Robust Empirical Risk Minimization with Tolerance

Robi Bhattacharjee RCBHATTA@ENG.UCSD.EDU
UCSD

Max Hopkins NMHOPKIN@ENG.UCSD.EDU
UCSD

Akash Kumar AKK002@UCSD.EDU
UCSD

Hantao Yu HY2751@COLUMBIA.EDU
Columbia

Kamalika Chaudhuri KAMALIKA@ENG.UCSD.EDU
UCSD

Editors: Shipra Agrawal and Francesco Orabona

Abstract

Developing simple, sample-efficient learning algorithms for robust classification is a pressing issue in today's tech-dominated world, and current theoretical techniques requiring exponential sample complexity and complicated improper learning rules fall far from answering the need. In this work we study the fundamental paradigm of (robust) *empirical risk minimization* (RERM), a simple process in which the learner outputs any hypothesis minimizing its training error. RERM famously fails to robustly learn VC classes (Montasser et al., 2019a), a bound we show extends even to 'nice' settings such as (bounded) halfspaces. As such, we study a recent relaxation of the robust model called *tolerant* robust learning (Ashtiani et al., 2022) where the output classifier is compared to the best achievable error over slightly larger perturbation sets. We show that under geometric niceness conditions, a natural tolerant variant of RERM is indeed sufficient for γ-tolerant robust learning VC classes over $\mathbb{R}^d$, and requires only $\tilde{O}\left(\frac{VC(H)d\log\frac{D}{\gamma\delta}}{\epsilon^2}\right)$ samples for robustness regions of (maximum) diameter D.

Keywords: empirical risk minimization, robust learning, vc dimension, tolerant learning

1. Introduction

Adversarially robust classification is a staple of modern machine learning. In the robust setting, along with meeting standard accuracy guarantees, predictions made by a learner at test time must additionally be robust to adversarial perturbations to the input, typically defined by a fixed family $\mathcal{U} = \{U_x\}_{x\in X}$ of possible perturbations. Developing robust algorithms with provable guarantees has been an important research direction in recent years, both for parametric Khim and Loh (2018); Attias et al. (2019); Montasser et al. (2019b); Yin et al. (2019); Ashtiani et al. (2020) and non-parametric Wang et al. (2018); Yang et al. (2019); Bhattacharjee and Chaudhuri (2020, 2021) classifiers, but understanding the performance of even the most basic algorithms in the setting remains open.

In this work, we study one of the simplest, most fundamental algorithmic paradigms in learning, a classical method called *empirical risk minimization* (ERM). In the robust setting, an algorithm is said to be an empirical risk minimizer (RERM) if it always outputs a hypothesis in the class with

minimal *robust* risk over its training data. In the standard setting, it is a classical result that any learnable class is learnable (near-optimally) by any ERM. Unfortunately, this is known to fail drastically in the robust setting—Montasser et al. (2019a) showed that there exist finite VC classes, $\mathcal{H}$, where no algorithm outputting hypotheses in $\mathcal{H}$ (called a *proper* learner) can converge towards the optimal classifier, even with arbitrary amounts of training data. Conversely, such classes *are* in fact robustly learnable, but require complicated improper learning rules and a potentially exponential number of samples.

The failure of Robust ERM for general classes raises an interesting question: *are there natural sufficient conditions for the success of RERM?* One obvious answer to this question is the notion of robust VC dimension, a combinatorial parameter promising the success of RERM. However, bounding robust VC is typically difficult, and such results are only known for very specialized examples of classifiers and robustness regions (e.g. linear classifiers under fixed-radius balls (Cullina et al., 2018) and other simple margin structures (Ashtiani et al., 2020), or VC-classes under finite perturbation sets (Attias et al., 2019)). To our knowledge there are no corresponding results for more general robustness regions and hypothesis classes beyond these special cases.

Given the current failure of combinatorial techniques in this setting, one might instead hope to show RERM works given sufficiently nice *geometric* conditions on the hypothesis class. Sadly, this is not the case. We show that there exist robustness regions for which RERM (indeed any proper algorithm) fails even for settings as simple as (bounded) linear classifiers.

Theorem 1 (Failure of RERM for Linear Classifiers) *For any $W > 0$ and $d > 1$, let $\mathcal{H}_W$ denote the set of linear classifiers with distance at most W from the origin. Then there exists a set of robustness regions U over $\mathbb{R}^d$ such that for any proper learning algorithm L there exists a distribution $\mathcal{D}$ for which the following hold:*

- *$\mathcal{D}$ **is realizable**: There exists $h^* \in \mathcal{H}_W$ such that $\ell_U(h^*, \mathcal{D}) = 0$.*
- *L **has high error**: With probability at least $\frac{1}{7}$ over $S \sim \mathcal{D}^m$, $\ell_U(L(S), \mathcal{D}) > \frac{1}{8}$.*

With this in mind, we turn our attention to a different approach: relaxing the notion of robustness itself. We'll consider a recent model of Ashtiani et al. (2022) called *tolerant* robust learning. In the tolerant setting, the learner is only required to compete with the best loss over a relaxed family of perturbation sets $\mathcal{U}^\gamma$ for a (potentially arbitrary) tolerance parameter $\gamma > 0$. Ashtiani et al. (2022) studied this setting in the special case of radius r balls, where the learner competes with robust error against $r(1+\gamma)$-balls. Under this framework, Ashtiani et al. (2022) give an algorithm with PAC-guarantees for VC classes using significantly fewer samples, but their techniques remain improper and only hold for the simplest robustness setting.

In this work, we show that a simple variant of RERM in the tolerant model indeed succeeds under natural geometric conditions on the hypothesis class. In particular, we study a notion of smoothness called *regularity*, which roughly promises that every point in the instance space should be contained in some ball of the same label. This captures many well-studied settings, such as cases where the decision boundaries are compact, differential manifolds in $\mathbb{R}^d$.

Theorem 2 (Tolerant RERM for Regular Classes) *Let $\mathcal{H}$ be a regular hypothesis class with VC dimension v over $\mathbb{R}^d$, and let $\mathcal{U}$ be any set of robustness regions. Then $TolRERM$ tolerantly*

PAC-learns $(\mathcal{H},\mathcal{U})$ *with tolerant sample complexity*

$$m(\epsilon,\delta,\gamma) = O\left(\frac{vd\log\frac{dDiam(U)}{\epsilon\gamma\delta}}{\epsilon^2}\right),$$

where $Diam(U)$ *denotes the maximum* ℓ_2 *diameter across robustness regions* U_x.

Theorem 2 matches the sample complexity given in Ashtiani et al. (2022) up to logarithmic factors and enjoys the additional benefits of applying to more general robustness regions along with its properness and general algorithmic simplicity. For completeness, we also analyze our algorithm's performance over non-regular classifiers in Appendix D, and show that it has a similar performance albeit at the cost of replacing the VC-dimension with v_{ball}, the robust VC dimension of $\mathcal{H}$ over balls of a fixed radius. Thus, for non-regular hypothesis classes, our algorithm gives a reduction from arbitrary robustness regions to the case where they are all balls of a fixed radius.

Finally it's worth noting that while Ashtiani et al. (2022) only requires sampling access to the perturbation sets, stronger access such as an empirical risk minimizer is inevitable in the general setting where $\mathcal{U}$ is unknown. We show that there exists hypothesis classes where $\Omega((\frac{D}{\gamma})^d)$ queries to a sampling oracle are required for robust learning with tolerance if no other interaction with U_x is permitted.

While Theorem 2 gives a natural sufficient condition for the success of RERM in relaxed settings, many questions in this direction remain wide open. It would be interesting to identify a necessary condition for the success of RERM, both in the tolerant and original robust models. Furthermore, it should be noted that while we prove RERM fails to learn nice classes in the latter, the perturbation family we use to achieve this is highly combinatorial. As such, there is still hope that RERM may be sufficient in the traditional setting under *joint* niceness conditions on $\mathcal{H}$ and $\mathcal{U}$, though the close interplay between the two families seems to make identifying such a condition difficult, if it is indeed possible at all.

2. Related Work

Much of the work on adversarial robustness (Carlini and Wagner, 2017; Liu et al., 2017; Papernot et al., 2017, 2016a; Szegedy et al., 2014; Hein and Andriushchenko, 2017; Katz et al., 2017; Papernot et al., 2016b; Raghunathan et al., 2018; Sinha et al., 2018) is done in the context of neural networks.

On the theoretical side, there has been a recent focus on developing algorithms with guarantees in convergence towards an optimal classifier. On the parametric side, several works (Khim and Loh, 2018; Attias et al., 2019; Montasser et al., 2019b; Yin et al., 2019; Ashtiani et al., 2020; Cullina et al., 2018) have focused on distribution agnostic bounds on the amount of data required to converge towards the optimal classifier in a given hypothesis class. For example, Montasser et al. (2019b) showed through an example that the VC dimension of robust learning may be much larger than standard or accurate learning indicating that the sample complexity bounds may be higher. There has also been some work considering the computation complexity required for robust learning such as Diakonikolas et al. (2020).

Aside from Ashtiani et al. (2022), there are several works which also consider variations on robust learning with tolerance. Yang et al. (2019) and Bhattacharjee and Chaudhuri (2020) show that

certain non-parametric algorithms exhibit a type of tolerant behavior when robustness regions are constrained to be balls of radius r. Montasser et al. (2022) considers robustness in the *transductive learning setting*. Their work employs a similar idea to Ashtiani et al. (2022) in that they consider expanded perturbation sets when giving their formal guarantees. However, their expansions are not based on tolerance $\gamma > 0$.

Finally, Awasthi et al. (2021), introduces a notion of pseudo-robustness which precisely matches our definition of a regular classifier (Definition 11). Their work focuses on using this notion to define a robust analog to the Bayes optimal classifier. By contrast, our work focuses on learning a robust hypothesis class that satisfies this condition.

3. Preliminaries

Let $\mathcal{H}$ be a family of binary classifiers $\{h : \mathbb{R}^d \to \{\pm 1\}\}$, and $U = \{U_x \subseteq \mathbb{R}^d : x \in \mathbb{R}^d\}$ any set of robustness regions. We define the robust loss function with respect to U as follows.

Definition 3 *Let $h \in \mathcal{H}$ be a classifier and $(x, y) \in \mathbb{R}^d \times \{\pm 1\}$ be a labeled point. Then the **robust loss** of h over (x, y), denoted $\ell_U(h, (x, y))$, is defined as*

$$\ell_U(h, (x, y)) = \begin{cases} 1 & \exists x' \in U_x \text{ such that } h(x') \neq y \\ 0 & \text{otherwise.} \end{cases}.$$

That is, h achieves a loss of 0 only if it labels all points in U_x as y.

For a distribution, $\mathcal{D}$ over $\mathbb{R}^d \times \{\pm 1\}$, we let $\ell_U(h, \mathcal{D})$ denote the expected loss h pays over a labeled point drawn from $\mathcal{D}$. That is, $\ell_U(h, \mathcal{D}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}[\ell_U(h, (x, y))]$.

Similarly, for a set of n labeled points, S, we let $\ell_U(h, S)$ denote the average robust loss h pays over S. that is, $\ell_U(h, S) = \frac{1}{n}\sum_{i=1}^{n} \ell_U(h, (x_i, y_i))$.

We will also use $||x - x'||$ to denote the ℓ_2 distance between x and x', and $B(x, r)$ to denote the (closed) ℓ_2 ball centered at x with radius r.

3.1. Robust PAC-learning

We now review a natural generalization of PAC learning to the robust setting called robust PAC-learning (Montasser et al., 2019a).

Definition 4 *Let $\mathcal{H}$ be a hypothesis class and U be a set of robustness regions. A learner L **robustly PAC-learns** $(\mathcal{H}, U)$ if for every $\epsilon, \delta > 0$, there exists $m(\epsilon, \delta)$ such that for all $n \geq m(\epsilon, \delta)$, for all data distributions, $\mathcal{D}$, with probability $1 - \delta$ over $S \sim \mathcal{D}^n$,*

$$\ell_U(\hat{h}, \mathcal{D}) \leq \min_{h\in\mathcal{H}} \ell_U(h, \mathcal{D}) + \epsilon,$$

*where $\hat{h} = L(S)$ denotes the classifier in $\mathcal{H}$ outputted by L from training sample S. $m(\epsilon, \delta)$ is said to be the **sample complexity** of L with respect to $(\mathcal{H}, U)$.*

Algorithms that are able to robustly PAC-learn a pair $(\mathcal{H}, U)$ are the natural robust analogs of standard learning algorithms, and thus an important question is understanding how the sample complexities, $m(\epsilon, \delta)$, for doing so are bounded.

4. Robust Empirical Risk Minimization on Linear Classifiers

Montasser et al. (2019a) showed that there exist hypothesis classes $\mathcal{H}$ with bounded VC dimension, and robustness regions U, such that proper robust PAC-learning is not possible, meaning no matter how much data one is allowed, there always exists a distribution where the learner will suffer high robust loss.

However, for many practical examples, this does not appear to be the case – for example, Cullina et al. (2018) showed that when $\mathcal{H}$ is the set of all linear classifiers and U is the set of robustness regions with $U_x = B(x, r)$, the sample complexity of robustly learning with RERM is at most $m(\epsilon, \delta) = \tilde{O}\left(\frac{d}{\epsilon^2}\right)$, matching the standard complexity for linear classification.

Motivated by recent interest in more general robustness regions than balls of a fixed radius, we consider the case where $\mathcal{H}$ is a natural hypothesis class, but U is a potentially arbitrary robustness region. That is, we ask the following question: are there examples of natural hypothesis classes for which there exist robustness regions leading to arbitrary high sample complexities?

Unfortunately, the answer turns out to be yes. To show this, we begin by defining the natural hypothesis class of *bounded* linear classifiers.

Definition 5 *A W-bounded linear classifier, $f : \mathbb{R}^d \to \mathbb{R}^d$, is a linear classifier h whose decision boundary has distance at most W from the origin. That is, there exist $w \in \mathbb{R}^d$ and , $b \in \mathbb{R}$ with $\frac{|b|}{||w||} \leq W$ such that*

$$h(x) = \begin{cases} 1 & \langle w, x \rangle + b \geq 0 \\ -1 & otherwise \end{cases}.$$

We let $\mathcal{H}_W$ denote the class of all W-bounded linear classifiers

The boundedness condition, W, can be thought of as a regularization term which is common during any kind of practical optimization.

We now show that there exist robustness regions, U, for which $(\mathcal{H}_W, U)$ is not robustly PAC-learnable, even in the realizable setting. For convenience, we restate Theorem 1 from the introduction.

Theorem 1 *For any $W > 0$, $d > 1$, and $m > 1$, there exists a set of robustness regions U over $\mathbb{R}^d$ such that for any learning algorithm L there exists a distribution $\mathcal{D}$ for which the following hold:*

- *$\mathcal{D}$ **is realizable**: There exists $h^* \in \mathcal{H}_W$ such that $\ell_U(h^*, \mathcal{D}) = 0$.*
- *L **has high error**: With probability at least $\frac{1}{7}$ over $S \sim \mathcal{D}^m$, $\ell_U(L(S), \mathcal{D}) > \frac{1}{8}$.*

Theorem 1 consequently shows that the observations made in Montasser et al. (2019a) hold even over practical hypothesis classes such as (bounded) linear classifiers.

To prove Theorem 1, we begin with the following critical lemma.

Lemma 6 *For every $M \in \mathbb{N}$ there exists a family of M subsets of $\mathbb{R}^d$*

$$Z^{(M)} := \left\{ Z_1^{(M)}, Z_2^{(M)}, \ldots, Z_M^{(M)} \right\}$$

satisfying the following conditions:

- *There exists* $1 \leq i \leq M$ *and* $z \in Z_i^{(M)}$ *with* $h(z) = 1$.
- *For every* $1 \leq i \leq M$*, there exists* $h_i \in \mathcal{H}_W$ *such that* $h_i(z) = -1$ *for all* $z \in \cup_{j \neq i} Z_j^{(M)}$.
- *For any distinct natural numbers* M *and* M'*, the sets* $\cup_{i=1}^{M} Z_i^M$ *and* $\cup_{i=1}^{M'} Z_i^{M'}$ *are disjoint. Thus, there is no point that is contained in subsets from both* Z^M *And* $Z^{M'}$.

Proof Let $\{\beta_i\}_{i \in \mathbb{N}} > 0$ be a strictly decreasing sequence of sufficiently small real numbers (that we will specify later). For notational simplicity, fix an $M \in \mathbb{N}$ and write $\beta = \beta_M$ and $W' = (1+\beta)W$. For any $r > 0$, let S_r^{d-1} denote the $(d-1)$-sphere centered at the origin of radius r.

Observe that for any $x \in S_W^{d-1}$, there exists a unique classifier $h \in \mathcal{H}_W$ whose decision boundary is tangent to S_W^{d-1} at x so that $h(x) = 1$. We denote this classifier as h_x. It follows that the set of all points on $S_{W'}^{d-1}$ that h_x classifies as 1 can be easily characterized in terms of x. In particular, by the definition of h_x, it follows from geometry that

$$\left\{z : h_x(z) = 1, z \in S_{W'}^{d-1}\right\} = \left\{z : ||z - (1+\beta)x|| \leq W\sqrt{2\beta(\beta+1)}, z \in S_{W'}^{d-1}\right\}. \quad (1)$$

Let $r_\beta = 2W\sqrt{2\beta(\beta+1)}$, and let $z_1, z_2, \ldots, z_{M_\beta}$ denote a a greedy r_β cover of $S_{W'}^{d-1}$, meaning that points are successively selected from $S_{W'}^{d-1}$ until no point with distance strictly greater than r_β from all other points can be selected. Finally, define $Z_i = Z_i^{(M)}$ as the set of elements in $S_{W'}^{d-1}$ with nearest neighbor z_i (ties broken arbitrarily).

We claim that this construction suffices for $M_\beta \geq M$. First, observe that $\lim_{\beta \to 0} r_\beta = 0$, which means that for sufficiently small β that M_β will be arbitrarily large (thus satisfying $M_\beta \geq M$). So select any β for which this hold, and merge enough regions so that we are left with exactly M regions (i.e. set $Z_M = \cup_{i=M}^{M_\beta} Z_i$). Note that we can always choose $0 < \beta < \beta_{M-1}$ since the naturals can be embedded into any interval. We now verify the two stipulations of Lemma 6.

The first stipulation clearly holds since $\{Z_i\}_{i=1}^M$ partition $S_{W'}^{d-1}$ and every halfspace $h \in \mathcal{H}_W$ intersects the latter by construction.

For the second stipulation, observe that for any i, the ball centered at z_i of radius $\frac{r_\beta}{2}$, $B\left(z_i, \frac{r_\beta}{2}\right)$, does not intersect Z_j for any $i \neq j$. This is because such an intersection would imply by the triangle inequality that $||z_i - z_j|| \leq r_\beta$, which is a contradiction. This observation allows us to find a classifier, h_i, as desired – we set h_i to be the previously defined classifier, $h_{\frac{z_i}{1+\beta}}$. Equation 1 implies that the only points in $S_{W'}^{d-1}$ that it will classify as 1 are precisely the points in $B\left(z_i, \frac{r_\beta}{2}\right) \cap S_{W'}^{d-1}$. Since this is a subset of Z_i, the second stipulation is met, as desired.

Finally, it is left to observe that over each choice of M these $Z^{(M)}$ are mutually disjoint. This is true so long as the choices of β themselves are disjoint, since $Z^{(M)}$ lies in the sphere of radius $W(1+\beta_M)$. As noted previously it is easy to see $\{\beta_M\}$ can be chosen in this manner in an inductive fashion. ■

We are now sketching a proof for Theorem 1, with the full proof deferred Appendix A.

Proof Sketch: (Theorem 1) Our goal is to show that for any $m \in \mathbb{N}$, any learner on m samples must fail with constant probability. Fix any m. The main idea will be to construct a set of robustness regions, $U_{x_1}, U_{x_2}, \ldots, U_{x_{3m}}$ such that any classifier in $\mathcal{H}_W$ will lack robustness on at least m of them. T

Toward this end, set $M = \binom{3m}{m}$, and let $Z_1^{(M)}, Z_2^{(M)}, \ldots, Z_M^{(M)}$ be subsets of $\mathbb{R}^d$ as described by Lemma 6 (we will drop the superscript in what follows). Let $\mathcal{M}$ denote the set of all subsets of $\{1, \ldots, 3m\}$ with exactly m elements. Associate with each Z_i a unique element of $\mathcal{M}$, thus allowing us to rename our subsets as $\{Z_T : T \in \mathcal{M}\}$. We now define

$$U_{x_i} = \cup_{T: i \in T} Z_T,$$

where x_i is an arbitrary point inside U_{x_i}.

Lemma 6 that if all x_i are given a label of -1, then any $h \in \mathcal{H}_W$ will label some (for some set T) some $z \in Z_T$ as $+1$, thus causing it to lack robustness on *all* $i \in T$. Conversely, we see that for any T, there is a classifier $h_T \in \mathcal{H}_W$ that is accurate and robust at all x_i with $i \notin T$.

With these observations, we are now prepared to show that for any learner L, there exists a distribution D for which L has large expected robust loss. To do this, we use a standard lower bound technique found in Shalev-Shwartz and Ben-David (2014) that was adapted to the robust setting in Montasser et al. (2019a). The idea will be to pick D to be the uniform distribution over a random subset of $2m$ points in $\{x_1, \ldots, x_{3m}\}$. We will then argue that because L only has access to m points from D, it won't be able to distinguish which subset D corresponds to, and this will lead to a large expected loss. □

As demonstrated in Lemma 6, the robustness regions U used in our lower bound are combinatorial in nature and unlikely to represent any practical kinds of robustness regions. Nevertheless, our lower bound does show that naturality assumptions on the hypothesis class alone are *not* sufficient for ensuring robust PAC-learnability.

A natural next step would be to fully characterizes pairs $(\mathcal{H}, U)$ for which proper robust PAC-learnability is possible, but we leave this as a direction for future work. We instead turn towards relaxing the requirements of the robust PAC-learning model in order to find algorithms that are able to succeed in the case that $\mathcal{H}$ is natural but U is arbitrary.

5. Tolerant PAC learning

Theorem 1 implies that for complex robustness region, robust PAC-learning (Definition 4) is not possible, even when $\mathcal{H}$ is a very simple hypothesis class. Thus, robust learning will require other ideas.

One such idea is Tolerant PAC-learning, introduced in Ashtiani et al. (2022). Here, the idea is to relax the goal of robust PAC-learning by introducing a tolerance parameter γ representing the amount of "slack" the learner gets with respect to the robustness regions U. We now expand their definition to arbitrary robustness regions by introducing *perturbed regions*, U^γ, which are defined as follows.

Definition 7 *Let U be a set of robustness regions and $\gamma > 0$ be a distance. For any point $x \in \mathbb{R}^d$, define U_x^γ as the set of all points with distance at most γ from U_x. That is,*

$$U_x^\gamma = \{x' : ||x' - U_x|| \leq \gamma\}.$$

*Finally, we let $U^\gamma = \{U_x^\gamma : x \in \mathbb{R}^d\}$ denote the set of γ-**perturbed regions** of U.*

Tolerant PAC-learning is then defined as follows

Definition 8 *Let $\mathcal{H}$ be a hypothesis class and U a set of robustness regions. A learner L **tolerantly PAC-learns** $(\mathcal{H}, U)$ if for every $\epsilon, \delta, \gamma > 0$, there exists $m(\epsilon, \delta, \gamma)$ such that for all $n \geq m(\epsilon, \delta, \gamma)$, for all data distributions, $\mathcal{D}$, with probability $1 - \delta$ over $S \sim \mathcal{D}^n$,*

$$\ell_U(\hat{h}, \mathcal{D}) \leq \min_{h \in \mathcal{H}} \ell_{U^\gamma}(h, \mathcal{D}) + \epsilon,$$

*where $\hat{h} = L(S)$ denotes the classifier outputted by L from training sample S. As before, we let $m(\epsilon, \delta, \gamma)$ denote the **tolerant sample complexity** of L with respect to $(\mathcal{H}, U)$.*

5.1. Tolerant RERM oracles

Because our robustness regions, U_x, are arbitrary subsets of $\mathbb{R}^d$, any learning algorithm will require some sort of access to U. We describe this access through an oracle for U.

Ashtiani et al. (2022) employs a *sampling oracle* for U which allows the learner to sample points at uniform from the set U_x for any point x. In their setting, U_x is constrained to be a closed ball of known radius centered at x, and consequently the sampling oracle selects points from the uniform distribution over the ball. We say that a robust learner is in the *sampling model* if its only way of interacting with the regions U_x is through a sampling oracle.

In our setting, where U_x can be an arbitrary regions, sampling oracles pose a significant challenge – there exists choices of U for which tolerant PAC learning requires an exponential number of queries to the sampling oracle. We state this as a proposition with the proof deferred to Appendix B.

Proposition 9 *For any $D > 10\gamma > 0$, there exists a hypothesis class $\mathcal{H}$ and a set of robustness regions, U such that the following holds. There exist constants ϵ and δ such that for any $n > 0$, any learner L on n samples that achieves*

$$\ell_U(L(S), \mathcal{D}) \leq \min_{h \in \mathcal{H}} \ell_{U^\gamma}(h, \mathcal{D}) + \epsilon$$

with probability at least $1 - \delta$ must make at least $\Omega\left(\left(\frac{D}{\gamma}\right)^d\right)$ calls to the sampling oracle on some valid data distribution $\mathcal{D}$, .

To circumvent this issue, we turn our attention to a different natural oracle first proposed in Montasser et al. (2019a) that is based on Robust Empirical Risk Minimization (RERM). An RERM oracle, $\mathcal{O}_{U,\mathcal{H}}(S)$, is a function that returns the classifier $h \in \mathcal{H}$ with minimal robust empirical risk over S. That is,

$$\mathcal{O}_{U,\mathcal{H}}(S) = \arg\min_{h \in \mathcal{H}} \ell_U(h, S).$$

In our work, we will assume access to a mild strengthening of this oracle that allows empirical risk minimization over any perturbed robustness region, U^r.

Definition 10 ***A tolerant RERM-oracle** for robustness regions U and hypothesis class $\mathcal{H}$ is a function $\mathcal{O}_{U,\mathcal{H}}(S, r)$ that maps any set of labeled points S and any distance $r > 0$ to the classifier with minimal empirical risk over S with respect to U^r. That is,*

$$\mathcal{O}_{U,\mathcal{H}}(S, r) = \arg\min_{h \in H} \ell_{U^r}(h, S).$$

Observe that in the case that U consists of balls of radius r, a tolerant oracle merely implies we can also minimize empirical risk for balls of larger radii.

6. Tolerant PAC learning for Regular Hypothesis Classes

Before presenting our algorithm, we first present a key assumption on our hypothesis class, $\mathcal{H}$, that we refer to as *regularity.*

6.1. Regular hypothesis classes

Definition 11 *We say that a hypothesis class, $\mathcal{H}$ is α-**regular** for $\alpha > 0$ if for all $h \in \mathcal{H}$ and for all $x \in \mathbb{R}^d$, there exists a closed ball B of radius α containing x such that $h(x') = h(x)$ for all $x' \in B$. We also say that $\mathcal{H}$ is **regular** if it is α-regular for some $\alpha > 0$.*

This notion was previously introduced in Awasthi et al. (2021) as *pseudo-robustness*.

One important type of classifiers satisfying this condition are hypothesis classes with relatively smooth manifolds as decision boundaries. In particular, the parameter α can be tied to the smoothness measure of a manifold known as its *reach*.

Definition 12 *Let M be a closed manifold embedded in $\mathbb{R}^d$. The **reach** of M is the largest $\alpha > 0$ such that for all $x \in \mathbb{R}^d$, if $||x - M|| \leq \alpha$, then x has a unique nearest neighbor in M.*

This parameter directly translates to regularity.

Proposition 13 *Let h be a classifier with decision boundary M. Suppose that M is a closed $(d-1)$-dimensional submanifold over $\mathbb{R}^d$ with reach α. Then h is $\alpha/2$-regular.*

Proof Let $h \in \mathcal{H}$ be a classifier with decision boundary M. Let x be an arbitrary point with $h(x) = y$. We desire to exhibit a ball B of radius $\alpha/2$ containing x for which h is uniformly y.

Let $\rho : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ be the distance function $\rho(x) = ||x - M||$. It is well known that this function is everywhere continuous and has a continuous derivative over $\{x : 0 < \rho(x) < \alpha\}$.

If $\rho(x) > \alpha/2$, then we can simply take $B = B(x, \alpha/2)$ as all points here must be classified as y by the definition of a decision boundary. Thus, assume $\rho(x) \leq \alpha/2$.

Let V be the gradient vector field of ρ defined over $\{x : \rho(x) < \alpha\}$. Since all points in this region have a unique nearest neighbor in M, it becomes clear that the gradient has magnitude 1 for all such points, and the direction is precisely opposite the straight line path from the point's nearest neighbor in M.

Since V is continuous, (and Lipshitz over a bounded region), there exists a unique curve τ starting at x of length $\frac{\alpha}{2}$ that is always tangent to V. It follows that the endpoint of this path, x' must satisfy $\rho(x') = \frac{\alpha}{2} + \rho(x) > \frac{\alpha}{2}$ and $||x - x'|| \leq \frac{\alpha}{2}$. This means that $B = B(x', \frac{\alpha}{2})$ suffices, as desired. ■

6.2. Our Algorithm

We now give a tolerant PAC learning algorithm called $TolRERM$ (Algorithm 1) which assumes access to a tolerant RERM oracle (Definition 10). $TolRERM$ is essentially robust empirically risk minimization with a slight modification: rather than using the original robustness regions, U, we use the perturbed regions, U^r where $0 < r < \gamma$ is chosen at random. $TolRERM$'s performance is given by Theorem 2, which is restated here for convenience.

Algorithm 1: $TolRERM(\mathcal{D}, \epsilon, \delta, \gamma, n)$
Sample $r \sim [\frac{\epsilon\delta\gamma}{7}, \gamma]$ at uniform
Sample $S \sim \mathcal{D}^n$
Output $\hat{h} = \mathcal{O}_{U,\mathcal{H}}(S, r)$

Theorem 2 *Let $\mathcal{H}$ be a regular hypothesis class with VC dimension v, and let U be any set of robustness regions. Then $TolRERM$ tolerantly PAC-learns $(\mathcal{H}, U)$ with tolerant sample complexity $m(\epsilon, \delta, \gamma) = O\left(\frac{vd \log \frac{dD}{\epsilon\gamma\delta}}{\epsilon^2}\right)$, where D denotes the maximum ℓ_2 diameter of any robustness region, U_x.*

Since the set of bounded linear classifiers, $\mathcal{H}_W$ (Definition 5) is clearly regular and has VC dimension $O(d)$, Theorem 2 immediately implies the following corollary.

Corollary 14 *For any set of robustness regions, U, $TolRERM$ tolerantly PAC-learns $(\mathcal{H}_W, U)$ with tolerant sample complexity $m(\epsilon, \delta, \gamma) = O\left(\frac{d^2 \log \frac{dD}{\epsilon\gamma\delta}}{\epsilon^2}\right)$, where D denotes the maximum ℓ_2 diameter of any robustness region, U_x.*

Observe that $TolRERM$ matches the known sample complexities for linear classifiers found in Montasser et al. (2019a) and Ashtiani et al. (2022). However, it enjoys the advantage of being simpler (as it is essentially an empirical risk minimization algorithm) and a *proper* learning algorithm (as it outputs a linear classifier).

Beyond regular hypothesis classes: It turns out that Algorithm 1 has bounded sample complexity for *any* hypothesis class with finite robust VC-dimension for balls (see Appendix D for a full description). Thus, Algorithm 1 can alternatively be thought of as a reduction from the sample complexity for learning robust classifiers over arbitrary robustness regions to the sample complexity for balls of fixed radii. This is expressed in the following result (proved in Appendix D).

Theorem 3 *Let $\mathcal{H}$ be any hypothesis class with maximal adversarial VC dimension v_{ball}, and let U be any set of robustness regions. Then $TolRERM$ tolerantly PAC-learns $(\mathcal{H}, U)$ with tolerant sample complexity $m(\epsilon, \delta, \gamma) = O\left(\frac{v_{ball} d \log \frac{dD}{\epsilon\gamma\delta}}{\epsilon^2}\right)$, where D denotes the maximum ℓ_2 diameter of any robustness region, U_x.*

6.3. Proof of Theorem 2

We begin by showing that randomly choosing r allows the optimal empirical loss U^r to change relatively smoothly with respect to r.

Lemma 15 *For $r \in [0, \gamma]$, let $OPT_S^r = \min_{h \in H} \ell_{U^r}(h, S)$. Then with probability at least $1 - \frac{\delta}{2}$ over $r \sim [\frac{\epsilon\delta\gamma}{7}, \gamma]$, $OPT_S^r \leq OPT_S^{r - \frac{\epsilon\delta\gamma}{7}} + \frac{\epsilon}{3}$.*

Proof. Let $\alpha = \frac{\epsilon\delta\gamma}{7}$. Our goal is to show that $OPT_S^r - OPT_S^{r-\alpha}$ is likely to be small. Our strategy is to bound the expected value of $OPT_S^r - OPT_S^{r-\alpha}$ and then apply Markov's inequality. As a

technical note, the function $r \mapsto OPT_S^r$ is monotonic and bounded, and consequently measurable, which ensures that our expectations are well defined. To this end, we have,

$$\begin{aligned}
\mathbb{E}[OPT_S^r - OPT_S^{r-\alpha}] &= \mathbb{E}[OPT_S^r] - \mathbb{E}[OPT_S^{r-\alpha}] \\
&= \frac{1}{\gamma-\alpha}\left(\int_\alpha^\gamma OPT_S^r dr - \int_\alpha^\gamma OPT_S^{r-\alpha} dr\right) \\
&= \frac{1}{\gamma-\alpha}\left(\int_\alpha^\gamma OPT_S^r dr - \int_0^{\gamma-\alpha} OPT_S^r dr\right) \\
&= \frac{1}{\gamma-\alpha}\left(\int_{\gamma-\alpha}^\gamma OPT_S^r dr - \int_0^{\alpha} OPT_S^r dr\right) \\
&\leq \frac{\alpha}{\gamma-\alpha} \\
&= \frac{\delta\epsilon\gamma}{7\gamma - \delta\epsilon\gamma} \\
&\leq \frac{\delta\epsilon}{6},
\end{aligned}$$

since $\epsilon, \delta \leq 1$. Applying Markov's inequality, with probability at least $1 - \frac{\delta}{2}$, $OPT_S^r - OPT_S^{r-\alpha} \leq \frac{\epsilon}{3}$. $\square$.

Next, we construct a set of robustness regions V^r that have similar robust loss to U^r and are also finite.

Lemma 16 *Suppose that $\mathcal{H}$ is γ-regular. For all $r \in [\frac{\epsilon\delta\gamma}{7}, \gamma]$, there exists a set of robustness regions $V^r = \{V_x^r : x \in \mathbb{R}^d\}$ satisfying the following two properties.*

1. *$|V_x^r| = O\left(\left(\frac{D}{\epsilon\delta\gamma}\right)^d\right)$, where D denotes the maximum diameter of U_x.*
2. *Let $\alpha = \frac{\epsilon\delta\gamma}{7}$. For all labeled points (x, y) and for all classifiers $h \in \mathcal{H}$,*

$$\ell_{U^{r-\alpha}}(h, (x, y)) \leq \ell_{V^r}(h, (x, y)) \leq \ell_{U^r}(h, (x, y)).$$

Proof For any $x \in \mathbb{R}^d$, we will show how to construct V_x so that it satisfies the two conditions above.

Observe that U_x^r is closed and bounded as it is a union of closed balls of radius r. Since each U_x has diameter at most D, this means that U_x^r is compact. Thus, there exists a finite set of balls of radius $\alpha/2$ that cover U_x^r. Note that these balls are *not* necessarily contained within U_x^r – only that U_x^r is a subset of their union. Let C_x denote the set of all centers of the smallest such cover. We claims that $V_x = C_x \cap U_x^r$ suffices.

First, $|C_x| \leq O\left((\frac{D}{\alpha})^d\right)$ because any ball of diameter D can be covered by $O\left((\frac{D}{\alpha})^d\right)$ balls of radius $\alpha/2$, and U_x^r is a subset of a ball of diameter $D + 2r$. This implies that the first condition holds.

Second, pick any labeled point (x, y) and any classifier $h \in \mathcal{H}$. If $\ell_{V^r}(h, (x, y)) = 1$, then we immediately have $\ell_{U^r}(h, (x, y) = 1$ since $V^r \subseteq U^r$. This implies that $\ell_{V^r}(h, (x, y)) \leq \ell_{U^r}(h, (x, y)$ giving the second half of the second condition.

If $\ell_{U^{r-\alpha}}(h,(x,y))=1$, then there exists $x' \in U_x^{r-\alpha}$ such that $h(x') \neq y$. It follows that since h is γ-regular, h must also be α-regular (as $\alpha < \gamma$). This means that there exists a ball B of radius $\alpha/2$ containing x' such that h does not output y for any point in B.

By the triangle inequality, $B \subseteq U_x^r$, and since C_x covers U_x^r, it follows that there exists $x^* \in C_x \cap B$. By definition, this also means $x^* \in V_x^r$. However, by the definition of B, we must have $h(x^*) \neq y$, and this means that $\ell_{V_x^r}(h,(x,y)) = 1$. Since (x,y) was arbitrary, this proves the second half of the second condition. ■

We are now prepared to prove Theorem 2.

Proof (**Theorem 2**) Let $\alpha = \frac{\epsilon\delta\gamma}{7}$. For all $s > 0$, let $h^s \in \mathcal{H}$ denote any fixed choice of classifier with minimal empirical loss with respect to U^s. That is,

$$h^s = \arg\min_{h\in\mathcal{H}} \ell_{U^s}(h,S).$$

It suffices to show that with probability at least $1-\delta$ over $S \sim \mathcal{D}^n$ and $r \sim [\alpha,\gamma]$,

$$\ell_U(h^r,\mathcal{D}) \leq \min_{h\in\mathcal{H}} \ell_{U^\gamma}(h,\mathcal{D}) + \epsilon.$$

Let $h^* = \arg\min_{h\in\mathcal{H}} \ell_{U^\gamma}(h,\mathcal{D})$, and let V^r be the set of robustness regions defined in Lemma 16. Then by Lemma 16 and the fact that $r \leq \gamma$,

$$\ell_{U^\gamma}(h^*,\mathcal{D}) \geq \ell_{U^r}(h^*,\mathcal{D}) \geq \ell_{V^r}(h^*,\mathcal{D}). \tag{2}$$

Next, since $|V_x| = O\left(\left(\frac{D}{\epsilon\delta\gamma}\right)^d\right)$, Proposition 20 (proved in the Appendix C) implies that the Robust VC dimension of $\mathcal{H}$ with respect to V_x is at most $O\left(vd\log\frac{Dv}{\epsilon\delta\gamma}\right)$, where v denotes the VC dimension of $\mathcal{H}$.

Because S is independent from r, there exists a constant C such that if $n \geq C\frac{vd\log\frac{Dv}{\epsilon\delta\gamma}+\log\frac{1}{\delta}}{\epsilon^2}$, then classical connections with uniform convergence Vapnik and Chervonenkis (1974) imply that with probability at least $1-\frac{\delta}{2}$ over $S \sim \mathcal{D}^n$, for all $h \in \mathcal{H}$, $|\ell_{V^r}(h,S) - \ell_{V^r}(h,\mathcal{D})| \leq \frac{\epsilon}{3}$. This implies,

$$\ell_{V^r}(h^*,\mathcal{D}) \geq \ell_{V^r}(h^*,S) - \frac{\epsilon}{3}. \tag{3}$$

Then, using the fact that $\ell_{U_x^{r-\alpha}} \geq \ell_{V_x^r}$ (Lemma 16) along with the definition of $h^{r-\alpha}$, we have

$$\ell_{V^r}(h^*,S) - \frac{\epsilon}{3} \geq \ell_{U^{r-\alpha}}(h^*,S) - \frac{\epsilon}{3} \geq \ell_{U^{r-\alpha}}(h^{r-\alpha},S) - \frac{\epsilon}{3}. \tag{4}$$

Applying Lemma 15, we have with probability at least $1-\frac{\delta}{2}$ over $r \sim [\alpha,\gamma]$,

$$\ell_{U^{r-\alpha}}(h^{r-\alpha},S) - \frac{\epsilon}{3} \geq \ell_{U^r}(h^r,S) - \frac{2\epsilon}{3}. \tag{5}$$

Using $V_x^r \subset U_x^r$ and uniform convergence over V_x one more time, we get that

$$\ell_{U^r}(h^r,S) - \frac{2\epsilon}{3} \geq \ell_{V^r}(h^r,S) - \frac{2\epsilon}{3} \geq \ell_{V^r}(h^r,\mathcal{D}) - \epsilon. \tag{6}$$

Finally, using Lemma 16 along with the fact that $U_x \subset U_x^{r-\alpha}$, we have

$$\ell_{V^r}(h^r, \mathcal{D}) - \epsilon \geq \ell_{U^{r-\alpha}}(h^r, \mathcal{D}) - \epsilon \geq \ell_U(h^r, \mathcal{D}) - \epsilon. \tag{7}$$

Combining Equations 2, 3, 4, 5, 6, and 7 with the transitive property, completes the proof, as a simple union bound shows that they all hold simultaneously with probability at least $1 - \delta$, as desired.

■

Acknowledgments

Robi Bhattacharjee thanks NSF under CNS 1804829 for research support.

Acknowledgments

We thank a bunch of people.

References

Hassan Ashtiani, Vinayak Pathak, and Ruth Urner. Black-box certification and learning under adversarial perturbations. *CoRR*, abs/2006.16520, 2020. URL https://arxiv.org/abs/2006.16520.

Hassan Ashtiani, Vinayak Pathak, and Ruth Urner. Adversarially robust learning with tolerance. *CoRR*, abs/2203.00849, 2022. doi: 10.48550/arXiv.2203.00849. URL https://doi.org/10.48550/arXiv.2203.00849.

Idan Attias, Aryeh Kontorovich, and Yishay Mansour. Improved generalization bounds for robust learning. In Aurélien Garivier and Satyen Kale, editors, *Algorithmic Learning Theory, ALT 2019, 22-24 March 2019, Chicago, Illinois, USA*, volume 98 of *Proceedings of Machine Learning Research*, pages 162–183. PMLR, 2019.

Pranjal Awasthi, Natalie Frank, and Mehryar Mohri. On the existence of the adversarial bayes classifier. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 2978–2990, 2021.

Robi Bhattacharjee and Kamalika Chaudhuri. When are non-parametric methods robust? *CoRR*, abs/2003.06121, 2020. URL https://arxiv.org/abs/2003.06121.

Robi Bhattacharjee and Kamalika Chaudhuri. Consistent non-parametric methods for maximizing robustness. In Marc'Aurelio Ranzato, Alina Beygelzimer, Yann N. Dauphin, Percy Liang, and Jennifer Wortman Vaughan, editors, *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 9036–9048, 2021.

Nicholas Carlini and David A. Wagner. Towards evaluating the robustness of neural networks. In *2017 IEEE Symposium on Security and Privacy, SP 2017, San Jose, CA, USA, May 22-26, 2017*, pages 39–57, 2017.

Daniel Cullina, Arjun Nitin Bhagoji, and Prateek Mittal. Pac-learning in the presence of adversaries. In Samy Bengio, Hanna M. Wallach, Hugo Larochelle, Kristen Grauman, Nicolò Cesa-Bianchi, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 31: Annual Conference on Neural Information Processing Systems 2018, NeurIPS 2018, December 3-8, 2018, Montréal, Canada*, pages 228–239, 2018.

Ilias Diakonikolas, Daniel M. Kane, and Pasin Manurangsi. The complexity of adversarially robust proper learning of halfspaces with agnostic noise. *CoRR*, abs/2007.15220, 2020. URL https://arxiv.org/abs/2007.15220.

Matthias Hein and Maksym Andriushchenko. Formal guarantees on the robustness of a classifier against adversarial manipulation. In I. Guyon, U. V. Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 2266–2276. Curran Associates, Inc., 2017.

Guy Katz, Clark W. Barrett, David L. Dill, Kyle Julian, and Mykel J. Kochenderfer. Towards proving the adversarial robustness of deep neural networks. In *Proceedings First Workshop on Formal Verification of Autonomous Vehicles, FVAV@iFM 2017, Turin, Italy, 19th September 2017.*, pages 19–26, 2017.

Justin Khim and Po-Ling Loh. Adversarial risk bounds for binary classification via function transformation. *CoRR*, abs/1810.09519, 2018. URL http://arxiv.org/abs/1810.09519.

Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. In *5th International Conference on Learning Representations, ICLR 2017, Toulon, France, April 24-26, 2017, Conference Track Proceedings*, 2017.

Omar Montasser, Steve Hanneke, and Nathan Srebro. VC classes are adversarially robustly learnable, but only improperly. In Alina Beygelzimer and Daniel Hsu, editors, *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, pages 2512–2530. PMLR, 2019a.

Omar Montasser, Steve Hanneke, and Nathan Srebro. VC classes are adversarially robustly learnable, but only improperly. In Alina Beygelzimer and Daniel Hsu, editors, *Conference on Learning Theory, COLT 2019, 25-28 June 2019, Phoenix, AZ, USA*, volume 99 of *Proceedings of Machine Learning Research*, pages 2512–2530. PMLR, 2019b.

Omar Montasser, Steve Hanneke, and Nathan Srebro. Transductive robust learning guarantees. In Gustau Camps-Valls, Francisco J. R. Ruiz, and Isabel Valera, editors, *International Conference on Artificial Intelligence and Statistics, AISTATS 2022, 28-30 March 2022, Virtual Event*, volume 151 of *Proceedings of Machine Learning Research*, pages 11461–11471. PMLR, 2022.

Nicolas Papernot, Patrick D. McDaniel, Somesh Jha, Matt Fredrikson, Z. Berkay Celik, and Ananthram Swami. The limitations of deep learning in adversarial settings. In *IEEE European Symposium on Security and Privacy, EuroS&P 2016, Saarbrücken, Germany, March 21-24, 2016*, pages 372–387, 2016a.

Nicolas Papernot, Patrick D. McDaniel, Xi Wu, Somesh Jha, and Ananthram Swami. Distillation as a defense to adversarial perturbations against deep neural networks. In *IEEE Symposium on Security and Privacy, SP 2016, San Jose, CA, USA, May 22-26, 2016*, pages 582–597, 2016b.

Nicolas Papernot, Patrick D. McDaniel, Ian J. Goodfellow, Somesh Jha, Z. Berkay Celik, and Ananthram Swami. Practical black-box attacks against deep learning systems using adversarial examples. *ASIACCS*, 2017.

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

Norbert Sauer. On the density of families of sets. *Journal of Combinatorial Theory, Series A*, 13 (1):145–147, 1972.

Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning - From Theory to Algorithms.* Cambridge University Press, 2014. ISBN 978-1-10-705713-5.

Saharon Shelah. A combinatorial problem; stability and order for models and theories in infinitary languages. *Pacific Journal of Mathematics*, 41(1):247–261, 1972.

Aman Sinha, Hongseok Namkoong, and John C. Duchi. Certifying some distributional robustness with principled adversarial training. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*, 2018.

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian J. Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *2nd International Conference on Learning Representations, ICLR 2014, Banff, AB, Canada, April 14-16, 2014, Conference Track Proceedings*, 2014.

Vladimir Vapnik and Alexey Chervonenkis. Theory of pattern recognition, 1974.

Yizhen Wang, Somesh Jha, and Kamalika Chaudhuri. Analyzing the robustness of nearest neighbors to adversarial examples. In *Proceedings of the 35th International Conference on Machine Learning, ICML 2018, Stockholmsmässan, Stockholm, Sweden, July 10-15, 2018*, pages 5120–5129, 2018.

Yao-Yuan Yang, Cyrus Rashtchian, Yizhen Wang, and Kamalika Chaudhuri. Adversarial examples for non-parametric methods: Attacks, defenses and large sample limits. *CoRR*, abs/1906.03310, 2019. URL http://arxiv.org/abs/1906.03310.

Dong Yin, Kannan Ramchandran, and Peter L. Bartlett. Rademacher complexity for adversarially robust generalization. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pages 7085–7094. PMLR, 2019. URL http://proceedings.mlr.press/v97/yin19b.html.

Appendix A. Details for the proof of Theorem 1

Proof We want to show that for any $m \in \mathbb{N}$, any learner on m samples must fail with constant probability. Toward this end, set $M = \binom{3m}{m}$, and let $Z_1^{(M)}, Z_2^{(M)}, \ldots, Z_M^{(M)}$ be subsets of $\mathbb{R}^d$ as described by Lemma 6 (we will drop the superscript in what follows). Let $\mathcal{M}$ denote the set of all subsets of $\{1, \ldots, 3m\}$ with exactly m elements. Associate with each Z_i a unique element of $\mathcal{M}$, thus allowing us to rename our subsets as $\{Z_T : T \in \mathcal{M}\}$. We will now construct a set of robustness regions U from these subsets. For $1 \leq i \leq 3m$, define

$$U_{x_i} = \cup_{T: i \in T} Z_T,$$

where x_i is an arbitrary point inside U_{x_i}. Note this is well-defined since the $Z^{(M)}$ are mutually disjoint.

By Lemma 6, it follows that if all x_i are given a label of -1, then any classifier $h \in \mathcal{H}_W$ satisfies that $h(z) = 1$ for some subset T and some $z \in Z_T$. However, this will imply that h lacks robustness on all $x \in \{x_i : i \in T\}$, meaning that there are at least m points among $\{x_1, \ldots, x_{3m}\}$ where h has robust loss 1. Furthermore, the second part of Lemma 6 implies that for any $T \in \mathcal{M}$, there exists a classifier h_T for which h_S is -1 over all $Z_{T'}$ for $T' \neq T$. This implies that h_T is robust at all x_i *except* for x_i with $i \in S$.

With these observations, we are now prepared to show that for any learner L, there exists a distribution D for which L has large expected robust loss. To do this, we use a standard lower bound technique found in Shalev-Shwartz and Ben-David (2014) that was adapted to the robust setting in Montasser et al. (2019a).

The idea will be to pick D to be the uniform distribution over a random subset of $2m$ points in $\{x_1, \ldots, x_{3m}\}$. We will then argue that because L only has access to m points from D, it won't be able to distinguish which subset D corresponds to, and this will lead to a large expected loss.

To this end, for any $T \in \mathcal{M}$, let D_T be the data distribution over $\mathbb{R}^d \times \{\pm 1\}$ where x is chosen at uniform from $\{x_i : i \notin T\}$ and y is always -1. We may assume without loss of generality that our learning algorithm, L, always outputs a classifier among the set $\{h_T : T \in \mathcal{M}\}$. This is because Lemma 6 implies that any classifier in $h \in \mathcal{H}_W$ has robust loss that is at least as bad some h_T (namely, if the decision boundary of h crosses Z_T).

Next, let $T, T' \in \mathcal{M}$ be arbitrary. By definition, h_T lacks robustness on all x_i with $i \in T$, and is perfectly accurate and robust at all other points. It follows that among the $2m$ points in the support of $D_{T'}$, there are $m - |T \cap T'|$ where h_T lacks robustness, implying the the loss of classifier h_T with respect to distribution $D_{T'}$ is $\frac{1}{2} - \frac{|T \cap T'|}{2m}$. Note that this implies that h_T has 0 robust loss over D_T (thus meeting the first stipulation of Theorem 1).

Finally, we bound the expected loss of the learner L with respect to a uniformly random choice of D_T. Let $\mathcal{M}$ also denote the uniform distribution over itself, and let $\mathcal{U}$ denote the uniform distribution over $\{1, 2, 3, \ldots, 3m\}$. Taking expectations over $T \sim \mathcal{M}$ and $S \sim D_T^m$, and letting $h_{L(S)}$ denote the classifier learned by L, we have that

$$\begin{aligned}
\mathbb{E}_{T \sim \mathcal{M}} \mathbb{E}_{S \sim D_T^m} \left[\ell_U(h_{L(S)}, D_T) \right] &= \mathbb{E}_{S \sim \mathcal{U}^m} \mathbb{E}_{T \sim (\mathcal{M}|S)} \left[\ell_U(h_{L(S)}, D_T) \right] \\
&= \mathbb{E}_{S \sim \mathcal{U}^m} \mathbb{E}_{T \sim \{T' : T' \in \mathcal{M}, S \cap T' = \emptyset\}} \left[\frac{1}{2} - \frac{|T \cap L(S)|}{2m} \right].
\end{aligned}$$

To bound the inner expectation, observe that since $|S| = m$, T' has a conditional distribution that is an arbitrary (at uniform) subset of at least $2m$ indices. Since $L(S)$ is fixed, it follows that the

probability that any element in $L(S)$ is an element of T' is at most $\frac{1}{2}$, meaning that the expected value of $|T \cap L(S)|$ is at most $\frac{|L(S)|}{2} = \frac{m}{2}$. Substituting this, we have that

$$\mathbb{E}_{T\sim\mathcal{M}}\mathbb{E}_{S\sim D_T^m}\left[\ell_U(h_{L(S)}, D_T)\right] \geq \mathbb{E}_{S\sim\mathcal{U}^m}\mathbb{E}_{T\sim\{T':T'\in\mathcal{M},S\cap T'=\emptyset\}}\left[\frac{1}{2} - \frac{m}{4m}\right] = \frac{1}{4}.$$

By Markov's inequality, any random variable between 0 and 1 with expectation $\frac{1}{4}$ is strictly larger than $\frac{1}{8}$ with probability at least $\frac{1}{7}$. Since the loss above is bounded between 0 and 1, it follows that $\Pr_{T\sim\mathcal{M}}\Pr_{S\sim D_T}[\ell_U(h_{L(S)}, D) > \frac{1}{8}] \geq \frac{1}{7}$. Thus, for some $D = D_T$, the desired claim holds, finish the proof. ■

Appendix B. Sample Oracle Lower Bounds

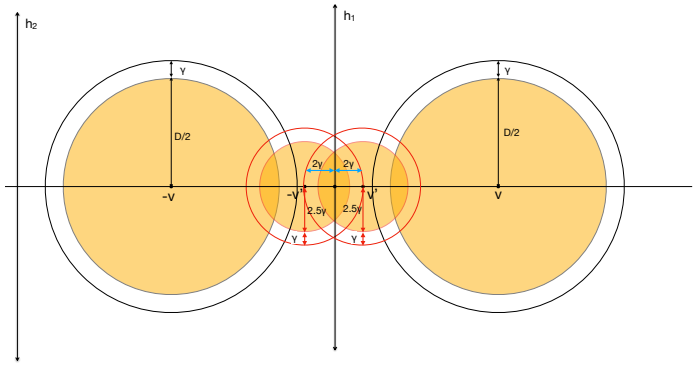

Figure 1: Illustration for the sampling oracle lower bound in Proposition 9 in $\mathbb{R}^2$.

We now show a lower bound on the number of oracle calls required for tolerant learning in Ashtiani et al. (2022)'s sample oracle model. We first recall the model itself for completeness, focusing on the case of $(\mathbb{R}^d, \ell_2)$ endowed with the standard Lebesgue measure for simplicity.

Definition 17 (Sampling Oracle Ashtiani et al. (2022)) *Let $U : \mathbb{R}^d \to P(\mathbb{R}^d)$ be any perturbation function such that $U(x)$ has finite Lebesgue measure for all $x \in \mathbb{R}^d$. The sampling oracle $\mathcal{O}_U$ inputs any $x \in \mathbb{R}^d$, and outputs a sample y from the induced distribution on $U(x)$ under the Lebesgue measure.*

We prove that tolerant learning requires exponentially many calls to the sampling oracle.

Proposition 6 *For any $D > 10\gamma > 0$, there exists a hypothesis class $\mathcal{H}$ and a set of robustness regions, U such that the following holds. There exist constants ϵ and δ, along with a data distribution $\mathcal{D}$, such that for any $n > 0$, any learner L that achieves*

$$\ell_U(L(S), \mathcal{D}) \leq \min_{h\in\mathcal{H}} \ell_{U^\gamma}(h, \mathcal{D}) + \epsilon$$

with probability at least $1 - \delta$ over $S \sim \mathcal{D}^n$ must make at least $\Omega\left(\left(\frac{D}{\gamma}\right)^d\right)$ calls to the sampling oracle.

Proof [Proof of Proposition 9]

Appealing to Yao's Minimax Principle, it is enough to find a class $\mathcal{H}$ and strategy for the adversary (over valid choices of perturbation sets and data distributions) such that any deterministic learner using at most $O((\frac{D}{\gamma})^d)$ oracle calls incurs at least constant error (ϵ) over the optimum in $\mathcal{H}$ with constant probability (δ).

With this in mind, fix $D_0 = D - 9\gamma$, let $r = 4\gamma$, and let e_1 denote the first canonical basis vector in $\mathbb{R}^d$. Our (marginal) data distribution will consist of two points in $\mathbb{R}^d$ $\left\{(\frac{D_0}{2} + 4\gamma)e_1, -(\frac{D_0}{2} + 4\gamma)e_1\right\}$. For the ease of notation, we denote $v := (\frac{D_0}{2} + 4\gamma)e_1$. Note, $||v - (-v)||_2 = D_0 + 2r$. We now define the underlying hypothesis class $\mathcal{H}$ which consists of two linear classifiers $\mathcal{H} := \{h_1, h_2\}$ such that $h_1 = sgn(\langle e_1, \cdot \rangle)$ and $h_2 = (\langle e_1, \cdot \rangle - D_0 - 4\gamma)$. Note that h_1 is a perpendicular bisector of the line segment joining v and $-v$, and h_2 is parallel to h_1 but biased to the left of v.

Finally, we construct two perturbation sets with bounded diameter U and V. Fix $v' = 2\gamma e_1$. We define balls of radius $r > 0$ for any given $x \in \mathbb{R}^d$ as $B_2(x, r) := \left\{x' \in \mathbb{R}^d : ||x' - x||_2 \leq r\right\}$. First, we define a perturbation U and its γ-perturbed region U^γ as follows:

$$U := \{U_v, U_{-v}\} \text{ where for any } x \in \{v, -v\}, U_x = B_2\left(x, \frac{D_0}{2}\right),$$
$$U^\gamma := \left\{U_v^\gamma, U_{-v}^\gamma\right\} \text{ where for any } x \in \{v, -v\}, U_x^\gamma = B_2\left(x, \frac{D_0}{2} + \gamma\right)$$

Similarly, we define another perturbation set V and its γ-perturbed region V^γ:

$$V := \{V_v, V_{-v}\} \text{ where for any } x \in \{v, -v\}, V_x = U_x \cup B_2\left(x', \frac{5\gamma}{2}\right),$$
$$V^\gamma := \left\{V_v^\gamma, V_{-v}^\gamma\right\} \text{ where for any } x \in \{v, -v\}, V_x^\gamma = U_x^\gamma \cup B_2\left(x', \frac{7\gamma}{2}\right)$$

where $x' = v'$ or $-v'$ if $x = v$ or $-v$ respectively. We assume that the perturbation set for $\mathbb{R}^d \setminus \{v, -v\}$ is null for simplicity. Observe that $\bigcap_{x' \in \{v,-v\}} U_{x'} = \emptyset$ and so is the intersection of perturbations in U^γ. But, we note that $\bigcap_{x' \in \{v,-v\}} V_{x'} \neq \emptyset$. This entire construction is illustrated in Figure 1.

We are now ready to describe the adversary's strategy, who chooses one of U or V independently with probability $1/2$, and employs a single fixed choice of data distribution $\mathcal{D}$ where $\Pr[Y = -1| - v] = \Pr[Y = 1|v] = 1$, and the marginal distribution is uniform over v and $-v$. Note that if the perturbation set is U, then h_1 is optimal as $\ell_U(h_1, \mathcal{D}) = 0$ whereas $\ell_U(h_2, \mathcal{D}) = 1/2$. On the other hand if V is chosen then h_2 is optimal as $\ell_V(h_2, \mathcal{D}) = \frac{1}{2}$ and $\ell_V(h_1, \mathcal{D}) = 1$. The idea is to show that the learner cannot distinguish between U and V with high probability, and thus cannot choose the right hypothesis. We note that since the data distribution is fixed and known to the learner, we only need to consider randomness over the sample oracle—labeled samples have no effect on the bound.

More formally, we split our analysis into two cases based on whether or not the learner draws an (oracle) sample in $V^\gamma \setminus U^\gamma$. First, note that conditioned on the fact that the learner draws no such sample, by construction the posterior probability of U is strictly higher than that of V. This means the learner's expected excess error is minimized by always outputting h_1 on such samples. On the

other hand, if the learner observes a sample in $V^\gamma \setminus U^\gamma$, they can always achieve optimal error by outputting h_2.

Since the above learning rule minimizes the learner's expected excess error, it is enough to bound the expected error of this rule:

$$\begin{aligned}\mathbb{E}_{Z,S\sim\mathcal{O}_Z}[OPT_Z - \ell_Z(\mathcal{A}(S),\mathcal{D})] &\geq \frac{1}{2}\Pr[S \subset U^\gamma \wedge Z = V] \\ &= \frac{1}{2}\Pr[Z = V]\Pr[S \subset U^\gamma | Z = V] \\ &= \frac{1}{4}\Pr[S \subset U^\gamma | Z = V]\end{aligned}$$

The key observation is then simply to notice that $\Pr[S \subset U^\gamma | Z = V]$ is constant whenever the learner draws at most $O((\frac{D}{\gamma})^d)$ oracle samples. This follows from the fact that under the induced distribution P_{V^γ} on V:

$$P_{V^\gamma}(V^\gamma \setminus U^\gamma) = \frac{\mu(V^\gamma \setminus U^\gamma)}{\mu(V^\gamma)} \leq \frac{\mu(B_2(v', \frac{7\gamma}{2}))}{\mu(U_v^\gamma) + \mu(B_2(v', \frac{7\gamma}{2}))} \leq \frac{(\frac{7}{2}\gamma)^d}{D_0^d}$$

where μ is the standard Lebesgue measure. Similarly we then have

$$P_{V^\gamma}(U^\gamma) \geq 1 - \frac{(\frac{7}{2}\gamma)^d}{D_0^d}$$

and finally that

$$\Pr[S \subset U^\gamma | Z = V] \geq \left(1 - \frac{(\frac{7}{2}\gamma)^d}{D_0^d}\right)^{|S|}$$

which is at least some constant when $|S| \leq c(\frac{D_0}{\gamma})^d$ for some sufficiently small absolute constant $c < 0$. Since $D_0 = D - 9\gamma$, there exists c' such that this holds when $|S| \leq c'(\frac{D}{\gamma})^d$ which implies the proposition.

■

We note that in Ashtiani et al. (2022), the sampling oracle is defined more generally for any *doubling-measure* μ, that is any measure for which there exists some "doubling-constant" $C > 0$ such that for all $x \in \mathbb{R}^d$ and $r \in \mathbb{R}^+$:

$$0 < \mu(B(x, 2r)) \leq C\mu(B(x, r)) < \infty.$$

In this more general setting, one can prove a lower bound that scales with the doubling-constant (typically exponential in the associated doubling-dimension of the metric space) simply by appropriately increasing the concentration of measure on U^γ.

Appendix C. Robust VC for k points

In this section, we prove that the size-k perturbation sets only cost a $\log(k)$ factor over the VC dimension of the original class. To formalize this, we first recall a few basic definitions standard to the (adversarially robust) learning literature.

Definition 18 (Robust Loss Class) *Given a hypothesis $h : \mathcal{X} \to \{0,1\}$ and perturbation function $U : \mathcal{X} \to P(\mathcal{X})$, let $h_U^\ell : X \times \{0,1\}$ be the function over labeled samples measuring the robust loss of h:*

$$h_U^\ell(x,y) = \begin{cases} 0 & \text{if } \forall x' \in U(x) : h(x') = y \\ 1 & \text{else.} \end{cases}$$

The robust loss class of $(\mathcal{X}, \mathcal{H})$ is the hypothesis class over $\mathcal{X} \times \{0,1\}$ given by:

$$\mathcal{L}_{\mathcal{H}}^U \coloneqq \{h_U^\ell : h \in \mathcal{H}\}.$$

We are interested in analyzing a standard complexity measure of the robust loss class called VC dimension

Definition 19 (VC Dimension) *The VC dimension of a hypothesis class $(\mathcal{X}, \mathcal{H})$ is the size of largest subset $S \subseteq \mathcal{X}$ such that $\mathcal{H}$ obtains all $2^{|S|}$ labelings on S. We say such a set is shattered by $\mathcal{H}$.*

We show the VC dimension of the robust loss class incurs at most $\log(k)$ blow-up over the original class.

Proposition 20 (Overhead of Robust VC) *Let $(\mathcal{X}, \mathcal{H})$ be a hypothesis class of VC-dimension d and $U : \mathcal{X} \to P(\mathcal{X})$ any perturbation function with support bounded by some $k \in \mathbb{N}$. Then the VC dimension of $\mathcal{L}_{\mathcal{H}}^U$ is at most $O(d \log(dk))$.*

This result was also independently communicated to us by Omar Montasser. The proof of Proposition 20 relies on the classical Sauer-Shelah-Perles lemma, which we recall here for completeness.

Lemma 21 (Sauer-Shelah-Perles Sauer (1972); Shelah (1972)) *Let $(\mathcal{X}, \mathcal{H})$ be a hypothesis class of VC-dimension d. Then for any finite subset $S \subseteq \mathcal{X}$, $\mathcal{H}$ obtains at most $O(|S|^d)$ distinct labelings on S.*

Proposition 20 simply follows from using Sauer-Shelah-Perles to bound the total number of permissible patterns across perturbation sets of a sample in the loss space.

Proof [Proof of Proposition 20] Let $m \in \mathbb{N}$ and assume there exists a sample $S = (x_1, y_1), \ldots, (x_m, y_m)$ that is shattered by $\mathcal{L}_{\mathcal{H}}^U$. We will show $m \leq O(d \log(kd))$. With this in mind, let $T = \bigcup_{i=1}^m U(x_i)$ denote the set of at most km points corresponding to the robustness regions of our sample. The key observation is the following (essentially trivial) claim:

Claim 1 *Any two $g_U^\ell, h_U^\ell \in \mathcal{L}_{\mathcal{H}}^U$ that give distinct labelings of S correspond to $g, h \in \mathcal{H}$ with distinct labelings of T.*

By robust shattering, there exist 2^m distinct labelings of S by $\mathcal{L}_{\mathcal{H}}^U$, so the above claim implies $\mathcal{H}$ must have 2^m distinct labelings of T. However the latter has at most $O((km)^d)$ labelings by VC dimension, so

$$2^m \leq O((km)^d) \Rightarrow m \leq O(d \log(dk))$$

by standard manipulations. Finally, we note the claim is immediate from definition, since the behavior of a function $h_U^\ell \in \mathcal{L}_{\mathcal{H}}^U$ on S depends only on the labels of its corresponding hypothesis $h \in \mathcal{H}$ on T by definition. ■

Appendix D. Proof of Theorem 3

We begin by defining v_{ball}, which is the adversarial VC dimension when the robustness regions are all balls of a fixed radius. We start by precisely defining these robustness regions.

Definition 22 *Let U^r be the set of robustness regions defined by $\{U^r_x = B(x,r)\}$, where $B(x,r)$ denotes the closed ball of ℓ_2-radius r centered at x.*

We now define the adversarial VC dimension of a set of classifiers $\mathcal{H}$ for a fixed set of regions, U^r.

Definition 23 *Let $\mathcal{H}$ be a set of classifier. Then the adversarial VC dimension of $\mathcal{H}$ with respect to U^r is the maximum integer v, for which there exist v labeled points, $(x_1,y_1),\ldots,(x_r,y_r)$ so that for any subset $S \subset \{(x_1,y_1),\ldots,(x_r,y_r)\}$, there exists $h_S \in \mathcal{H}$ with*

$$\ell(h_S,(x_i,y_i) = \begin{cases} 0 & i \in S \\ 1 & i \notin S \end{cases}.$$

We denote this by v^r_{ball}.

Finally, we define v_{ball} as the maximum value of v^r_{ball} over all $r > 0$. Note that this quantity has been well studied – for example Cullina et al. (2018) shows that for linear classifiers, $v_{ball} = O(d)$.

Proving Theorem 3 We now turn our attention towards the proof. The key observation is that the main steps from the proof of Theorem 2 *perfectly* carry over. In particular, Lemma 15 exactly holds in this setting, and the argument given in the proof of Theorem 2 also holds provided that an appropriate choice of V exists. The only issue arises from Lemma 16, which requires that $\mathcal{H}$ be regular. To remedy this, we now state and prove a different version of this lemma that uses a union of balls (of fixed radius) for V_x rather than a finite set of points.

Lemma 24 *Let $\mathcal{H}$ by an arbitrary hypothesis class. For all $r \in [\frac{\epsilon\delta\gamma}{7},\gamma]$, let α, U^r and $U^{r-\alpha}$ be as described in the proof of Theorem 2. Then there exists a set of robustness regions $V^r = \{V^r_x : x \in \mathbb{R}^d\}$ satisfying the following two properties.*

1. *V^r_x is a union of $O\left(\left(\frac{D}{\epsilon\delta\gamma}\right)^d\right)$ balls of radius , where D denotes the maximum diameter of U_x.*
2. *Let $\alpha = \frac{\epsilon\delta\gamma}{7}$. For all labeled points (x,y) and for all classifiers $h \in \mathcal{H}$,*

$$\ell_{U^{r-\alpha}}(h,(x,y)) \le \ell_{V^r}(h,(x,y)) \le \ell_{U^r}(h,(x,y)).$$

Proof Since $U^{r-\alpha}_x$ has diameter at most D, it follows that it can be covered with $O\left(\left(\frac{D}{\epsilon\delta\gamma}\right)^d\right)$ balls of radius $\frac{\alpha}{2}$. We let V^r_x be any such cover that is minimal (meaning that (1.) is satisfied), meaning that each ball intersects $U^{r-\alpha}_x$. It follows that for all x, $U^{r-\alpha}_x \subseteq V^r_x \subseteq U^r_x$, which immediately implies (2.) and completes the proof. ■

Finally, to prove Theorem 3, we note that the proof of Theorem 2 essentially works. The only differences are that instead of bounding the robust VC dimension of $\mathcal{H}$ with respect to V_X in terms of v, we must use v_{ball} as we are now considering unions of balls rather than points. As a detail, note that we are using the following minor modification of Proposition 20 to bound the robust VC dimension of unions of balls using the robust VC dimension for balls.

Proposition 25 *Let $(\mathcal{X}, \mathcal{H})$ be a hypothesis class whose robust loss class with respect to r-balls has VC dimension v^r_{ball}. Then the loss class of $(\mathcal{X}, \mathcal{H})$ with respect to perturbations that are a union of at most k r-balls has VC dimension at most $O(v^r_{ball} \log(v^r_{ball} k))$.*

Proof The proof is largely the same as 20. Denote the original perturbation family as U, and the k-union perturbation family by U^k. Given a sample $S = (x_1, y_1), \ldots, (x_m, y_m)$, let C_i denote the centers of the at most k balls appearing in the perturbation set of x_i. It is enough to observe that any two distinct labelings of $S = (x_1, y_1), \ldots, (x_m, y_m)$ by $\mathcal{L}^{U^k}_{\mathcal{H}}$ correspond to distinct labelings of the extended sample $T = \bigcup_{i=1}^m (C_i, y_i)$ with respect to $\mathcal{L}^{U}_{\mathcal{H}}$, where (C_i, y_i) denotes the sample $\bigcup_{c \in C_i} (c, y_1)$. The bound then follows from the same double counting argument as in Proposition 20. ■