Lower Bounds for Adversarially Robust PAC Learning

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

Saeed Mahloujifar[*]
University of Virginia
saeed@virginia.edu

Mohammad Mahmoody
University of Virginia
mohammad@virginia.edu

Abstract

In this work, we study probably approximately correct (PAC) learning under general perturbation-based evasion attacks. Here the adversary's goal is to misclassify an adversarially perturbed sample point $\widetilde{x}$, i.e., $h(\widetilde{x}) \neq c(\widetilde{x})$, where c is the ground truth concept, h is the learned hypothesis, and x is the original honestly sampled point. The only limitation on the adversary is that $\widetilde{x}$ is not "too far" from x, controlled by a metric measure. Previous work on PAC learning of adversarial examples have all modeled adversarial examples as *corrupted inputs* in which the goal of the adversary is to achieve $h(\widetilde{x}) \neq c(x)$, where x is the original untampered instance. These two definitions of adversarial risk coincide as long as the ground truth c does not change under the allowed perturbations. However, our work is more general and allows arbitrary perturbations, bounded by a metric.

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* $o(\|x\|)$ amount of perturbations on the test instances, PAC learning requires sample complexity that is *exponential* in n. This is in contrast with results proved using the corrupted-input framework, in which the sample complexity of robust learning is only polynomially more.

We then formalize *hybrid* attacks in which the evasion attack is preceded by a poisoning attack. This is perhaps reminiscent of "trapdoor attacks" in which a poisoning phase is involved as well, but again we focus on the general setting in which adversary's evasion attack is only controlled by a specified amount of perturbation based on input's length, and thus we again (have to) use the error-region definition of risk that aims at misclassifying the perturbed instances in general settings. In this case, we show PAC learning is sometimes *impossible* all together, even when it is possible without the attack (e.g., due to the bounded VC dimension).

1 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; Biggio et al. 2013; Goodfellow, Shlens, and Szegedy 2015) 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 against them and more attacks are introduced afterwards (Biggio et al. 2013; Biggio, Fumera, and Roli 2014; Goodfellow, Shlens, and Szegedy 2015; Papernot et al. 2016b; Carlini and Wagner 2017; Xu, Evans, and Qi 2017; Madry et al. 2017).

A fundamental question in robust learning is whether one can design learning algorithms that achieve "uniform converegence" 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

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

where $\mathsf{AdvRisk}$ is the "adversarial" risk, and $c(\cdot)$ 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, Mahloujifar, and Mahmoody 2018; Degwekar and Vaikuntanathan 2019; Ford et al. 2019) and was formalized by Diochnos, Mahloujifar, and Mahmoody (2018) as the "error-region" adversarial risk, because the adversary's goal here is to push $\widetilde{x}$ into

[*] Authors have contributed equally.

the error region

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

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, Mansour, and Schapire 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.

Why PAC Learning under General 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 right definition (for the general setting) because if one happens to learn the concept class c completely 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 a natural question 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.

Previous Work Several works have already studied PAC learning with provable guarantees under adversarial perturbations (Bubeck, Price, and Razenshteyn 2018; Cullina, Bhagoji, and Mittal 2018; Feige, Mansour, and Schapire 2018; Attias, Kontorovich, and Mansour 2018; Khim and Loh 2018; Yin, Ramchandran, and Bartlett 2018; Montasser, Hanneke, and Srebro 2019). However, all these works use the *corrupted-input* notion of adversarial risk. In particular, it is proved by Attias, Kontorovich, and Mansour (2018) that robust learning might require more data, but it was also shown by Attias, Kontorovich, and Mansour; Bubeck, Price, and Razenshteyn (2018; 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. 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?*

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 definition 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 affect 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.

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. 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.

Putting our Work into Perspective Our work in (Mahloujifar, Diochnos, and Mahmoody 2018b) dealt with clean-label "poisoning" attacks in situations where the adversary has the opportunity to substitute $\approx p$ *randomly selected* fraction of the training examples, with some "adversarial" ones of their choosing but the labels of the injected training examples need to respect the ground truth c (and hence the term "clean-label"). Such attacks are called p-tampering. In particular, the adversary can also effectively reduce the sample size by repeating training examples at the randomly selected p fraction of the changed examples. Our work in (Mahloujifar, Diochnos, and Mahmoody 2018b) is connected to the second part of our work here, where we formalize and study hybrid attacks.

In (Diochnos, Mahloujifar, and Mahmoody 2018) we provided a taxonomy of definitions that are used for the computation of adversarial examples and ultimately for the computation of the adversarial risk and robustness of learned classifiers. In addition, we showed that when misclassification is really the goal of an adversarial perturbation, then there is a natural problem (under the uniform distribution over $\{0, 1\}^n$) where only the error region definition computes the adversarial risk and robustness correctly. As a result we decided to use the error-region adversarial risk and robustness by default. Finally in that work, we computed inherent bounds that classifiers have on risk and robustness (based on the error-region) when again the distribution is uniform over $\{0,1\}^n$ – these bounds were information-theoretic.

In (Mahloujifar, Diochnos, and Mahmoody 2018a) extended the previous (information-theoretic) inherent bounds that classifiers have on adversarial risk and robustness, from the uniform distribution over $\{0,1\}^n$, to information-theoretic bounds on any Normal Lévy families (which, for example, include product distributions over $\{0,1\}^n$ and many more examples), using the phenomenon of concentration of measure. In the same work, we showed that the same phenomenon of concentration of measure allows an adversary to substitute a sublinear amount of training examples (that is, in a poisoning attack) and increase the probability of any bad property (e.g., misclassifying a particular test instance) from some non-negligible value (say 1%) to almost certainty (say 99%) by changing only $\approx \sqrt{n}$ of the examples, using correct labels.

The works of (Mahloujifar and Mahmoody 2019; Etesami, Mahloujifar, and Mahmoody 2019) extended the above information-theoretic results on poisoning and evasion attacks by explicitly providing efficient (polynomial-time) attacks on product distributions, so that the perturbation budget used in the attack scheme matches the information-theoretic bounds from the previous work of (Mahloujifar, Diochnos, and Mahmoody 2018a).

In Mahloujifar et al. (2019), it was shown how to empirically approximate (more specifically, upper bound) the concentration of a distribution of inputs given only (black-box) samples from the distribution. This is relevant to the line of work in which concentration of measure plays a key role in the hardness of adversarially robust learning, because one would need to know whether specific input distributions of interest (e.g., MNIST) are concentrated or not.

As the description above shows, our previous work on adversarial examples has focused on the power of an attacker. However, once one fixes the perturbation budget for the attacker, a natural question to ask is to what extent a learner can *defend* the hypothesis that it forms – that is, flip the table of the point of view of the analysis. Indeed, our first result in this work shows that a PAC learner needs exponentially many training examples in order to form a robust hypothesis when the attacker can substitute only a sublinear amount of the coordinates of the test instance. In the second part of the paper we introduce hybrid attacks and see that essentially a learner is helpless to form a robust hypothesis when the attacker has access both to the training as well as to the testing phase.

We are now ready to provide more details for the results of this current work.

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 a strong lower bound for the sample complexity of PAC learning in this setting. We 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, Kontorovich, and Mansour 2018; Bubeck, Price, and Razenshteyn 2018; Montasser, Hanneke, and Srebro 2019; Cullina, Bhagoji, and Mittal 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 defined over them, 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 almost all measure ≈ 1 of the input space.
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, Srebro, and Tishby 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 3.3).

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 the works (e.g., see (Mahloujifar, Diochnos, and Mahmoody 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 1.1 (Approximation error in error-region robust learning). *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, Nelson, and Laskov 2012; Papernot et al. 2016a). This attack model bears similarities to "trapdoor attacks" (Gu, Dolan-Gavitt, and Garg 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, Eiron, and Kushilevitz 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 the poisoning attack used here is a clean/correct label attack (Mahloujifar, Diochnos, and Mahmoody 2018b; Shafahi et al. 2018).

2 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 $\ell oss(y', y) = \mathbb{1}[y' = y]$. 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* sequence $\mathcal{S} = (z_1, \ldots, z_m)$ is a sequence of m examples. As is usual, sometimes we might refer to 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 that has empirical and true risk zero; though finding such h might be challenging.

Evasion Attacks An evasion attacker A is one that 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. For example, $\mathcal{A}$ 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,

[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.

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, Diochnos, and Mahmoody 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.

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, Mahloujifar, and Mahmoody 2018) adversary aims to misclassify the perturbed instance $\widetilde{x}$.

Definition 2.1 (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.*

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 2.2 (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*

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

PAC learning under (pure) poisoning attacks or evasion attacks could be derived from Definition 2.2 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 2.2 by using other forms of adversarial risk, e.g., corrupted-input adversarial risk (Feige, Mansour, and Schapire 2015; 2018; Madry et al. 2017; Schmidt et al. 2018; Attias, Kontorovich, and Mansour 2018)

3 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 3.1 (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]

$$\boldsymbol{\alpha}_n(b) \leq k_1 \cdot \mathrm{e}^{-k_2 \cdot b^2 / n}$$

Examples. Many natural metric probability spaces are Normal Lévy families. For example, all the following examples under normalized distance (to make the typical norms $\approx n$) are normal Lévy families as stated in Definition 3.1: the unit n-sphere with uniform distribution under the Euclidean or geodesic distance, $\mathbb{R}^n$ under Gaussian distribution and Euclidean distance, $\mathbb{R}^n$ under Gaussian distribution and Euclidean distance, the unit n-cube and unit n-ball under the uniform distribution and Euclidean distance, any product distribution of dimension n under the Hamming distance. See (Ledoux 2001; Giannopoulos and Milman 2001; Milman and Schechtman 1986) for more examples.

The following lemma was proved in (Mahloujifar, Diochnos, and Mahmoody 2018a) when Normal Lévy input spaces.

Lemma 3.2. *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 3.3 (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* $\Pr_{x \leftarrow D}[c_1(x) \neq c_2(x)] = \alpha$*.*

[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).

[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.

Examples. The set of homogeneous half spaces in $\mathbb{R}^n$ are α-close for all $\alpha \in (0, 1]$ under any of the following natural distributions: uniform over the unit sphere, uniform inside the unit ball, and isotropic Gaussian. This can be proved by picking two half spaces that their disagreement region under the mentioned distributions is exactly α. The set of (monotone, or not necessarily monotone) conjunctions are α-close for $\alpha = 2^{-k}$ for all $k \in \{2, \ldots, n\}$ under the uniform distribution over $\{0, 1\}^n$. This can be proved by looking at $c_1 = x_1 \wedge \ldots \wedge x_{k-1}$ and $c_2 = x_1 \wedge \ldots \wedge x_{k-1} \wedge x_k = c_1 \wedge x_k$. Since all the variables that appear in c_1 also appear in c_2, we have that $\Pr_{x \leftarrow \{0,1\}^n}[c_1(x) \neq c_2(x)]$ is equal to $\Pr_{x \leftarrow \{0,1\}^n}[(c_1(x) = 1) \wedge (c_2(x) = 0)]$, and as a consequence this is equal to $2^{-(k-1)} - 2^{-k} = 2^{-k}$.

We now state and prove our main results. Theorem 3.4 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 3.5).

Theorem 3.4 (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 3.1) 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 of PAC learning robust fo evasion attacks:

(a) **Exponential lower bound:** *Any PAC learning algorithm that is robust against* all *attacks with a sublinear tampering $b = o(n)$ budget under the metric* d *requires exponential sample complexity $m \geq 2^{\Omega(n)}$.*

(b) **Super-polynomial lower bound:** *PAC learning that is robust against against all tampering attacks with budget $b = \widetilde{O}(\sqrt{n})$, requires at least $m \geq n^{\omega(1)}$ 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 $\mathcal{R}$.*

As we will see, Part 1a and Part 1b of Theorem 3.4 are special cases of the following more quantitative lower bound that might be of independent interest.

Lemma 3.5. *For the setting of Theorem 3.4, 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*

$$m(n) \geq 2^{\Omega(\rho^2 \cdot n)}.$$

Examples. Here we list some natural scenarios that fall into the conditions of Theorem 3.4. All examples of Normal Lévy families listed after Definition 3.1 together with the concept class of half spaces satisfy the conditions of Theorem 3.4 and hence cannot be PAC learned using a $\operatorname{poly}(n)$ number of samples. The reason is that one can always find two half spaces whose symmetric difference has measure exactly ε. Moreover, as discussed in examples following Definition 3.3, even discrete problems such as learning monotone-conjunctions under the uniform distribution (and Hamming distance as perturbation metric) fall into the conditions of Theorem 3.4, for which a lower bound on their sample complexity (or even impossibility) of robust PAC learning could be obtained.

Remark 3.6 (Evasion-robust PAC learning in the RAM computing model with real numbers)**.** *We remark that if we allow (truly) real numbers represent the concept and hypothesis classes, one can even* rule out *PAC learning (not just lower bounds on sample complexity) under similar perturbations describe in Part 1. Indeed, by inspecting the same proof of Theorem 3.4 for Part 1 one can get such results, e.g., for learning half-spaces in dimension n when inputs come from isotropic Gaussian. However, we emphasize that such (seemingly) stronger lower bounds are not realistic, as in real settings, we eventually work with* finite *precision to represent the concept functions (of half spaces). This makes the set of concept functions* finite*, in which case the test error eventually reaches* zero*, using perhaps exponentially many samples. Theorem 3.4, however, has the useful feature that it applies even in those settings, as long as the concept functions are rich enough to allow the sufficiently close (but not too close) pairs under the distribution D according to Definition 3.3.*

In what follows, we will first prove Lemma 3.5. We will then use Lemma 3.5 to prove Theorem 3.4.

Proof of Lemma 3.5. Let $m = \mathsf{m}(0.9, 0.49, n)$ be the sample complexity of the (presumed) learner L that achieves (ε, δ)-PAC learning for $\varepsilon = 0.9, \delta = 0.49$. If $m = 2^{\Omega(n)}$ already, we are done, as it is even larger than what Lemma 3.5 states, so let $m = 2^{o(n)}$, and we will derive a contradiction. Since the distribution D is fixed, in the discussion below, we simply denote $\mathsf{Risk}(D, h, c)$ as $\mathsf{Risk}(h, c)$.

Recall that, by assumption, for all $\varepsilon \in [2^{-\Theta(n)}, 1]$, there are $c_1, c_2 \in \mathcal{C}$ that are $\Theta(\varepsilon)$-close under the distribution D. Because $m = 2^{o(n)}$, it holds that $1/m \geq \omega(2^{-\Theta(n)})$, and so there are $c_1, c_2 \in \mathcal{C}$ such that for $\Delta(c_1, c_2) = \{x \in \mathcal{X} \mid c_1(x) \neq c_2(x)\}$ we have

$$\Omega\left(\frac{1}{m}\right) \leq \Pr_{x \leftarrow D}[x \in \Delta(c_1, c_2)] \leq \frac{1}{100m}.$$

Now, consider m i.i.d. samples that are given to the learner L as a training set $\mathcal{S}$. With probability at least 0.99 of the sampling of $\mathcal{S}$, all $x \in \mathcal{S}$ would be outside $\Delta(c_1, c_2)$, in which case L would have no way to distinguish c_1 from c_2. So, if we pick $c \leftarrow \{c_1, c_2\}$ at random and pick test instance $x \leftarrow (D \mid \Delta(c_1, c_2))$, the hypothesis $h = L(\mathcal{S})$ fails with probability at least $0.99/2$. Thus, we can fix the choice of

$c \in \{c_1, c_2\}$, such that with probability $0.99/2 > 0.49$ we get a $h \leftarrow L(\mathcal{S})$ where

$$\begin{aligned}\mathsf{Risk}(h,c) = \Pr_{x \leftarrow D}[h(x) \neq c(x)] &\geq \frac{1}{2} \cdot \Pr_{x \leftarrow D}[x \in \Delta(c_1, c_2)] \\ &\geq \Omega\left(\frac{1}{m}\right).\end{aligned}$$

For this fixed c and any such learned hypothesis h with $\mathsf{Risk}(h,c) = \Omega(1)/m$, by Lemma 3.2, the adversarial risk reaches $\mathsf{AdvRisk}_{\mathcal{A}_b}(h,c) \geq 0.99$ by an attack $\mathsf{A} \in \mathcal{A}_b$ that has tampering budget:

$$b = O(\sqrt{n}) \cdot \left(\sqrt{\ln(O(m))} + \sqrt{O(1)}\right) \leq t \cdot (\sqrt{n \cdot \ln m})$$

for universal constant t. But, we said at the beginning that the tampering budget of the adversary is $\rho(n) \cdot n$. Therefore, it should be that

$$\rho(n) \cdot n < t \cdot (\sqrt{n \cdot \ln m}),$$

as otherwise the evasion-robust PAC learner is not actually robust as stated. Thus, we get

$$m \geq e^{\rho(n)^2 \cdot n/t} = 2^{\Omega(\rho(n)^2 \cdot n)}$$

which finishes the proof of Lemma 3.5. □

We now prove Theorem 3.4 using Lemma 3.5.

Proof of Theorem 3.4. Using Lemma 3.5, we will first prove Part 1a, then Part 1b, and then Part 2. Throughout, $\varepsilon = 0.9, \delta = 0.49$ are fixed, so the sample complexity $m = m(n)$ is a function of n.

Proving Part 1a. We claim that PAC learning resisting all $b = o(n)$-tampering attacks requires sample complexity $m \geq 2^{\Omega(n)}$. The reason is that, otherwise, there will be an infinite sequence of values $n_1 < n_2 < \ldots$ for n for which $m = m(n_i) \leq 2^{\gamma(n_i) \cdot (n_i)}$ for $\gamma(n) = o(1)$. However, in that case, if we let $\rho(n) = \gamma(n)^{1/3}$, because $\rho(n) = o(n)$, by Lemma 3.5, the sample complexity is

$$m(n_i) \geq 2^{\Omega(\rho(n_i)^2 \cdot n_i)} = \omega\left(2^{\gamma(n_i) \cdot n_i}\right).$$

However, this is a contradiction as we previously assumed $m(n_i) \leq 2^{\gamma(n_i) \cdot (n_i)}$.

Proving Part 1b. Suppose the adversary can tamper instances with budget $b(n) = \kappa(n) \cdot \sqrt{n}$ for $\kappa(n) \in \operatorname{polylog}(n)$. Since we can rewrite $b(n) = \rho(n) \cdot n$ for $\rho(n) = \kappa(n)/\sqrt{n}$, then by Lemma 3.5, the sample complexity of L should be at least

$$m(n) \geq 2^{\Omega(\rho(n)^2 \cdot n)} = 2^{\Omega(\kappa(n)^2)}.$$

Therefore, if we choose $\kappa(n) = \log(n)^2$, the sample complexity of L becomes $m \geq n^{\log n} \geq n^{\omega(1)}$.

Proving Part 2. Let be $c_1, c_2 \in \mathcal{C}$ be such that for $\Delta(c_1, c_2) = \{x \in \mathcal{X} \mid c_1(x) \neq c_2(x)\}$ we have

$$\Omega(\lambda) \leq \Pr_{x \leftarrow D(c_1, c_2)}[x \in \Delta(c_1, c_2)] \leq \lambda.$$

Consider a poisoning attacker A_1 that given a data set $\mathcal{S}$, it removes any (x, y) from $\mathcal{S}$ such that $x \in \Delta(c_1, c_2)$. Note that the (expected) number of such examples is $\Pr[x \in \Delta(c_1, c_2)] \leq \lambda$. Let $\widetilde{\mathcal{S}}$ be the modified training set. The learner $L(\widetilde{\mathcal{S}})$ now has now way to distinguish between c_1 and c_2. Thus, like in Lemma 3.5, we can fix $c \in \{c_1, c_2\}$, such that $L(\widetilde{\mathcal{S}})$ always produces h where

$$\begin{aligned}\mathsf{Risk}(h,c) = \Pr_{x \leftarrow D}[h(x) \neq c(x)] &\geq \frac{1}{2} \cdot \Pr_{x \leftarrow D}[x \in \Delta(c_1, c_2)] \\ &\geq \Omega(\lambda).\end{aligned}$$

For this fixed c and any such learned hypothesis h with $\mathsf{Risk}(h,c) = \Omega(\lambda)$, by Lemma 3.2, the adversarial risk (under attacks) reaches $\mathsf{AdvRisk}_{\mathcal{A}_b}(h,c) \geq 0.99$ by an attack $\mathsf{A} \in \mathcal{A}_b$ that changes test instances x by at most b for

$$b = O(\sqrt{n}) \cdot \left(\sqrt{\ln(O(1/\lambda))} + \sqrt{O(1)}\right) \leq O(\sqrt{n \cdot \ln(1/\lambda)}).$$

Since $\lambda = 1/\operatorname{poly}(n)$, it holds that $b = \widetilde{O}(\sqrt{n})$. □

4 Extensions

In this section, we describe some extensions to Theorem 3.4 in various directions.

Extension to Randomized Predictors In Theorem 3.4, we ruled out PAC learning (or its small sample complexity) even for very large values $\varepsilon = 0.9, \delta = 0.49$. One might argue that proving such lower bound could not be impossible because a trivial hypothesis (for the setting where $\mathcal{Y} = \{0, 1\}$) can achieve $\varepsilon = 0.5$ by outputting random bits. However, this trivial predictor is *randomized*, while Theorem 3.4 is proved for deterministic hypotheses. For the case of randomized hypotheses, one can adjust the proof of Theorem 3.4 to get similar lower bounds for $\varepsilon = 0.49, \delta = 0.49$ as follows.

In the proof of Theorem 3.4 we first showed that small sample complexity implies the existence of c that with probability > 0.49 it will have an error region with a non-negligible measure. When the hypothesis is randomized, however, we cannot work with the traditional notion of error region, because on every point $x \in \mathcal{X}$, the hypothesis could be wrong $h(x) \neq c(x)$ with some probability in $[0, 1]$. We can, however, work with the relaxed notion of "approximate error" region, defined as $\mathcal{AE}(h, c) = \{x \mid \Pr_h[h(x) \neq c(x)] \geq 1/2\}$, where the probability is over the randomness of h.

In proofs of both Lemma 3.5 and Theorem 3.4 we deal with two close concept functions c_1, c_2 that are "indistinguishable" for the hypothesis h and then conclude that for each point $x \in \Delta(c_1, c_2)$, h makes a mistake on at least one of c_1, c_2. If h is randomized, we cannot say this anymore, but we can still say that for each such point $x \in \Delta(c_1, c_2)$, for at least one of c_1, c_2, $h(x)$ is wrong with probability at least 0.5. Therefore, we get the same lower bound on the size of the $\mathcal{AE}$ as we got in Lemma 3.5 and Theorem 3.4. However, expanding the set $\mathcal{AE}$ instead of an actual error-region, implies that the adversarially perturbed points $\widetilde{x}$ that fall into $\mathcal{AE}$ are now misclassified with probability 0.5. Thus, at least 0.99 fraction of inputs can be perturbed into $\mathcal{AE}$ to be misclassified with probability > 0.49.

Lower Bound for PAC Learning of a "Typical" Concept Function Theorem 3.4 only proves the *existence* of at least *one* concept function $c \in \mathcal{C}$ for which the (presumed) robust PAC learner will either fail (to PAC learn) or will need large sample complexity. Now, suppose concept functions themselves come from a (natural) distribution and we only want to robustly PAC learn *most* of them. Indeed, we can extend the proof of Theorem 3.4 to show that for natural settings, the impossibility result extends to at least *half* of the concept functions, not just a few pathological cases.

To extend Theorem 3.4 to the more general "typical" failure over $c \leftarrow \mathcal{C}$ (stated as Claim 4.2 below) we need the following definition as an extension to Definition 3.3.

Definition 4.1 (Uniformly 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 $\mathcal{C}$* uniformly α-close *with respect to D, if there is a joint distribution $(\mathbf{c}_1, \mathbf{c}_2)$ where both coordinates are* uniformly *distributed over $\mathcal{C}$, and that for all $(c_1, c_2) \leftarrow (\mathbf{c}_1, \mathbf{c}_2)$, it both holds that $c_1, c_2 \in \mathcal{C}$ and that $\Pr_{x \leftarrow D}[c_1(x) \neq c_2(x)] = \alpha$.*

Claim 4.2. *In Theorem 3.4 and Lemma 3.5, make the only change in the setting as follows. The concept class $\mathcal{C}$ now satisfies the stronger condition of being* uniform *α-close with respect to D. Then, the same limitations of PAC learning hold for at least measure half of $c \leftarrow \mathcal{C}$.*

Here we sketch why Claim 4.2 holds. The difference is that now, instead of knowing the *existence* of an α-close pair (c_1, c_2), we have *distribution* $(\mathbf{c}_1, \mathbf{c}_2)$ samples from which satisfy the α-close property. Therefore, for all samples $(c_1, c_2) \leftarrow (\mathbf{c}_1, \mathbf{c}_2)$, at least one of c_1 or c_2 is "bad" for the (presumed) PAC learner L (with the same proof before). But, since each of the coordinates in $(\mathbf{c}_1, \mathbf{c}_2)$ is marginally uniform, therefore, at least measure $1/2$ of $c \leftarrow \mathcal{C}$ is bad for L.

Example Consider the uniform measure over homogeneous half spaces in dimension n as the set of concept functions $\mathcal{C}$: choose a point w in the unit sphere and select the half space $\{x \mid \langle x, w \rangle \geq 0\}$. It is easy to see that $\mathcal{C}$ with such measure is uniformly α-close with respect to the isotropic Gaussian distribution (or uniform distribution over the unit sphere). Thus, Claim 4.2 applies to this case.

5 Conclusion and Open Questions

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 the data dimension 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 $o(n)$ (or even $\widetilde{O}(\sqrt{n})$).

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.

Open Questions Our Theorem 3.4 relies on a level of tampering to be at least $\widetilde{O}(\sqrt{n})$ to imply the super-polynomial lower bounds. One natural question is to find the exact threshold of perturbations needed that triggers super-polynomial lower bounds on sample complexity.

Another important direction is to study the sample complexity of PAC learning (with concrete parameters ε, δ) for practical distributions such as images or voice. Our lower bounds of this work are only proved for theoretically natural distributions that are provably concentrated in high dimension. Mahloujifar et al. (2019), presents a method for empirically approximating the concentration of such distributions given i.i.d. samples from them.

Finally, we ask if similar results could be proved for corrupted-input adversarial risk. Note that previous work studying learning under corrupted-input adversarial risk (Bubeck, Price, and Razenshteyn 2018; Cullina, Bhagoji, and Mittal 2018; Feige, Mansour, and Schapire 2018; Attias, Kontorovich, and Mansour 2018; Khim and Loh 2018; Yin, Ramchandran, and Bartlett 2018; Montasser, Hanneke, and Srebro 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

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

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

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

Biggio, B.; Fumera, G.; and Roli, F. 2014. Security evaluation of pattern classifiers under attack. *IEEE transactions on knowledge and data engineering* 26(4):984–996.

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

Bshouty, N. H.; Eiron, N.; and Kushilevitz, E. 2002. PAC learning with nasty noise. *Theoretical Computer Science* 288(2):255–275.

Bubeck, S.; Lee, Y. T.; Price, E.; and Razenshteyn, I. 2018. Adversarial examples from cryptographic pseudo-random generators. *arXiv preprint arXiv:1811.06418*.

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

Carlini, N., and Wagner, D. A. 2017. 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*, 39–57.

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

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

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

Etesami, O.; Mahloujifar, S.; and Mahmoody, M. 2019. Computational concentration of measure: Optimal bounds, reductions, and more. *CoRR* abs/1907.05401.

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

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

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

Giannopoulos, A. A., and Milman, V. D. 2001. Euclidean structure in finite dimensional normed spaces. *Handbook of the geometry of Banach spaces* 1:707–779.

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

Goodfellow, I.; Shlens, J.; and Szegedy, C. 2015. Explaining and Harnessing Adversarial Examples. In *ICLR*.

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

Kearns, M. J., and Li, M. 1993. Learning in the Presence of Malicious Errors. *SIAM Journal on Computing* 22(4):807–837.

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

Ledoux, M. 2001. *The Concentration of Measure Phenomenon*. Number 89 in Mathematical Surveys and Monographs. American Mathematical Society.

Long, P. M. 1995. On the sample complexity of pac learning half-spaces against the uniform distribution. *IEEE Transactions on Neural Networks* 6(6):1556–1559.

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

Mahloujifar, S., and Mahmoody, M. 2019. Can adversarially robust learning leverage computational hardness? *Algorithmic Learning Theory (ALT)*.

Mahloujifar, S.; Zhang, X.; Mahmoody, M.; and Evans, D. 2019. Empirically measuring concentration: Fundamental limits on intrinsic robustness. *Safe Machine Learning workshop at ICLR*.

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

Mahloujifar, S.; Diochnos, D. I.; and Mahmoody, M. 2018b. Learning under p-Tampering Attacks. In *ALT*, 572–596.

Milman, V. D., and Schechtman, G. 1986. *Asymptotic theory of finite dimensional normed spaces*, volume 1200. Springer Verlag.

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

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

Papernot, N.; McDaniel, P. D.; Wu, X.; Jha, S.; and Swami, A. 2016b. 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*, 582–597.

Sabato, S.; Srebro, N.; and Tishby, N. 2013. Distribution-dependent sample complexity of large margin learning. *The Journal of Machine Learning Research* 14(1):2119–2149.

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

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

Szegedy, C.; Zaremba, W.; Sutskever, I.; Bruna, J.; Erhan, D.; Goodfellow, I.; and Fergus, R. 2014. Intriguing properties of neural networks. In *ICLR*.

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

Valiant, L. G. 1984. A Theory of the Learnable. *Communications of the ACM* 27(11):1134–1142.

Valiant, L. G. 1985. Learning disjunctions of conjunctions. In *IJCAI*, 560–566.

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

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