Sample Complexity Bounds for Robustly Learning Decision Lists against Evasion Attacks *

Pascale Gourdeau , **Varun Kanade** , **Marta Kwiatkowska** and **James Worrell**
University of Oxford
{pascale.gourdeau, varunk, marta.kwiatkowska, james.worrell}@cs.ox.ac.uk,

Abstract

A fundamental problem in adversarial machine learning is to quantify how much training data is needed in the presence of evasion attacks. In this paper we address this issue within the framework of PAC learning, focusing on the class of decision lists. Given that distributional assumptions are essential in the adversarial setting, we work with probability distributions on the input data that satisfy a Lipschitz condition: nearby points have similar probability. Our key results illustrate that the adversary's budget (that is, the number of bits it can perturb on each input) is a fundamental quantity in determining the sample complexity of robust learning. Our first main result is a sample-complexity lower bound: the class of monotone conjunctions (essentially the simplest non-trivial hypothesis class on the Boolean hypercube) and any superclass has sample complexity at least exponential in the adversary's budget. Our second main result is a corresponding upper bound: for every fixed k the class of k-decision lists has polynomial sample complexity against a $\log(n)$-bounded adversary. This sheds further light on the question of whether an efficient PAC learning algorithm can always be used as an efficient $\log(n)$-robust learning algorithm under the uniform distribution.

1 Introduction

Adversarial machine learning has been extensively studied in recent years, first with spam filtering in [Dalvi *et al.*, 2004; Lowd and Meek, 2005a; Lowd and Meek, 2005b], and then when the notion of *adversarial examples* was introduced by [Szegedy *et al.*, 2013], and independently noticed by [Biggio *et al.*, 2013]. Various settings to study adversarial machine learning guarantees (and impossibility results) have emerged in the literature since. The most common distinction, presented in [Biggio and Roli, 2017], differentiates between attacks at training time, known as *poisoning attacks*, and attacks at test time, called *evasion attacks*.

In the context of evasion attacks, a misclassification by a model has been defined in various ways, and sometimes regrettably referred to by the same terminology. [Dreossi *et al.*, 2019; Diochnos *et al.*, 2018; Gourdeau *et al.*, 2021] offer thorough discussions on the subject. We will focus on the *exact-in-the-ball* notion of robustness (also known as *error region* risk in [Diochnos *et al.*, 2018]), which necessitates a ground truth function. Briefly, the exact-in-the-ball notion of robustness requires a hypothesis to be correct with respect to the ground truth in a perturbation region around each test point. Note that, in this case, the ground truth must be specified on all input points in the perturbation region. By contrast, the constant-in-the-ball notion of robustness (which is also known as *corrupted input* robustness) is concerned with the stability of the hypothesis to perturbations in the input, and requires that the label produced by the hypothesis remain constant in the perturbation region, meaning that we only need access to the test point labels.

The hardness of robust classification has been explored from both a computational complexity and a statistical viewpoint, see for e.g., [Bubeck *et al.*, 2019; Montasser *et al.*, 2019]. In this paper, we focus on the Boolean hypercube $\{0,1\}^n$ as our input space and study the information-theoretic complexity of robust learning by exhibiting sample complexity upper and lower bounds that depend on an *adversarial budget*, i.e., the number of bits an adversary is allowed to flip at test time, thus illustrating that the adversarial budget is a fundamental quantity in determining the sample complexity of robustly learning important concept classes.

1.1 Our Contributions

Our work builds on [Gourdeau *et al.*, 2019] and its extended version [Gourdeau *et al.*, 2021]. Our results hold for the *exact-in-the-ball* robustness to evasion attacks.

Robust Learning of Decision Lists: As shown in [Gourdeau *et al.*, 2021], *efficient, exact-in-the-ball* robust learning is not possible without distributional assumptions on the training data.[1] We follow their line of work and establish the sample-efficient robust learnability of decision lists against a $\log(n)$-bounded adversary under log-Lipschitz distributions, which include the uniform and product distributions. The algorithms we use to show such upper bounds are called ρ-

*Full paper: http://www.fun2model.org/papers/gkkw22.pdf.

[1] This is in contrast to PAC learning, which is distribution-free.

robust learning algorithms, where ρ is the allowed perturbation budget for an adversary. In proving our first result we obtain an isoperimetric bound that may be of independent interest: for a CNF formula φ we give an upper bound on the number of points in the Boolean hypercube within a given Hamming distance to a satisfying assignment of φ. An analogue result was shown only for *monotone* decision lists in [Gourdeau *et al.*, 2021]. More importantly, [Gourdeau *et al.*, 2021] suggested the following open problem:

Let $\mathcal{A}$ be a sample-efficient (potentially proper) PAC-learning algorithm for concept class $\mathcal{C}$. Is $\mathcal{A}$ also a sample-efficient $\log(n)$-robust learning algorithm for $\mathcal{C}$ under the uniform distribution?

So far, all the concept classes that have been studied point towards a positive answer to this question. As log-Lipschitz distributions subsume the uniform distribution, our result thus adds to the body of positive evidence for this problem.

An Adversarial Sample Complexity Lower Bound: To complement the above result, we show that any ρ-robust learning algorithm for monotone conjunctions must have a sample complexity that is exponential in the number ρ of bits an adversary is allowed to flip during an evasion attack. Previously, [Gourdeau *et al.*, 2021] showed that there does not exist such an algorithm with polynomial sample complexity against an adversary that can perturb $\omega(\log(n))$ bits of the input.

1.2 Related Work

The inevitability of adversarial examples under the constant-in-the-ball definition of robustness has been extensively studied, see for e.g., [Fawzi *et al.*, 2016; Fawzi *et al.*, 2018a; Fawzi *et al.*, 2018b; Gilmer *et al.*, 2018; Shafahi *et al.*, 2018; Tsipras *et al.*, 2019; Ilyas *et al.*, 2019]. We first outline related work on sample complexity lower bounds for robust learning. [Bhagoji *et al.*, 2019] work with the constant-in-the-ball definition of robustness and use an optimal transport cost function to derive lower bounds for learning classes with labels that come from a mixture of Gaussian distributions. [Montasser *et al.*, 2019] also use this notion of robustness to show a lower bound that depends on a complexity measure adapted to robustness from the shattering dimension of a concept class. Closer to our work, [Diochnos *et al.*, 2019; Diochnos *et al.*, 2020] exhibit lower bounds for the exact-in-the-ball robust risk. They focus on a family of concentrated distributions, Normal Lévy families, which include, for e.g., the Gaussian distribution on $\mathbb{R}^n$ and product distribution of dimension n under the Hamming distance.[2] Instead of looking at a specific class of functions, they allow any concept class that contain concepts that have small enough ($2^{-\Theta(n)}$) standard error with respect to each other, and so would be indistinguishable for sufficiently small samples. Note that monotone conjunctions satisfy this property. When considering the Boolean hypercube and an adversary that can perturb ρ bits of the input, they get that any robust PAC learning algorithm for their robust learning setting requires a sample of size $2^{\Omega(\rho^2/n)}$. Note that this lower bound is non trivial only when considering adversaries that can perturb $\sqrt{n}$ bits or more, while we show a lower bound that is strictly *exponential* in the adversary's budget (though for slightly more restricted concept classes), and thus meaningful for a wider class of adversaries.

In terms of sample complexity upper bounds, [Montasser *et al.*, 2019] show sample complexity upper bounds that are linear (ignoring log factors) in the VC dimension and the dual VC dimension of a concept class under the constant-in-the-ball notion of robustness, yielding an exponential upper bound in the VC dimension. As noted in [Gourdeau *et al.*, 2021], their techniques do not apply to the exact-in-the-ball setting, which is studied for evasion attacks notably in [Diochnos *et al.*, 2018; Mahloujifar and Mahmoody, 2019; Mahloujifar *et al.*, 2019; Gourdeau *et al.*, 2019; Gourdeau *et al.*, 2021]. The work of [Diochnos *et al.*, 2018] addresses the ability of an adversary to cause a blow up the adversarial error with respect to the standard error. For instance, they show that, under the uniform distribution, a $O(\sqrt{n})$-bounded adversary can cause the probability of a misclassification to be $1/2$ given that the standard error is 0.01 for any learning problem. These results are extended in [Mahloujifar *et al.*, 2019] for a wider family of distributions. Finally, [Gourdeau *et al.*, 2021] exhibit sample complexity upper bounds for the robust learnability of a variety of concept classes (parities, *monotone* decision lists, and decision trees) under log-Lipschitz distributions for various adversarial budgets.

2 Problem Set Up

In this section, we will first recall two definitions of robustness that have been widely used in the literature, and formalize the notion of robustness thresholds in the robust PAC-learning framework. We will then review relevant concept classes for this paper, as well as log-Lipschitz distributions, which were introduced in [Awasthi *et al.*, 2013] and will be the focus of our results.

2.1 Robust Learning

We work in the PAC learning framework of [Valiant, 1984] (see Appendix A.1), but where the (standard) risk function is replaced by a *robust* risk function. Since we focus on the Boolean hypercube $\{0,1\}^n$ as the input space, the only relevant notion of distance between points is the Hamming distance (denoted d_H), i.e., the number of bits that differ between two points. Thus, the adversary's perturbation budget will be the number of bits of the input the adversary is allowed to flip to cause a misclassification. We will use the *exact-in-the-ball* definition of robust risk (which is called *error-region* risk in [Diochnos *et al.*, 2018]). Given respective hypothesis and target functions $h, c : \mathcal{X} \to \{0,1\}$, distribution D on $\mathcal{X}$, and robustness parameter $\rho \geq 0$, the exact-in-the-ball robust risk of h with respect to c is defined as $\mathsf{R}^E_\rho(h,c) = \Pr_{x\sim D}(\exists z \in B_\rho(x) : h(z) \neq c(z))$, where $B_\rho(x) = \{z \in \{0,1\}^n \mid d_H(x,z) \leq \rho\}$. This is in contrast to the more widely-used *constant-in-the-ball* risk function (also called *corrupted-instance* risk from the work of [Feige *et al.*, 2015]) $\mathsf{R}^C_\rho(h,c) = \Pr_{x\sim D}(\exists z \in B_\rho(x) : h(z) \neq c(x))$

[2] We work with the uniform distribution, which is a special case of product distributions.

where the hypothesis is required to be constant in the perturbation region in addition to being correct with respect to the unperturbed point's label $c(x)$.

Both [Diochnos *et al.*, 2018] and [Gourdeau *et al.*, 2021] offer a thorough discussion on the advantages and drawbacks of the two notions of robust risk. We will study the *exact-in-the-ball* robust risk, as our learning problems have considerable probability mass near the decision boundary. Thus it makes sense to consider the faithfulness of the hypothesis with respect to the target function. The exact-in-the-ball robust risk also has various advantages: if the distribution is supported on the whole input space (e.g., the uniform distribution), exact learnability implies robust learnability and the target concept is always the robust risk minimizer.[3] We have from [Gourdeau *et al.*, 2021] the following definition of robust learnability with respect to the exact-in-the-ball robust risk. Note that we will henceforth drop the superscript and simply use R_ρ to denote the exact-in-the-ball robust risk.

Definition 1 ([Gourdeau *et al.*, 2021])**.** *Fix a function $\rho : \mathbb{N} \to \mathbb{N}$. We say that an algorithm $\mathcal{A}$* efficiently ρ-robustly learns *a concept class $\mathcal{C}$ with respect to distribution class $\mathcal{D}$ if there exists a polynomial poly$(\cdot,\cdot,\cdot,\cdot)$ such that for all $n \in \mathbb{N}$, all target concepts $c \in \mathcal{C}_n$, all distributions $D \in \mathcal{D}_n$, and all accuracy and confidence parameters $\epsilon, \delta > 0$, if $m \geq$ poly$(n, 1/\epsilon, 1/\delta,$ size$(c))$, whenever $\mathcal{A}$ is given access to a sample $S \sim D^m$ labelled according to c, it outputs a polynomially evaluable function $h : \{0,1\}^n \to \{0,1\}$ such that $\Pr_{S\sim D^m}(\mathsf{R}_\rho(h,c) < \epsilon) > 1 - \delta$.*

2.2 Concept Classes and Distribution Families

Our work uses formulas in the conjunctive normal form (CNF) to show the robust learnability of decision lists. This concept class was proposed and shown to be PAC learnable in [Rivest, 1987]. Formally, given the maximum size k of a conjunctive clause, a decision list $f \in k$-DL is a list $(K_1, v_1), \ldots, (K_r, v_r)$ of pairs where K_j is a term in the set of all conjunctions of size at most k with literals drawn from $\{x_1, \bar{x_1}, \ldots, x_n, \bar{x_n}\}$, v_j is a value in $\{0,1\}$, and K_r is `true`. The output $f(x)$ of f on $x \in \{0,1\}^n$ is v_j, where j is the least index such that the conjunction K_j evaluates to `true`.

Given $k, n \in \mathbb{N}$, we denote by φ a k-CNF on n variables, where k refers to the size of the largest clause in φ. Note that the class MON-CONJ of monotone conjunctions, where each variable appears as a positive literal, is a subclass of 1-CNF formulas. Moreover, since decision lists generalize formulas in disjunctive normal form (DNF) and conjunctive normal form, in the sense that k-CNF $\cup$ k-DNF $\subseteq$ k-DL, a robust learnability result for k-DL holds for k-CNF and k-DNF as well. We refer the reader to Appendix A.2 for more background on conjunctions and k-CNF formulas.

For a formula φ, we will denote by $\mathsf{S}_0(\varphi)$ the probability $\Pr_{x\sim D}(x \models \varphi)$ that x drawn from distribution D results in a satisfying assignment of φ. We will also denote the probability mass $\Pr_{x\sim D}(\exists z \in B_\rho(x) \,.\, z \models \varphi)$ of the ρ-expansion of a satisfying assignment by $\mathsf{S}_\rho(\varphi)$.

[3] This is not necessarily the case with the constant-in-the-ball definition [Gourdeau *et al.*, 2021].

Our robust learnability results will hold for a class of sufficiently smooth distributions, called log-Lipschitz distributions, originally introduced in [Awasthi *et al.*, 2013]:

Definition 2. *A distribution D on $\{0,1\}^n$ is said to be* α-log-Lipschitz *if for all input points $x, x' \in \{0,1\}^n$, if $d_H(x,x') = 1$, then $|\log(D(x)) - \log(D(x'))| \leq \log(\alpha)$.*

Neighbouring points in $\{0,1\}^n$ have probability masses that differ by at most a multiplicative factor of α under α-log-Lipschitz distributions. The decay of probability mass along a chain of neighbouring points is thus at most exponential; not having sharp changes to the underlying distribution is a very natural assumption, and one weaker than many often make in the literature. Note that features are allowed a small dependence between each other and, by construction, log-Lipschitz distributions are supported on the whole input space. Notable examples of log-Lipschitz distributions are the uniform distribution (with parameter $\alpha = 1$) and the class of product distributions with bounded means.

3 The $\log(n)$-Expansion of Satisfying Assignments for k-CNF Formulas

In this section, we show that, under log-Lipschitz distributions, the probability mass of the $\log(n)$-expansion of the set of satisfying assignments of a k-CNF formula can be bounded above by an arbitrary constant $\varepsilon > 0$, given an upper bound on the probability of a satisfying assignment. The latter bound is polynomial in ε and $1/n$. While this result is of general interest, our goal is to prove the efficient robust learnability of decision lists against a $\log(n)$-bounded adversary. Here the relevant fact is that, given two decision lists $c, h \in k$-DL, the set of inputs in which c and h differ can be written as a disjunction of quadratically many (in the combined length of c and h) k-CNF formulas. The $\log(n)$-expansion of this set is then the set of inputs where a $\log(n)$-bounded adversary can force an error at test time. This is the main technical contribution of this paper, and the theorem is stated below. The combinatorial approach, below, vastly differs from the approach of [Gourdeau *et al.*, 2021] in the special case of monotone k-DL, which relied on facts about propositional logic.

Theorem 3. *Suppose that $\varphi \in k$-CNF and let D be an α-log-Lipschitz distribution on the valuations of φ. Then there exist constants $C_1, C_2, C_3, C_4 \geq 0$ that depend on α and k such that if the probability of a satisfying assignment satisfies $\mathsf{S}_0(\varphi) < C_1\varepsilon^{C_2}\min\{\varepsilon^{C_3}, n^{-C_4}\}$, then the $\log(n)$-expansion of the set of satisfying assignments has probability mass bounded above by ε.*

Corollary 4. *The class of k-decision lists is efficiently $\log(n)$-robustly learnable under log-Lipschitz distributions.*

The proof of Corollary 4 is similar to Theorem 24 in [Gourdeau *et al.*, 2021], and is included in Appendix B. We note that it is imperative that the constants C_i do not depend on the learning parameters or the input dimension, as the quantity $C_1\varepsilon^{C_2}\min\{\varepsilon^{C_3}, n^{-C_4}\}$ is directly used as the accuracy parameter in the (proper) PAC learning algorithm for decision lists, which is used as a black box.

To prove Theorem 3, we will need several lemmas outlined below, which are either taken directly or slightly adapted from [Gourdeau *et al.*, 2021]. The first is an adaptation of Lemma 17 in [Gourdeau *et al.*, 2021] for conjunctions, which was originally stated for decision lists:

Lemma 5. *Let φ be a conjunction and let D be an α-log-Lipschitz distribution. If $\Pr_{x\sim D}(x \models \varphi) < (1+\alpha)^{-d}$, then φ is a conjunction on at least d variables.*

The second result, which states an upper bound on the expansion of satisfying assignments for conjunctions, will be used for the base case of the induction proof.

Lemma 6. *Let D be an α-log-Lipschitz distribution on the n-dimensional Boolean hypercube and let φ be a conjunction of d literals. Set $\eta = \frac{1}{1+\alpha}$. Then for all $0 < \varepsilon < 1/2$, if $d \geq \max\left\{\frac{4}{\eta^2}\log\left(\frac{1}{\varepsilon}\right), \frac{2\rho}{\eta}\right\}$, then $\Pr_{x\sim D}((\exists y \in B_\rho(x)\cdot y \models \varphi)) \leq \varepsilon$.*

Finally, we will use the following lemma, which will be used in the inductive step of the induction proof.

Lemma 7. *Let φ be a k-CNF formula that has a set of variable-disjoint clauses of size M. Let D be an α-log-Lipschitz distribution on valuations for φ. Let $0 < \varepsilon < 1/2$ be arbitrary and set $\eta := (1+\alpha)^{-k}$. If $M \geq \max\left\{\frac{4}{\eta^2}\log\left(\frac{1}{\varepsilon}\right), \frac{2\rho}{\eta}\right\}$ then $\Pr_{x\sim D}(\exists y \in B_\rho(x)\cdot y \models \varphi) \leq \varepsilon$.*

We are now ready to prove Theorem 3. The main idea behind the proof is to consider a given k-CNF formula φ and distinguish two cases: (i) either φ contains a sufficiently-large set of variable-disjoint clauses, in which case the adversary is not powerful enough to make φ satisfied by Lemma 7; or (ii) we can rewrite φ as the disjunction of a sufficiently small number of $(k-1)$-CNF formulas, which allows us to use the induction hypothesis to get the desired result. The final step of the proof is to derive the constants mentioned in the statement of Theorem 3.

Proof of Theorem 3. We will use the lemmas above and restrictions on φ to show the following.

Induction hypothesis: Suppose that $\varphi \in (k-1)$-CNF and let D be an α-log-Lipschitz distribution on the valuations of φ. Then there exists constants $C_1, C_2, C_3, C_4 \geq 0$ that depend on α and k and satisfy $C_3 \geq \frac{\eta}{2}C_4$ such that if $\mathsf{S}_0(\varphi) < C_1\varepsilon^{C_2}\min\left\{\varepsilon^{C_3}, n^{-C_4}\right\}$, then $\mathsf{S}_{\log(n)}(\varphi) \leq \varepsilon$.

Base case: This follows from Lemmas 5 and 6. Set η to $(1+\alpha)^{-1}$, and $C_1 = 1$, $C_2 = 0$, $C_3 = \frac{4}{\eta^2}$ and $C_4 = \frac{2}{\eta}$. Note that $C_3 \geq \frac{\eta}{2}C_4$.

Inductive step: Suppose $\varphi \in k$-CNF and let D be an α-log-Lipschitz distribution on the valuations of φ. Set $\eta = (1+\alpha)^{-k}$. Let C_1', C_2', C_3', C_4' be the constants in the induction hypothesis for $\varphi' \in (k-1)$-CNF. Set the following constants:

$$C_1 = C_1' 2^{-k(C_2'+C_3')} \qquad C_2 = C_2' + C_3'$$
$$C_3 = \frac{8}{\eta^2}\max\{C_2', C_3'\} \qquad C_4 = \frac{2}{\eta}\max\{C_2', C_3'\} \ ,$$

and note that these are all constants that depend on k and α by the induction hypothesis, and that $C_3 \geq \frac{\eta}{2}C_4$.

Let $\mathsf{S}_0(\varphi) < C_1\varepsilon^{C_2}\min\left\{\varepsilon^{C_3}, n^{-C_4}\right\}$. Let $\mathcal{M}$ be a maximal set of clauses of φ such that no two clauses contain the same variable. Denote by $I_{\mathcal{M}}$ the indices of the variables in $\mathcal{M}$ and let $M = \max\left\{\frac{4}{\eta^2}\log\frac{1}{\varepsilon}, \frac{2}{\eta}\log n\right\}$.

We distinguish two cases:

(i) $|\mathcal{M}| \geq M$:
We can then invoke Lemma 7 and guarantee that $\mathsf{S}_{\log(n)} \leq \varepsilon$, and we get the required result.

(ii) $|\mathcal{M}| < M$:
Then let $\mathcal{A}_{\mathcal{M}}$ be the set of assignments of variables in $\mathcal{M}$, i.e. $a \in \mathcal{A}_{\mathcal{M}}$ is a function $a : I_{\mathcal{M}} \to \{0,1\}$, which represents a partial assignment of variables in φ. We can thus rewrite φ as follows:

$$\varphi \equiv \bigvee_{a\in\mathcal{A}_{\mathcal{M}}}\left(\varphi_a \wedge \bigwedge_{i\in I_{\mathcal{M}}} l_i\right) \ ,$$

where φ_a is the restriction of φ under assignment a and l_i is x_i in case $a(i) = 1$ and $\bar{x}_i$ otherwise. For short, denote by φ_a' the formula $\varphi_a \wedge \bigwedge_{i\in I_{\mathcal{M}}} l_i$. By the maximality of $\mathcal{M}$ every clause in φ mentions some variable in $\mathcal{M}$, and hence φ_a' is $(k-1)$-CNF. Moreover, the formulas φ_a' are disjoint, in the sense that if some assignment x satisfies φ_a', it will not satisfy another φ_b' for a distinct index b. Note also that

$$A_{n,\varepsilon} := |\mathcal{A}_{\mathcal{M}}| \leq 2^k \max\left\{\left(\frac{1}{\varepsilon}\right)^{4/\eta^2}, n^{2/\eta}\right\} \ .$$

Thus,

$$\mathsf{S}_0(\varphi) = \sum_{a\in\mathcal{A}_{\mathcal{M}}} \Pr_{x\sim D}(x \models \varphi_a') = \sum_{a\in\mathcal{A}_{\mathcal{M}}} \mathsf{S}_0(\varphi_a') \ . \tag{1}$$

By the induction hypothesis, we can guarantee that if

$$\mathsf{S}_0(\varphi_a') < C_1'\left(\frac{\varepsilon}{A_{n,\varepsilon}}\right)^{C_2'} \min\left\{\left(\frac{\varepsilon}{A_{n,\varepsilon}}\right)^{C_3'}, n^{-C_4'}\right\} \tag{2}$$

for all φ_a' then the $\log(n)$-expansion $\mathsf{S}_{\log(n)}(\varphi)$ can be bounded as follows:

$$\begin{aligned}
\mathsf{S}_{\log(n)}(\varphi) &= \Pr_{x\sim D}(\exists z \in B_{\log n}(x) \,.\, z \models \varphi) \\
&= \sum_{a\in\mathcal{A}_{\mathcal{M}}} \Pr_{x\sim D}(\exists z \in B_{\log n}(x) \,.\, z \models \varphi_a') \\
&\leq \sum_{a\in\mathcal{A}_{\mathcal{M}}} \frac{\varepsilon}{A_{n,\varepsilon}} \qquad \text{(I.H.)} \\
&= \varepsilon \ .
\end{aligned}$$

By Equation 1, the upper bound $\mathsf{S}_0(\varphi) < C_1\varepsilon^{C_2}\min\left\{\varepsilon^{C_3}, n^{-C_4}\right\}$ on the probability of a satisfying assignment for φ implies an upper bound $\mathsf{S}_0(\varphi_a') < C_1\varepsilon^{C_2}\min\left\{\varepsilon^{C_3}, n^{-C_4}\right\}$ on the probability

of the restrictions φ'_a. Thus it only remains to show that the condition on $\mathsf{S}_0(\varphi)$ implies that Equation 2 holds.

Let us rewrite the RHS of Equation 2 as follows, where each of the equations is a stricter condition on $\mathsf{S}_0(\varphi'_a)$ than its predecessor:

$$
\begin{aligned}
&C'_1 \left(\frac{\varepsilon}{A_{n,\varepsilon}}\right)^{C'_2} \min\left\{\left(\frac{\varepsilon}{A_{n,\varepsilon}}\right)^{C'_3}, n^{-C'_4}\right\} \\
&\geq C'_1 \left(\frac{\varepsilon}{2^k}\right)^{C'_2} \min\left\{\varepsilon^{4C'_2/\eta^2}, n^{-2C'_2/\eta}\right\} \\
&\qquad \times \min\left\{\left(\frac{\varepsilon^{1+4/\eta^2}}{2^k}\right)^{C'_3}, \left(\frac{\varepsilon n^{-2/\eta}}{2^k}\right)^{C'_3}, n^{-C'_4}\right\} \\
&= C'_1 \left(\frac{\varepsilon}{2^k}\right)^{C'_2} \min\left\{\varepsilon^{4C'_2/\eta^2}, n^{-2C'_2/\eta}\right\} \\
&\qquad \times \min\left\{\left(\frac{\varepsilon^{1+4/\eta^2}}{2^k}\right)^{C'_3}, \left(\frac{\varepsilon n^{-2/\eta}}{2^k}\right)^{C'_3}\right\} \\
&= C'_1 2^{-k(C'_2+C'_3)} \varepsilon^{C'_2+C'_3} \min\left\{\varepsilon^{4C'_2/\eta^2}, n^{-2C'_2/\eta}\right\} \\
&\qquad \times \min\left\{\varepsilon^{4C'_3/\eta^2}, n^{-2C'_3/\eta}\right\} \\
&\geq C'_1 2^{-k(C'_2+C'_3)} \varepsilon^{C'_2+C'_3} \\
&\qquad \times \min\left\{\varepsilon^{8C'_2/\eta^2}, n^{-4C'_2/\eta}, \varepsilon^{8C'_3/\eta^2}, n^{-4C'_3/\eta}\right\} \\
&= C'_1 2^{-k(C'_2+C'_3)} \varepsilon^{C'_2+C'_3} \\
&\qquad \times \min\left\{\varepsilon^{8\max\{C'_2,C'_3\}/\eta^2}, n^{-4\max\{C'_2,C'_3\}/\eta}\right\} \\
&= C_1 \varepsilon^{C_2} \min\left\{\varepsilon^{C_3}, n^{-C_4}\right\} \ ,
\end{aligned}
$$

where the first step is by definition of $A_{n,\varepsilon}$, the second from the induction hypothesis, which guarantees $C'_3 \geq \frac{\eta}{2} C'_4$, and the fourth from the property $\min\{a,b\} \cdot \min\{c,d\} \geq \min\{a^2,b^2,c^2,d^2\}$. Finally, the last equality follows by the definition of the C_i's.

Note that we set $\eta = (1+\alpha)^{-k}$ to be able to apply Lemma 7 in the first part of the inductive step. Then, $A_{n,\epsilon}$ is a function of $\eta = (1+\alpha)^{-k}$. When we consider the distribution on the valuations of the restriction φ'_a, we still operate with an α-log-Lipschitz distribution on its valuations, by log-Lipschitz facts (see Appendix A.3).

Constants. We want to get explicit constants C_1, C_2, C_3 and C_4 as a function of k and η. Note that $\eta = (1+\alpha)^{-k}$ is dependent on k. Let us recall the recurrence system from the inductive step:

$$
\begin{aligned}
C_1^{(k)} &= C_1^{(k-1)} 2^{-k(C_2^{(k-1)} + C_3^{(k-1)})} \\
C_2^{(k)} &= C_2^{(k-1)} + C_3^{(k-1)} \\
C_3^{(k)} &= \frac{8}{\eta^2} \max\left\{C_2^{(k-1)}, C_3^{(k-1)}\right\} \\
C_4^{(k)} &= \frac{2}{\eta} \max\left\{C_2^{(k-1)}, C_3^{(k-1)}\right\} \ .
\end{aligned}
$$

It is easy to see that $C_3^{(k)} \geq C_2^{(k)}$ for all $k \in \mathbb{N}$. If we fix $\eta = (1+\alpha)^{-k}$ at each level of the recurrence, we can now consider the following recurrence system, which dominates the previous one:

$$
\begin{aligned}
C_1^{(k)} &= C_1^{(k-1)} 2^{-2kC_3^{(k-1)}} & C_2^{(k)} &= 2C_3^{(k-1)} \\
C_3^{(k)} &= \frac{8}{\eta^2} C_3^{(k-1)} & C_4^{(k)} &= \frac{2}{\eta} C_3^{(k-1)} \ .
\end{aligned}
$$

We can now see that

$$
\begin{aligned}
C_2^{(k)} &= 2\left(\frac{8}{\eta^2}\right)^{k-1} = 2(8(1+\alpha)^{2k})^{k-1} \\
C_3^{(k)} &= \left(\frac{8}{\eta^2}\right)^{k} = (8(1+\alpha)^{2k})^{k} \\
C_4^{(k)} &= \frac{2}{\eta}\left(\frac{8}{\eta^2}\right)^{k-1} = 2(1+\alpha)^k (8(1+\alpha)^{2k})^{k-1} \ .
\end{aligned}
$$

Finally, we can get a lower bound on the value of $C_1^{(k)}$ as follows:

$$
\begin{aligned}
C_1^{(k)} &= \prod_{i=2}^{k} 2^{-2iC_3^{(i-1)}} \\
&= 2^{-2\sum_{i=2}^{k} i \cdot \left(\frac{8}{\eta^2}\right)^{(i-1)}} \\
&\geq 2^{-2k^2\left(\frac{8}{\eta^2}\right)^{(k-1)}} \\
&= 2^{-2k^2(8(1+\alpha)^{2k})^{k-1}} \ .
\end{aligned}
$$

□

4 An Adversarial Sample Complexity Lower Bound

In this section, we will show that any robust learning algorithm for monotone conjunctions under the uniform distribution must have an exponential sample-complexity dependence on an adversary's budget ρ. This result extends to any superclass of monotone conjunctions, such as CNF formulas, decision lists and halfspaces. It is a generalization of Theorem 13 in [Gourdeau *et al.*, 2021], which shows that no sample-efficient robust learning algorithm exists for monotone conjunctions against adversaries that can perturb $\omega(\log(n))$ bits of the input under the uniform distribution.

The idea behind the proof is to show that, for a fixed constant $\kappa < 2$, and sufficiently large input dimension, a sample of size $2^{\kappa\rho}$ from the uniform distribution won't be able to distinguish between two disjoint conjunctions of length 2ρ. However, the robust risk between these two conjunctions can be lower bounded by a constant. Hence, there does not exist a robust learning algorithm with sample complexity $2^{\kappa\rho}$ that works for the uniform distribution, and arbitrary input dimension and confidence and accuracy parameters.

Recall that the sample complexity of PAC learning conjunctions is $\Theta(n)$ in the non-adversarial setting. On the other hand, our adversarial lower bound in terms of the robust parameter is super linear in n as soon as the adversary can perturb more than $\log(\sqrt{n}))$ bits of the input.

Theorem 8. *Fix a positive increasing robustness function $\rho : \mathbb{N} \to \mathbb{N}$. For $\kappa < 2$ and sufficiently large input dimensions n, any $\rho(n)$-robust learning algorithm for* MON-CONJ *has a sample complexity lower bound of $2^{\kappa\rho(n)}$ under the uniform distribution.*

The proof of the theorem follows similar reasoning as Theorem 13 in [Gourdeau *et al.*, 2021], and is included in Appendix C. The main difference in the proof is its reliance on the following lemma, which shows that, for sufficiently large input dimensions, a sample of size $2^{\kappa\rho}$ from the uniform distribution will look constant with probability $1/2$ if labelled by two disjoint monotone conjunctions of length 2ρ. As shown in Lemma 14, which can be found in Appendix C, these two conjunctions have a robust risk bounded below by a constant against each other.

Lemma 9. *For any constant $\kappa < 2$, for any robustness parameter $\rho \leq n/4$, for any disjoint monotone conjunctions c_1, c_2 of length 2ρ, there exists n_0 such that for all $n \geq n_0$, a sample S of size $2^{\kappa\rho}$ sampled i.i.d. from D will have that $c_1(x) = c_2(x) = 0$ for all $x \in S$ with probability at least $1/2$.*

Proof. We begin by bounding the probability that c_1 and c_2 agree on an i.i.d. sample of size m. We have

$$\Pr_{S\sim D^m}\left(\forall x \in S \cdot c_1(x) = c_2(x) = 0\right) = \left(1 - \frac{1}{2^{2\rho}}\right)^{2m} . \tag{3}$$

In particular, if

$$m \leq \frac{\log(2)}{2\log(2^{2\rho}/(2^{2\rho}-1))} , \tag{4}$$

then the RHS of Equation 3 is at least $1/2$.

Now, let us consider the following limit, where ρ is a function of the input parameter n:

$$\begin{aligned}
\lim_{n\to\infty} 2^{\kappa\rho} \log\left(\frac{2^{2\rho}}{2^{2\rho}-1}\right) &= \frac{-\log(4)}{\kappa\log(2)} \lim_{n\to\infty} \frac{2^{\kappa\rho}}{1-2^{2\rho}} \\
&= \frac{-\log(4)}{\kappa\log(2)} \frac{\kappa\log(2)}{-2\log(2)} \lim_{n\to\infty} \frac{2^{\kappa\rho}}{2^{2\rho}} \\
&= \lim_{n\to\infty} 2^{(\kappa-2)\rho} \\
&= \begin{cases} 0 & \text{if } \kappa < 2 \\ 1 & \text{if } \kappa = 2 \\ \infty & \text{if } \kappa > 2 \end{cases} ,
\end{aligned}$$

where the first two equalities follow from l'Hôpital's rule.

Thus if $\kappa < 2$ then $2^{\kappa\rho}$ is $o\left(\left(\log\left(\frac{2^{2\rho}}{2^{2\rho}-1}\right)\right)^{-1}\right)$.

□

Remark 10. *Note that for a given $\kappa < 2$, the lower bound $2^{\kappa\rho}$ holds only for sufficiently large $\rho(n)$. By looking at Equation 3, and letting $m = 2^\rho$, we get that $\rho(n) \geq 2$ is a sufficient condition for it to hold. If we want a lower bound for robust learning that is larger than that of standard learning (where the dependence is $\Theta(n)$) for a $\log(n)$ adversary, setting $m = 2^{1.7\rho}$ and requiring $\rho(n) \geq 6$, for e.g., would be sufficient.*

5 Conclusion

We have shown that the class k-DL is efficiently robustly learnable against a logarithmically-bounded adversary, thus making progress on the open problem of [Gourdeau *et al.*, 2021] of whether PAC-learnable classes are always robust in general against a logarithmically-bounded adversary. The main technical tool was an isoperimetric result concerning CNF formulas. Moreover, we have shown that, for monotone conjunctions and any superclass thereof, any ρ-robust learning algorithm must have a sample complexity that is exponential in the adversarial budget ρ.

Deriving sample complexity bounds for the robust learnability of halfspaces under the uniform distribution is perhaps the most natural next step towards resolving the abovementioned open problem. Another direction of further research concerns improving the sample complexity bounds for k-DL in the present paper. Here we have used a proper PAC-learning algorithm as a black box in our robust learning procedure (see Corollary 4). By controlling the accuracy parameter of the standard PAC-learning algorithm, we are able to get a robust learning algorithm. From this, we get polynomial sample complexity upper bounds for k-DL in terms of the *robustness* accuracy parameter ε, the distribution parameter α, and the input dimension n. The resulting polynomial has degree $O(8^k(1+\alpha)^{2k^2})$ in the term $1/\varepsilon$ and degree $O(k8^k(1+\alpha)^{2k^2})$ in the dimension n. It is natural to ask whether these bounds can be improved in a significant way, e.g., by adapting the learning procedure to directly take robustness into account, rather than using a PAC-learning algorithm as a black box. Connected to this, we note that our lower bound focuses on establishing the exponential dependence of the number of samples on the robustness parameter. The bound is derived from the case of monotone conjunctions (a special case of 1-DL) under the uniform distribution and so does not mention k, nor the distribution parameter α. Likewise, it does not mention the desired accuracy ε. Deriving sample complexity lower bounds with a dependence on these parameters, potentially through other techniques, would help give a complete picture of the robust learnability of k-DL.

Acknowledgments

MK and PG received funding from the ERC under the European Union's Horizon 2020 research and innovation programme (FUN2MODEL, grant agreement No. 834115).

References

[Awasthi *et al.*, 2013] Pranjal Awasthi, Vitaly Feldman, and Varun Kanade. Learning using local membership queries. In *COLT*, volume 30, pages 1–34, 2013.

[Bhagoji *et al.*, 2019] Arjun Nitin Bhagoji, Daniel Cullina, and Prateek Mittal. Lower bounds on adversarial robustness from optimal transport. *arXiv preprint arXiv:1909.12272*, 2019.

[Biggio and Roli, 2017] Battista Biggio and Fabio Roli. Wild patterns: Ten years after the rise of adversarial machine learning. *arXiv preprint arXiv:1712.03141*, 2017.

[Biggio *et al.*, 2013] Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

[Bubeck *et al.*, 2019] Sebastien Bubeck, Yin Tat Lee, Eric Price, and Ilya Razenshteyn. Adversarial examples from computational constraints. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 831–840, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

[Dalvi *et al.*, 2004] Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, Deepak Verma, et al. Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 99–108. ACM, 2004.

[Diochnos *et al.*, 2018] Dimitrios Diochnos, Saeed Mahloujifar, and Mohammad Mahmoody. Adversarial risk and robustness: General definitions and implications for the uniform distribution. In *Advances in Neural Information Processing Systems*, 2018.

[Diochnos *et al.*, 2019] Dimitrios I Diochnos, Saeed Mahloujifar, and Mohammad Mahmoody. Lower bounds for adversarially robust pac learning. *arXiv preprint arXiv:1906.05815*, 2019.

[Diochnos *et al.*, 2020] Dimitrios I. Diochnos, Saeed Mahloujifar, and Mohammad Mahmoody. Lower bounds for adversarially robust PAC learning under evasion and hybrid attacks. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 717–722, 2020.

[Dreossi *et al.*, 2019] Tommaso Dreossi, Shromona Ghosh, Alberto Sangiovanni-Vincentelli, and Sanjit A Seshia. A formalization of robustness for deep neural networks. *arXiv preprint arXiv:1903.10033*, 2019.

[Fawzi *et al.*, 2016] Alhussein Fawzi, Seyed-Mohsen Moosavi-Dezfooli, and Pascal Frossard. Robustness of classifiers: from adversarial to random noise. In *Advances in Neural Information Processing Systems*, pages 1632–1640, 2016.

[Fawzi *et al.*, 2018a] Alhussein Fawzi, Hamza Fawzi, and Omar Fawzi. Adversarial vulnerability for any classifier. *arXiv preprint arXiv:1802.08686*, 2018.

[Fawzi *et al.*, 2018b] Alhussein Fawzi, Omar Fawzi, and Pascal Frossard. Analysis of classifiers? robustness to adversarial perturbations. *Machine Learning*, 107(3):481–508, 2018.

[Feige *et al.*, 2015] Uriel Feige, Yishay Mansour, and Robert Schapire. Learning and inference in the presence of corrupted inputs. In *Conference on Learning Theory*, pages 637–657, 2015.

[Gilmer *et al.*, 2018] Justin Gilmer, Luke Metz, Fartash Faghri, Samuel S Schoenholz, Maithra Raghu, Martin Wattenberg, and Ian Goodfellow. Adversarial spheres. *arXiv preprint arXiv:1801.02774*, 2018.

[Gourdeau *et al.*, 2019] Pascale Gourdeau, Varun Kanade, Marta Kwiatkowska, and James Worrell. On the hardness of robust classification. In *Advances in Neural Information Processing Systems*, pages 7444–7453, 2019.

[Gourdeau *et al.*, 2021] Pascale Gourdeau, Varun Kanade, Marta Kwiatkowska, and James Worrell. On the hardness of robust classification. *Journal of Machine Learning Research*, 22, 2021.

[Ilyas *et al.*, 2019] Andrew Ilyas, Shibani Santurkar, Dimitris Tsipras, Logan Engstrom, Brandon Tran, and Aleksander Madry. Adversarial examples are not bugs, they are features. *arXiv preprint arXiv:1905.02175*, 2019.

[Lowd and Meek, 2005a] Daniel Lowd and Christopher Meek. Adversarial learning. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 641–647. ACM, 2005.

[Lowd and Meek, 2005b] Daniel Lowd and Christopher Meek. Good word attacks on statistical spam filters. In *CEAS*, volume 2005, 2005.

[Mahloujifar and Mahmoody, 2019] Saeed Mahloujifar and Mohammad Mahmoody. Can adversarially robust learning leveragecomputational hardness? In *Algorithmic Learning Theory*, pages 581–609. PMLR, 2019.

[Mahloujifar *et al.*, 2019] Saeed Mahloujifar, Dimitrios I Diochnos, and Mohammad Mahmoody. The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure. *AAAI Conference on Artificial Intelligence*, 2019.

[Montasser *et al.*, 2019] Omar Montasser, Steve Hanneke, and Nathan Srebro. Vc classes are adversarially robustly learnable, but only improperly. In *Conference on Learning Theory*, pages 2512–2530. PMLR, 2019.

[Rivest, 1987] Ronald L Rivest. Learning decision lists. *Machine learning*, 2(3):229–246, 1987.

[Shafahi *et al.*, 2018] Ali Shafahi, W Ronny Huang, Christoph Studer, Soheil Feizi, and Tom Goldstein. Are adversarial examples inevitable? *arXiv preprint arXiv:1809.02104*, 2018.

[Szegedy *et al.*, 2013] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2013.

[Tsipras *et al.*, 2019] Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. In *International Conference on Learning Representations*, 2019.

[Valiant, 1984] Leslie G Valiant. A theory of the learnable. In *Proceedings of the sixteenth annual ACM symposium on Theory of computing*, pages 436–445. ACM, 1984.