When are Local Queries Useful for Robust Learning?

Pascale Gourdeau, Varun Kanade, Marta Kwiatkowska, and James Worrell

University of Oxford

Abstract

Distributional assumptions have been shown to be necessary for the robust learnability of concept classes when considering the exact-in-the-ball robust risk and access to random examples by Gourdeau et al. (2019). In this paper, we study learning models where the learner is given more power through the use of *local* queries, and give the first *distribution-free* algorithms that perform robust empirical risk minimization (ERM) for this notion of robustness. The first learning model we consider uses local membership queries (LMQ), where the learner can query the label of points near the training sample. We show that, under the uniform distribution, LMQs do not increase the robustness threshold of conjunctions and any superclass, e.g., decision lists and halfspaces. Faced with this negative result, we introduce the local *equivalence* query (LEQ) oracle, which returns whether the hypothesis and target concept agree in the perturbation region around a point in the training sample, as well as a counterexample if it exists. We show a separation result: on one hand, if the query radius λ is strictly smaller than the adversary's perturbation budget ρ, then distribution-free robust learning is impossible for a wide variety of concept classes; on the other hand, the setting $\lambda = \rho$ allows us to develop robust ERM algorithms. We then bound the query complexity of these algorithms based on online learning guarantees and further improve these bounds for the special case of conjunctions. We finish by giving robust learning algorithms for halfspaces with margins on both $\{0,1\}^n$ and $\mathbb{R}^n$.

1 Introduction

Adversarial examples have been widely studied since the work of (Dalvi et al., 2004; Lowd and Meek, 2005a,b), and later (Biggio et al., 2013; Szegedy et al., 2013), the latter having coined the term. As presented in Biggio and Roli (2017), two main settings exist for adversarial machine learning: *evasion* attacks, where an adversary perturbs data at test time, and *poisoning* attacks, where the data is modified at training time.

The majority of the guarantees and impossibility results for evasion attacks are based on the existence of adversarial examples, potentially crafted by an all-powerful adversary. However, what is considered to be an adversarial example has been defined in two different, and in some respects contradictory, ways in the literature. The *exact-in-the-ball* notion of robustness (also known as *error region* risk in Diochnos et al. (2018)) requires that the hypothesis and the ground truth agree in the perturbation region around each test point; the ground truth must thus be specified on all input points in the perturbation region. On the other hand, the constant-in-the-ball notion of robustness (which is also known as *corrupted input* robustness from the work of Feige et al. (2015)) requires that the unperturbed point be correctly classified and that the points in the perturbation region share its label, meaning that we only need access to the test point labels; see, e.g., (Diochnos et al., 2018; Dreossi et al., 2019; Gourdeau et al., 2021; Pydi and Jog, 2021) for thorough discussions on the subject.

We study the problem of robust classification against evasion attacks under the exact-in-the-ball definition of robustness. Previous work for this problem, e.g., (Diochnos et al., 2020; Gourdeau et al., 2021), has considered the setting where the learner only has access to random examples. However, many defences against evasion attacks have used adversarial training, the practice by which a dataset is augmented with previously misclassified points. Moreover, in the learning theory literature, some learning models give more power to the learner, e.g., by using membership and equivalence queries. Our work studies the robust learning problem mentioned above from a learning theory point of view, and investigates the power of *local queries* in this setting.

1.1 Our Contributions

We outline our contributions below. All our results use the *exact-in-the-ball* definition of robustness. Conceptually, we study the powers and limitations of robust learning with access to oracles that only reveal information nearby the training sample. Our results are particularly relevant as they contrast with the impossibility of robust learning in the *distribution-free* setting when only random examples are given, as demonstrated in Gourdeau et al. (2019).

Limitations of the Local Membership Query Model. In the local membership query (LMQ) model, the learner is allowed to query the label of points in the vicinity of the training sample. This model was introduced by Awasthi et al. (2013) and shown to guarantee the PAC learnability of various concept classes (which are believed or known to be hard to learn with only random examples) under distributional assumptions. However, we show that LMQs do not improve the robustness threshold of the class of conjunctions under the uniform distribution. Indeed, any ρ-robust learning algorithm will need a joint sample and query complexity that is exponential in ρ, and thus superpolynomial in the input dimension n against an adversary that can flip $\rho = \omega(\log n)$ input bits.

The Local Equivalence Query Model. Faced with the query lower bound for LMQ above, one may consider giving a different power to the learner to improve robust learning guarantees. We thus introduce the local equivalence query (LEQ) model, where the learner is allowed to query whether a hypothesis and the ground truth agree in the vicinity of points in the training sample. The LEQ oracle is the natural exact-in-the-ball analogue of the Perfect Attack Oracle introduced in Montasser et al. (2021), which was developed for the constant-in-the-ball robustness. It is also a variant of Angluin's equivalence query oracle (Angluin, 1987).

Distribution-Free Robust ERM with an LEQ Oracle. We show that having access to a *robustly* consistent learner (i.e., one that can get zero robust risk on the training sample) gives sample complexity upper bounds that are logarithmic in the size of the hypothesis class or linear in its *robust* VC dimension–a complexity measure adapted from Cullina et al. (2018) for our notion of robustness, which we develop in this paper. We study the setting where the learner has access to random examples and an LEQ oracle. In the case where the query radius λ of the LEQ oracle is strictly smaller than the adversarial perturbation budget ρ, we show that, for a wide variety of concept classes, distribution-free robust learning is impossible, regardless of the training sample size. In contrast, when $\lambda = \rho$ we exhibit robustly consistent learners that use an LEQ oracle. This separation result further validates the need for an LEQ oracle in the distribution-free setting. We furthermore use online learning setting results to exhibit upper bounds on the LEQ oracle query complexity and then improve these bounds in the specific case of conjunctions. Finally, we study the sample and query complexity of halfspaces on both $\{0,1\}^n$ and $\mathbb{R}^n$. To our knowledge, the results presented in this paper feature the first robust empirical risk minimization (ERM) algorithms for the *exact-in-the-ball* robust risk in the literature.[1]

1.2 Related Work

Learning with Membership and Equivalence Queries. Membership and equivalence queries (MQ and EQ, respectively) have been widely used in learning theory. Membership queries allow the learner to

[1] Note that previous work, e.g., Gourdeau et al. (2021), used PAC learning algorithms as black boxes, which are not in general robust risk minimizers, unless they also happen to be exact learning algorithms, and that (Montasser et al., 2019, 2021) use the constant-in-the-ball definition of robustness.

query the label of any point in the input space $\mathcal{X}$, namely, if the target concept is c, MQ returns $c(x)$ when queried with $x \in \mathcal{X}$. The goal is usually to learn the target c exactly. Recall that, in the probabilistically approximately correct (PAC) learning model of Valiant (1984), the learner has access to the example oracle $\mathsf{EX}(c, D)$, which upon being queried returns a point $x \sim D$ sampled from the underlying distribution and its label $c(x)$, and the goal is to output h such that with high probability h has low error.[2] The EQ oracle takes as input a hypothesis h and returns whether $h = c$, and provides a counterexample z such that $h(z) \neq c(z)$ otherwise. The seminal work of Angluin (1987) showed that deterministic finite automata (DFA) are exactly learnable with a polynomial number of queries to MQ and EQ in the size of the DFA. Many classes were then showed to be learnable in this setting as well as others, see e.g., (Bshouty, 1993; Angluin, 1988; Jackson, 1997). Moreover, the $\mathsf{MQ} + \mathsf{EQ}$ model has recently been used for recurrent and binarized neural networks (Weiss et al., 2018, 2019; Okudono et al., 2020; Shih et al., 2019), and interpretability (Camacho and McIlraith, 2019). But even these powerful learning models have limitations: learning DFAs only with EQ is hard (Angluin, 1990) and, under cryptographic assumptions, they are also hard to learn solely with the MQ oracle (Angluin and Kharitonov, 1995). It is also worth noting that the MQ learning model has been criticized by the applied machine learning community, as labels can be queried in the whole input space, irrespective of the distribution that generates the data. In particular, (Baum and Lang, 1992) observed that query points generated by a learning algorithm on the handwritten characters oftentimes appeared meaningless to human labellers. Awasthi et al. (2013) thus offered an alternative learning model to Valiant's original model, the PAC and local membership query ($\mathsf{EX} + \mathsf{LMQ}$) model, where the learning algorithm is only allowed to query the label of points that are close to examples from the training sample. Bary-Weisberg et al. (2020) later showed that many concept classes, including DFAs, remain hard to learn in the $\mathsf{EX} + \mathsf{LMQ}$.

Existence of Adversarial Examples. It has been shown that, in many instances, the vulnerability of learning models to adversarial examples is inevitable due to the nature of the learning problem. The majority of the results have been shown for the constant-in-the-ball notion of robustness, see e.g., (Fawzi et al., 2016, 2018a,b; Gilmer et al., 2018; Shafahi et al., 2018; Tsipras et al., 2019). As for the exact-in-the-ball definition of robustness, Diochnos et al. (2018) consider the robustness of monotone conjunctions under the uniform distribution. Using the isoperimetric inequality for the boolean hypercube, they show that an adversary that can perturb $O(\sqrt{n})$ bits can increase the misclassification error from 0.01 to 1/2. Mahloujifar et al. (2019) then generalize this result to Normal Lévy families and a class of well-behaved classification problems (i.e., ones where the error regions are measurable and average distances exist).

Sample Complexity of Robust Learning. Our work uses a similar approach to Cullina et al. (2018), who define the notion of adversarial VC dimension to derive sample complexity upper bounds for robust ERM algorithms, with respect to the constant-in-the-ball robust risk. Montasser et al. (2019) use the same notion of robustness and show sample complexity upper bounds for robust ERM algorithms that are polynomial in the VC and dual VC dimensions of concept classes, giving general upper bounds that are exponential in the VC dimension–though they sometimes must be achieved by an improper learner. Ashtiani et al. (2020) build on their work and delineate when proper robust learning is possible. On the other hand, (Khim et al., 2019; Yin et al., 2019; Awasthi et al., 2020) study *adversarial* Rademacher complexity bounds for robust learning, giving results for linear classifiers and neural networks when the robust risk can be minimized (in practice, this is approximated with adversarial training). Viallard et al. (2021) derive PAC-Bayesian generalization bounds for the averaged risk on the perturbations, rather than working in a worst-case scenario. As for the exact-in-the-ball definition of robustness, Diochnos et al. (2020) show that, for a wide family of concept classes, any learning algorithm that is robust against all $\rho = o(n)$ attacks must have a sample complexity that is at least an exponential in the input dimension n. They also show a superpolynomial lower bound in case $\rho = \Theta(\sqrt{n})$. Gourdeau et al. (2019) show that distribution-free robust learning is generally impossible. They also show that monotone conjunctions have a robustness threshold of $\Theta(\log n)$ under log-Lipschitz

[2] This is known as the realizable setting. It is also possible to have a distribution over the labels, in which case we are working in the *agnostic* setting.

distributions, meaning that this class is efficiently robustly learnable against an adversary that can perturb $\log n$ bits of the input, but if an adversary is allowed to perturb $\rho = \omega(\log n)$ bits of the input, there does not exist a sample-efficient learning algorithm for this problem. Gourdeau et al. (2021) extended this result to the class of monotone decision lists and Gourdeau et al. (2022) showed a sample complexity lower bound for monotone conjunctions that is exponential in ρ and that the robustness threshold of decision lists is also $\Theta(\log n)$. Finally, Diakonikolas et al. (2020) and Bhattacharjee et al. (2021) have used online learning algorithms for robust learning with respect to the constant-in-the-ball notion of robustness.

Restricting the Power of the Learner and the Adversary. Most adversarial learning guarantees and impossibility results in the literature have focused on all-powerful adversaries. Recent work has studied learning problems where the adversary's power is curtailed. E.g, Mahloujifar and Mahmoody (2019) and Garg et al. (2020) study the robustness of classifiers to polynomial-time attacks. Closest to our work, Montasser et al. (2020, 2021) study the sample and query complexity of robust learning with respect to the constant-in-the-ball robust risk when the learner has access to a Perfect Attack Oracle (PAO). For a perturbation type $\mathcal{U} : \mathcal{X} \to 2^{\mathcal{X}}$, hypothesis h and labelled point (x, y), the PAO returns the constant-in-the-ball robust loss of h in the perturbation region $\mathcal{U}(x)$ and a counterexample z where $h(z) \neq y$ if it exists. Our LEQ oracle is the natural analogue of the PAO oracle for our notion of robustness. In the constant-in-the-ball *realizable* setting,[3] the authors use online learning results to show sample and query complexity bounds that are linear and quadratic in the Littlestone dimension of concept classes, respectively (Montasser et al., 2020). Montasser et al. (2021) moreover use the algorithm from (Montasser et al., 2019) to get a sample complexity of $\tilde{O}\left(\frac{\mathsf{VC}(\mathcal{H})\mathsf{VC}^{*2}(\mathcal{H})+\log(1/\delta)}{\epsilon}\right)$ and query complexity of $\tilde{O}(2^{\mathsf{VC}(\mathcal{H})^2\mathsf{VC}^*(\mathcal{H})^2\log^2(\mathsf{VC}^*(\mathcal{H}))}\mathsf{Lit}(\mathcal{H}))$. Finally, they extend their results to the agnostic setting and derive lower bounds. As in the setting with having only access to the example oracle, different notions of robustness have vastly different implications in terms of robust learnability of certain concept classes. Whenever relevant, we will draw a thorough comparison in the next sections between our work and that of Montasser et al. (2021).

2 Problem Set Up

We work in the PAC learning framework (see Appendix A.1), with the distinction that a robust risk function is used instead of the standard risk. We will study metric spaces $(\mathcal{X}_n, d)$ of input dimension n with a perturbation budget function $\rho : \mathbb{N} \to \mathbb{R}$ defining the perturbation region $B_\rho(x) := \{z \in \mathcal{X}_n \mid d(x, z) \leq \rho(n)\}$. When the input space is the boolean hypercube $\mathcal{X}_n = \{0, 1\}^n$, the metric is the Hamming distance.

We use the exact-in-the-ball robust risk, which is defined w.r.t. a hypothesis h, target c and distribution D as the probability $\mathsf{R}_\rho^D(h, c) := \Pr_{x\sim D}(\exists z \in B_\rho(x) \;.\; c(z) \neq h(z))$ that h and c disagree in the perturbation region. On the other hand, the constant-in-the-ball robust risk is defined as $\Pr_{x\sim D}(\exists z \in B_\rho(x) \;.\; c(x) \neq h(z))$. Note that it is possible to adapt the latter to a joint distribution on the input and label spaces, but that there is an implicit *realizability assumption* in the former as the prediction on perturbed points' labels are compared to the ground truth c. We emphasize that choosing a robust risk function should depend on the learning problem at hand. The constant-in-the-ball notion of robustness requires a certain form of *stability*: the hypothesis should be correct on a random example and not change label in the perturbation region; this robust risk function may be more appropriate in settings with a strong margin assumption. In contrast, the exact-in-the-ball notion of robustness speaks to the *fidelity* of the hypothesis to the ground truth, and may be more suitable when a considerable portion of the probability mass is in the vicinity of the decision boundary. Diochnos et al. (2018); Dreossi et al. (2019); Gourdeau et al. (2021); Pydi and Jog (2021) offer a thorough comparison between different notions of robustness.

In the face of the impossibility or hardness of robustly learning certain concept classes, either through statistical or computational limitations, it is natural to study whether these issues can be circumvented by giving more power to the learner. The λ-local membership query (λ-LMQ) set up of Awasthi et al. (2013),

[3] I.e., there exists a hypothesis that has zero constant-in-the-ball robust loss.

which is formally defined in Appendix A.3, allows the learner to query the label of points that are at distance at most λ from a sample S drawn randomly from D. Inspired by this learning model, we define the λ-local equivalence query (λ-LEQ) model where, for a point x in a sample S drawn from the underlying distribution D, the learner is allowed to query an oracle that returns whether h agrees with the ground truth c in the ball $B_\lambda(x)$ of radius λ around x.[4] If they disagree, a counterexample in $B_\lambda(x)$ is returned as well. Clearly, by setting $\lambda = n$, we recover the EQ oracle.[5] Note moreover that when $\lambda = \rho$, this is equivalent to querying the (exact-in-the-ball) robust loss around a point. We will show a separation result for robust learning algorithms between models that only allow random examples and ones that allow random examples and access to LEQ.

Definition 1 (λ-LEQ Robust Learning)**.** *Let $\mathcal{X}_n$ be the instance space, $\mathcal{C}$ a concept class over $\mathcal{X}_n$, and $\mathcal{D}$ a class of distributions over $\mathcal{X}_n$. We say that $\mathcal{C}$ is ρ-robustly learnable using λ-local equivalence queries with respect to distribution class, $\mathcal{D}$, if there exists a learning algorithm, $\mathcal{A}$, such that for every $\epsilon > 0$, $\delta > 0$, for every distribution $D \in \mathcal{D}$ and every target concept $c \in \mathcal{C}$, the following hold:*[6]

1. *$\mathcal{A}$ draws a sample S of size $m = poly(n, 1/\delta, 1/\epsilon)$ using the example oracle $\mathsf{EX}(c, D)$*

2. *Each query made by $\mathcal{A}$ at $x \in S$ and for a candidate hypothesis h to λ-LEQ either confirms that c and h coincide on $B_\lambda(x)$ or returns $z \in B_\lambda(x)$ such that $c(z) \neq h(z)$. $\mathcal{A}$ is allowed to update h after seeing a counterexample*

3. *$\mathcal{A}$ outputs a hypothesis h that satisfies $\mathsf{R}_\rho^D(h, c) \leq \epsilon$ with probability at least $1 - \delta$*

4. *The running time of $\mathcal{A}$ (hence also the number of oracle accesses) is polynomial in n, $1/\epsilon$, $1/\delta$ and the output hypothesis h is polynomially evaluable.*

We remark that this model evokes the online learning setting, where the learner receives counterexamples after making a prediction, but with a few key differences. Contrary to the online setting (and the exact learning framework with MQ and EQ), there is an underlying distribution with which the performance of the hypothesis is evaluated in both the LMQ and LEQ models. Moreover, in online learning, when receiving a counterexample, the only requirement is that there is a concept that correctly classifies all the data given to the learner up until that point, and so the counterexamples can be given in an *adversarial* fashion, in order to maximize the regret. However, both the LMQ and LEQ models require that a target concept be chosen a priori. Note though that the LEQ oracle can give any counterexample for the robust loss at a given point.

In practice, one always has to find a way to approximately implement oracles studied in theory. A possible way to generate counterexamples with respect to the exact-in-the-ball notion of robustness is as follows. Suppose that there is an adversary that can generate points $z \in B_\rho(x)$ such that $h(z) \neq c(z)$. Provided such an adversary can be simulated, there is a way to (imperfectly) implement the LEQ oracle in practice.

Both the LMQ and LEQ models are particularly well-suited for the standard and exact-in-the-ball risks, as they address *information-theoretic* limitations of learning with random examples only. On the other hand, while information-theoretic limitations of robust learning with respect to the *constant-in-the-ball* notion of robustness arise when the perturbation function $\mathcal{U}$ is unknown to the learner, *computational* obstacles can also occur even when the definition of $\mathcal{U}$ is available. Indeed, determining whether the hypothesis changes label in the perturbation region could be intractable. In these cases, the Perfect Attack Oracle of Montasser et al. (2021) can be used to remedy these limitations for robust learning with respect to the constant-in-the-ball robust risk. Crucially, in their setting, counterexamples could have a different label to the ground truth: a counterexample $z \in \mathcal{U}(x)$ for x is such that $h(z) \neq c(x)$, not necessarily $h(z) \neq c(z)$. This could compromise the standard accuracy of the hypothesis (see e.g., Tsipras et al. (2019) for a learning problem where robustness and accuracy are at odds). Finally, an LMQ analogue for the constant-in-the-ball risk is not needed: the only information we need for a perturbed point $z \in B_\rho(x)$ is the label of x (given by the example oracle) and $h(z)$.

[4] Similarly to ρ, we implicitly consider λ as a function of the input dimension n. It is also possible to extend this definition to an arbitrary perturbation function $\mathcal{U} : \mathcal{X} \to 2^{\mathcal{X}}$.

[5] This is evidently not the case for the Perfect Attack Oracle of Montasser et al. (2021).

[6] We implicitly assume that a concept $c \in \mathcal{C}$ can be represented in size polynomial in n, where n is the input dimension; otherwise a parameter $size(c)$ can be introduced in the sample and query complexity requirements.

Given that one of the requirements of PAC learning is that the hypothesis is efficiently evaluatable, we can easily compute $h(z)$.

3 Distribution-Free Robust Learning with Local Equivalence Queries

In this section, we show that having access to a local equivalence query oracle can guarantee the efficient *distribution-free* robust learnability of certain concept classes. We start with a negative result which shows that for a wide variety of concept classes, if $\lambda < \rho$, then *distribution-free* robust learnability is impossible with $\mathsf{EX} + \lambda\text{-}\mathsf{LEQ}$ – regardless of how many queries are allowed. However, the regime $\lambda = \rho$, which implies giving similar power to the learner as the adversary, enables robust learnability guarantees. Indeed, Section 3.2 exhibits upper bounds on sample sizes that will guarantee *robust* generalization. These bounds are logarithmic in the size of the hypothesis class (finite case) and linear in the *robust* VC dimension of a concept class (infinite case). Section 3.3 draws a comparison between our framework and the online learning setting, and exhibits robustly consistent learners. Section 3.4 studies conjunctions and presents a robust learning algorithm that is *both* statistically and computationally efficient. Finally, Section 3.5 looks at linear classifiers in the discrete and continuous cases, and adapts the Winnow and Perceptron algorithms to both settings.

3.1 Impossibility of Distribution-Free Robust Learning When $\lambda < \rho$

We start with a negative result, saying that whenever the local query radius is strictly smaller than the adversary's budget, monotone conjunctions are not distribution-free robustly learnable, which is in contrast to the standard PAC setting where guarantees hold *for any distribution.* Note that our result goes beyond efficiency: no query can distinguish between two potential targets. Choosing the target uniformly at random lower bounds the expected robust risk, and hence renders robust learning impossible in this setting. The proof of this theorem can be found in Appendix C.5.

Theorem 2. *For locality and robustness parameters $\lambda, \rho \in \mathbb{N}$ with $\lambda < \rho$, monotone conjunctions (and any superclass) are not distribution-free ρ-robustly learnable with access to a λ-*LEQ *oracle.*

The result holds for monotone conjunctions and all superclasses (e.g., decision lists and halfspaces), but, in fact, we can generalize this reasoning for any concept class that has a certain form of stability: if we can find concepts c_1 and c_2 in $\mathcal{C}$ and points $x, x' \in \mathcal{X}$ such that c_1 and c_2 agree on $B_\lambda(x)$ but disagree on x', then if $\lambda < \rho$, the concept class $\mathcal{C}$ is not distribution-free ρ-robustly learnable with access to a λ-LEQ oracle. It suffices to "move" the center of the ball x until we find a point in the set $B_\rho(x) \setminus B_\lambda(x)$ where c_1 and c_2 disagree, which is guaranteed to happen by the existence of x'.

3.2 General Sample Complexity Bounds for Robustly-Consistent Learners

In this section, we show that we can derive sample complexity upper bounds for *robustly* consistent learners, i.e., learning algorithms that return a *robust* loss of zero on a training sample. Note that, crucially, the exact-in-the-ball notion of robustness and its realizability imply that any robust ERM algorithm will achieve zero empirical robust loss on a given training sample. As we will see in the next sections, the challenge is to find a *robustly* consistent learning algorithm that uses queries to ρ-LEQ. The first bound is for finite classes, where the dependency is logarithmic in the size of the hypothesis class. The proof is a simple application of Occam's razor and is included in Appendix C.2 for completeness. The reasoning is similar to Bubeck et al. (2019).

Lemma 3. *Let $\mathcal{C}$ be a concept class and $\mathcal{H}$ a hypothesis class. Any ρ-robust ERM algorithm using $\mathcal{H}$ on a sample of size $m \geq \frac{1}{\epsilon}\left(\log|\mathcal{H}_n| + \log\frac{1}{\delta}\right)$ is a ρ-robust learner for $\mathcal{C}$.*

For the infinite case, we cannot immediately use the VC dimension as a tool for bounding the sample complexity of robust learning. To this end, we define the *robust* VC dimension of a concept class, which is the

VC dimension of the class of functions representing the ρ-expansion of the error region between any possible target and hypothesis. This definition is analogous to the adversarial VC dimension defined by Cullina et al. (2018) for the constant-in-the-ball definition of robustness.

Definition 4 (Robust VC dimension)**.** *Given a target concept class $\mathcal{C}$, a hypothesis class $\mathcal{H}$ and a robustness parameter ρ, the robust VC dimension is defined as $\mathsf{RVC}_\rho(\mathcal{C},\mathcal{H}) = \mathsf{VC}((\mathcal{C}\oplus\mathcal{H})_\rho)$, where $(\mathcal{C}\oplus\mathcal{H})_\rho = \{(c\oplus h)_\rho : x \mapsto \mathbf{1}[\exists z\in B_\rho(x)\ .\ c(z)\neq h(z)] \mid c\in\mathcal{C}, h\in\mathcal{H}\}$. Whenever $\mathcal{C}=\mathcal{H}$, we simply write $\mathsf{RVC}_\rho(\mathcal{C})$.*

We now show that we can use the robust VC dimension to upper bound the sample complexity of robustly-consistent learning algorithms. We will use this result in Section 3.5 when dealing with an infinite concept class: halfspaces on $\mathbb{R}^n$.

Lemma 5. *Let $\mathcal{C}$ be a concept class and $\mathcal{H}$ a hypothesis class. Any ρ-robust ERM algorithm using $\mathcal{H}$ on a sample of size $m \geq \frac{1}{\epsilon}\left(\mathsf{RVC}_\rho(\mathcal{C},\mathcal{H})\log(1/\epsilon)+\log\frac{1}{\delta}\right)$ is a ρ-robust learner for $\mathcal{C}$.*

Proof Sketch of Lemma 5. The proof is very similar to the VC dimension upper bound in PAC learning. The main distinction is that, instead of looking at the error region of the target and any function in $\mathcal{H}$, we look at its ρ-expansion. Namely, let the target $c\in\mathcal{C}$ be fixed and, for $h\in\mathcal{H}$, consider the function $(c\oplus h)_\rho : x\mapsto \mathbf{1}[\exists z\in B_\rho(x)\ .\ c(z)\neq h(z)]$ and define a new concept class $\Delta_{c,\rho}(\mathcal{H}) = \{(c\oplus h)_\rho \mid h\in\mathcal{H}\}$. It is easy to show that $\mathsf{VC}(\Delta_{c,\rho}(\mathcal{H}))\leq \mathsf{RVC}_\rho(\mathcal{C},\mathcal{H})$, as any sign pattern achieved on the LHS can be achieved on the RHS. The rest of the proof follows from the definition of an ϵ-net and the bound on the growth function of $\Delta_{c,\rho}(\mathcal{H})$; see Appendix C.3 for details. □

Remark 6. Note that, as $\rho(n)/n$ tends to 1, we move towards the exact and online learning settings, and the underlying distribution becomes less important. In this case, the robust VC dimension starts to decrease. Indeed, say if $\rho=n$, then $(\mathcal{C}\oplus\mathcal{C})_\rho$ only contains the constant functions 0 and 1. We thus only need a single example to query the LEQ oracle (which has become the EQ oracle). However, this comes at a cost: the *query complexity* upper bounds presented in the next sections could be tight. Understanding the behaviour of the robust VC dimension as a function of ρ and deriving joint sample and query complexity bounds are both avenues for future research.

3.3 Query Complexity Bounds Using Online Learning Results

In the previous section, we derived sample complexity upper bounds for robustly consistent learners. The challenge is thus to create algorithms that perform robust empirical risk minimization, as we are operating in the realizable setting. We begin by showing that, if one can ignore computational limitations, then online learning results can be used to guarantee robust learnability. We recall the online learning setting in Appendix A.5. We denote by $\mathsf{Lit}(\mathcal{C})$ the Littlestone dimension of a concept class $\mathcal{C}$, which is defined in Appendix A.4 and appears in the query complexity bound in the theorem below, whose proof can be found in Appendix C.4.

Theorem 7. *A concept class $\mathcal{C}$ is ρ-robustly learnable with the Standard Optimal Algorithm (SOA) (Littlestone, 1988) using the EX and ρ-LEQ oracles with sample complexity $m(n,\epsilon,\delta)=\frac{1}{\epsilon}\left(\mathsf{RVC}_\rho(\mathcal{C})+\log\frac{1}{\delta}\right)$ and query complexity $r(n,\epsilon,\delta)=m(n,\epsilon,\delta)\cdot\mathsf{Lit}(\mathcal{C})$. Furthermore, if $\mathcal{C}$ is a finite concept class on $\{0,1\}^n$, then $\mathcal{C}$ is ρ-robustly learnable with sample complexity $m(n,\epsilon,\delta)=\frac{1}{\epsilon}\left(\log(|\mathcal{C}|)+\log\frac{1}{\delta}\right)$ and query complexity $r(n,\epsilon,\delta,\rho)=m(n,\epsilon,\delta)\cdot\min\{\mathsf{Lit}(\mathcal{C}),\rho\log(en)\}$.*

Of course, some concept classes, e.g., thresholds, have infinite Littlestone dimension, so our bounds are not useful in these settings. In Section 3.5, we study distributional assumptions that give reasonable query upper bounds for linear classifiers, using the theorem below. It exhibits a query upper bound for robustly learning with an online algorithm $\mathcal{A}$ with a given mistake upper bound M. This is moreover particularly useful in case M is polynomial in the input dimension and $\mathcal{A}$ is *computationally* efficient (which is not the case for the Standard Optimal Algorithm in Theorem 7).

Lemma 8. *Let $\mathcal{C}$ be a concept class learnable in the online setting with mistake bound $M(n)$. Then $\mathcal{C}$ is ρ-robustly learnable using the* EX *and* ρ-LEQ *oracles with sample complexity* $m(n,\epsilon,\delta) = \frac{1}{\epsilon}\left(\mathsf{RVC}_\rho(\mathcal{H},C) + \log\frac{1}{\delta}\right)$ *and query complexity* $r(n,\epsilon,\delta) = m(n,\epsilon,\delta)\cdot M(n)$.

Proof. The sample complexity bound is obtained from Lemma 5 and, for each point in the sample, a query to LEQ can either return a robust loss of 0 or 1 and give a counterexample. Since the mistake bound is M_U, we have a query upper bound of $r = m\cdot M$, as required. □

3.4 Improved Query Complexity Bounds: Conjunctions

In this section, we show how to improve the query upper bound from the previous section in the special case of conjunctions. Moreover, the algorithm used to robustly learn conjunctions is both statistically and *computationally* efficient, which is not the case of the Standard Optimal Algorithm. The proof of the following theorem can be found in Appendix C.5.

Theorem 9. *The class CONJUNCTIONS is efficiently ρ-robustly learnable in the distribution-free setting using the* EX *and* ρ-LEQ *oracles with at most* $O\left(\frac{1}{\epsilon}\left(n+\log\frac{1}{\delta}\right)\right)$ *random examples and* $O\left(\frac{1}{\epsilon}\left(n+\log\frac{1}{\delta}\right)\right)$ *queries to* ρ-LEQ.

Note that the query upper bound that we get is of the form $m+M$, as opposed to $m\cdot M$ from Lemma 5 (where m is the sample complexity and M the mistake bound). This is because we have adapted the PAC learning algorithm for conjunctions to our setting. Any update to its hypothesis will not affect the consistency of previously queried points with robust loss of zero, and thus once zero robust loss is achieved on a point, it does not need to be queried again.

3.5 Linear Classifiers

In this section, we derive sample and query complexity upper bounds for restricted subclasses of linear classifiers. We start with linear classifiers on $\{0,1\}^n$ with bounded weights, and continue with linear classifiers on $\mathbb{R}^n$ with a margin condition. We use the well-known Winnow and Perceptron algorithms. Note that the robustness threshold[7] of linear classifiers on $\{0,1\}^n$ *without* access to the LEQ oracle remains an open problem (Gourdeau et al., 2022).

Let $\mathsf{LTF}^W_{\{0,1\}^n}$ be the class of linear threshold functions on $\{0,1\}^n$ with integer weights such that the sum of the absolute values of the weights and the bias is bounded above by W. We have the following theorem, whose proof relies on bounding the size of $\mathsf{LTF}^W_{\{0,1\}^n}$ and using the mistake bound for Winnow (Littlestone, 1988). The proof can be found in Appendix C.6.

Theorem 10. *The class $\mathsf{LTF}^W_{\{0,1\}^n}$ is ρ-robustly learnable with access to the* EX *and* ρ-LEQ *oracles by using the Winnow algorithm with sample complexity* $m(n,\epsilon,\delta) = O\left(\frac{1}{\epsilon}\left(n+\min\{n,W\}\log(W+n)+\log\frac{1}{\delta}\right)\right)$ *and query complexity* $O(m(n,\epsilon,\delta)\cdot W^2\log(n))$.

Now, we derive sample and query complexity upper bounds for the robust learnability of linear classifiers $\mathsf{LTF}_{\mathbb{R}^n}$ on $\mathbb{R}^n$. Note that, unlike in previous results, the distribution family is restricted to guarantee the existence of a margin for each concept and distribution pair, and so we cannot guarantee distribution-free robust learning in this case. This is because the Littlestone dimension of thresholds, and thus halfspaces, is infinite if there are no distributional assumptions on this concept class. We remark that whenever the margin γ is greater than $\rho/2$, the constant and exact-in-the-ball notions of robustness could coincide,[8] in which case the results from (Diakonikolas et al., 2020; Montasser et al., 2021) apply. However, unlike in Diakonikolas et al. (2020); Montasser et al. (2021), under our notion of robustness, if $\gamma < \rho/2$, we may still be in the realizable setting (there exists at least one concept that is robustly consistent with the data), while when considering the constant-in-the-ball risk, we are necessarily in the non-realizable/agnostic setting. As

[7] With respect to the exact-in-the-ball definition of robustness.

[8] Given that the choice of target implies constant-in-the-ball realizability.

mentioned earlier, guarantees obtained in the latter do not necessarily translate to the former. The full proof of the theorem below appears in Appendix C.7.

Theorem 11. *Fix constants $B, \gamma > 0$. Let $\mathcal{L} = \{(c, D) \mid c \in \mathsf{LTF}_{\mathbb{R}^n}, D \in \mathcal{D}\}$ be a family of halfspace and distribution pairs, where each pair (c, D) with $c(x) = a^\top x + a_0$ is such that if $x \in supp(D)$, then (i) $\|x\|_2 \leq B$ and (ii) $\gamma \leq \frac{c(x)(a^\top x)}{\|x\|_2}$, i.e., D has support bounded by B and induces a margin of γ w.r.t. c. Let the adversary's budget be measured by the ℓ_2 norm. Then, $\mathcal{L}$ is ρ-robustly learnable using the* EX *and ρ-*LEQ *oracles with sample complexity $m = O(\frac{1}{\epsilon}(n^3 + \log(1/\delta)))$ and query complexity $r = \frac{mB^2}{\gamma^2}$. Note that this is query-efficient if $\frac{B^2}{\gamma^2} = poly(n)$.*

Proof Sketch. The first step is to derive the sample complexity bound. To this end, we use Lemma 5 and bound the robust VC dimension of linear classifiers on $\mathbb{R}^n$. We do this using a result of Goldberg and Jerrum (1995) (Theorem 28 in Appendix C.7), which bounds the VC dimension of concept classes expressible as boolean combinations of polynomial inequalities. We first express the ρ-expansion of the error region, i.e., the robust loss, between two linear classifiers as a first order logical formula ψ over the reals where the atomic predicates are polynomial inequalities. We then use the quantifier-elimination method from Renegar (1992) to transform ψ into a quantifier-free formula φ. This method allows us to show an upper bound on the number of atomic predicates, their degree, and the number of variables in φ. We can apply the result of Goldberg and Jerrum (1995) on φ to get a robust VC dimension of $O(n^3)$.

The second step is to derive the query upper bound, which follows from Lemma 8 and the mistake bound for the Perceptron algorithm, which appears in Appendix B. □

4 A Local Membership Query Lower Bound for Conjunctions

In this section, we show that the amount of data needed to ρ-robustly learn conjunctions under the uniform distribution has an exponential dependence on the adversary's budget ρ when the learner only has access to the EX and LMQ oracles. Here, the lower bound on the sample drawn from the example oracle is 2^ρ, which is the same as the lower bound for *monotone* conjunctions derived in Gourdeau et al. (2022), and the local membership query lower bound is $2^{\rho-1}$. The result relies on showing there there exists a family of conjunctions that remain indistinguishable from each other on any sample of size 2^ρ and any sequence of $2^{\rho-1}$ LMQs with constant probability.

Theorem 12. *Fix a monotone increasing robustness function $\rho : \mathbb{N} \to \mathbb{N}$ satisfying $2 \leq \rho(n) \leq n/4$ for all n. Then, for any query radius λ, any $\rho(n)$-robust learning algorithm for the class* CONJUNCTIONS *with access to the* EX *and λ-*LMQ *oracles has joint sample and query complexity lower bounds of 2^ρ and $2^{\rho-1}$ under the uniform distribution.*

Proof. Let D be the uniform distribution and WLOG let $\rho \geq 2$. Fix two disjoint sets I_1 and I_2 of 2ρ indices in $[n]$, which will be the set of variables appearing in potential target conjunctions c_1 and c_2, respectively (i.e., their support). We have $2^{4\rho}$ possible pairs of such conjunctions, as each variable can appear as a positive or negative literal.

Let us consider a randomly drawn sample S of size 2^ρ. We will first consider what happens when all the examples in S and the queried inputs S' are negatively labelled. Each negative example $x \in S$ allows us to remove at most $2^{2\rho+1}$ pairs from the possible set of pairs of conjunctions, as each component x_{I_1} and x_{I_2} removes at most one conjunction from the possible targets. By the same reasoning, each LMQ that returns a negative example can remove at most $2^{2\rho+1}$ pairs of conjunctions. Note that the parameter λ is irrelevant in this setting as each LMQ can only test one concept pair. Thus, after seeing any random sample of size 2^ρ and querying any $2^{\rho-1}$ points, there remains

$$\frac{2^{4\rho} - 2^{3\rho+1} - 2^{3\rho}}{2^{4\rho}} \geq 1/4 \tag{1}$$

of the initial conjunction pairs that label all points in S and S' negatively. Then, choosing a pair (c_1, c_2) of possible target conjunctions uniformly at random and then choosing c uniformly at random gives at least a $1/4$ chance that S and S' only contain negative examples (both conjunctions are consistent with this).

Moreover, note that any two conjunctions in a pair will have a robust risk lower bounded by $15/32$ against each other under the uniform distribution (see Lemma 23 in Appendix B). Thus, any learning algorithm $\mathcal{A}$ with LMQ query budget $m' = 2^{\rho-1}$ and strategy $\sigma : (\{0,1\}^n \times \{0,1\})^m \to (\{0,1\}^n \times \{0,1\})^{m'}$ (note that the queries can be adaptive) can do no better than to guess which of c_1 or c_2 is the target if they are both consistent on the augmented sample $S \cup \sigma(S)$, giving an expected robust risk lower bounded by a constant. Letting $\mathcal{E}$ be the event that all points in both S and $\sigma(S)$ are labelled zero, we get

$$\begin{aligned}
\mathop{\mathbb{E}}_{c,S}\left[\mathsf{R}_\rho^D(\mathcal{A}(S\cup\sigma(S)),c)\right] &\geq \Pr_{c,S}(\mathcal{E}) \mathop{\mathbb{E}}_{c,S}\left[\mathsf{R}_\rho^D(\mathcal{A}(S\cup\sigma(S)),c) \mid \mathcal{E}\right] && \text{(Law of Total Expectation)}\\
&\geq \frac{1}{4} \mathop{\mathbb{E}}_{c,S}\left[\mathsf{R}_\rho^D(\mathcal{A}(S\cup\sigma(S)),c) \mid \mathcal{E}\right] && \text{(Equation 1)}\\
&= \frac{1}{4}\cdot\frac{1}{2} \mathop{\mathbb{E}}_{S}\left[\mathsf{R}_\rho^D(\mathcal{A}(S\cup\sigma(S)),c_1) + \mathsf{R}_\rho^D(\mathcal{A}(S\cup\sigma(S)),c_2) \mid \mathcal{E}\right] && \text{(Random choice of } c\text{)}\\
&\geq \frac{1}{8} \mathop{\mathbb{E}}_{S}\left[\mathsf{R}_\rho^D(c_1,c_2) \mid \mathcal{E}\right] && \text{(Lemma 22)}\\
&> \frac{1}{8}\cdot\frac{15}{32} && \text{(Lemma 23)}\\
&= \frac{15}{256}\ ,
\end{aligned}$$

which completes the proof. □

We use the term *robustness threshold* from Gourdeau et al. (2021) to denote an adversarial budget function $\rho : \mathbb{N} \to \mathbb{R}$ of the input dimension n such that, if the adversary is allowed perturbations of magnitude $\rho(n)$, then there exists a sample-efficient $\rho(n)$-robust learning algorithm, and if the adversary's budget is $\omega(\rho(n))$, then there does not exist such an algorithm. Robustness thresholds are distribution-dependent when the learner only has access to the example oracle EX, as seen in (Gourdeau et al., 2021, 2022). Now, since the local membership query lower bound above has an exponential dependence on ρ, any perturbation budget $\omega(\log n)$ will require a sample and query complexity that is superpolynomial in n, giving the following corollary.

Corollary 13. *The robustness threshold of the class* $\mathsf{CONJUNCTIONS}$ *under the uniform distribution with access to* EX *and an* LMQ *oracle is* $\Theta(\log(n))$.

The robustness threshold above is the same as when only using the EX oracle (Gourdeau et al., 2021). Finally, since decision lists and halfspaces both subsume conjunctions, the lower bound of Theorem 12 also holds for these classes.

5 Conclusion

We have shown that local membership queries do not change the robustness threshold of conjunctions, or any superclass, under the uniform distribution. However, access to a ρ-local *equivalence* query oracle allows us to develop robust ERM algorithms. We have introduced the notion of robust VC dimension to determine sample complexity bounds and have used online learning results to derive query complexity bounds. We have moreover adapted the PAC learning algorithm for conjunctions for this setting and have greatly improved its query complexity compared to the general case. Finally, we have studied halfspaces, both in the boolean hypercube and continuous settings. The latter is, to our knowledge, the first robust learning algorithm with respect to the exact-in-the-ball notion of robustness for a non-trivial concept class in $\mathbb{R}^n$. Overall, we have shown that the LEQ oracle is *essential* to ensure the *distribution-free* robust learning of commonly studied concept classes in our setting. Note that this is in contrast with standard PAC learning with the EX and EQ oracles, where EQs don't give more power to learner.

We finally outline various avenues for future research:

1. Can we give a more fine-grained picture of the sample and query complexity tradeoff outlined in Remark 6, e.g., by improving LEQ query upper bounds when ρ is small?
2. Can we derive sample and query lower bounds for robust learning with an LEQ oracle?
3. The LMQ lower bound from Section 4 was derived for conjunctions. The technique does not work for monotone conjunctions.[9] Can we get a similar LMQ lower bound where the dependence on ρ is exponential for monotone conjunctions, or it is possible to robustly learn them with $o(2^\rho)$ local membership queries?

Acknowledgements

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

Angluin, D. (1987). Learning regular sets from queries and counterexamples. *Information and computation*, 75(2):87–106.

Angluin, D. (1988). Queries and concept learning. *Machine learning*, 2(4):319–342.

Angluin, D. (1990). Negative results for equivalence queries. *Machine Learning*, 5(2):121–150.

Angluin, D. and Kharitonov, M. (1995). When won't membership queries help? *Journal of Computer and System Sciences*, 50(2):336–355.

Ashtiani, H., Pathak, V., and Urner, R. (2020). Black-box certification and learning under adversarial perturbations. In *International Conference on Machine Learning*, pages 388–398. PMLR.

Awasthi, P., Feldman, V., and Kanade, V. (2013). Learning using local membership queries. In *COLT*, volume 30, pages 1–34.

Awasthi, P., Frank, N., and Mohri, M. (2020). Adversarial learning guarantees for linear hypotheses and neural networks. In *International Conference on Machine Learning*, pages 431–441. PMLR.

Bary-Weisberg, G., Daniely, A., and Shalev-Shwartz, S. (2020). Distribution free learning with local queries. In *Algorithmic Learning Theory*, pages 133–147. PMLR.

Baum, E. B. and Lang, K. (1992). Query learning can work poorly when a human oracle is used. In *International joint conference on neural networks*, volume 8, page 8. Beijing China.

Bhattacharjee, R., Jha, S., and Chaudhuri, K. (2021). Sample complexity of robust linear classification on separated data. In *International Conference on Machine Learning*, pages 884–893. PMLR.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. (2013). Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer.

Biggio, B. and Roli, F. (2017). Wild patterns: Ten years after the rise of adversarial machine learning. *arXiv preprint arXiv:1712.03141*.

[9] For a given set of indices I, there exists only one monotone conjunction using all indices in I.

Bshouty, N. H. (1993). Exact learning via the monotone theory. In *Proceedings of 1993 IEEE 34th Annual Foundations of Computer Science*, pages 302–311. IEEE.

Bubeck, S., Lee, Y. T., Price, E., and Razenshteyn, I. (2019). 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. PMLR.

Camacho, A. and McIlraith, S. A. (2019). Learning interpretable models expressed in linear temporal logic. In *Proceedings of the International Conference on Automated Planning and Scheduling*, volume 29, pages 621–630.

Cullina, D., Bhagoji, A. N., and Mittal, P. (2018). PAC-learning in the presence of evasion adversaries. *Advances in Neural Information Processing Systems*.

Dalvi, N., Domingos, P., Sanghai, S., Verma, D., et al. (2004). Adversarial classification. In *Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 99–108. ACM.

Diakonikolas, I., Kane, D. M., and Manurangsi, P. (2020). The complexity of adversarially robust proper learning of halfspaces with agnostic noise. *Advances in Neural Information Processing Systems*, 33:20449–20461.

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*.

Diochnos, D. I., Mahloujifar, S., and Mahmoody, M. (2020). 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.

Dreossi, T., Ghosh, S., Sangiovanni-Vincentelli, A., and Seshia, S. A. (2019). A formalization of robustness for deep neural networks. *arXiv preprint arXiv:1903.10033*.

Fawzi, A., Fawzi, H., and Fawzi, O. (2018a). Adversarial vulnerability for any classifier. *arXiv preprint arXiv:1802.08686*.

Fawzi, A., Fawzi, O., and Frossard, P. (2018b). Analysis of classifiers? robustness to adversarial perturbations. *Machine Learning*, 107(3):481–508.

Fawzi, A., Moosavi-Dezfooli, S.-M., and Frossard, P. (2016). Robustness of classifiers: from adversarial to random noise. In *Advances in Neural Information Processing Systems*, pages 1632–1640.

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

Garg, S., Jha, S., Mahloujifar, S., and Mohammad, M. (2020). Adversarially robust learning could leverage computational hardness. In *Algorithmic Learning Theory*, pages 364–385. PMLR.

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

Goldberg, P. W. and Jerrum, M. R. (1995). Bounding the vapnik-chervonenkis dimension of concept classes parameterized by real numbers. *Machine Learning*, 18(2-3):131–148.

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

Gourdeau, P., Kanade, V., Kwiatkowska, M., and Worrell, J. (2021). On the hardness of robust classification. *Journal of Machine Learning Research*, 22.

Gourdeau, P., Kanade, V., Kwiatkowska, M., and Worrell, J. (2022). Sample complexity bounds for robustly learning decision lists against evasion attacks. In *International Joint Conference in Artificial Intelligence*.

Jackson, J. C. (1997). An efficient membership-query algorithm for learning dnf with respect to the uniform distribution. *Journal of Computer and System Sciences*, 55(3):414–440.

Khim, J., Jog, V., and Loh, P.-L. (2019). Adversarial influence maximization. In *2019 IEEE International Symposium on Information Theory (ISIT)*, pages 1–5. IEEE.

Littlestone, N. (1988). Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm. *Machine learning*, 2(4):285–318.

Lowd, D. and Meek, C. (2005a). Adversarial learning. In *Proceedings of the eleventh ACM SIGKDD international conference on Knowledge discovery in data mining*, pages 641–647. ACM.

Lowd, D. and Meek, C. (2005b). Good word attacks on statistical spam filters. In *CEAS*, volume 2005.

Mahloujifar, S., Diochnos, D. I., and Mahmoody, M. (2019). The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure. *AAAI Conference on Artificial Intelligence*.

Mahloujifar, S. and Mahmoody, M. (2019). Can adversarially robust learning leveragecomputational hardness? In *Algorithmic Learning Theory*, pages 581–609. PMLR.

Mohri, M., Rostamizadeh, A., and Talwalkar, A. (2012). *Foundations of machine learning*. MIT press.

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

Montasser, O., Hanneke, S., and Srebro, N. (2020). Reducing adversarially robust learning to non-robust pac learning. *Advances in Neural Information Processing Systems*, 33:14626–14637.

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

Okudono, T., Waga, M., Sekiyama, T., and Hasuo, I. (2020). Weighted automata extraction from recurrent neural networks via regression on state spaces. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 5306–5314.

Pydi, M. S. and Jog, V. (2021). The many faces of adversarial risk. *Advances in Neural Information Processing Systems*, 34.

Renegar, J. (1992). On the computational complexity and geometry of the first-order theory of the reals. part i: Introduction. preliminaries. the geometry of semi-algebraic sets. the decision problem for the existential theory of the reals. *Journal of symbolic computation*, 13(3):255–299.

Shafahi, A., Huang, W. R., Studer, C., Feizi, S., and Goldstein, T. (2018). Are adversarial examples inevitable? *arXiv preprint arXiv:1809.02104*.

Shih, A., Darwiche, A., and Choi, A. (2019). Verifying binarized neural networks by angluin-style learning. In *International Conference on Theory and Applications of Satisfiability Testing*, pages 354–370. Springer.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. (2013). Intriguing properties of neural networks. In *International Conference on Learning Representations*.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. (2019). Robustness may be at odds with accuracy. In *International Conference on Learning Representations.*

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

Viallard, P., VIDOT, E. G., Habrard, A., and Morvant, E. (2021). A pac-bayes analysis of adversarial robustness. *Advances in Neural Information Processing Systems*, 34.

Weiss, G., Goldberg, Y., and Yahav, E. (2018). Extracting automata from recurrent neural networks using queries and counterexamples. In *International Conference on Machine Learning*, pages 5247–5256. PMLR.

Weiss, G., Goldberg, Y., and Yahav, E. (2019). Learning deterministic weighted automata with queries and counterexamples. *Advances in Neural Information Processing Systems*, 32.

Yin, D., Kannan, R., and Bartlett, P. (2019). Rademacher complexity for adversarially robust generalization. In *International conference on machine learning*, pages 7085–7094. PMLR.

A Preliminaries

A.1 The PAC Framework

Definition 14 (PAC Learning, Valiant (1984))**.** *Let $\mathcal{C}_n$ be a concept class over $\mathcal{X}_n$ and let $\mathcal{C} = \bigcup_{n\in\mathbb{N}} \mathcal{C}_n$. We say that $\mathcal{C}$ is* PAC learnable *using hypothesis class $\mathcal{H}$ and sample complexity function $p(\cdot,\cdot,\cdot,\cdot)$ if there exists an algorithm $\mathcal{A}$ that satisfies the following: for all $n \in \mathbb{N}$, for every $c \in \mathcal{C}_n$, for every D over $\mathcal{X}_n$, for every $0 < \epsilon < 1/2$ and $0 < \delta < 1/2$, if whenever $\mathcal{A}$ is given access to $m \geq p(n, 1/\epsilon, 1/\delta, size(c))$ examples drawn i.i.d. from D and labeled with c, $\mathcal{A}$ outputs a polynomially evaluatable $h \in \mathcal{H}$ such that with probability at least $1 - \delta$,*

$$\Pr_{x\sim D}\left(c(x) \neq h(x)\right) \leq \epsilon \ .$$

We say that $\mathcal{C}$ is statistically efficiently PAC learnable if p is polynomial in $n, 1/\epsilon$, $1/\delta$ and $size(c)$, and computationally efficiently PAC learnable if $\mathcal{A}$ runs in polynomial time in $n, 1/\epsilon$, $1/\delta$ and $size(c)$.

The setting where $\mathcal{C} = \mathcal{H}$ is called *proper learning*, and *improper learning* otherwise. The PAC setting where the guarantees hold for any distribution is called *distribution-free.*

A.2 Robust Learnability

Definition 15 (Efficient Robust Learnability, 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 $\Pr_{S\sim D^m}\left(\mathsf{R}_\rho(h, c) < \epsilon\right) > 1 - \delta$.*

A.3 Local Membership Queries and Robust Learning

We recall the formal definition of the LMQ model from (Awasthi et al., 2013), but where we have changed the standard risk to the robust risk. Here, given a sample S drawn from the example oracle, a membership query for a point x is λ-local if there exists $x' \in S$ such that $x \in B_\lambda(x')$.

Definition 16 (λ-LMQ Robust Learning)**.** *Let $\mathcal{X}$ be the instance space, $\mathcal{C}$ a concept class over $\mathcal{X}$, and $\mathcal{D}$ a class of distributions over $\mathcal{X}$. We say that $\mathcal{C}$ is ρ-robustly learnable using λ-local membership queries with respect to $\mathcal{D}$ if there exists a learning algorithm $\mathcal{A}$ such that for every $\epsilon > 0$, $\delta > 0$, for every distribution $D \in \mathcal{D}$ and every target concept $c \in \mathcal{C}$, the following hold:*

1. *$\mathcal{A}$ draws a sample S of size $m = poly(n, 1/\delta, 1/\epsilon, size(c))$ using the example oracle $\mathsf{EX}(c, D)$*
2. *Each query x' made by $\mathcal{A}$ to the LMQ oracle is λ-local with respect to some example $x \in S$*
3. *$\mathcal{A}$ outputs a hypothesis h that satisfies $\mathsf{R}_\rho^D(h, c) \leq \epsilon$ with probability at least $1 - \delta$*
4. *The running time of $\mathcal{A}$ (hence also the number of oracle accesses) is polynomial in n, $1/\epsilon$, $1/\delta$, $size(c)$ and the output hypothesis h is polynomially evaluable.*

A.4 Complexity Measures

For a more in-depth introduction to these concepts, we refer the reader to Mohri et al. (2012).

Definition 17 (Shattering)**.** *Given a class of functions $\mathcal{F}$ from input space $\mathcal{X}$ to $\{0, 1\}$, we say that a set $S \subseteq \mathcal{X}$ is* shattered by *$\mathcal{F}$ if all the possible dichotomies of S (i.e., all the possible ways of labelling the points in S) can be realized by some $f \in \mathcal{F}$.*

Definition 18 (VC Dimension)**.** *The VC dimension of a hypothesis class $\mathcal{H}$, denoted $\mathsf{VC}(\mathcal{H})$, is the size d of the largest set that can be shattered by $\mathcal{H}$. If no such d exists then $\mathsf{VC}(\mathcal{H}) = \infty$.*

Definition 19 (Littlestone Tree)**.** *A Littlestone tree for a hypothesis class $\mathcal{H}$ on $\mathcal{X}$ is a complete binary tree T of depth d whose internal nodes are instances $x \in \mathcal{X}$. Each edge is labeled with $-$ or $+$ and corresponds to the potential labels of the parent node. Each path from the root to a leaf must be consistent with some $h \in \mathcal{H}$, i.e. if $x_1, \ldots, x_d$ with labelings $y_1, \ldots, y_d$ is a path in T, there must exist $h \in \mathcal{H}$ such that $h(x_i) = y_i$ for all i.*

Definition 20 (Littlestone Dimension)**.** *The Littlestone dimension of a hypothesis class $\mathcal{H}$, denoted $\mathsf{Lit}(\mathcal{H})$, is the depth d of the largest Littlestone tree for $\mathcal{H}$. If no such d exists then $\mathsf{Lit}(\mathcal{H}) = \infty$.*

A.5 Online Learning

In online learning, the learner is given access to examples *sequentially.* At each time step, the learner receives an example x, predicts its label using its hypothesis h, receives the true label y and updates its hypothesis if $h(x) \neq y$. A fundamental difference between PAC learning and online learning is that, in the latter, there are no distributional assumptions. Examples can be given adversarially, and the performance of the learner is evaluated with respect to the number of mistakes it makes compared to the ground truth.

Definition 21 (Mistake Bound)**.** *For a given hypothesis class $\mathcal{C}$ and instance space $\mathcal{X} = \bigcup_n \mathcal{X}_n$, we say that an algorithm $\mathcal{A}$ learns $\mathcal{C}$ with mistake bound M if A makes at most M mistakes on any sequence of samples consistent with a concept $c \in \mathcal{C}$. In the mistake bound model, we usually require that M be polynomial in n and $size(c)$.*

We now recall the Standard Optimal Algorithm (Littlestone, 1988), which has a mistake bound $M = \mathsf{Lit}(\mathcal{C})$ when given concept class $\mathcal{C}$.

B Useful Results

B.1 Robust Risk Bounds

Lemma 22 (Lemma 6 in Gourdeau et al. (2019))**.** *Let $c_1, c_2 \in \{0, 1\}^{\mathcal{X}}$ and fix a distribution D on $\mathcal{X}$. Then, for all $h : \{0, 1\}^n \to \{0, 1\}$,*

$$R_\rho^D(c_1, c_2) \leq R_\rho^D(c_1, h) + R_\rho^D(c_2, h) \ .$$

Algorithm 1 Standard Optimal Algorithm from Littlestone (1988)

```
Input: A hypothesis class $\mathcal{H}$
  for $t = 1, 2, \ldots$ do
      Receive example $x_t$
      $V_t^{(b)} \leftarrow \{h \in V_t \mid h(x_t) = b\}$
      $\hat{y_t} = \arg\max_b \mathsf{Lit}(V_t^{(b)})$      ▷ Predict label acc. to subclass with larger Littlestone dimension
      Receive true label $y_t$
      $V_{t+1} \leftarrow V_t^{(y_t)}$
  end for
```

Lemma 23 (Lemma 14 in Gourdeau et al. (2022)). *Under the uniform distribution, for any $n \in \mathbb{N}$, disjoint $c_1, c_2 \in$ MON-CONJ of even length $3 \leq l \leq n/2$ on $\{0,1\}^n$ and robustness parameter $\rho = l/2$, we have that $\mathsf{R}_\rho^D(c_1, c_2)$ is bounded below by a constant that can be made arbitrarily close to $\frac{1}{2}$ as l (and thus ρ) increases.*

Remark 24. Note that the statement and proof of the above lemma remains unchanged if considering disjoint conjunctions, as opposed to monotone conjunctions.

B.2 Mistake Bounds for Winnow and Perceptron

Now, we recall the mistake upper bound for Winnow in the special case of $\mathsf{LTF}_{\{0,1\}^n}^{W+}$, where the weights are positive integers[10] and the mistake bound for the Perceptron algorithm.

Theorem 25 (Winnow). *The Winnow algorithm for learning the class $\mathsf{LTF}_{\{0,1\}^n}^{W+}$ makes at most $O(W^2 \log(n))$ mistakes.*

Theorem 26 (Mistake Bound for Perceptron, Margin Condition; Theorem 7.8 in Mohri et al. (2012)). *Let $\mathbf{x}_1, \ldots, \mathbf{x}_T \in \mathbb{R}^n$ be a sequence of T points with $\|\mathbf{x}_t\| \leq r$ for all $1 \leq t \leq T$ for some $r > 0$. Assume that there exists $\gamma > 0$ and $\mathbf{v} \in \mathbb{R}^n$ such that for all $1 \leq t \leq T$, $\gamma \leq \frac{y_t(\mathbf{v} \cdot \mathbf{x}_t)}{\|\mathbf{v}\|}$. Then, the number of updates made by the Perceptron algorithm when processing $\mathbf{x}_1, \ldots, \mathbf{x}_T$ is bounded by r^2/γ^2.*

B.3 Quantifier Elimination

Theorem 27 (Theorem 1.2 in Renegar (1992)). *Let Ψ be a formula in the first-order theory of the reals of the form*

$$(Q_1 x^{[1]} \in \mathbb{R}^{n_1}) \ldots (Q_\omega x^{[\omega]} \in \mathbb{R}^{n_\omega}) P(x^{[1]}, \ldots, x^{[n_\omega]}, y) \ ,$$

with free variables $y = (y_1, \ldots, y_l)$, quantifiers Q_i ($\exists$ or $\forall$) and quantifier-free Boolean formula $P(x^{[1]}, \ldots, x^{[n_\omega]}, y)$ with m atomic predicates consisting of polynomial inequalities of degree at most d. There exists a quantifier elimination method which constructs a quantifier-free formula Φ of the form

$$\bigvee_{i=1}^{I} \bigwedge_{j=1}^{J_i} (h_{ij}(y) \Delta_{ij} 0) \ ,$$

where

$$\begin{aligned} I &\leq (md)^{2^{O(\omega)} l \prod_k n_k} \\ J_i &\leq (md)^{2^{O(\omega)} \prod_k n_k} \\ \deg(h_{ij}) &\leq (md)^{2^{O(\omega)} \prod_k n_k} \\ \Delta_{ij} &\in \{\leq, \geq, =, \neq, >, <\} \ . \end{aligned}$$

[10] See https://www.cs.utexas.edu/ klivans/05f7.pdf for a full derivation.

C Proofs from Section 3

C.1 Proof of Theorem 2

Proof. Fix $\lambda, \rho \in \mathbb{N}$ such that $\lambda < \rho$, and consider the following monotone conjunctions: $c_1(x) = \bigwedge_{1\leq i\leq \rho} x_i$ and $c_2(x) = \bigwedge_{1\leq i\leq \rho+1} x_i$. Let D be the distribution on $\{0,1\}^n$ which puts all the mass on $\mathbf{0}$. Then, the target concept is drawn at random between c_1 and c_2. Now, c_1 and c_2 will both give all points in $B_\lambda(\mathbf{0})$ the label 0, so the learner has to choose a hypothesis that is consistent with both c_1 and c_2 (otherwise the robust risk is 1 and we are done). However, the learner has no way of distinguishing which of c_1 or c_2 is the target concept, while these two functions have a ρ-robust risk of 1 against each other under D. Formally,

$$\begin{aligned} \mathsf{R}_\rho^D(c_1, c_2) &= \Pr_{x\sim D}\left(\exists z \in B_\rho(x) \;.\; c_1(z) \neq c_2(z)\right) \\ &= \mathbf{1}[\exists z \in B_\rho(\mathbf{0}) \;.\; c_1(z) \neq c_2(z)] \\ &= 1 \ , \end{aligned} \tag{2}$$

where such $z = \mathbf{1}_\rho \mathbf{0}_{n-\rho}$. To lower bound the expected robust risk, letting $\mathcal{A}$ be any learning algorithm and $\mathcal{E}$ be the event that all points in a randomly drawn sample S are all labeled 0, we have

$$\begin{aligned} \mathop{\mathbb{E}}_{c,S}\left[\mathsf{R}_\rho^D(\mathcal{A}(S), c)\right] &= \mathop{\mathbb{E}}_{c,S}\left[\mathsf{R}_\rho^D(\mathcal{A}(S), c) \mid \mathcal{E}\right] && \text{(By construction of } D\text{)} \\ &= \frac{1}{2}\mathop{\mathbb{E}}_{S}\left[\mathsf{R}_\rho^D(\mathcal{A}(S), c_1) + \mathsf{R}_\rho^D(\mathcal{A}(S), c_2) \mid \mathcal{E}\right] && \text{(Random choice of } c\text{)} \\ &\geq \frac{1}{2}\mathop{\mathbb{E}}_{S}\left[\mathsf{R}_\rho^D(c_1, c_2) \mid \mathcal{E}\right] && \text{(Lemma 22)} \\ &= \frac{1}{2} \ . && \text{(Equation 2)} \end{aligned}$$

□

C.2 Proof of Lemma 3

Proof. Fix a target concept $c \in \mathcal{C}$ and the target distribution D over $\mathcal{X}$. Define a hypothesis h to be "bad" if $R_\rho^D(c,h) \geq \epsilon$. Note that any robust ERM algorithm will be robustly consistent on the training sample by the realizability assumption. Let $\mathcal{E}_h$ be the event that m independent examples drawn from $\mathsf{EX}(c, D)$ are all robustly consistent with h. Then, if h is bad, we have that $\Pr(\mathcal{E}_h) \leq (1-\epsilon)^m \leq e^{-\epsilon m}$. Now consider the event $\mathcal{E} = \bigcup_{h\in\mathcal{H}} \mathcal{E}_h$. We have that, by the union bound,

$$\Pr(\mathcal{E}) \leq \sum_{h\in\mathcal{H}} \Pr(\mathcal{E}_h) \leq |\mathcal{H}|\, e^{-\epsilon m} \ .$$

Then, bounding the RHS by δ, we have that whenever $m \geq \frac{1}{\epsilon}\left(\log|\mathcal{H}_n| + \log\frac{1}{\delta}\right)$, no bad hypothesis is *robustly* consistent with m random examples drawn from $\mathsf{EX}(c, D)$. If a hypothesis is not bad, it has robust risk bounded above by ϵ, as required. □

C.3 Proof of Lemma 5

Proof. The proof is very similar to the VC dimension upper bound in PAC learning. The main distinction is that instead of looking at the error region of the target and any function in $\mathcal{H}$, we must look at its ρ-expansion. Namely, we let the target $c \in \mathcal{C}$ be fixed and, for $h \in \mathcal{H}$, we consider the function $(c \oplus h)_\rho : x \mapsto \mathbf{1}[\exists z \in B_\rho(x) \;.\; c(z) \neq h(z)]$ and define a new concept class $\Delta_{c,\rho}(\mathcal{H}) = \{(c\oplus h)_\rho \mid h \in \mathcal{H}\}$. It is easy to show that $\mathsf{VC}(\Delta_{c,\rho}(\mathcal{H})) \leq \mathsf{RVC}_\rho(\mathcal{C},\mathcal{H})$, as any sign pattern achieved on the LHS can be achieved on the RHS.

The rest of the proof follows from the definition of an ϵ-net and the bound on the growth function of $\Delta_{c,\rho}(\mathcal{H})$.

First, define the class $\Delta_{c,\rho,\epsilon}(\mathcal{H})$ as $\left\{\tilde{c} \in \Delta_{c,\rho}(\mathcal{H}) \mid \Pr_{x\sim D}(\tilde{c}(x)=1) \geq \epsilon\right\}$, i.e., the set of functions in $\Delta_{c,\rho}(\mathcal{H})$ which have a robust risk greater than ϵ. Recall that a set S is an ϵ-net for $\Delta_{c,\rho}(\mathcal{H})$ if for every $\tilde{c} \in \Delta_{c,\rho,\epsilon}(\mathcal{H})$, there exists $x \in S$ such that $\tilde{c}(x) = 1$. We want to bound the probability that a sample $S \sim D^m$ fails to be an ϵ-net for the class $\Delta_{c,\rho}(\mathcal{H})$, as if S is an ϵ-net, then any robustly consistent $h \in \mathcal{H}$ on S will have robust risk bounded above by ϵ. As with the standard VC dimension, a sample S will be drawn in two phases. First draw a sample $S_1 \sim D^m$ and let $\mathcal{E}_1$ be the event that S_1 is not an ϵ-net for $\Delta_{c,\rho}(\mathcal{H})$. Now, suppose $\mathcal{E}_1$ occurs. This means there exists $\tilde{c} \in \Delta_{c,\rho,\epsilon}(\mathcal{H})$ such that $\tilde{c}(x) = 0$ for all the points $x \in S_1$. Fix such a $\tilde{c}$ and draw a second sample $S_2 \sim D^m$. Then, letting X be the random variable representing the number of points in S_2 that are such that $\tilde{c}(x) = 1$, we can use Chernoff bound to show that

$$\Pr(X < \epsilon m/2) \leq 2\exp\left(-\frac{\epsilon m}{12}\right) \quad , \tag{3}$$

ensuring that whenever $\epsilon m \geq 24$, the probability that at least $\epsilon m/2$ points in S_2 satisfy $\tilde{c}(x) = 1$ is bounded below by 1/2.

Now, consider the event $\mathcal{E}_2$ where a sample $S = S_1 \cup S_2$ of size $2m$ such that $|S_1| = |S_2| = m$ is drawn from $\mathsf{EX}(c, D)$ and there exists a concept $\tilde{c} \in \Pi_{\Delta_{c,\rho,\epsilon}(\mathcal{H})}(S)$ such that $|\{x \in S \mid \tilde{c}(x) = 1| \geq \epsilon m/2\}$ and $\tilde{c}(x) = 0$ for all $x \in S_1$, where $\Pi_{\Delta_{c,\rho,\epsilon}(\mathcal{H})}(S)$ is the set all possible dichotomies on S induced by $\Delta_{c,\rho,\epsilon}(\mathcal{H})$. Then $\Pr(\mathcal{E}_2) \geq \frac{1}{2}\Pr(\mathcal{E}_1)$ from Equation 3. Now, the probability that $\mathcal{E}_2$ happens for a fixed $\tilde{c} \in \Delta_{c,\rho,\epsilon}(\mathcal{H})$ is

$$\frac{\binom{m}{\epsilon m/2}}{\binom{2m}{\epsilon m/2}} \leq 2^{-\epsilon m/2} \quad .$$

Finally, letting $d = \mathsf{RVC}_\rho(\mathcal{C}, \mathcal{H})$ we can bound the probability of $\mathcal{E}_1$ using the union bound:

$$\begin{aligned}
\Pr(\mathcal{E}_1) &\leq 2\Pr(\mathcal{E}_2) \\
&\leq 2\left|\Pi_{\Delta_{c,\rho,\epsilon}(\mathcal{H})}(S)\right| 2^{-\epsilon m/2} \\
&\leq 2\left|\Pi_{\Delta_{c,\rho}(\mathcal{H})}(S)\right| 2^{-\epsilon m/2} \\
&\leq 2\left(\frac{2em}{d}\right)^d 2^{-\epsilon m/2} \quad . && \text{(Sauer's Lemma)}
\end{aligned}$$

Thus, there exists a universal constant such that provided m is larger than the bound given in the statement of the theorem, $\Pr(\mathcal{E}_1) < \delta$, as required. □

C.4 Proof of Theorem 7

Proof. The sample complexity bounds come from Lemmas 3 and 5 and the fact that the Standard Optimal Algorithm (SOA) is a consistent learner, as it will be given counterexamples in the perturbation region until a robust loss of zero is achieved.

For each query to LEQ, a counterexample is returned, or the robust loss is zero. Then, using the mistake upper bound of SOA, which is $\mathsf{Lit}(\mathcal{C})$, we get the first query upper bound. The second query upper bound comes from the observation that if we restrict the nodes in the Littlestone trees (see Appendix A.4) to be in $B_\rho(x)$, their depth must be bounded above by

$$\log(|B_\rho(x)|) = \log\left(\sum_{i=1}^{\rho}\binom{n}{i}\right) \leq \log\left(\rho\left(\frac{en}{\rho}\right)^\rho\right) \leq \rho\log(en) \quad .$$

□

C.5 Proof of Theorem 9

Proof. Let c be the target conjunction and let D be an arbitrary distribution. We describe an algorithm $\mathcal{A}$ with polynomial sample and query complexity with access to a ρ-LEQ oracle. By Lemma 3, if we can get guarantee that $\mathcal{A}$ returns a hypothesis with zero robust loss on a i.i.d. sample of size $m = O\left(\frac{1}{\epsilon}\left(n + \log\frac{1}{\delta}\right)\right)$ with a polynomial number of queries to the ρ-LEQ oracle, we are done.

The algorithm is similar to the standard PAC learning algorithm, in that it only learns from positive examples. Indeed, the original hypothesis h is a conjunction of all of the $2n$ literals. After seeing a positive example x, $\mathcal{A}$ removes from h the literals $\bar{x_i}$ for $i = 1, \dots, n$, as they cannot be in c. Note that, by construction, any hypothesis h returned by $\mathcal{A}$ always satisfies $c \subseteq h$[11]. Thus, any counter example returned by the LEQ oracle will have that $c(z) = 1$ and $h(z) = 0$. This allows us to remove at least one literal from the hypothesis set for every counter example. Now, it is easy to see that, for $c \subseteq h' \subseteq h$, if the robust loss $\mathbf{1}[\exists z \in B_\lambda(x) \;.\; c(z) \neq h(z)]$ on x w.r.t. h is zero, so will be the robust loss on x w.r.t. the updated hypothesis h'. Hence, $\mathcal{A}$ makes at most $m + 2n$ queries to the LEQ oracle. □

C.6 Proof of Theorem 10

Proof. The sample complexity bound uses Lemma 3. Note the class $\mathsf{LTF}^W_{\{0,1\}^n}$ has size $O(2^n(n+W)^{\min\{n,W\}})$. This is a simple application of the stars and bars identity, where W is the number of stars and $n+1$ the number of bars (as we are considering the bias term as well): $\binom{n+W}{W} = O((n+W)^{min\{n,W\}})$. The 2^n term comes from the fact that each weight can be positive or negative. The query complexity uses the fact that the mistake bound for Winnow for $\mathsf{LTF}^W_{\{0,1\}^n}$ is $O(W^2\log(n))$ in the case of positive weights (the full statement can be found in Appendix. B). Littlestone (1988) outlines how to use the Winnow algorithm when the linear classifier's weights can vary in sign, at the cost of doubling the input dimension and weight bound (see Theorem 10 and Example 6 therein). □

C.7 Proof of Theorem 11

The proof of this theorem mainly relies on deriving an upper bound on the robust VC dimension of halfspaces. This will help us bound the sample complexity needed to guarantee robust accuracy. The query complexity upper bound follows from this upper bound and the mistake bound for the Perceptron algorithm. To bound the robust VC dimension of linear classifiers, we will need the following theorem from Goldberg and Jerrum (1995):

Theorem 28 (Theorem 2.2 in Goldberg and Jerrum (1995))**.** *Let $\{\mathcal{C}_{k,n}\}_{k,n\in\mathbb{N}}$ be a family of concept classes where concepts in $\mathcal{C}_{k,n}$ and instances are represented by k and n real values, respectively. Suppose that the membership test for any instance α in any concept C of $\mathcal{C}_{k,n}$ can be expressed as a boolean formula $\Phi_{k,n}$ containing $s = s(k,n)$ distinct atomic predicates, each predicate being a polynomial inequality or equality over $k+n$ variables (representing C and α) of degree at most $d = d(k,n)$. Then $\mathsf{VC}(\mathcal{C}_{k,n}) \leq 2k\log(8eds)$.*

We will now translate the ρ-expansion of the error region (i.e., the robust loss function) between two halfspaces as a boolean formula using a result from Renegar (1992). This will allow us to use the theorem above from Goldberg and Jerrum (1995) to bound the robust VC dimension of $\mathsf{LTF}_{\mathbb{R}^n}$.

Lemma 29. *Let $a, b \in \mathbb{R}^n, a_0, b_0 \in \mathbb{R}$, and define the map $\varphi : x \mapsto \mathbf{1}[\exists z \in B_\rho(x) \;.\; sgn(a^\top z + a_0) \neq sgn(b^\top z + b_0)]$. Then φ can be represented as a boolean formula Φ with $s = 10^{Cn^2}$ distinct atomic predicates, each predicate being a polynomial inequality over $2n+2$ variables of degree at most $10^{C'n}$ for some constants $C, C' > 0$.*

Proof. First not that the predicate $\text{sgn}(a\top z + a_0) \neq \text{sgn}(b^\top z + b_0)$ can be represented as the following formula:

$$\left(a^\top z + a_0 \geq 0 \land b^\top z + b_0 < 0\right) \lor \left(a^\top z + a_0 < 0 \land b^\top z + b_0 \geq 0\right) \;\;,$$

[11] We overload c, h to mean both the functions and the set of literals in the conjunction, as it will be unambiguous to distinguish them from context.

which contains $n + (2n + 2)$ variables and 4 predicates. Moreover, given a perturbation $\zeta \in \mathbb{R}^n$, the constraint $\|\zeta\|_2 \leq \rho$ on its magnitude is a polynomial inequality of degree 2:

$$\sum_i \zeta_i^2 \leq \rho^2 \ .$$

Now, consider the following formula:

$$\Psi(x) = \exists \zeta \in \mathbb{R}^n \ . \ \big(\mathrm{sgn}(a\top z + a_0) \neq \mathrm{sgn}(b^\top z + b_0) \wedge \|\zeta\|_2 \leq \rho\big) \ .$$

This is a formula of first-order logic over the reals. Using the notation of Theorem 27, we have $\omega = 1$ quantifier, and thus $\prod_k n_k = n$, one Boolean formula with $m = 5$ polynomial inequalities of degree d at most 2, and $l = n$. Thus, $\Psi(x)$ can be expressed as a quantifier-free formula $\Phi(x) = \bigvee_{i=1}^{I} \bigwedge_{j=1}^{J_i} (h_{ij}(y) \Delta_{ij} 0)$ of size

$$I \max_i J_i \leq (md)^{2^{O(\omega)} l \prod_k n_k + 2^{O(\omega)} \prod_k n_k} \leq 10^{Cn^2}$$

for some constant C, where the polynomial inequalities are of degree at most $(md)^{2^{O(\omega)} \prod_k n_k} \leq 10^{C'n}$ for some constant C'. □

We thus get the following corollary.

Corollary 30. *The robust VC dimension of* $\mathsf{LTF}_{\mathbb{R}^n}$ *is* $O(n^3)$.

Proof. We let $s = 10^{Cn^2}$, $k = 2n+2$ and $d = 10^{C'n}$ from the proof above and use Definition 4 and Theorem 28 to get a robust VC dimension upper bound of $O(k \log(sd)) = O(n^3)$.[12] □

Proving Theorem 11 is now a straightforward application of the results above.

Proof of Theorem 11. The sample complexity upper bound is a consequence of Corollary 30 and Lemma 5. The query complexity upper bound follows from Lemma 8 and the mistake bound for the Perceptron algorithm, which appears in Appendix B. □

[12] Note that Corollary 2.4 in Goldberg and Jerrum (1995) uses this reasoning.