Lower Bounds for Adversarially Robust PAC Learning under Evasion and Hybrid Attacks

Dimitrios I. Diochnos*
University of Oklahoma
diochnos@ou.edu

Saeed Mahloujifar*
Princeton
sfar@princeton.edu

Mohammad Mahmoody
University of Virginia
mohammad@virginia.edu

***Abstract*—In this work, we study probably approximately correct (PAC) learning under general perturbation-based adversarial attacks. In the most basic setting, referred to as an evasion attack, the adversary's goal is to misclassify an honestly sampled point x by adversarially perturbing it into $\widetilde{x}$, i.e., $h(\widetilde{x}) \neq c(\widetilde{x})$, where c is the ground truth concept and h is the learned hypothesis. The only limitation on the adversary is that $\widetilde{x}$ is not "too far" from x, controlled by a metric measure.**

We first prove that for many theoretically natural input spaces of high dimension n (e.g., isotropic Gaussian in dimension n under ℓ_2 perturbations), if the adversary is allowed to apply up to a *sublinear* amount of perturbations in the expected norm, PAC learning requires sample complexity that is *exponential* in the data dimension n. We then formalize *hybrid* attacks in which the evasion attack is preceded by a poisoning attack in which a poisoning phase is followed by specific evasion attacks. *Special* forms of hybrid attacks include so-called "backdoor attacks" but here we focus on the general setting in which adversary's evasion attack is only controlled by a pre-specified amount of perturbation based on data dimension and aim to misclassifying the perturbed instances. We show that PAC learning is sometimes *impossible* under such hybrid attacks, while it is possible without the attack (e.g., due to the bounded VC dimension).

I. Introduction

Learning predictors is the task of outputting a hypothesis h using a training set $\mathcal{S}$ in such a way that h can predict the correct label $c(x)$ of unseen instances with high probability. A successful learner, however, could be vulnerable to adversarial perturbations. In particular, it was shown (Szegedy et al., 2014) that deep neural nets (DNNs) are vulnerable to so-called adversarial examples that are the result of small (even imperceptible to human eyes) perturbations on the original input x. Since the introduction of such attacks, many works have studied defenses as well as newer attacks (Biggio et al., 2013; Goodfellow et al., 2015; Papernot et al., 2016a; Carlini and Wagner, 2017; Xu et al., 2017; Madry et al., 2017).

A fundamental question in robust learning is whether one can design learning algorithms that achieve "uniform convergence" even under such adversarial perturbations. Namely, we want to know when we can learn a robust classifier h that still correctly classifies its inputs even if they are adversarially perturbed in a limited way. Indeed, one can ask when (ε, δ) PAC (probably approximately correct) learning (Valiant, 1984) is possible in adversarial settings. More formally, the goal here is to learn a robust h from the data set $\mathcal{S}$ consisting of m independently sampled labeled (non-adversarial) instances in such a way that, with probability $1-\delta$ over the learning process, the produced h has error at most ε even under "limited" adversarial perturbations of the input. This limitation is carefully defined by some metric d defined over the input space $\mathcal{X}$ and some upper bound "budget" b on the amount of perturbations that the adversary can introduce. That is, we would like to minimize "adversarial" risk defined as follows

$$\mathsf{AdvRisk}(h) = \Pr_{x \leftarrow D}[\exists\ \widetilde{x}\colon d(x, \widetilde{x}) \leq b, h(\widetilde{x}) \neq c(\widetilde{x})] \leq \varepsilon$$

where $c(\cdot)$ is the ground truth (i.e., the concept function).

Error-region adversarial risk. The above notion of adversarial risk has been used implicitly or explicitly in previous work (Gilmer et al., 2018; Diochnos et al., 2018; Degwekar and Vaikuntanathan, 2019; Ford et al., 2019) and here we use the version as formalized by Diochnos et al. (2018) known as the "error-region" adversarial risk, because the adversary's goal here is to push $\widetilde{x}$ into the error region

$$\mathcal{E} = \{x \mid h(x) \neq c(x)\}$$

regardless of whether or not the ground truth $c(x)$ is robust under b perturbations or not. In other words, the above notion captures the vulnerability of a classifier against adversaries of budget b for all values of b.

Hence we can define the PAC Learning counterpart of this definition. This leads us to our central question:

> *What problems are PAC learnable under evasion attacks that perturb instances into the error region? If PAC learnable, what is their sample complexity?*

Hybrid attacks. Another attack model are Hybrid attacks that we formalize in this work. Hybrid attacks are closely related to the notion of poisoning attacks where adversary interferes the training phase with the goal of hurting the performance of the resulting classifier. Hybrid attacks consist of two adversaries that work together. A poisoning adversary that adds a few training examples to the training set and an evasion adversary who perturbs the instances fed to the resulting classifier. This type of attacks that are also known as backdoor attacks , could potentially be more devastating and increase the required sample complexity for PAC Learning. We ask the following question in presence of Hybrid attacks.

> *How much data is required to PAC learn a classifier that is robust to adversarial hybrid attacks?*

* First two authors have contributed equally.

A. Our Contribution

In this work, we initiate a formal study of PAC learning under adversarial perturbations, where the goal of the adversary is to increase the error-region adversarial risk using small (sublinear $o(\|x\|)$) perturbations of the inputs x. Therefore, in what follows, whenever we refer to adversarial risk, by default it means the error-region variant (In next section we discuss another mainstream definiton of adversarial risk called corrupted input definition). Bellow, we will shortly describe our two lower bound on PAC learning using error-region risk. Before we proceed, so that we can better put our work into perspective, we first give a short description explaining our main contributions in previous work that we have done that is related to the work of this paper.

Result 1: exponential lower bound on sample complexity. Suppose the instances of a learning problem come from a metric probability space $(\mathcal{X}, D, \mathsf{d})$ where D is a distribution and d is a metric defining some norm $\|\cdot\|$. Suppose the input instances have norms $\|x\| \approx n$ where n is a parameter related (or is in fact equal) to the data dimension. One natural setting of study for PAC learning is to study attackers that can only perturb x by a *sublinear* amount $o(\|x\|) = o(n)$ (e.g., $\sqrt{n}$).

Our first result is to prove that for many theoretically natural input spaces of high dimension n (e.g., isotropic Gaussian in dimension n under ℓ_2 perturbations), PAC learning of certain problems under sublinear perturbations of the test instances requires *exponentially* many samples in n, even though the problem in the no-attack setting is PAC learnable using polynomially many samples. This holds e.g., when we want to learn half spaces in dimension n under such distributions (which is possible in the no-attack setting).

We note that even though PAC learning is defined for all distributions, proving such lower bound for a specific input distribution D over $\mathcal{X}$ only makes the negative result *stronger*. Our lower bound is in contrast with previously proved results (Attias et al., 2018; Bubeck et al., 2018; Montasser et al., 2019; Cullina et al., 2018) in which the gap between the sample complexity of the normal and robust learning is only *polynomial*. However, as mentioned before, all these previous results are proved using the *corrupted-input* variant of adversarial risk.

Our result extends to any learning problem where input space $\mathcal{X}$, the metric d and the distribution D, and the class of concept functions $\mathcal{C}$ have the following two conditions.

1) The inputs $\mathcal{X}$ under the distribution D and small perturbations measured by the metric d forms a *concentrated* metric probability space (Ledoux, 2001; Milman and Schechtman, 1986). A concentrated space has the property that relatively small events (e.g., of measure 0.1) under small (e.g., smaller than the diameter of the space) perturbations expand to cover ≈ 1 measure of inputs.
2) The set of concept functions $\mathcal{C}$ is complex enough to allow proving lower bounds for the sample complexity for (distribution-dependent) PAC learners in the *no-attack* setting under the *same distribution* D. Distribution-dependent sample complexity lower bounds are known for certain settings (Long, 1995; Balcan and Long, 2013; Sabato et al., 2013), however, we use a more relaxed condition that can be applied to broader settings. In particular, we require that for a sufficiently small ε, there are two concept functions c_1, c_2 that are equal for $1-\varepsilon$ fraction of inputs sampled from D (see Definition IV.3).

Having the above two conditions, our proof proceeds as follows **(I)** We show that the (normal) risk $\mathsf{Risk}(h)$ of a hypothesis produced by *any* learning algorithm with sub-exponential sample complexity cannot be as large as an inverse polynomial over the dimension. **(II)** We then use ideas from (Mahloujifar et al., 2018a) to show that such sufficiently large risk will expand into a large *adversarial* risk of almost all inputs, due to the measure concentration the input space.

Remark: realizablity under the error-region definition: If a learning problem is *realizable* in the no-attack setting, i.e., there is a hypothesis h that has risk zero over the test instances, it means that the same hypothesis h will have adversarial (true) risk zero over the test instances as well, because any perturbed point is still going to be correctly classified. This is in contrast with corrupted-input notion of adversarial risk that even in realizable problems, the smallest corrupted-input (true) adversarial risk could still be large, and even at odds with correctness (Tsipras et al., 2018). This means that our results rule out (efficient) PAC learning even in the *agnostic* setting as well, because in the realizable setting there is at least one hypothesis with error-region adversarial risk zero while (as we prove), in some settings learning a model with adversarial risk (under sublinear perturbations) close to zero requires exponentially many samples.

Result 2: ruling out PAC learning under hybrid attacks. We then study PAC learning under adversarial perturbations that happen during *both* training and testing phases. We formalize *hybrid* attacks in which the final evasion attack is preceded by a poisoning attack (Biggio et al., 2012; Papernot et al., 2016b). This attack model bears similarities to "trapdoor attacks" (Gu et al., 2017) in which a poisoning phase is involved before the evasion attack, and here we give a formal definition for PAC learning under such attacks. Our definition of hybrid attacks is general and can incorporate any notion of adversarial risk, but our results for hybrid attacks use the *error-region* adversarial risk.

Under hybrid attacks, we show that PAC learning is sometimes *impossible* all together, even though it is possible without such attacks. For example, even if the VC dimension of the concept class is bounded by n, if the adversary is allowed to poison only $1/n^{10}$ fraction of the m training examples, then it can do so in such a way that a subsequent evasion attack could then increase the adversarial risk to ≈ 1. This means that PAC learning is in fact impossible under such hybrid attacks.

We also note that classical results about malicious noise (Valiant, 1985; Kearns and Li, 1993) and nasty noise (Bshouty et al., 2002) could be interpreted as ruling out PAC learning under poisoning attacks. However, there are two differences: **(I)** The adversary in these previous works needs to change a *constant* fraction of the training examples, while our attacker

changes only an *arbitrarily small* inverse polynomial fraction of them. **(II)** Our poisoning attacker only *removes* a fraction of the training set, and hence it does *not* add any misclassified examples to the pool. Thus this poisoning attack uses clean labels only (Mahloujifar et al., 2018b; Shafahi et al., 2018).

II. Comparison with Other Definitions of Adversarial Risk and Related Work

Corrupted-input adversarial risk. Another notion of adversarial risk (that is similar, but still different from the error-region adversarial risk explained above) has been used in many works such as (Feige et al., 2015; Madry et al., 2017; Bubeck et al., 2018) in which the perturbed $\widetilde{x}$ is interpreted as a "corrupted input". Namely, here the goal of the learner is to find the label of the original *untampered* point x by only having its corrupted version $\widetilde{x}$, and thus adversary's success criterion is to reach $d(x,\widetilde{x}) \leq b, h(\widetilde{x}) \neq c(x)$. Hence, in that setting, the goal of the learner is to find an h that minimizes

$$\Pr_{x \leftarrow D}[\exists\ \widetilde{x} \colon d(x,\widetilde{x}) \leq b, h(\widetilde{x}) \neq c(x)].$$

It is easy to see that, if the ground truth $c(x)$ does not change under b-perturbations, $c(x) = c(\widetilde{x})$, the two notions of error-region and corrupted-input adversarial risk will be equal. In particular, this is the case for practical distributions of interest, such as images or voice, where sufficiently-small perturbations do not change human's judgment about the true label. However, if b-perturbations can change the ground truth, $c(x) \neq c(\widetilde{x})$, the two definitions are incomparable.

Several works have already studied PAC learning under adversarial perturbations (Bubeck et al., 2018; Cullina et al., 2018; Feige et al., 2018; Attias et al., 2018; Khim and Loh, 2018; Yin et al., 2018; Montasser et al., 2019). However, all these works use the *corrupted-input* notion of adversarial risk. In particular, it is proved by Attias et al. (2018) that robust learning might require more data, but it was also shown by Attias et al. (2018); Bubeck et al. (2018) that in natural settings, if robust classification is feasible, robust classifiers could be found with a sample complexity that is only *polynomially* larger than that of normal learning.

Comparison of the error-region definition and corrupted input definition has been the focus of multiple studies. Diochnos et al. (2018) compares different definitions and picks the error-region definition as the right notion as it guarantees misclassification of the adversarial examples. Gourdeau et al. (2019) also study these two definitions from a PAC learning perspective, for multiple problems. They show that in many interesting scenarios, the behavior of the corrupted input definition is not as expected. For instance, even if a learning algorithm manages to learn the exact same classifier, the corrupted input risk will not be 0 and there are other classifiers with lower corrupted input risk. It was also shown in Chen et al. (2020) that the corrupted input definition can have surprising behaviors when more data is provided to the learning algorithm. In particular, when training a half-space to separate a mixture of Gaussian using the corrupted input risk, the error of the resulting half-space would drop at first and then start increasing, and potentially start to drop again, for some adversarial budgets. This is an evidence of why corrupted instance might not be the right definition for certain learning problems. There are also other results that show trade-off between accuracy and robustness when working with the corrupted-input definition (Tsipras et al., 2018). However, for the case of error-region adversarial risk, these types of trade-off do not exist and the ground-truth is the most robust classifier. As argued by Suggala et al. (2019), this is another evidence that studying PAC learning with the error-region definition is more meaningful.

Note that previous positive (or negative) results about PAC learning under the corrupted-input definition do not answer our question above, as we study general arbitrary perturbation budgets allowed to the adversary. Also, when the ground truth can also change under that amount of perturbation we have to use the error-region definition. More technically, we note that positive results about adversarial PAC learning (cited above) do not answer our question for the following reason. When the allowed perturbation is limited to keep the ground truth c robust, then the two definitions are equivalent, yet, when the budget gets larger, then a positive result proved using the corrupted-input definition would simply mean that there is a way to learn a hypothesis h that has only ε adversarial risk more than the "best possible" h^*. However, this could be just a side-effect that any h^* under the corrupted-input definition (and certain amount of allowed perturbations) could have very large (even $1-\varepsilon$) adversarial risk, making the job of agnostic learning trivial (to output anything). That is why, when we work with arbitrary perturbation budget, we need to employ the error-region definition, which still allows $c = h$ to have small adversarial risk, which is the intuitive decision as well.

III. Adversarially Robust PAC Learning

Notation. By $\widetilde{O}(f(n))$ we refer to the set of all functions of the form $O(f(n)\log(f(n))^{O(1)})$. We use capital calligraphic letters (e.g., $\mathcal{D}$) for sets and capital non-calligraphic letters (e.g., D) for distributions. $x \leftarrow D$ denotes sampling x from D. For an event $\mathcal{S}$, we let $D(\mathcal{S}) = \Pr_{x \leftarrow D}[x \in \mathcal{S}]$.

A classification problem $\mathcal{P} = (\mathcal{X}, \mathcal{Y}, \mathcal{C}, \mathcal{D}, \mathcal{H})$ is specified by the following components. The set $\mathcal{X}$ is the set of possible *instances*, $\mathcal{Y}$ is the set of possible *labels*, $\mathcal{D}$ is a class of distributions over instances $\mathcal{X}$. In the standard setting of PAC learning, $\mathcal{D}$ includes all distributions, but since we deal with *negative* results, we sometimes work with fixed $\mathcal{D} = \{D\}$ distributions, and show that even *distribution-dependent* robust PAC learning is sometimes hard. In that case, we represent the problem as $\mathcal{P} = (\mathcal{X}, \mathcal{Y}, \mathcal{C}, D, \mathcal{H})$. The set $\mathcal{C} \subseteq \mathcal{Y}^{\mathcal{X}}$ is the *concept class* and $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is the *hypothesis class*. In general, we can allow *randomized* concept and hypothesis functions to model, in order, label uncertainly (usually modeled by a joint distribution over instances and labels) and randomized predictions. All of our results extend to randomized learners and randomized hypothesis functions, but for simplicity of presentation, we treat them as deterministic

mappings. By default, we consider 0-1 *loss functions* where $\ell oss(y', y) = \mathbb{1}[y' \neq y]$. For a given distribution $D \in \mathcal{D}$ and a concept function $c \in \mathcal{C}$, the *risk* of a hypothesis $h \in \mathcal{H}$ is the expected loss of h with respect to D, namely $\mathsf{Risk}(D, c, h) = \Pr_{x \leftarrow D}[\ell oss(h(x), c(x))]$. An *example* z is a pair $z = (x, y)$ where $x \in \mathcal{X}$ and $y \in \mathcal{Y}$. An example is usually sampled by first sampling $x \leftarrow D$ for some $D \in \mathcal{D}$ followed by letting $y = c(x)$ for some $c \in \mathcal{C}$. A *sample* $\mathcal{S} = (z_1, \ldots, z_m)$ is a sequence of m examples; sometimes we may refer to such a sample sequence as the training *set*. By $\mathcal{S} \leftarrow (D, c(D))^m$ we denote the process of obtaining $\mathcal{S}$ by sampling m iid samples from D and labeling them by c.

Our learning problems $\mathcal{P}_n = (\mathcal{X}_n, \mathcal{Y}_n, \mathcal{C}_n, \mathcal{D}_n, \mathcal{H}_n)$ are usually parameterized by n where n denotes the "data dimension" or (closely) capture the bit length of the instances. Thus, the "efficiency" of the algorithms could depend on n. Even in this case, for simplicity of notation, we might simply write $\mathcal{P} = (\mathcal{X}, \mathcal{Y}, \mathcal{C}, \mathcal{D}, \mathcal{H})$. By default, we will have $\mathcal{C} \subseteq \mathcal{H}$, in which case we call $\mathcal{P}$ *realizable*. This means that for any training set for $c \in \mathcal{C}, D \in \mathcal{D}$, there is a hypothesis $h \in \mathcal{H}$ that has empirical and true risk zero.

Evasion attacks. An evasion attacker A changes the test instance x, denoted as $\widetilde{x} \leftarrow \mathsf{A}(x)$. The behavior and actions taken by A could, in general, depend on the choices of $D \in \mathcal{D}, c \in \mathcal{C}$, and $h \in \mathcal{H}$. As a result, in our notation, we provide A with access to D, c, h by giving them as special inputs to A,[1] denoting the process as $\widetilde{x} \leftarrow \mathsf{A}[D, c, h](x)$. We use calligraphic font $\mathcal{A}$ to denote a *class/set* of attacks; $\mathcal{A}$ could contain all attackers who could change test instance x by at most b perturbations under a metric defined over $\mathcal{X}$.

Poisoning attacks. A poisoning attacker A is one that changes the training sequence as $\widetilde{\mathcal{S}} \leftarrow \mathsf{A}(\mathcal{S})$. Such attacks, in general, might add examples to $\mathcal{S}$, remove examples from $\mathcal{S}$, or do both. The behavior and actions taken by A could, in general, depend on the choices of $D \in \mathcal{D}, c \in \mathcal{C}$ (but not on $h \in \mathcal{H}$, as it is not produced by the learner at the time of the poisoning attack)[2]. As a result, we provide implicit access to D, c by giving them as special inputs to A, denoting the process as $\widetilde{\mathcal{S}} \leftarrow \mathsf{A}[D, c](\mathcal{S})$. We use calligraphic font $\mathcal{A}$ to denote a *class/set* of attacks. For example, $\mathcal{A}$ could contain attacks that change $1/n$ fraction of $\mathcal{S}$ only using clean labels (Mahloujifar et al., 2018a; Shafahi et al., 2018).

Hybrid attacks. A hybrid attack $\mathsf{A} = (\mathsf{A}_1, \mathsf{A}_2)$ is a two phase attack in which A_1 is a poisoning attacker and A_2 is an evasion attacker. One subtle point is that A_2 is also aware of the internal state of A_1, as they are a pair of coordinating attacks. More formally, A_1 outputs an extra "state" information st which will be given as an extra input to A_2. As discussed above, A_1 can depend on D, c, and A_2 can depend on D, c, h as defined for evasion and poisoning attacks.

[1] This dependence is information theoretic, and for example, A might want to find $\widetilde{x}$ that is misclassified, in which case its success is defined as $h(\widetilde{x}) \neq c(\widetilde{x})$ which depends on both h, c.

[2] For example, an attack model might require A to choose its perturbed instances still using *correct/clean* labels, in which case the attack is restricted based on the choice of c)

We now define PAC learning under adversarial perturbation attacks. To do so, we need to first define our notion of adversarial risk. We will do so by employing the *error-region* notion adversarial risk as formalized in Diochnos et al. (2018) adversary aims to misclassify the perturbed instance $\widetilde{x}$.

Definition III.1 (Error-region (adversarial) risk)**.** *Suppose* A *is an evasion adversary and let* D, c, h *be fixed. The* error-region *(adversarial) risk is defined as follows.*

$$\mathsf{AdvRisk}_{\mathsf{A}}(D, c, h) = \Pr_{x \leftarrow D, \widetilde{x} \leftarrow \mathsf{A}[D,c,h](x)}[h(\widetilde{x}) \neq c(\widetilde{x})].$$

For randomized h*, the above probability is also over the randomness of* h *chosen after* $\widetilde{x}$ *is selected.*

Why PAC learning under perturbation is meaningful. We emphasize that, even if the b-perturbation *could* change the ground truth's judgement, asking whether a learning problem is PAC learnable or not is very meaningful. In fact, the problem is still "realizable" under the error region definition of adversarial risk (this is not correct for the other mainstream definition of adversarial risk that we discuss in next section) because if one happens to learn the concept class c exactly and output the hypothesis $h = c$, then h will have adversarial risk *zero* under the error-region definition. In other words, the ground truth can still be *predicted* robustly. Thus, it is natural to ask whether one can learn a hypothesis h that has small adversarial risk even under perturbations that are still small in magnitude compared to the size of the original sample x.

We now define PAC learning under hybrid attacks, from which one can derive also the definition of PAC learning under evasion attacks and under poisoning attacks.

Definition III.2 (PAC learning under hybrid attacks)**.** *Suppose* $\mathcal{P}_n = (\mathcal{X}_n, \mathcal{Y}_n, \mathcal{C}_n, \mathcal{D}_n, \mathcal{H}_n)$ *is a realizable classification problem, and suppose* $\mathcal{A}$ *is a class of hybrid attacks for* $\mathcal{P}_n$*.* $\mathcal{P}_n$ *is PAC learnable with sample complexity* $\mathsf{m}(\varepsilon, \delta, n)$ *under hybrid attacks of* $\mathcal{A}$ *, if there is a learning algorithm* L *such that for every* n*,* $0 < \varepsilon, \delta < 1, c \in \mathcal{C}, D \in \mathcal{D}$*, and* $(\mathsf{A}_1, \mathsf{A}_2) \in \mathcal{A}$*, if* $m = \mathsf{m}(\varepsilon, \delta, n)$*, then*

$$\Pr_{\substack{\mathcal{S} \leftarrow (D, c(D))^m, \\ (\widetilde{S}, \mathsf{st}) \leftarrow \mathsf{A}_1[D,c](\mathcal{S}), \\ h \leftarrow L(\widetilde{S})}} \left[\mathsf{AdvRisk}_{\mathsf{A}_2[D,c,h,\mathsf{st}]}(h, c, D) > \varepsilon\right] \leq \delta.$$

PAC learning under (pure) poisoning attacks or evasion attacks could be derived from Definition III.2 by letting either of A_1 *or* A_2 *be a trivial attack that does no tampering at all.*

We also note that one can obtain other definitions of PAC learning under evasion or hybrid attacks in Definition III.2 by using other forms of adversarial risk, e.g., corrupted-input adversarial risk (Feige et al., 2015, 2018; Madry et al., 2017; Schmidt et al., 2018; Attias et al., 2018)

IV. Lower Bounds for PAC Learning under Evasion and Hybrid Attacks

Before proving our main results, we need to recall the notion of Normal Lévy families, and define a desired and

common property of set of concept functions with respect to the distribution of inputs.

Notation. Let $(\mathcal{X}, \mathsf{d})$ be a metric space. For $\mathcal{S} \subseteq \mathcal{X}$, by $\mathsf{d}(x, \mathcal{S}) = \inf\{\mathsf{d}(x, y) \mid y \in \mathcal{S}\}$ we denote the distance of a point x from $\mathcal{S}$. We also let $\mathcal{S}_b = \{y \mid \mathsf{d}(x, y) \leq b, x \in \mathcal{S}\}$ be the *b-expansion* of $\mathcal{S}$. When there is also a measure D defined over the metric space $(\mathcal{X}, \mathsf{d})$, the *concentration function* is defined and denoted as $\boldsymbol{\alpha}(b) = 1 - \inf\{\Pr_D[\mathcal{E}_b] \mid \Pr_D[\mathcal{E}] \geq 1/2\}$.

Definition IV.1 (Normal Lévy families)**.** *A* family *of metric probability spaces $(\mathcal{X}_n, \mathsf{d}_n, D_n)_{i\in\mathbb{N}}$ with concentration function $\boldsymbol{\alpha}_n(\cdot)$ is called a* normal Lévy family *if there are k_1, k_2, such that*[3] $\boldsymbol{\alpha}_n(b) \leq k_1 \cdot \mathrm{e}^{-k_2 \cdot b^2/n}$.

The following lemma was proved in Mahloujifar et al. (2018a) for Normal Lévy input spaces.

Lemma IV.2. *Let the input space of a hypothesis classifier h be a Normal Lévy family $(\mathcal{X}_n, \mathsf{d}_n, D_n)_{i\in\mathbb{N}}$. If the risk of h with respect to the ground truth concept function c is bigger than α, $\mathsf{Risk}(D_n, c, h) \geq \alpha$, and if an adversary A can perturb instances by up to b in metric d_n for*

$$b = \sqrt{n/k_2} \cdot \left(\sqrt{\ln(k_1/\alpha)} + \sqrt{\ln(k_1/\beta)}\right),$$

then the adversarial risk is $\mathsf{AdvRisk}_{\mathsf{A}}(D, h, c) \geq 1 - \beta$.

Definition IV.3 (α-close function families)**.** *Suppose D is a distribution over $\mathcal{X}$, and let $\mathcal{C}$ be a set of functions from $\mathcal{X}$ to some set $\mathcal{Y}$. We call $\mathcal{C}$* α-close *with respect to D, if there are $c_1, c_2 \in \mathcal{C}$ such that $\Pr_{x\leftarrow D}[c_1(x) \neq c_2(x)] = \alpha$.*

We now state our main results. Theorem IV.4 is stated in the *asymptotic* form considering attack families that attack the problem for sufficiently large index $n \in \mathbb{N}$ of the problem. We describe a quantitative variant afterwards (Lemma IV.5).

Theorem IV.4 (Limits of adversarially robust PAC learning)**.** *Suppose $\mathcal{P}_n = (\mathcal{X}, \mathcal{Y}, \mathcal{C}, \mathcal{D}, \mathcal{H})$ is a realizable classification problem and that $\mathcal{X}$ is a Normal Lévy Family (Definition IV.1) over D and a metric d, and that $\mathcal{C}$ is $\Theta(\alpha)$-close with respect to D for all $\alpha \in [2^{-\Theta(n)}, 1]$. Then, the following hold even for PAC learning with parameters $\varepsilon = 0.9, \delta = 0.49$.*

1) Sample complexity under evasion attacks:
   - *a)* **Exponential lower bound:** *Any PAC learning algorithm that is robust against tampering attacks of budget $b = o(n)$ requires $m \geq 2^{\Omega(n)}$ many samples.*
   - *b)* **Super-polynomial lower bound:** *Any PAC learning algorithm that is robust against tampering attacks of budget $b = \widetilde{O}(\sqrt{n})$ requires $m \geq n^{\omega(1)}$ many samples.*

2) Ruling out PAC learning robust to hybrid attacks: *Suppose the tampering budget of the evasion adversary can be any $b = \widetilde{O}(\sqrt{n})$, and let $\mathcal{B}_\lambda$ be any class of poisoning attacks that can remove $\lambda = \lambda(n)$ fraction of the training examples for an (arbitrary small) inverse polynomial $\lambda(n) \geq 1/\operatorname{poly}(n)$. Let $\mathcal{R}$ be the class of hybrid attacks that first do a poisoning by some $\mathsf{B} \in \mathcal{B}_\lambda$ and then an evasion by some adversary of budget $b = \widetilde{O}(\sqrt{n})$. Then, $\mathcal{P}_n$ is* not *PAC learnable (regardless of sample complexity) under hybrid attacks in $\mathcal{R}$.*

[3]Another common formulation of Normal Lévy families uses $\boldsymbol{\alpha}_n(b) \leq k_1 \cdot \mathrm{e}^{-k_2 \cdot b^2 \cdot n}$, but here we scale the distances up by n to achieve "typical norms" to be $\approx n$, which is the dimension.

In fact, Part 1a and Part 1b of Theorem IV.4 are special cases of the more quantitative lower of the lemma below.

Lemma IV.5. *For the setting of Theorem IV.4, if the tampering budget is $b = \rho \cdot n$, for a fixed function $\rho = \rho(n) = o(1)$, then any PAC learning algorithm for $\mathcal{P}_n$ under evasion attacks of tampering budget $b = b(n)$, even for parameters $\varepsilon = 0.9, \delta = 0.49$ requires sample complexity at least $m(n) \geq 2^{\Omega(\rho^2 \cdot n)}$.*

V. Conclusion

We examined evasion attacks, where the adversary can perturb instances during test time, as well as hybrid attacks where the adversary can perturb instances during both training and test time. For evasion attacks we gave an exponential lower bound on the sample complexity even when the adversary can perturb instances by an amount of $o(n)$, where n is capturing the "typical" norm of an input. For hybrid attacks, PAC learning is ruled out altogether when the adversary can poison a small fraction of the training examples and still perturb the test instance by a sublinear amount $o(n)$.

Our result shows a different behavior when it comes to PAC learning for error-region adversarial risk compared to previously used notions of adversarial robustness based on corrupted inputs. In particular, in the error-region variant of adversarial risk, realizable problems stay realizable, as normal risk zero for a hypothesis h also implies (error-region) adversarial risk zero for the same h. This makes our results more striking, as they apply to agnostic learning as well.

One natural question is if similar results could be proved for corrupted-input adversarial risk. Note that previous work studying learning under corrupted-input adversarial risk (Bubeck et al., 2018; Cullina et al., 2018; Feige et al., 2018; Attias et al., 2018; Khim and Loh, 2018; Yin et al., 2018; Montasser et al., 2019) focus on agnostic learning, by aiming to get close to the "best" robust classifier. However, it is not clear how good the best classifier is. It remains open to find out when we can learn robust classifiers (under corrupted-input risk) in which the *total* adversarial risk is small.

References

C. Szegedy, W. Zaremba, I. Sutskever, J. Bruna, D. Erhan, I. Goodfellow, and R. Fergus, "Intriguing properties of neural networks," in *ICLR*, 2014. [Online]. Available: http://arxiv.org/abs/1312.6199

B. Biggio, I. Corona, D. Maiorca, B. Nelson, N. Srndic, P. Laskov, G. Giacinto, and F. Roli, "Evasion Attacks against Machine Learning at Test Time," in *ECML/PKDD*, 2013, pp. 387–402.

I. Goodfellow, J. Shlens, and C. Szegedy, "Explaining and Harnessing Adversarial Examples," in *ICLR*, 2015. [Online]. Available: http://arxiv.org/abs/1412.6572

N. Papernot, P. D. McDaniel, X. Wu, S. Jha, and A. Swami, "Distillation as a Defense to Adversarial Perturbations Against Deep Neural Networks," in *IEEE Symposium on Security and Privacy, SP 2016, San Jose, CA, USA, May 22-26, 2016*, 2016, pp. 582–597.

N. Carlini and D. A. Wagner, "Towards Evaluating the Robustness of Neural Networks," in *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, 2017, pp. 39–57.

W. Xu, D. Evans, and Y. Qi, "Feature Squeezing: Detecting Adversarial Examples in Deep Neural Networks," *CoRR*, vol. abs/1704.01155, 2017.

A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu, "Towards deep learning models resistant to adversarial attacks," *arXiv preprint arXiv:1706.06083*, 2017.

L. G. Valiant, "A Theory of the Learnable," *Communications of the ACM*, vol. 27, no. 11, pp. 1134–1142, 1984.

J. Gilmer, L. Metz, F. Faghri, S. S. Schoenholz, M. Raghu, M. Wattenberg, and I. Goodfellow, "Adversarial spheres," *arXiv preprint arXiv:1801.02774*, 2018.

D. Diochnos, S. Mahloujifar, and M. Mahmoody, "Adversarial risk and robustness: General definitions and implications for the uniform distribution," in *Advances in Neural Information Processing Systems*, 2018, pp. 10 359–10 368.

A. Degwekar and V. Vaikuntanathan, "Computational limitations in robust classification and win-win results," *arXiv preprint arXiv:1902.01086*, 2019.

N. Ford, J. Gilmer, N. Carlini, and D. Cubuk, "Adversarial examples are a natural consequence of test error in noise," *arXiv preprint arXiv:1901.10513*, 2019.

I. Attias, A. Kontorovich, and Y. Mansour, "Improved generalization bounds for robust learning," *arXiv preprint arXiv:1810.02180*, 2018.

S. Bubeck, E. Price, and I. Razenshteyn, "Adversarial examples from computational constraints," *arXiv preprint arXiv:1805.10204*, 2018.

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

D. Cullina, A. N. Bhagoji, and P. Mittal, "Pac-learning in the presence of evasion adversaries," *arXiv preprint arXiv:1806.01471*, 2018.

M. Ledoux, *The Concentration of Measure Phenomenon*, ser. Mathematical Surveys and Monographs. American Mathematical Society, 2001, no. 89.

V. D. Milman and G. Schechtman, *Asymptotic theory of finite dimensional normed spaces*. Springer Verlag, 1986.

P. M. Long, "On the sample complexity of pac learning halfspaces against the uniform distribution," *IEEE Transactions on Neural Networks*, vol. 6, no. 6, pp. 1556–1559, 1995.

M.-F. Balcan and P. Long, "Active and passive learning of linear separators under log-concave distributions," in *Conference on Learning Theory*, 2013, pp. 288–316.

S. Sabato, N. Srebro, and N. Tishby, "Distribution-dependent sample complexity of large margin learning," *The Journal of Machine Learning Research*, vol. 14, no. 1, 2013.

S. Mahloujifar, D. I. Diochnos, and M. Mahmoody, "The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure," *arXiv preprint arXiv:1809.03063*, 2018.

D. Tsipras, S. Santurkar, L. Engstrom, A. Turner, and A. Madry, "Robustness may be at odds with accuracy," *stat*, vol. 1050, p. 11, 2018.

B. Biggio, B. Nelson, and P. Laskov, "Poisoning attacks against support vector machines," in *Proceedings of the 29th International Coference on International Conference on Machine Learning*. Omnipress, 2012, pp. 1467–1474.

N. Papernot, P. McDaniel, A. Sinha, and M. Wellman, "Towards the science of security and privacy in machine learning," *arXiv preprint arXiv:1611.03814*, 2016.

T. Gu, B. Dolan-Gavitt, and S. Garg, "Badnets: Identifying vulnerabilities in the machine learning model supply chain," *arXiv preprint arXiv:1708.06733*, 2017.

L. G. Valiant, "Learning disjunctions of conjunctions," in *IJCAI*, 1985, pp. 560–566.

M. J. Kearns and M. Li, "Learning in the Presence of Malicious Errors," *SIAM Journal on Computing*, vol. 22, no. 4, pp. 807–837, 1993.

N. H. Bshouty, N. Eiron, and E. Kushilevitz, "PAC learning with nasty noise," *Theoretical Computer Science*, vol. 288, no. 2, pp. 255–275, 2002.

S. Mahloujifar, D. I. Diochnos, and M. Mahmoody, "Learning under p-Tampering Attacks," in *ALT*, 2018, pp. 572–596.

A. Shafahi, W. R. Huang, M. Najibi, O. Suciu, C. Studer, T. Dumitras, and T. Goldstein, "Poison frogs! targeted clean-label poisoning attacks on neural networks," *arXiv preprint arXiv:1804.00792*, 2018.

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

U. Feige, Y. Mansour, and R. E. Schapire, "Robust inference for multiclass classification," in *Algorithmic Learning Theory*, 2018, pp. 368–386.

J. Khim and P.-L. Loh, "Adversarial risk bounds for binary classification via function transformation," *arXiv preprint arXiv:1810.09519*, 2018.

D. Yin, K. Ramchandran, and P. Bartlett, "Rademacher complexity for adversarially robust generalization," *arXiv preprint arXiv:1810.11914*, 2018.

P. Gourdeau, V. Kanade, M. Kwiatkowska, and J. Worrell, "On the hardness of robust classification," in *Advances in Neural Information Processing Systems*, 2019, pp. 7446–7455.

L. Chen, Y. Min, M. Zhang, and A. Karbasi, "More data can expand the generalization gap between adversarially robust and standard models," *preprint arXiv:2002.04725*, 2020.

A. S. Suggala, A. Prasad, V. Nagarajan, and P. Ravikumar, "Revisiting adversarial risk," in *The 22nd International Conference on Artificial Intelligence and Statistics*, 2019, pp. 2331–2339.

L. Schmidt, S. Santurkar, D. Tsipras, K. Talwar, and A. Madry, "Adversarially Robust Generalization Requires More Data," *arXiv preprint arXiv:1804.11285*, 2018.