ as subsets of $X \times Y$, it is sometimes natural to identify them with their projection on X, thus simply as subsets of X.

Remark 5. *There is more than one way to formulate a loss function that captures both classification accuracy and robustness to (small) adversarial perturbations. The notion we adopt has the property that even the true labeling function can have positive robust loss, if the true labels themselves change within the adversarial neighbourhoods. A natural alternative is to say an adversarial point is a point in the neighbourhood of an instance that is misclassified by the classifier. However, note that such a notion cannot be phrased as a loss function $\ell(h, x, y)$ (as it depends on the true label of the perturbed instance). Previous studies have provided excellent discussions of the various options (Diochnos et al., 2018; Gourdeau et al., 2019).*

Semi-Supervised Learning (SSL) Since the margin areas $\mathrm{mar}_h^{\mathcal{U}}$ can naturally be viewed as subsets of X, their weights $P(\mathrm{mar}_h^{\mathcal{U}})$ under the data generating distribution can potentially be estimated with samples from P_X, that is, from *unlabeled data*. A learner that takes in both a labeled sample S from P and an unlabeled sample T from P_X, is called a *semi-supervised learner*. For scenarios where robust learning has been shown to be hard, we explore whether this hardness can be overcome by SSL. We consider semi-supervised learners that take in labeled and unlabeled samples, and also *idealized semi-supervised learners* that, in addition to a labeled samples have oracle access to probability weights of certain subsets of X (Göpfert et al., 2019).

3. Robust Learning of VC Classes

It has been shown that there is a class $\mathcal{H}$ of bounded VC-dimension ($\mathrm{VC}(\mathcal{H}) = 1$ in fact) and a perturbation type $\mathcal{U}$ such that $\mathcal{H}$ is not robustly properly learnable (Montasser et al., 2019), even if the distribution is realizable with respect to $\mathcal{H}$ under the $\mathcal{U}$-robust loss. The perturbation type $\mathcal{U}$ in that lower bound construction can actually chosen to be balls with respect to some metric over $X = \mathbb{R}^d$ (for any d, even $d = 1$). The same work also shows that if a class has bounded VC-dimension, then it is (improperly) robustly learnable with respect to any perturbation type $\mathcal{U}$.

Theorem 6 ((Montasser et al., 2019)). *(1) There is a class $\mathcal{H}$ over $X = \mathbb{R}^d$ with $\mathrm{VC}(\mathcal{H}) = 1$, and a set of distributions $\mathcal{P}$ with $\mathcal{L}^r_P(\mathcal{H}) = 0$ for all $P \in \mathcal{P}$, such that $\mathcal{H}$ is not proper learnable over $\mathcal{P}$ with respect to loss function ℓ^r.*
(2) Let X be any domain and $\mathcal{U} : X \to 2^X$ be any type of perturbation, and let $\mathcal{H} \subseteq \{0,1\}^X$ be a hypothesis class with finite VC-dimension. Then $\mathcal{H}$ is distribution-free agnostic learnable with respect to loss function $\ell^{\mathcal{U}}$.

While the second part of the above theorem seems to settle adversarially robust learnability for binary hypothesis classes, the positive result is achieved with a compression-based learner, which has potentially much higher sample complexity than what suffices for the binary loss. In fact, the size of the best known general compression scheme (Moran & Yehudayoff, 2016) depends on the VC-dimension of the dual class of $\mathcal{H}$, making the sample complexity of this approach generally exponential in VC-dimension of $\mathcal{H}$.

We first show that the impossibility part of the above theorem crucially depends on the combination of the class $\mathcal{H}$ and a pertubation type $\mathcal{U}$ (despite these being balls in a Euclidian space) so that the margin class $\mathcal{H}^{\mathcal{U}}_{\mathrm{mar}}$ has infinite VC-dimension. We prove that, if both $\mathcal{H}$ and $\mathcal{H}^{\mathcal{U}}_{\mathrm{mar}}$ have finite VC-dimension then $\mathcal{H}$ is (distribution-free) learnable with respect to the robust loss, with a proper learner.

Theorem 7 (Proper learnability for finite VC and finite margin-VC). *Let X be any domain and $\mathcal{H} \subseteq \mathcal{F}$ be a hypothesis class with finite VC-dimension. Further, let $\mathcal{U} : X \to 2^X$ be any perturbation type such that $\mathcal{H}^{\mathcal{U}}_{\mathrm{mar}}$ has finite VC-dimension. We set $D = \mathrm{VC}(\mathcal{H}) + \mathrm{VC}(\mathcal{H}^{\mathcal{U}}_{\mathrm{mar}})$. Then $\mathcal{H}$ is distribution-free (agnostically) properly learnable with respect to the robust loss $\ell^{\mathcal{U}}$, and the sample complexity is $O\left(\frac{D\log(D)+\log(1/\delta)}{\epsilon^2}\right)$.*

Proof Sketch. We provide the more detailed argument in the appendix. Recall that a set $S \subseteq X \times Y$ is said to be an ϵ-approximation of P with respect to $\mathcal{H} \subseteq X \times Y$ if for all $h \in \mathcal{H}$ we have $\left|P[h] - \frac{|h\cap S|}{|S|}\right| \leq \epsilon$, that is, if the empirical estimates with respect to S of the sets in h are ϵ-close to their true probability weights. Consider the class of subsets $\mathcal{G} = \{(\mathrm{err}_h \cup \mathrm{mar}^{\mathcal{U}}_h) \subseteq X \times Y \mid h \in \mathcal{H}\}$ of point-wise unions of error and margin regions. A simple counting argument shows that $\mathrm{VC}(\mathcal{G}) \leq D\log(D)$, where $D = \mathrm{VC}(\mathcal{H}) + \mathrm{VC}(\mathcal{H}^{\mathcal{U}}_{\mathrm{mar}})$. Thus, by basic VC-theory, a sample of size $\Theta\left(\frac{D\log D+\log(1/\delta)}{\epsilon^2}\right)$ will be an ϵ-approximation of $\mathcal{G}$ with respect to P with probability at least $1-\delta$. Thus any empirical risk minimizer with respect to $\ell^{\mathcal{U}}$ is a successful proper and agnostic robust learner for $\mathcal{H}$. □

Observation 8. *We believe the conditions of Theorem 7 hold for most natural classes and perturbation types $\mathcal{U}$. Eg. if $\mathcal{H}$ is the class of linear predictors in $\mathbb{R}^d$ and $\mathcal{U}$ are sets of balls with respect to some ℓ_p-norm, then both $\mathcal{H}$ and $\mathcal{H}^{\mathcal{U}}_{\mathrm{mar}}$ have finite VC-dimension (see also (Yin et al., 2019)).*

3.1. Using unlabeled data for robust proper learning

In light of the above two general results, we turn to investigate whether unlabeled data can help in overcoming the discrepancy between the two setups. In particular, under various additional assumptions, we consider the case of $\mathrm{VC}(\mathcal{H})$ being finite but $\mathrm{VC}(\mathcal{H}^{\mathcal{U}}_{\mathrm{mar}})$ (potentially) being infinite and a learner having additional access to P_X.

We model knowledge of P_X as the learner having access to an oracle that returns the probability weights of various subsets of X. We say that the learner has access to a *margin oracle for class* $\mathcal{H}$ if, for every $h \in \mathcal{H}$, it has access (can query) the probability weight of the margin set of h, that is $P(\mathrm{mar}^{\mathcal{U}}_h)$. Since the margin areas can be viewed as subsets of X, if the margin class of $\mathcal{H}$ under perturbation type $\mathcal{U}$ has finite VC-dimension, a margin oracle can be approximated using an unlabeled sample from the distribution P.

Similarly, one could define an *error oracle for* $\mathcal{H}$ as an oracle, that, for every $h \in \mathcal{H}$ would return the weight of the error sets $P(\mathrm{err}_h)$. This is typically approximated with a labeled sample from the data-generating distribution, if the class has finite VC-dimension. This is similar to the settings of learning by distances (Ben-David et al., 1995) or learning with statistical queries (Kearns, 1998; Feldman, 2017).

To minimize the adversarial loss however, the learner needs to find (through oracle access or through approximations by samples) a minimizer of the weights $P(\mathrm{err}_h \cup \mathrm{mar}^{\mathcal{U}}_h)$. We now first show that having access to both an exact error oracle and an exact margin oracle does not suffice for this.

Theorem 9. *There is a class $\mathcal{H}$ with $\mathrm{VC}(\mathcal{H}) = 1$ over a domain X with $|X| = 7$, a perturbation type $\mathcal{U} : X \to 2^X$, and two distributions P^1 and P^2 over $X \times \{0,1\}$, that are indistinguishable with error and margin oracles for $\mathcal{H}$, while their robust loss minimizers in $\mathcal{H}$ differ.*

Proof. Let $X = \{x_1, x_2, \ldots, x_7\}$ be the domain. We consider two distributions P^1 and P^2 over $X \times \{0,1\}$. Both have true label 0 on all points, that is $P(y = 1|x) = 0$ for all $x \in X$. However their marginals P^1 and P^2 differ:

$$\begin{aligned}
P^1_X(x_1) &= P^1_X(x_3) = 0,\ P^1_X(x_2) = 2/6, \text{and} \\
P^1_X(x_i) &= 1/6 \text{ for } i \in \{4,5,6,7\}. \\
P^2_X(x_4) &= P^2_X(x_6) = 0,\ P^2_X(x_5) = 2/6, \text{and} \\
P^2_X(x_i) &= 1/6 \text{ for } i \in \{1,2,3,7\}.
\end{aligned}$$

The class $\mathcal{H}$ consists of two functions: $h_1 =$

$\mathbb{1}[x = x_2 \vee x = x_3]$ and $h_2 = \mathbb{1}[x = x_5 \vee x = x_6]$. Further, we consider the following perturbation sets (for readability, we first state them without the points themselves):

$$\tilde{\mathcal{U}}(x_1) = \{x_2\},\ \tilde{\mathcal{U}}(x_2) = \{x_1, x_3\},\ \tilde{\mathcal{U}}(x_3) = \{x_2\},$$
$$\tilde{\mathcal{U}}(x_4) = \{x_5\},\ \tilde{\mathcal{U}}(x_5) = \{x_4, x_6\},\ \tilde{\mathcal{U}}(x_6) = \{x_5\},$$
$$\tilde{\mathcal{U}}(x_7) = \emptyset$$

Now we set $\mathcal{U}(x_i) = \tilde{\mathcal{U}}(x_i) \cup \{x_i\}$, so that each point is included in its own perturbation set. Now, both h_1 and h_2 have $0/1$-loss $2/6 = 1/3$ on both P^1 and P^2. And for both h_1 and h_2 the margin areas have weight $2/6 = 1/3$ on both P^1 and P^2. However, the adversarial loss minimizer for P^1 is h_1 and for P^2 is h_2 (by a gap of $1/6$ each). □

While the impossibility result in the above example, of course, can be overcome by estimating the weights of the seven points in the domain, the construction exhibits that merely estimating classification error and weights of margin sets does not suffice for proper learning with respect to the adversarial loss. The example shows, that the learner also needs to take into account the interactions (intersections between the sets) of the two components of the adversarial loss. However the weights of the intersection sets $\mathrm{err}_h \cap \mathrm{mar}_h^{\mathcal{U}}$, inherently involve label information.

In the following subsection we show that realizability with respect to the robust loss implies that robust learning becomes possible with access to a (bounded size) labeled sample from the distribution and additional access to a margin oracle or a (bounded size) unlabeled sample. In the appendix Section C.3, we further explore weakening this assumption to only require $0/1$-reazability with access to stronger version of the margin oracle.

3.1.1. Robust realizability: $\exists h^* \in \mathcal{H}$ with $\mathcal{L}_P^{\mathcal{U}}(h^*) = 0$

This is the setup of the impossibility result for proper learning (Montasser et al., 2019). We show that proper learning becomes possible with access to a margin oracle for $\mathcal{H}$.

Theorem 10. *Let X be some domain, $\mathcal{H}$ a hypothesis class with finite VC-dimension and $\mathcal{U} : X \to 2^X$ any perturbation type. If a learner is given additional access to a margin oracle for $\mathcal{H}$, then $\mathcal{H}$ is properly learnable with respect to the robust loss $\ell^{\mathcal{U}}$ and the class of distributions P that are robust-realizable by $\mathcal{H}$, $\mathcal{L}_P^{\mathcal{U}}(\mathcal{H}) = 0$, with labeled sample complexity $\tilde{O}(\frac{\mathrm{VC}(\mathcal{H}) + \log(1/\delta)}{\epsilon})$*

Proof Sketch. By the robust realizability, there is an $h^* \in \mathcal{H}$ with $\mathcal{L}_P^{\mathcal{U}}(h^*) = 0$ implying that $\mathcal{L}_P^{0/1}(h^*) = 0$, that is, the distribution is (standard) realizable by $\mathcal{H}$. Basic VC-theory tells us that an iid sample S of size $\tilde{\Theta}\left(\frac{\mathrm{VC}(\mathcal{H}) + \log(\frac{1}{\delta})}{\epsilon}\right)$ guarantees that all functions in the *version space* of S (that is all h with $\mathcal{L}_S^{0/1}(h) = 0$) have true binary loss at most ϵ (with probability at least $1 - \delta$). Now, with access to a margin oracle for $\mathcal{H}$ a learner can remove all hypotheses with $P(\mathrm{mar}_h^U) > 0$ from the version space and return any remaining hypothesis (at least h^* will remain). □

Note that the above procedure crucially depends on actual access to a margin oracle. The weights $P(\mathrm{mar}_h^U)$ cannot be generally estimated if $\mathcal{H}_{\mathrm{mar}}^{\mathcal{U}}$ has infinite VC-dimension, as the impossibility result for proper learning from finite samples shows. Thus proper learnability even under these (strong) assumptions cannot always be manifested by a semi-supervised proper learner that has access only to finite amounts of unlabeled data. We also note that the above result (even with access to P_X) does not allow for an extension to the agnostic case via the type of reductions known from compression-based bounds (Montasser et al., 2019; Moran & Yehudayoff, 2016).

On the other hand, if the margin class has finite, but potentially much larger VC-dimension than $\mathcal{H}$, then we can use unlabeled data to approximate the margin oracle in Theorem 10. The following result thus provides an improved bound on the number of *labeled samples* that suffice for robust proper learning under the assumptions of Theorem 7.

Theorem 11. *Let X be some domain, $\mathcal{H}$ a hypothesis class with finite VC-dimension and let $\mathcal{U} : X \to 2^X$ be a perturbation type such that the margin class $\mathcal{H}_{\mathrm{mar}}^{\mathcal{U}}$ also has finite VC-dimension. If a learner is given additional access to an (unlabeled) sample T from P_X, then $\mathcal{H}$ is properly learnable with respect to the robust loss $\ell^{\mathcal{U}}$ and the class of distributions P that are robust-realizable by $\mathcal{H}$, $\mathcal{L}_P^{\mathcal{U}}(\mathcal{H}) = 0$, with labeled sample complexity $\tilde{O}(\frac{\mathrm{VC}(\mathcal{H}) + \log(1/\delta)}{\epsilon})$ and unlabeled sample complexity $\tilde{O}(\frac{\mathrm{VC}(\mathcal{H}_{\mathrm{mar}}^{\mathcal{U}}) + \log(1/\delta)}{\epsilon})$*

Proof. The stated sample sizes imply that all functions in $\mathcal{H}$ in the version space of the labeled sample S have true binary loss at most ϵ and all functions in $\mathcal{H}$ whose margin areas are not hit by T have true margin weight at most ϵ. The learner can thus output any function h with 0 classification error on S and 0 margin weight under T (at least h^* will satisfy these conditions), and we get $\mathcal{L}_h^{\mathcal{U}} = P(\mathrm{err}_h \cup \mathrm{mar}_h^{\mathcal{U}}) \leq 2\epsilon$. □

The assumption in the above theorems states that there exists one function h^* in the class that has both perfect classification accuracy and no weight in its margin area. The proof of the impossibility construction of Theorem 9 employs a class and distributions where no function in the class has perfect margin or perfectly classifies the task. We can modify that construction to show that the "double realizability" in Theorem 10 is necessary if the access to the marginal should be restricted to a margin oracle for $\mathcal{H}$. The proof of the follwing result can be found in Appendix C.2.

Theorem 12. *There is a class $\mathcal{H}$ with $\mathrm{VC}(\mathcal{H}) = 1$ over a domain X with $|X| = 8$, a perturbation type $\mathcal{U} : X \to 2^X$, and two distributions P^1 and P^2 over $X \times \{0,1\}$, such that there are functions $h_r, h_c \in \mathcal{H}$ with $\mathcal{L}^{0/1}_{P^i}(h_r) = 0$ and $P^i(\mathrm{mar}^{\mathcal{U}}_{h_c}) = 0$ for both $i \in \{1,2\}$, while P^1 and P^2 are indistinguishable with error and margin oracles for $\mathcal{H}$ and their robust loss minimizers in $\mathcal{H}$ differ.*

4. Black-box Certification and the Query Complexity of Adversarial Attacks

Given a fixed hypothesis h, a basic concentration inequality (e.g., Hoeffding's inequality) indicates that the empirical loss of h on a samples $S \sim P^m$, $\mathcal{L}^{0/1}_S(h)$, gives an $O(m^{-1/2})$-accurate estimate of the true loss with respect to P, $\mathcal{L}^{0/1}_P(h)$. In fact, in order to compute $\mathcal{L}^{0/1}_S(h)$, we do not need to know h directly; it would suffice to be able to query $h(x)$ on the given sample. Therefore, we can say it is possible to estimate the true binary loss of h up to additive error ϵ using $O(1/\epsilon^2)$ samples from P and $O(1/\epsilon^2)$ queries to $h(.)$.

The high-level question that we ask in this section is whether and when we can do the same for the adversarial loss, $\mathcal{L}^{\mathcal{U}}_P(h)$. If possible, it would mean that we can have a third-party that "certifies" the robustness of a given black-box predictor (e.g., without relying on the knowledge of the learning algorithm that produced it)

Definition 13 (Label Query Oracle)**.** *We call an oracle $\mathcal{O}_h$ a* label query oracle for a hypothesis *h, if for all $x \in X$, upon querying for x, the oracle returns the label $\mathcal{O}_h(x) = h(x)$.*

Definition 14 (Query-based Algorithm)**.** *We call an algorithm $\mathcal{A} : (\bigcup_{i=1}^{\infty} X^i, \mathcal{O}_h) \to \mathbb{R}$ a* query-based algorithm*, if $\mathcal{A}$ has access to a label query oracle $\mathcal{O}_h$.*

Definition 15 (Certifiablility)**.** *A class $\mathcal{H}$ is* certifiable *with respect to $\mathcal{U}$ if there exists a query based algorithm $\mathcal{A}$ and there are functions $q, m : (0,1)^2 \to \mathbb{N}$ such that for every $\epsilon, \delta \in (0,1]$, every distribution P over $X \times Y$, and every $h \in \mathcal{H}$, we have that with probability at least $1 - \delta$ over an iid sample $S \sim P^m_X$ of size $m \geq m(\epsilon, \delta)$*

$$|\mathcal{A}(S, \mathcal{O}_h) \; - \; \mathcal{L}^{\mathcal{U}}_P(h)| \; < \; \epsilon$$

with a query budget of $q(\epsilon, \delta)$ for $\mathcal{A}$. In this case, we say that $\mathcal{H}$ admits (m, q) blackbox query certification.

In light of Section 2.1, the task of robust certification is to estimate the probability weight of the set $\mathrm{err}_h \cup \mathrm{mar}^{\mathcal{U}}_h$.

Observation 16. *Let $\mathcal{H}$ be the set of all half-spaces in $\mathbb{R}^2$ and let $\mathcal{U}(x) = \{z : \|x - z\|_1 \leq 1\}$ be the unit ball wrt ℓ_1-norm centred at x. Then $\mathcal{H}$ admits (m, q)-certification under $\mathcal{U}$ for functions $m, q \in O(1/\epsilon^2)$.*

Proof. Say we have a sample $S \sim P^m_X$. For each point $x \in S$ define the set $w(x) = \{x + (0,1), x + (1,0), x + (-1,0), x + (0,-1)\}$, i.e., the four corner points of $\mathcal{U}(x)$. The certifier can determine whether $x \in \mathrm{err}_h$ by querying the label of x; further it can determine whether $x \in \mathrm{mar}^{\mathcal{U}}_h$ by querying all points in $w(x)$. Let $W = \cup_{x \in S} w(x)$. By querying all points in $S \cup W$, the certifier can calculate the robust loss of h on S. This will be an ϵ-accurate estimate of $L^{\mathcal{U}}_P(h)$ when $m = O(1/\epsilon^2)$. □

We immediately see that certification is non-trivial, in that there are cases where robust certification is impossible. The proof can be found in Appendix D.

Observation 17. *Let $\mathcal{H}$ be the set of all half-spaces in $\mathbb{R}^2$ and let $\mathcal{U}(x) = \{z : \|x - z\|_2 \leq 1\}$ be the unit ball wrt ℓ_2-norm centred at x. Then $\mathcal{H}$ is not certifiable under $\mathcal{U}$.*

This motivates us to define a *tolerant certification* version.

Definition 18 (Restriction of a perturbation type)**.** *Let $\mathcal{U}, \mathcal{V} : X \to 2^X$ be a perturbation types. We say that $\mathcal{U}$* is a restriction of *$\mathcal{V}$ if $\mathcal{U}(x) \subseteq \mathcal{V}(x)$ for all $x \in X$.*

Note that if $\mathcal{U}$ is a restriction of $\mathcal{V}$, then, for all distributions P and predictors h we have $\mathcal{L}^{\mathcal{U}}_P \leq \mathcal{L}^{\mathcal{V}}_P$.

Definition 19 (Tolerant Certification)**.** *A class $\mathcal{H}$ is* tolerantly certifiable *with respect to $\mathcal{U}$ and $\mathcal{V}$, where $\mathcal{U}$ is a restriction of $\mathcal{V}$, if there exists a query based algorithm $\mathcal{A}$, and there are functions $q, m : (0,1)^2 \to \mathbb{N}$ such that for every $\epsilon, \delta \in (0,1]$, every distribution P over $X \times Y$, and every $h \in \mathcal{H}$, we have that with probability at least $1 - \delta$ over an i.i.d. sample $S \sim P^m_X$ of size $m \geq m(\epsilon, \delta)$*

$$\mathcal{A}(S, \mathcal{O}_h) \; \in \; [\mathcal{L}^{\mathcal{U}}_P(h) - \epsilon, \mathcal{L}^{\mathcal{V}}_P(h) + \epsilon]$$

with a query budget of $q(\epsilon, \delta)$ for $\mathcal{A}$. In this case, we say that $\mathcal{H}$ admits tolerant (m, q) blackbox query certification.

Observation 20. *Let $\mathcal{U}(x) = \{z : \|x - z\|_2 \leq 1\}$ and $\mathcal{V}(x) = \{z : \|x - z\|_2 \leq 1 + \gamma\}$. Let $\mathcal{H}$ be the set of all half-spaces in $\mathbb{R}^2$. Then $\mathcal{H}$ is $(O(1/\epsilon^2), O(1/\sqrt{\gamma}\epsilon^2))$ tolerantly certifiable with respect to $\mathcal{U}$ and $\mathcal{V}$.*

Proof sketch. For each x, we can always find a regular polygon with $O(\pi/\sqrt{\gamma})$ vertices that "sits" between the $\mathcal{U}(x)$ and $\mathcal{V}(x)$. Therefore, in order to find out whether x is adversarially vulnerable or not, it would suffice to make $O(\pi/\sqrt{\gamma})$ queries. Combining this with Hoeffding's inequality shows that if we sample $1/\epsilon^2$ points from P and make $O(\pi/\sqrt{\gamma})$ queries for each, we can estimate $\mathcal{L}^{\mathcal{U},\mathcal{V}}_P(h)$ within error ϵ. □

Though more realistic, even the tolerant notion of certifiability does not make all seemingly simple classes certifiable.

Observation 21. *Let $\mathcal{U}(x) = \{z : \|x - z\|^2 \leq 1\}$ and $\mathcal{V}(x) = \{z : \|x - z\|^2 \leq 1 + \gamma\}$. There exists a hypothesis class $\mathcal{H}$ with VC-dimension 1, such that $\mathcal{H}$ is not tolerantly certifiable with respect to $\mathcal{U}$ and $\mathcal{V}$.*

Proof sketch. For any $p \in \mathbb{R}^2$, let $h_p(x) = \mathbb{1}[x = p]$. Let $\mathcal{H} = \{h_p : p \in \mathbb{R}^2\}$. $\mathcal{H}$ clearly has a VC-dimension of 1, but we claim that it is not tolerantly certifiable. We construct an argument similar to Observation 17. The idea is that no matter what queries the certifier chooses, we can always set p to be a point that was not queried and is either inside $\mathcal{U}$ or outside $\mathcal{V}$ depending on the certifier's answer. □

4.1. Witness Sets for Certification

A common observation in the previous examples was that if the certifier could identify a set of points whose labels determined the points in S that were in the margin of h, then querying those points was enough for robust certification. This motivates the following definition.

Definition 22 (Witness sets). *Given a hypothesis class $\mathcal{H}$ and a perturbation type $\mathcal{U}$, for any point $x \in X$, we say that $w(x) \subset X$ is a* witness set *for x if there exists a mapping $f : \{0,1\}^{w(x)} \to \{0,1\}$ such that for any hypothesis $h \in \mathcal{H}$, $f(h|_{w(x)}) = 1$ if and only if x lies in the margin of h (where $h|_{w(x)}$ denotes the restriction of hypothesis h to set $w(x)$).*

Clearly, all positive examples above were created using witness sets. The following theorem identifies a large class of $\mathcal{H}$,$\mathcal{U}$ pairs that exhibit finite witness sets.

Theorem 23. *For any $x \in X$, consider two partial orderings $\prec_0^x$ and $\prec_1^x$ over the elements of $\mathcal{H}$ where for $h_1, h_2 \in \mathcal{H}$, we say $h_1 \prec_1^x h_2$ if $\mathcal{U}(x) \cap h_1 \subset \mathcal{U}(x) \cap h_2$, and $h_1 \prec_0^x h_2$ if $\mathcal{U}(x) \setminus h_1 \subset \mathcal{U}(x) \setminus h_2$. For both partial orderings we identify (as equivalent) hypotheses where these intersections co-incide and further we remove all hypotheses where the intersections are empty.* [3] *If both partial orders have finite number of minima for each x, then the pair $\mathcal{H}$,$\mathcal{U}$ exhibits a finite witness set and hence is certifiable.*

Proof. For this proof we will identify hypotheses with their equivalence classes in each partial ordering. Let $\mathcal{M}_0(x) \subset \mathcal{H}$ be the set of minima for $\prec_0^x$ and $\mathcal{M}_1(x) \subset \mathcal{H}$ for $\prec_1^x$. For each $h \in \mathcal{M}_0(x)$, we pick a point $x' \in \mathcal{U}(x)$ such that $x' \in h$ but $x' \notin h'$ for any $h' \succ h$, thus forming a set $w_0(x)$. Similarly, we define the set $w_1(x)$. We claim that $w(x) = w_0(x) \cup w_1(x) \cup \{x\}$ is a witness set for x, i.e., we can determine whether x is in the margin of any hypothesis $h \in \mathcal{H}$ by looking at labels that h assigns to points in $w(x)$.

We only consider the case where $h(x) = 0$, since the $h(x) = 1$ case is similar. We claim that x is in the margin of h if and only if there exists a point in $w_1(x)$ that is assigned the label 1 by h. Indeed, suppose there exists such a point. Then since the point lies in $\mathcal{U}(x)$ and is assigned the opposite label as x by h, x must lie in the margin of h. For the other direction, suppose x lies in the margin of h. Then there must exist a point $x' \in \mathcal{U}(x)$ such that $h(x') = 1$, which means there must be a hypothesis $\hat{h} \in \mathcal{M}_1(x)$ such that $\hat{h} \prec_1^x h$, which means there must exist $\hat{x} \in w_1(x)$ such that $h(\hat{x}) = 1$. □

We can easily verify, for example, that for the $(\mathcal{H},\mathcal{U})$ pair defined in Observation 16, the set of minima defined by the partial orderings above is finite. Indeed, the (equivalence class of) half-spaces corresponding to the four corners of the unit cube constitute the minima.

4.2. Query complexity of adversarial attacks and its connection to robust PAC learning

Even though in the literature on (practical) adversarial attacks an adversary is often modelled as an actual algorithm, in the theoretical literature the focus has been on whether adversarial examples merely exist[4]. However, one can say a robust learner is successful if it merely finds a hypothesis that is potentially non-robust in the conventional sense yet whose adversarial examples are hard to find for the adversary. To formalize this idea, one needs to define some notion of "bounded adversary" in a way that enables the study of the complexity of finding adversarial examples. Attempts have been made at studying computationally bounded adversaries in certain scenarios (Garg et al., 2019) but not in the distribution-free setting. Here we study an adversary's *query complexity*. We start by formally defining an adversary and discussing a few different properties of adversaries.

Definition 24. *For an $(\mathcal{H},\mathcal{U})$ pair, an* adversary *is an algorithm $\mathcal{A}$ tasked with the following: given a set S of n points from the domain, and query access to a hypothesis $h \in \mathcal{H}$, return a set S' such that (i) each point $x' \in S'$ is an adversarial point to some point in S and (ii) for every $x \in S$ that has an adversarial point, there exists $x' \in S'$ such that x' is an adversarial point for x. If the two conditions hold we call S' an* admissible attack *on S w.r.t. $(\mathcal{H},\mathcal{U})$. We call an adversary* perfect *for $(\mathcal{H},\mathcal{U})$ if for every $S \subset X$ it outputs an admissible attack on S. We say that the adversary is* proper *if all its queries are in the set $\bigcup_{x \in S} \mathcal{U}(x)$.*

There have been (successful) attempts (Papernot et al., 2017; Brendel et al., 2018) at attacking trained neural network models where the adversary was not given any information about the gradients, and had to rely solely on black-box queries to the model. Our definition of the adversary fits those scenarios. Next, we define the query complexity of the adversary.

Definition 25. *If, for $(\mathcal{H},\mathcal{U})$, there is a function $f : \mathbb{N} \to \mathbb{N}$ such that, for any $h \in \mathcal{H}$ and any set S, the adversary $\mathcal{A}$ will produce an admissible attack S' after at most $f(|S|)$ queries,*

[3] Here, we think of hypotheses h_1 and h_2 as the pre-image of 1 (as noted in Section 2), and hence subsets of X.

[4] E.g., the definition of adversarial loss in Section 2 is only concerned with whether an adversarial point exists.

we say that adversary $\mathcal{A}$ has query complexity *bounded by f on $(\mathcal{H},\mathcal{U})$. We say that the adversary is* efficient *if $f(n)$ is linear in n.*

Note that it is possible that the adversary's queries are adaptive, i.e., the i^{th} point it queries depends on the output of its first $i-1$ queries. A weaker version of an adversary is one where that is not the case.

Definition 26. *An adversary is called* non-adaptive *if the set of points it queries is uniquely determined by the set S before making any queries to h.*

Intuitively, there is a connection between perfect adversaries and witness sets because a witness set merely helps identify the points in S that have adversarial points, whereas an adversary finds those adversarial points.

Observation 27. *If the $(\mathcal{H},\mathcal{U})$ pair exhibits a perfect, non-adaptive adversary with query complexity $f(n)$, then it also has witness sets of size $f(n)$.*

Finally, we tie everything together by showing that, for *properly compressible classes*, the existence of a proper, perfect adversary implies that the robust learning problem has a small sample complexity. We say a class $\mathcal{H}$ is properly compressible if *(i)* it admits a sample compression scheme (Littlestone & Warmuth, 1986) of size $O(\mathrm{VC}\cdot\log(n))$—where VC is the VC-dimension of $\mathcal{H}$ and n is the number of samples that are being compressed—, and *(ii)* the hypothesis outputted by the scheme is always a member of $\mathcal{H}$. Note that (i) holds for all hypothesis classes (by a boosting-based compression scheme (Schapire & Freund, 2013; Moran & Yehudayoff, 2016)). Furthermore, many natural classes are shown to have proper compression schemes, and it is open if this is true for all VC-classes. (see (Floyd & Warmuth, 1995; Ben-David & Litman, 1998)).

Theorem 28. *Assume $\mathcal{H}$ is properly compressible. If the robust learning problem defined by $(\mathcal{H},\mathcal{U})$ has a perfect, proper, and efficient adversary, then in the robust realizable-case ($\mathcal{L}^{\mathcal{U}}_{P}(\mathcal{H})=0$) it can be robustly learned with $O(t\log^2(t))$ samples, where $t=\mathrm{VC}(\mathcal{H})/\epsilon^2$.*

Proof Sketch. We adapt the compression-based approach of (Montasser et al., 2019) to prove the result. Let us assume that we are given a sample S of size $|S|=m$ that is labeled by some $h\in\mathcal{H}$, and want to "compress" this sample using a small subset $K\subseteq S$. Let us assume that $\hat{h}$ is the hypothesis that is reconstructed using K. For the compression to succeed, we need to have $\ell^{\mathcal{U}}(h,x,y)=\ell^{\mathcal{U}}(\hat{h},x,y)$ for every $(x,y)\in S$. Given the perfect proper efficient adversary, we can find all the adversarial points in S using C queries per point in S, for some constant $C\geq 0$. In fact, we can amend the points corresponding to these queries to S to create an inflated set, which we call T. Note that $|T|\leq C\cdot|S|$. Furthermore, we can now replace the condition $\forall(x,y)\in S, \ell^{\mathcal{U}}(h,x,y)=\ell^{\mathcal{U}}(\hat{h},x,y)$ with $\forall(x,y)\in T, \ell^{0/1}(h,x,y)=\ell^{0/1}(\hat{h},x,y)$ (the latter implies the former because of the definition of perfect adversary). Therefore, our task becomes compressing T with respect to the standard binary loss, for which we will invoke the assumption that $\mathcal{H}$ has a proper compression scheme. Assume this compression scheme compressed T to a subset $K\subset T$ of size k. We argue that, by invoking the adversary we can convert the compressed set to only include points form the original sample S, and some additional bits as side information. For each point $x\in K$ in the compressed set that is an original sample point from S, we know that $x\in\mathcal{U}(x_S)$ for some $x_S\in S$, since the adversary is proper. We construct a new compressed set $K'\subset S$, by replacing such points x with their corresponding points x_S and bits b to encode the rank of the point x among the queries that the adversary would make for x_S. Now, the decompressor can first recover the set K by invoking the adversary, and then use the standard decompression. Finally, the size of the compression is $O(\log(m))VC(\mathcal{H})$ and the results follows from the classic connection of compression and learning (Littlestone & Warmuth, 1986; Montasser et al., 2019). □

This result shows an interesting connection between the difficulty of finding adversarial examples and that of robust learning. In particular, if the adversarial points can be found easily (at least when measured by query complexity), then robust learning is almost as easy as non-robust learning (in the sense of agnostic sample complexity). Or, stated in the contrapositive, if robust learning is hard, then even if adversarial points exist, finding them is going to be hard. It is possible to further extend the result to the agnostic learning scenario, using the same reduction from agnostic learning to realizable learning that was proposed by (David et al., 2016) and used in (Montasser et al., 2019).

5. Conclusion

We formalized the problem of black-box certification and its relation to an adversary with bounded query budget. We showed the existence of an adversary with small query complexity implies small sample complexity for robust learning. This suggests that the apparent hardness of robust learning – compared to standard PAC learning – in terms of sample complexity may not actually matter as long as we are dealing with bounded adversaries. It would be interesting to explore other types of adversaries (e.g., non-proper and/or non-perfect) to see if they lead to efficient robust learners as well. Another interesting direction is finding scenarios where finite unlabeled data can substitute the knowledge of the marginal distribution discussed in Section 3.

Acknowledgements

We thank the Vector Institute for providing us with the meeting space in which this work was developed! Ruth Urner and Hassan Ashtiani were supported by NSERC Discovery Grants.

References

Akhtar, N. and Mian, A. Threat of adversarial attacks on deep learning in computer vision: A survey. *IEEE Access*, 6:14410–14430, 2018.

Alayrac, J., Uesato, J., Huang, P., Fawzi, A., Stanforth, R., and Kohli, P. Are labels required for improving adversarial robustness? In *Advances in Neural Information Processing Systems 32, NeurIPS*, pp. 12192–12202, 2019.

Attias, I., Kontorovich, A., and Mansour, Y. Improved generalization bounds for robust learning. In *Algorithmic Learning Theory, ALT*, pp. 162–183, 2019.

Awasthi, P., Dutta, A., and Vijayaraghavan, A. On robustness to adversarial examples and polynomial optimization. In *Advances in Neural Information Processing Systems, NeurIPS*, pp. 13760–13770, 2019.

Ben-David, S. and Litman, A. Combinatorial variability of vapnik-chervonenkis classes with applications to sample compression schemes. *Discrete Applied Mathematics*, 86 (1):3–25, 1998.

Ben-David, S., Itai, A., and Kushilevitz, E. Learning by distances. *Inf. Comput.*, 117(2):240–250, 1995.

Blumer, A., Ehrenfeucht, A., Haussler, D., and Warmuth, M. K. Learnability and the vapnik-chervonenkis dimension. *Journal of the ACM (JACM)*, 36(4):929–965, 1989.

Brendel, W., Rauber, J., and Bethge, M. Decision-based adversarial attacks: Reliable attacks against black-box machine learning models. 2018.

Bubeck, S., Lee, Y. T., Price, E., and Razenshteyn, I. P. Adversarial examples from computational constraints. In *Proceedings of the 36th International Conference on Machine Learning, ICML*, pp. 831–840, 2019.

Carmon, Y., Raghunathan, A., Schmidt, L., Duchi, J. C., and Liang, P. Unlabeled data improves adversarial robustness. In *Advances in Neural Information Processing Systems 32, NeurIPS*, pp. 11190–11201, 2019.

Chakraborty, A., Alam, M., Dey, V., Chattopadhyay, A., and Mukhopadhyay, D. Adversarial attacks and defences: A survey. *CoRR*, abs/1810.00069, 2018.

Chen, P.-Y., Zhang, H., Sharma, Y., Yi, J., and Hsieh, C.-J. Zoo: Zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 15–26, 2017.

Cohen, J. M., Rosenfeld, E., and Kolter, J. Z. Certified adversarial robustness via randomized smoothing. In *Proceedings of the 36th International Conference on Machine Learning, ICML*, pp. 1310–1320, 2019.

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

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

Diochnos, D., Mahloujifar, S., and Mahmoody, M. Adversarial risk and robustness: General definitions and implications for the uniform distribution. In *Advances in Neural Information Processing Systems 31, NeurIPS*, pp. 10359–10368, 2018.

Diochnos, D. I., Mahloujifar, S., and Mahmoody, M. Lower bounds for adversarially robust PAC learning. *CoRR*, abs/1906.05815, 2019.

Dong, Y., Liao, F., Pang, T., Su, H., Zhu, J., Hu, X., and Li, J. Boosting adversarial attacks with momentum. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 9185–9193, 2018.

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

Feldman, V. A general characterization of the statistical query complexity. In *Proceedings of the 30th Conference on Learning Theory, COLT*, pp. 785–830, 2017.

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

Garg, S., Jha, S., Mahloujifar, S., and Mahmoody, M. Adversarially robust learning could leverage computational hardness. *CoRR*, abs/1905.11564, 2019.

Goodfellow, I. J., McDaniel, P. D., and Papernot, N. Making machine learning robust against adversarial inputs. *Commun. ACM*, 61(7):56–66, 2018.

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

Gourdeau, P., Kanade, V., Kwiatkowska, M., and Worrell, J. On the hardness of robust classification. In *Advances in Neural Information Processing Systems 32, NeurIPS*, pp. 7444–7453, 2019.

Haussler, D. and Welzl, E. epsilon-nets and simplex range queries. *Discret. Comput. Geom.*, 2:127–151, 1987.

Kearns, M. J. Efficient noise-tolerant learning from statistical queries. *J. ACM*, 45(6):983–1006, 1998.

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

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *6th International Conference on Learning Representations, ICLR*, 2018.

Montasser, O., Hanneke, S., and Srebro, N. VC classes are adversarially robustly learnable, but only improperly. In *Conference on Learning Theory, COLT*, pp. 2512–2530, 2019.

Montasser, O., Goel, S., Diakonikolas, I., and Srebro, N. Efficiently learning adversarially robust halfspaces with noise. *arXiv preprint arXiv:2005.07652*, 2020.

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

Narodytska, N. and Kasiviswanathan, S. Simple black-box adversarial attacks on deep neural networks. In *2017 IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pp. 1310–1318. IEEE, 2017.

Papernot, N., McDaniel, P., Goodfellow, I., Jha, S., Celik, Z. B., and Swami, A. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia conference on computer and communications security*, pp. 506–519, 2017.

Salman, H., Li, J., Razenshteyn, I. P., Zhang, P., Zhang, H., Bubeck, S., and Yang, G. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems 32, NeurIPS*, pp. 11289–11300, 2019.

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

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. In *Advances in Neural Information Processing Systems, NeurIPS*, pp. 5014–5026, 2018.

Shalev-Shwartz, S. and Ben-David, S. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge University Press, 2014.

Sinha, A., Namkoong, H., and Duchi, J. C. Certifying some distributional robustness with principled adversarial training. In *6th International Conference on Learning Representations, ICLR*, 2018.

Su, J., Vargas, D. V., and Sakurai, K. One pixel attack for fooling deep neural networks. *IEEE Transactions on Evolutionary Computation*, 23(5):828–841, 2019.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I. J., and Fergus, R. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR*, 2014.

Valiant, L. G. A theory of the learnable. *Commun. ACM*, 27 (11):1134–1142, 1984.

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

Wang, Y., Jha, S., and Chaudhuri, K. Analyzing the robustness of nearest neighbors to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning, ICML*, pp. 5120–5129, 2018.

Wong, E. and Kolter, J. Z. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *Proceedings of the 35th International Conference on Machine Learning, ICML*, pp. 5283–5292, 2018.

Yang, Y., Rashtchian, C., Wang, Y., and Chaudhuri, K. Adversarial examples for non-parametric methods: Attacks, defenses and large sample limits. *CoRR*, abs/1906.03310, 2019.

Yin, D., Ramchandran, K., and Bartlett, P. L. Rademacher complexity for adversarially robust generalization. In *Proceedings of the 36th International Conference on Machine Learning,ICML*, pp. 7085–7094, 2019.