On Proper Learnability between Average- and Worst-case Robustness

Vinod Raman
Department of Statistics
University of Michigan
Ann Arbor, MI 48104
vkraman@umich.edu

Unique Subedi
Department of Statistics
University of Michigan
Ann Arbor, MI 48104
subedi@umich.edu

Ambuj Tewari
Department of Statistics
University of Michigan
Ann Arbor, MI 48104
tewaria@umich.edu

Abstract

Recently, Montasser et al. [2019] showed that finite VC dimension is not sufficient for *proper* adversarially robust PAC learning. In light of this hardness, there is a growing effort to study what type of relaxations to the adversarially robust PAC learning setup can enable proper learnability. In this work, we initiate the study of proper learning under relaxations of the adversarially robust loss. We give a family of robust loss relaxations under which VC classes are properly PAC learnable with sample complexity close to what one would require in the standard PAC learning setup. On the other hand, we show that for an existing and natural relaxation of the adversarially robust loss, finite VC dimension is not sufficient for proper learning. Lastly, we give new generalization guarantees for the adversarially robust empirical risk minimizer.

1 Introduction

As deep neural networks become increasingly ubiquitous, their susceptibility to test-time adversarial attacks has become more and more apparent. Designing learning algorithms that are robust to these test-time adversarial perturbations has garnered increasing attention by machine learning researchers and practitioners alike. Prior work on adversarially robust learning has mainly focused on learnability under the worst-case *adversarially* robust risk [Montasser et al., 2019, Attias et al., 2021, Cullina et al., 2018],

$$R_{\mathcal{U}}(h;\mathcal{D}) := \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}}\left[\sup_{z\in\mathcal{U}(x)} \mathbb{1}\{h(z)\neq y\}\right],$$

where $\mathcal{U}(x) \subset \mathcal{X}$ is an arbitrary but fixed perturbation set (for example ℓ_p balls). In practice, worst-case robustness is commonly achieved via Empirical Risk Minimization (ERM) of the adversarially robust loss or some convex surrogate [Madry et al., 2017, Wong and Kolter, 2018, Raghunathan et al., 2018, Bao et al., 2020]. However, a seminal result by Montasser et al. [2019] shows that any proper learning rule, including ERM, even when trained on an arbitrarily large number of samples, may not return a classifier with small adversarially robust risk. These high generalization gaps for the adversarially robust loss have also been observed in practice [Schmidt et al., 2018]. Even worse, empirical studies have shown that classifiers trained to achieve worst-case adversarial robustness exhibit degraded *nominal* performance [Dobriban et al., 2020, Raghunathan et al., 2019, Su et al., 2018, Tsipras et al., 2018, Yang et al., 2020, Zhang et al., 2019, Robey et al., 2022].

In light of these difficulties, there has been a recent push to study when proper learning, and more specifically, when learning via ERM is possible for achieving adversarial robustness. The ability to achieve test-time robustness via proper learning rules is important from a practical standpoint. It aligns better with the current approaches used in practice (e.g. (S)GD-trained deep nets), and proper

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

learning algorithms are often simpler to implement than improper ones. In this vain, Ashtiani et al. [2022] and Bhattacharjee et al. [2022] consider adversarial robust learning in the *tolerant* setting, where the error of the learner is compared with the best achievable error with respect to a slightly *larger* perturbation set. They show that the sample complexity of tolerant robust learning can be significantly lower than the current known sample complexity for adversarially robust learning and that proper learning via ERM can be possible under certain assumptions. Additionally, Ashtiani et al. [2020] provide some sufficient conditions under which proper robust learnability of VC classes becomes possible under a PAC-type framework of semi-supervised learning. More recently, Attias et al. [2022] show that finite VC dimension is sufficient for proper semi-supervised adversarially robust PAC learning if the support of the marginal distribution on instances is known. On the other hand, Attias and Hanneke [2023] study the adversarially robust PAC learnability of real-valued functions and show that every convex function class is properly learnable.

In a different direction, several works have studied the consistency of various *surrogates* of the adversarially robust loss $\ell_{\mathcal{U}}(h,(x,y)) = \sup_{z \in \mathcal{U}(x)} \mathbb{1}\{h(z) \neq y\}$ [Awasthi et al., 2022a,b, 2023, Mao et al., 2023], while others have considered relaxing its worst-case nature [Robey et al., 2022, Li et al., 2020, 2021, Laidlaw and Feizi, 2019, Rice et al., 2021]. However, the PAC learnability of these relaxed notions of the adversarially robust loss has not been well-studied.

In this paper, we study relaxations of the worst-case adversarially robust learning setup from a learning-theoretic standpoint. We classify existing relaxations of worst-case adversarially robust learning into two approaches: one based on relaxing the loss function and the other based on relaxing the benchmark competitor. Much of the existing learning theory work studying relaxations of adversarial robustness focus on the latter approach. These works answer the question of whether proper PAC learning is feasible if the learner is evaluated against a stronger notion of robustness. In contrast, we focus on the *former* relaxation and pose the question: *can proper PAC learning be feasible if we relax the adversarially robust loss function itself?* In answering this question, we make the following main contributions:

- We show that the finiteness of the VC dimension is *not sufficient* for properly learning a natural relaxation of the adversarially robust loss proposed by Robey et al. [2022]. Our proof techniques involve constructing a VC class that is not properly learnable.
- We give a family of adversarially robust loss relaxations that interpolate between average- and worst-case robustness. For these losses, we use Rademacher complexity arguments relying on the Ledoux-Talagrand contraction to show that all VC classes are learnable via ERM.
- We extend a property implicitly appearing in margin theory (e.g., see Mohri et al. [2018, Section 5.4]), which we term "Sandwich Uniform Convergence" (SUC), to show new generalization guarantees for the adversarially robust empirical risk minimizer.

2 Preliminaries and Notation

Throughout this paper we let $[k]$ denote the set of integers $\{1, ..., k\}$, $\mathcal{X}$ denote an instance space, $\mathcal{Y} = \{-1, 1\}$ denote our label space, and $\mathcal{D}$ be any distribution over $\mathcal{X} \times \mathcal{Y}$. Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ denote a hypothesis class mapping examples in $\mathcal{X}$ to labels in $\mathcal{Y}$.

2.1 Problem Setting

In the standard adversarially robust learning setting, the learner, during training time, picks a set $\mathcal{G} \subseteq \mathcal{X}^{\mathcal{X}}$ [1] of perturbation functions $g : \mathcal{X} \to \mathcal{X}$ against which they wish to be robust. At test time, an adversary intercepts the labeled example (x, y), exhaustively searches over the perturbation set to find *the worst* $g \in \mathcal{G}$, and then passes the perturbed instance $g(x)$ to the learner. The learner makes a prediction $\hat{y}$ and then suffers the loss $\mathbb{1}\{\hat{y} \neq y\}$. The goal of the learner is to find a hypothesis $h \in \mathcal{Y}^{\mathcal{X}}$ that minimizes the adversarially robust risk $R_{\mathcal{G}}(h; \mathcal{D}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell_{\mathcal{G}}(h,(x,y))\right]$ where $\ell_{\mathcal{G}}(h,(x,y)) := \sup_{g \in \mathcal{G}} \mathbb{1}\{h(g(x)) \neq y\}$ is the adversarially robust loss.

[1] We highlight that our use of perturbation functions $\mathcal{G}$ instead of perturbation sets $\mathcal{U}$ is without loss of generality (see Appendix A for an equivalence).

In practice, however, such a worst-case adversary may be too strong and unnatural, especially in high-dimension. Accordingly, we relax this model by considering a *lazy*, computationally-bounded adversary that is unable to exhaustively search over $\mathcal{G}$, but can randomly sample a perturbation function $g \sim \mu$ given access to a measure μ over $\mathcal{G}$. In this setup, the learner picks both a perturbation set $\mathcal{G}$ and a measure μ over $\mathcal{G}$. At test-time, the lazy adversary intercepts the labeled example (x, y), *randomly samples* a perturbation function $g \sim \mu$, and then passes the perturbed instance $g(x)$ to the learner. From this perspective, the goal of the learner is to output a hypothesis such that the *probability* that the lazy adversary succeeds in sampling a bad perturbation function, for any labeled example in the support of $\mathcal{D}$, is small. In other words, the learner strives to be robust to most, but not all, perturbation functions in $\mathcal{G}$.

We highlight that the set $\mathcal{G}$ and the measure μ are fixed and chosen by the learner at training time. As a result, the measure μ does not depend on the unperturbed point x. Allowing the measure to depend on the unperturbed point x is an interesting future direction that lies between our model and adversarial robustness. Nevertheless, fixing the measure μ is still relevant from a practical standpoint and non-trivial from a theoretical standpoint. Indeed, a popular choice for $\mathcal{G}$ and μ are ℓ_p balls and the uniform measure respectively. These choices have been shown experimentally to strike a better balance between robustness and nominal performance [Robey et al., 2022]. Moreover, in Section 3, we show that even when the measure μ is fixed apriori, there are learning problems for which ERM does not always work. This parallels the hardness result from Montasser et al. [2019], who prove an analogous result for adversarially robust PAC learning.

2.2 Probabilistically Robust Losses and Proper Learnability

Given a perturbation set $\mathcal{G}$ and measure μ, we quantify the probabilistic robustness of a hypothesis h on a labeled example (x, y) by considering losses that are a function of $\mathbb{P}_{g\sim\mu}[h(g(x)) \neq y]$. For a labeled example $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $\mathbb{P}_{g\sim\mu}\left[h(g(x)) \neq y\right]$ measures the fraction of perturbations in $\mathcal{G}$ for which the classifier h is non-robust. Observe that $\mathbb{P}_{g\sim\mu}\left[h(g(x)) \neq y\right] = \frac{1-y\mathbb{E}_{g\sim\mu}[h(g(x))]}{2}$ is an affine transformation of quantity $y\mathbb{E}_{g\sim\mu}\left[h(g(x))\right]$, the probabilistically robust *margin* of h on (x, y) with respect to $(\mathcal{G}, \mu)$. Thus, we focus on loss functions that operate over the margin $y\mathbb{E}_{g\sim\mu}\left[h(g(x))\right]$. One important loss function is the ρ-probabilistically robust loss,

$$\ell^{\rho}_{\mathcal{G},\mu}(h, (x, y)) := \mathbb{1}\{\mathbb{P}_{g\sim\mu}\left(h(g(x)) \neq y\right) > \rho\},$$

where $\rho \in [0, 1)$ is selected apriori. The ρ-probabilistically robust loss was first introduced by Robey et al. [2022] for the case when $\mathcal{X} = \mathbb{R}^d$, $g_c(x) = x + c$, and the set of perturbations $\mathcal{G} = \{g_c : c \in \Delta\}$ for some $\Delta \subset \mathbb{R}^d$. In this paper, we generalize this loss to an arbitrary instance space $\mathcal{X}$ and perturbation set $\mathcal{G}$. As highlighted by Robey et al. [2022], this notion of robustness is desirable as it nicely interpolates between worst- and average-case robustness via an interpretable parameter ρ, while being more computationally tractable compared to existing relaxations.

In this work, we are primarily interested in understanding whether probabilistic relaxations of the adversarially robust loss enable *proper learning*. That is, given a hypothesis class $\mathcal{H}$, perturbation set and measure $(\mathcal{G}, \mu)$, loss function $\ell_{\mathcal{G},\mu}(h, (x, y)) = \ell(y\mathbb{E}_{g\sim\mu}\left[h(g(x))\right])$, and labeled samples from an unknown distribution $\mathcal{D}$, our goal is to design a learning algorithm $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{H}$ such that for any distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, the algorithm $\mathcal{A}$, given a sample of labeled examples from $\mathcal{D}$, finds a hypothesis $h \in \mathcal{H}$ with low risk with regards to $\ell_{\mathcal{G},\mu}(h, (x, y))$.

Definition 1 (Proper Probabilistically Robust PAC Learnability)**.** *Let $\ell_{\mathcal{G},\mu}(h, (x, y))$ denote an arbitrary probabilistically robust loss function. For any $\epsilon, \delta \in (0, 1)$, the sample complexity of probabilistically robust (ϵ, δ)-learning $\mathcal{H}$ with respect to $\ell_{\mathcal{G},\mu}$, denoted $n(\epsilon, \delta; \mathcal{H}, \ell_{\mathcal{G},\mu})$ is the smallest number $m \in \mathbb{N}$ for which there exists a* proper *learning rule $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{H}$ such that for every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, with probability at least $1 - \delta$ over $S \sim \mathcal{D}^m$,*

$$\mathbb{E}_{\mathcal{D}}\left[\ell_{\mathcal{G},\mu}(\mathcal{A}(S), (x, y))\right] \leq \inf_{h\in\mathcal{H}} \mathbb{E}_{\mathcal{D}}\left[\ell_{\mathcal{G},\mu}(h, (x, y))\right] + \epsilon.$$

We say that $\mathcal{H}$ is proper probabilistically robustly PAC learnable with respect to $\ell_{\mathcal{G},\mu}$ if $\forall\epsilon, \delta \in (0, 1)$, $n(\epsilon, \delta; \mathcal{H}, \ell_{\mathcal{G},\mu})$ is finite.

An important class of proper learning rules is ERMs which simply output the hypothesis $h \in \mathcal{H}$ that minimizes the loss $\ell_{\mathcal{G},\mu}$ over the training sample. In this paper, we ultimately show that for a wide family of probabilistically robust loss functions, ERM is a proper learner according to Definition 1. On the other hand, in Section 3, we show that proper probabilistically robust PAC learning is not always possible for the loss function $\ell^{\rho}_{\mathcal{G},\mu}(h,(x,y))$.

2.3 Complexity Measures

Under the standard 0-1 risk, the Vapnik-Chervonenkis dimension (VC dimension) plays an important role in characterizing PAC learnability, and more specifically, when ERM is possible. A hypothesis class $\mathcal{H}$ is PAC learnable with respect to the 0-1 loss if and only if its VC dimension is finite [Vapnik and Chervonenkis, 1971].

Definition 2 (VC dimension)**.** *A set $\{x_1,...,x_n\} \in \mathcal{X}$ is shattered by $\mathcal{H}$, if $\forall y_1,...,y_n \in \mathcal{Y}$, $\exists h \in \mathcal{H}$, such that $\forall i \in [n]$, $h(x_i) = y_i$. The VC dimension of $\mathcal{H}$, denoted* VC$(\mathcal{H})$*, is defined as the largest natural number $n \in \mathbb{N}$ such that there exists a set $\{x_1,...,x_n\} \in \mathcal{X}$ that is shattered by $\mathcal{H}$.*

One *sufficient* condition for proper learning, based on Vapnik's "General Learning" [Vapnik, 2006], is the finiteness of the VC dimension of a binary loss class

$$\mathcal{L}^{\mathcal{H}} := \{(x,y) \mapsto \ell(h,(x,y)) : h \in \mathcal{H}\}$$

where $\ell(h,(x,y))$ is some loss function mapping to $\{0,1\}$. In particular, if the VC dimension of the loss class $\mathcal{L}^{\mathcal{H}}$ is finite, then $\mathcal{H}$ is PAC learnable with respect to ℓ using ERM with sample complexity that scales linearly with VC$(\mathcal{L}^{\mathcal{H}})$. In this sense, if one can upper bound VC$(\mathcal{L}^{\mathcal{H}})$ in terms of VC$(\mathcal{H})$, then finite VC dimension is sufficient for proper learnability. Unfortunately, for adversarially robust learning, Montasser et al. [2019] show that there can be an arbitrary gap between the VC dimension of the adversarially robust loss class $\mathcal{L}^{\mathcal{H}}_{\mathcal{G}} := \{(x,y) \mapsto \ell_{\mathcal{G}}(h,(x,y)) : h \in \mathcal{H}\}$ and the VC dimension of $\mathcal{H}$. Likewise, in Section 3, we show that for the ρ-probabilistically robust loss $\ell^{\rho}_{\mathcal{G},\mu}$, there can also be an arbitrarily large gap between the VC dimension of the loss class and the VC dimension of the hypothesis class.

As many of the loss functions we consider will actually map to values in $\mathbb{R}$, the VC dimension of the loss class will not be well-defined. Instead, we can capture the complexity of the loss class via the *empirical* Rademacher complexity.

Definition 3 (Empirical Rademacher Complexity of Loss Class)**.** *Let ℓ be a loss function, $S = \{(x_1,y_1),...,(x_n,y_n)\} \in (\mathcal{X}\times\mathcal{Y})^*$ be any sequence of examples, and $\mathcal{L}^{\mathcal{H}} = \{(x,y) \mapsto \ell(h,(x,y)) : h \in \mathcal{H}\}$ be a loss class. The empirical Rademacher complexity of $\mathcal{L}^{\mathcal{H}}$ is defined as*

$$\hat{\mathfrak{R}}_m(\mathcal{L}^{\mathcal{H}}) = \mathbb{E}_\sigma\left[\sup_{f\in\mathcal{L}^{\mathcal{H}}}\left(\frac{1}{n}\sum_{i=1}^{m}\sigma_i f(x_i,y_i)\right)\right]$$

where $\sigma_1,...,\sigma_m$ are independent Rademacher *random variables.*

A standard result relates the empirical Rademacher complexity to the generalization error of hypotheses in $\mathcal{H}$ with respect to a real-valued bounded loss function $\ell(h,(x,y))$ [Bartlett and Mendelson, 2002].

Theorem 2.1 (Rademacher-based Uniform Convergence)**.** *Let $\mathcal{D}$ be a distribution over $\mathcal{X}\times\mathcal{Y}$ and $\ell(h,(x,y)) \leq c$ be a bounded loss function. With probability at least $1-\delta$ over the sample $S \sim \mathcal{D}^m$, for all $h \in \mathcal{H}$ simultaneously,*

$$\left|\mathbb{E}_{\mathcal{D}}[\ell(h(x),y)] - \hat{\mathbb{E}}_S[\ell(h(x),y)]\right| \leq 2\hat{\mathfrak{R}}_m(\mathcal{F}) + O\left(c\sqrt{\frac{\ln(\frac{1}{\delta})}{n}}\right)$$

where $\hat{\mathbb{E}}_S[\ell(h(x),y)] = \frac{1}{|S|}\sum_{(x,y)\in S}\ell(h(x),y)$ is the empirical average of the loss over S.

3 Not All Robust Loss Relaxations Enable Proper Learning

Recall the ρ-probabilistically robust loss

$$\ell^{\rho}_{\mathcal{G},\mu}(h,(x,y)) := \mathbb{1}\{\mathbb{P}_{g\sim\mu}\left(h(g(x)) \neq y\right) > \rho\},$$

and its corresponding risk $R^{\rho}_{\mathcal{G},\mu}(h;\mathcal{D}) := \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\ell^{\rho}_{\mathcal{G},\mu}(h,(x,y))\right]$. At a high-level, proper learning under the $\ell^{\rho}_{\mathcal{G},\mu}$ requires finding a hypothesis $h \in \mathcal{H}$ that is robust to at least a $1-\rho$ fraction of the perturbations in $\mathcal{G}$ for each example in the support of the data distribution $\mathcal{D}$.

Which hypothesis classes are properly learnable with respect to $\ell^{\rho}_{\mathcal{G},\mu}$ according to Definition 1? In Appendix B.1, we show that if $\mathcal{G}$ is finite, then finite VC dimension is sufficient for proper learning with respect to $\ell^{\rho}_{\mathcal{G},\mu}$. On the other hand, in this section, we show that if $\mathcal{G}$ is allowed to be arbitrary, VC dimension is not sufficient for *proper* learning with respect to $\ell^{\rho}_{\mathcal{G},\mu}$, let alone learning via ERM.

Theorem 3.1. *There exists $(\mathcal{G},\mu)$ such that for every $\rho \in [0,1)$, there exists a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ with* $\text{VC}(\mathcal{H}) \leq 1$ *such that $\mathcal{H}$ is not properly probabilistically robust PAC learnable with respect to $\ell^{\rho}_{\mathcal{G},\mu}$.*

To prove Theorem 3.1, we fix $\mathcal{X} = \mathbb{R}^d$, $\mathcal{G} = \{g_\delta : \delta \in \mathbb{R}^d, ||\delta||_p \leq \gamma\}$ such that $g_\delta(x) = x + \delta$ for all $x \in \mathcal{X}$ for some $\gamma > 0$, and μ to be the uniform measure over $\mathcal{G}$. In other words, we are picking our perturbation sets to be ℓ_p balls of radius γ and our perturbation measures to be uniform over each perturbation set. Note that by construction of $\mathcal{G}$, a uniform measure μ over $\mathcal{G}$ also induces a uniform measure μ_x over $\mathcal{G}(x) := \{g_\delta(x) : g_\delta \in \mathcal{G}\} \subset \mathbb{R}^d$. We start by showing that for every $\rho \in [0,1)$, there can be an arbitrary gap between the VC dimension of $\mathcal{H}$ and the loss class $\mathcal{L}^{\mathcal{H},\rho}_{\mathcal{G},\mu} := \{(x,y) \mapsto \ell^{\rho}_{\mathcal{G},\mu}(h,(x,y)) : h \in \mathcal{H}\}$.

Lemma 3.2. *For every $\rho \in [0,1)$ and $m \in \mathbb{N}$, there exists a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ such that* $\text{VC}(\mathcal{H}) \leq 1$ *but* $\text{VC}(\mathcal{L}^{\mathcal{H},\rho}_{\mathcal{G},\mu}) \geq m$.

The proof of Lemma 3.2 is found in Appendix B.2. We highlight two key differences between Lemma 3.2 and its analog, Lemma 2, in Montasser et al. [2019]. First, we need to provide *both* a perturbation set and a perturbation measure. The interplay between these two objects is not present in Montasser et al. [2019] and, apriori, it is not clear that these would indeed be ℓ_p balls and the uniform measure. Second, in order for a hypothesis to be probabilistically non-robust there needs to exist a large enough *region* of perturbations over which it makes mistakes. This is in contrast to Montasser et al. [2019], where a hypothesis is adversarially non-robust as long as there exists *one* non-robust perturbation. Constructing a hypothesis class that achieves all possible probabilistically robust loss behaviors while also having low VC dimension is non-trivial - we need hypotheses to be expressive enough to have large regions of non-robustness while not being too expressive such that VC dimension increases.

Next, we show that the hypothesis class construction in Lemma 3.2 can be used to show the existence of a hypothesis class that cannot be learned properly. Lemma 3.3 is similar to Lemma 3 in Montasser et al. [2019] and is proved in Appendix B.3.

Lemma 3.3. *For every $\rho \in [0,1)$ and $m \in \mathbb{N}$ there exists $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ with* $\text{VC}(\mathcal{H}) \leq 1$ *such that for any proper learner $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{H}$:* (1) *there is a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ and a hypothesis $h^* \in \mathcal{H}$ where $R^{\rho}_{\mathcal{G},\mu}(h^*;\mathcal{D}) = 0$ and* (2) *with probability at least* $1/7$ *over $S \sim D^m$, $R^{\rho}_{\mathcal{G},\mu}(\mathcal{A}(S);\mathcal{D}) > 1/8$.*

Finally, the proof of Theorem 3.1 uses Lemma 3.3 and follows a similar idea as its analog in Montasser et al. [2019] (Theorem 1). However, since our hypothesis class construction in Lemma 3.2 is different, some subtle modifications need to be made. We include a complete proof in Appendix B.4.

4 Proper Learnability Under Relaxed Losses

Despite the fact that VC classes are not ρ-probabilistically robust learnable using proper learning rules, our framework still enables us to capture a wide range of relaxations to the adversarially robust loss for which proper learning is possible.

In particular, consider robust loss relaxations of the form $\ell_{\mathcal{G},\mu}(h,(x,y)) = \ell(y\mathbb{E}_{g\sim\mu}\left[h(g(x))\right])$ where $\ell(t) : \mathbb{R} \to \mathbb{R}$ is a L-Lipschitz function. This class of loss functions is general, capturing many natural robust loss relaxations like the hinge loss $1 - y\mathbb{E}_{g\sim\mu}\left[h(g(x))\right]$, squared loss $(y -$

$\mathbb{E}_{g\sim\mu}[h(g(x))])^2 = (1 - y\mathbb{E}_{g\sim\mu}[h(g(x))])^2$, and exponential loss $e^{-y\mathbb{E}_{g\sim\mu}[h(g(x))]}$. Furthermore, the class of Lipschitz functions $\ell : \mathbb{R} \to \mathbb{R}$ on the margin $y\mathbb{E}_{g\sim\mu}[h(g(x))]$ enables us to capture levels of robustness between the average- and worst-case. For example, taking $\ell(t) = \frac{1-t}{2}$ results in the loss $\ell_{\mathcal{G},\mu}(h,(x,y)) = \ell(y\mathbb{E}_{g\sim\mu}[h(g(x))]) = \mathbb{P}_{g\sim\mu}[h(g(x)) \neq y]$, corresponding to average-case robustness, or *data augmentation*. On the other hand, taking $\ell(t) = \min(\frac{1-t}{2\rho}, 1)$ for some $\rho \in (0,1)$, results in the loss

$$\ell_{\mathcal{G},\mu}(h,(x,y)) = \ell(y\mathbb{E}_{g\sim\mu}[h(g(x))]) = \min\left(\frac{\mathbb{P}_{g\sim\mu}[h(g(x)) \neq y]}{\rho}, 1\right)$$

which corresponds to a notion of robustness that becomes stricter as ρ approaches 0. We note that some of the losses in our family were studied by Rice et al. [2021]. However, their focus was on evaluating robustness, while ours is about (proper) learnability.

Lemma 4.1 shows that for hypothesis classes $\mathcal{H}$ with finite VC dimension, for any $(\mathcal{G},\mu)$, all L-Lipschitz loss functions $\ell_{\mathcal{G},\mu}(h,(x,y))$ enjoy the uniform convergence property.

Lemma 4.1 (Uniform Convergence of Lipschitz Loss)**.** *Let $\mathcal{H}$ be a hypothesis class with finite VC dimension, $(\mathcal{G},\mu)$ be a perturbation set and measure, and $\ell_{\mathcal{G},\mu}(h,(x,y)) = \ell\left(y\mathbb{E}_{g\sim\mu}[h(g(x))]\right)$ such that $\ell : \mathbb{R} \to \mathbb{R}$ is a L-Lipschitz function. With probability at least $1-\delta$ over a sample $S \sim \mathcal{D}^n$ of size $n = O\left(\frac{\mathrm{VC}(\mathcal{H})L^2\ln(\frac{L}{\epsilon})+\ln(\frac{1}{\delta})}{\epsilon^2}\right)$, for all $h \in \mathcal{H}$ simultaneously,*

$$\left|\mathbb{E}_{\mathcal{D}}[\ell_{\mathcal{G},\mu}(h,(x,y))] - \hat{\mathbb{E}}_S[\ell_{\mathcal{G},\mu}(h,(x,y))]\right| \leq \epsilon.$$

Proof. Let $\mathrm{VC}(\mathcal{H}) = d$ and $S = \{(x_1,y_1),...,(x_m,y_m)\}$ be a set of examples drawn i.i.d from $\mathcal{D}$. Define the loss class $\mathcal{L}^{\mathcal{H}}_{\mathcal{G},\mu} = \{(x,y) \mapsto \ell_{\mathcal{G},\mu}(h,(x,y)) : h \in \mathcal{H}\}$. Observe that we can reparameterize $\mathcal{L}^{\mathcal{H}}_{\mathcal{G},\mu}$ as the composition of a L-Lipschitz function $\ell(x)$ and the function class $\mathcal{F}^{\mathcal{H}}_{\mathcal{G},\mu} = \{(x,y) \mapsto y\mathbb{E}_{g\sim\mu}[h(g(x))] : h \in \mathcal{H}\}$. By Proposition 2.1, to show the uniform convergence property of $\ell_{\mathcal{G},\mu}(h,(x,y))$, it suffices to upper bound $\hat{\mathfrak{R}}_m(\mathcal{L}^{\mathcal{H}}_{\mathcal{G},\mu}) = \hat{\mathfrak{R}}_m(\ell \circ \mathcal{F}^{\mathcal{H}}_{\mathcal{G},\mu})$, the empirical Rademacher complexity of the loss class. Since ℓ is L-Lipschitz, by Ledoux-Talagrand's contraction principle [Ledoux and Talagrand, 1991], it follows that $\hat{\mathfrak{R}}_m(\mathcal{L}^{\mathcal{H}}_{\mathcal{G},\mu}) = \hat{\mathfrak{R}}_m(\ell \circ \mathcal{F}^{\mathcal{H}}_{\mathcal{G},\mu}) \leq L \cdot \hat{\mathfrak{R}}_m(\mathcal{F}^{\mathcal{H}}_{\mathcal{G},\mu})$. Thus, it actually suffices to upperbound $\hat{\mathfrak{R}}_m(\mathcal{F}^{\mathcal{H}}_{\mathcal{G},\mu})$ instead. Starting with the definition of the empirical Rademacher complexity:

$$\begin{aligned}
\hat{\mathfrak{R}}_m(\mathcal{F}^{\mathcal{H}}_{\mathcal{G},\mu}) &= \frac{1}{m}\mathbb{E}_{\sigma\sim\{\pm1\}^m}\left[\sup_{h\in\mathcal{H}}\left(\sum_{i=1}^{m}\sigma_i y_i \mathbb{E}_{g\sim\mu}[h(g(x_i))]\right)\right] \\
&= \frac{1}{m}\mathbb{E}_{\sigma\sim\{\pm1\}^m}\left[\sup_{h\in\mathcal{H}}\left(\mathbb{E}_{g\sim\mu}\left[\sum_{i=1}^{m}\sigma_i h(g(x_i))\right]\right)\right] \\
&\leq \mathbb{E}_{g\sim\mu}\left[\frac{1}{m}\mathbb{E}_{\sigma\sim\{\pm1\}^m}\left[\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i h(g(x_i))\right]\right],
\end{aligned}$$

where the last inequality follows from Jensen's inequality and Fubini's Theorem. Note that the quantity $\frac{1}{m}\mathbb{E}_{\sigma\sim\{\pm1\}^m}[\sup_{h\in\mathcal{H}}\sum_{i=1}^{m}\sigma_i h(g(x_i))]$ is the empirical Rademacher complexity of the hypothesis class $\mathcal{H}$ over the sample $\{g(x_1),...,g(x_m)\}$ drawn i.i.d from the distribution defined by first sampling from the marginal data distribution, $x \sim \mathcal{D}_{\mathcal{X}}$, and then applying the transformation $g(x)$. By standard VC arguments, $\hat{\mathfrak{R}}_m(\mathcal{H}) \leq O\left(\sqrt{\frac{d\ln(\frac{m}{d})}{m}}\right)$, which implies that $\hat{\mathfrak{R}}_m(\mathcal{F}^{\mathcal{H}}_{\mathcal{G},\mu}) \leq \mathbb{E}_{g\sim\mu}\left[\hat{\mathfrak{R}}_m(\mathcal{H})\right] \leq O\left(\sqrt{\frac{d\ln(\frac{m}{d})}{m}}\right)$. Putting things together, we get $\hat{\mathfrak{R}}_m(\mathcal{L}^{\mathcal{H}}_{\mathcal{G},\mu}) = \hat{\mathfrak{R}}_m(\ell \circ \mathcal{F}^{\mathcal{H}}_{\mathcal{G},\mu}) \leq O\left(\sqrt{\frac{dL^2\ln(\frac{m}{d})}{m}}\right)$. Proposition 2.1 then implies that with probability $1-\delta$ over a sample $S \sim \mathcal{D}^m$ of size $m = O\left(\frac{dL^2\ln(\frac{L}{\epsilon})+\ln(\frac{1}{\delta})}{\epsilon^2}\right)$, we have

$$\left|\mathbb{E}_{\mathcal{D}}[\ell_{\mathcal{G},\mu}(h,(x,y))] - \hat{\mathbb{E}}_S[\ell_{\mathcal{G},\mu}(h,(x,y))]\right| \leq \epsilon$$

for all $h \in \mathcal{H}$ simultaneously. □

Uniform convergence of Lipschitz-losses immediately implies proper learning via ERM.

Theorem 4.2. *Let* $\ell_{\mathcal{G},\mu}(h,(x,y)) = \ell(y\mathbb{E}_{g\sim\mu}[h(g(x))])$ *such that* $\ell : \mathbb{R} \to \mathbb{R}$ *is a L-Lipschitz function. For every hypothesis class* $\mathcal{H}$*, perturbation set and measure* $(\mathcal{G},\mu)$*, and* $(\epsilon,\delta) \in (0,1)^2$*, the proper learning rule* $\mathcal{A}(S) = \arg\min_{h\in\mathcal{H}} \hat{\mathbb{E}}_S[\ell_{\mathcal{G},\mu}(h,(x,y))]$*, for any distribution* $\mathcal{D}$ *over* $\mathcal{X}\times\mathcal{Y}$*, achieves, with probability at least* $1-\delta$ *over a sample* $S\sim\mathcal{D}^n$ *of size* $n \geq O\left(\frac{\mathrm{VC}(\mathcal{H})L^2\ln(\frac{L}{\epsilon})+\ln(\frac{1}{\delta})}{\epsilon^2}\right)$*, the guarantee*

$$\mathbb{E}_{\mathcal{D}}[\ell_{\mathcal{G},\mu}(\mathcal{A}(S),(x,y))] \leq \inf_{h\in\mathcal{H}} \mathbb{E}_{\mathcal{D}}[\ell_{\mathcal{G},\mu}(h,(x,y))] + \epsilon.$$

At a high-level, Theorem 4.2 shows finite VC dimension is sufficient for achieving robustness *between* the average- and worst-case using ERM. In fact, the next theorem, whose proof can be found in Appendix C.2, shows that finite VC dimension may not even be necessary for this to be true. The proof of Theorem 4.3 involves considering the well-known infinite VC class $\mathcal{H} = \{x \mapsto \text{sign}(\sin(wx)) : w \in \mathbb{R}\}$ and picking $(\mathcal{G},\mu)$ such that $\mathbb{E}_{g\sim\mu}[h(g(x))]$ is essentially constant in x for all hypothesis $h \in \mathcal{H}$.

Theorem 4.3. *Let* $\ell_{\mathcal{G},\mu}(h,(x,y)) = \ell(y\mathbb{E}_{g\sim\mu}[h(g(x))])$ *such that* $\ell : \mathbb{R} \to \mathbb{R}$ *is a L-Lipschitz function. There exists* $\mathcal{H}$ *and* $(\mathcal{G},\mu)$ *such that* $\mathrm{VC}(\mathcal{H}) = \infty$ *but* $\mathcal{H}$ *is still properly learnable with respect to* $\ell_{\mathcal{G},\mu}(h,(x,y))$*.*

Together, Theorems 4.2 and 4.3 showcase an interesting trade-off. Theorem 4.3 indicates that by carefully choosing $(\mathcal{G},\mu)$, the complexity of $\mathcal{H}$ can be essentially smoothed out. On the other hand, Theorem 4.2 shows that any complexity in $(\mathcal{G},\mu)$ can be smoothed out if $\mathcal{H}$ has finite VC dimension. This interplay between the complexities of $\mathcal{H}$ and $(\mathcal{G},\mu)$ closely matches the intuition of Chapelle et al. [2000] in their work on Vicinal Risk Minimization. Note that the results in this section do not contradict that of Section 3 because $\ell^{\rho}_{\mathcal{G},\mu}(h,(x,y))$ is a *non-Lipschitz* function of $y\mathbb{E}_{g\sim\mu}[h(g(x))]$.

We end this section by noting that Lipschitzness of $\ell_{\mathcal{G},\mu}$ is sufficient but, in full generality, not necessary for proper learnability. For example, the loss function that completely ignores $(\mathcal{G},\mu)$ and just computes the 0-1 loss is not Lipschitz, however, it is learnable via ERM when the VC dimension of $\mathcal{H}$ is finite.

5 Proper Learnability Under Relaxed Competition

The results of Section 3 show that relaxing the adversarially robust loss may not always enable proper learning, even for very natural robust loss relaxations. In this section, we show that this bottleneck can be alleviated if we also allow the learner to compete against a slightly stronger notion of robustness. Furthermore, we expand on this idea by exploring other robust learning settings where allowing the learner to compete against a stronger notion of robustness enables proper learnability. We denote this type of modification to the standard adversarially and probabilistically robust leaning settings as robust learning under *relaxed competition*. Prior works on Tolerantly Robust PAC Learning [Ashtiani et al., 2022, Bhattacharjee et al., 2022] mentioned in the introduction fit under this umbrella.

Our main tool in this section is Lemma 5.1, which we term as Sandwich Uniform Convergence (SUC). Roughly speaking, SUC provides a sufficient condition under which ERM outputs a predictor that generalizes well with respect to a stricter notion of loss. A special case of SUC has implicitly appeared in margin theory (e.g., see Mohri et al. [2018, Section 5.4]), where one evaluates the 0-1 risk of the output hypothesis against the optimal *margin* 0-1 risk.

Lemma 5.1 (Sandwich Uniform Convergence)**.** *Let* $\ell_1(h,(x,y))$ *and* $\ell_2(h,(x,y))$ *be bounded, non-negative loss functions such that for all* $h \in \mathcal{H}$ *and* $(x,y) \in \mathcal{X}\times\mathcal{Y}$*, we have* $\ell_1(h,(x,y)) \leq \ell_2(h,(x,y)) \leq 1$*. If* there exists *a loss function* $\tilde{\ell}(h,(x,y))$ *such that* $\ell_1(h,(x,y)) \leq \tilde{\ell}(h,(x,y)) \leq \ell_2(h,(x,y))$ *and* $\tilde{\ell}(h,(x,y))$ *enjoys the* uniform convergence *property with sample complexity* $n(\epsilon,\delta)$*, then the learning rule* $\mathcal{A}(S) = \inf_{h\in\mathcal{H}} \hat{\mathbb{E}}_S[\ell_2(h,(x,y))]$ *achieves, with probability* $1-\delta$ *over a sample* $S\sim\mathcal{D}^m$ *of size* $m \geq n(\epsilon/2,\delta/2) + O\left(\frac{\ln(\frac{1}{\delta})}{\epsilon^2}\right)$*, the guarantee*

$$\mathbb{E}_{\mathcal{D}}[\ell_1(\mathcal{A}(S),(x,y))] \leq \inf_{h\in\mathcal{H}} \mathbb{E}_{\mathcal{D}}[\ell_2(h,(x,y))] + \epsilon.$$

Proof. Let $\mathcal{A}(S) = \inf_{h \in \mathcal{H}} \hat{\mathbb{E}}_S \left[\ell_2(h,(x,y))\right]$. By uniform convergence of $\tilde{\ell}(h,(x,y))$, we have that for sample size $m = n(\frac{\epsilon}{2}, \frac{\delta}{2})$, with probability at least $1 - \frac{\delta}{2}$, over a sample $S \sim \mathcal{D}^m$, for every hypothesis $h \in \mathcal{H}$ simultaneously,

$$\mathbb{E}_{\mathcal{D}}\left[\tilde{\ell}(h,(x,y))\right] \leq \hat{\mathbb{E}}_S\left[\tilde{\ell}(h,(x,y))\right] + \frac{\epsilon}{2}.$$

In particular, this implies that for $\hat{h} = \mathcal{A}(S)$, we have

$$\mathbb{E}_{\mathcal{D}}\left[\tilde{\ell}(\hat{h},(x,y))\right] \leq \hat{\mathbb{E}}_S\left[\tilde{\ell}(\hat{h},(x,y))\right] + \frac{\epsilon}{2}.$$

Since, $\ell_1(h,(x,y)) \leq \tilde{\ell}(h,(x,y)) \leq \ell_2(h,(x,y))$, we have that

$$\mathbb{E}_{\mathcal{D}}\left[\ell_1(\hat{h},(x,y))\right] \leq \hat{\mathbb{E}}_S\left[\ell_2(h^*,(x,y))\right] + \frac{\epsilon}{2}$$

where $h^* = \inf_{h \in \mathcal{H}} \mathbb{E}_{\mathcal{D}}\left[\ell_2(h,(x,y))\right]$. It now remains to upper bound $\hat{\mathbb{E}}_S\left[\ell_2(h^*,(x,y))\right]$ with high probability. However, a standard Hoeffding bound tells us that with probability $1 - \frac{\delta}{2}$ over a sample S of size $O(\frac{\ln(\frac{1}{\delta})}{\epsilon^2})$, $\hat{\mathbb{E}}_S\left[\ell_2(h^*,(x,y))\right] \leq \mathbb{E}_{\mathcal{D}}\left[\ell_2(h^*,(x,y))\right] + \frac{\epsilon}{2}$. Thus, by union bound, we get that with probability at least $1-\delta$, $\mathbb{E}_{\mathcal{D}}\left[\ell_1(\hat{h},(x,y))\right] \leq \mathbb{E}_{\mathcal{D}}\left[\ell_2(h^*,(x,y))\right] + \epsilon$, using a sample of size $n(\epsilon/2, \delta/2) + O(\frac{\ln(\frac{1}{\delta})}{\epsilon^2})$. □

Lemma 5.1 only requires the *existence* of such a sandwiched loss function that enjoys uniform convergence—we do not actually require it to be computable. In the next two sections, we exploit this fact to give three new generalization guarantees for the empirical risk minimizer over the adversarially robust loss $\ell_{\mathcal{G}}(h,(x,y))$ and ρ-probabilistically robust loss $\ell^{\rho}_{\mathcal{G},\mu}(h,(x,y))$, hereafter denoted by $\text{RERM}(S;\mathcal{G}) := \arg\min_{h \in \mathcal{H}} \hat{\mathbb{E}}_S\left[\ell_{\mathcal{G}}(h,(x,y))\right]$ and $\text{PRERM}(S;(\mathcal{G},\mu),\rho) := \arg\min_{h \in \mathcal{H}} \hat{\mathbb{E}}_S\left[\ell^{\rho}_{\mathcal{G},\mu}(h,(x,y))\right]$ respectively.

5.1 (ρ, ρ^*)-Probabilistically Robust PAC Learning

In light of the hardness result of Section 3, we slightly tweak the learning setup in Definition 1 by allowing $\mathcal{A}$ to compete against the hypothesis minimizing the probabilistic robust risk at a level $\rho^* < \rho$. Under this further relaxation, we show that proper learning becomes possible, and that too, via PRERM. In particular, Theorem 5.2 shows that while VC classes are not properly ρ-probabilistically robust PAC learnable, they are properly (ρ, ρ^*)-probabilistically robust PAC learnable.

Theorem 5.2 (Proper (ρ, ρ^*)-Probabilistically Robust PAC Learner)**.** *Let* $0 \leq \rho^* < \rho$. *Then, for every hypothesis class* $\mathcal{H}$, *perturbation set and measure* $(\mathcal{G}, \mu)$, *and* $(\epsilon, \delta) \in (0,1)^2$, *the proper learning rule* $\mathcal{A}(S) = \text{PRERM}(S;(\mathcal{G},\mu),\rho^*)$, *for any distribution* $\mathcal{D}$ *over* $\mathcal{X} \times \mathcal{Y}$, *achieves, with probability at least* $1-\delta$ *over a sample* $S \sim \mathcal{D}^n$ *of size* $n \geq O\left(\frac{\frac{\text{VC}(\mathcal{H})}{(\rho-\rho^*)^2}\ln(\frac{1}{(\rho-\rho^*)\epsilon}) + \ln(\frac{1}{\delta})}{\epsilon^2}\right)$, *the guarantee*

$$R^{\rho}_{\mathcal{G},\mu}(\mathcal{A}(S);\mathcal{D}) \leq \inf_{h \in \mathcal{H}} R^{\rho^*}_{\mathcal{G},\mu}(h;\mathcal{D}) + \epsilon.$$

In contrast to Section 3, where proper learning is not always possible, Theorem 5.2 shows that if we compare our learner to the best hypothesis for a *slightly* stronger level of probabilistic robustness, then not only is proper learning possible for VC classes, but it is possible via ERM. Our main technique to prove Theorem 5.2 is to consider a *different* probabilistically robust loss function that is (1) a Lipschitz function of $y\mathbb{E}_{g \sim \mu}\left[h(g(x)) \neq y\right]$ and (2) can be sandwiched in between $\ell^{\rho^*}_{\mathcal{G},\mu}$ and $\ell^{\rho}_{\mathcal{G},\mu}$. Then, Theorem 5.2 follows from Lemma 5.1. The full proof is in Appendix D.1.

5.2 $(\rho, \mathcal{G})$-Probabilistically Robust PAC Learning

Can measure-independent learning guarantees be achieved if we instead compare the learner's probabilistically robust risk $R^{\rho}_{\mathcal{G},\mu}$ to the best *adversarially robust risk* $R_{\mathcal{G}}$ over $\mathcal{H}$? We answer this in the affirmative by using SUC. We show that if one wants to compete against the best hypothesis for the worst-case adversarially robust risk, it is sufficient to run RERM.

Theorem 5.3 (Proper $(\rho, \mathcal{G})$-Probabilistically Robust PAC Learner)**.** *For every hypothesis class $\mathcal{H}$, perturbation set $\mathcal{G}$, and $(\epsilon, \delta) \in (0,1)^2$, the proper learning rule $\mathcal{A}(S) = \mathrm{RERM}(S;\mathcal{G})$, for any measure μ over $\mathcal{G}$ and any distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, achieves, with probability at least $1-\delta$ over a sample $S \sim \mathcal{D}^n$ of size $n \geq O\left(\frac{\frac{\mathrm{VC}(\mathcal{H})}{\rho^2}\ln(\frac{1}{\rho\epsilon})+\ln(\frac{1}{\delta})}{\epsilon^2}\right)$, the guarantee*

$$R^{\rho}_{\mathcal{G},\mu}(\mathcal{A}(S);\mathcal{D}) \leq \inf_{h \in \mathcal{H}} R_{\mathcal{G}}(h;\mathcal{D}) + \epsilon.$$

The proof of Theorem 5.3 can be found in Appendix D.2, which follows directly from Lemma 5.1 by a suitable choice of the sandwiched loss ℓ. We make a few remarks about the practical importance of Theorem 5.3. Theorem 5.3 implies that for any pre-specified perturbation function class $\mathcal{G}$ (for example ℓ_p balls), running RERM is sufficient to obtain a hypothesis that is probabilistically robust with respect to *any* fixed measure μ over $\mathcal{G}$. Moreover, the level of robustness of the predictor output by RERM, as measured by $1-\rho$, scales directly with the sample size - the more samples one has, the smaller ρ can be made. Alternatively, for a fixed sample size m, desired error ϵ and confidence δ, one can use the sample complexity guarantee in Theorem 5.3 to back-solve the robustness guarantee ρ.

We highlight that the generalization bound in Theorem 5.3, is a *measure-independent* guarantee. This means that ρ does not quantify the level of robustness of the output hypothesis with respect to any one particular measure, but for *any* measure. This is desirable, as in contrast to the previous section, the ρ here more succinctly quantifies the level of robustness achieved by the output classifier. Lastly, we highlight that while the sample complexity of adversarially robust PAC learning can be exponential in the VC dimension of $\mathcal{H}$ [Montasser et al., 2019], this is not the case for $(\rho, \mathcal{G})$-probabilistically robust PAC learning, where we only get a linear dependence on VC dimension.

5.3 Tolerantly Robust PAC Learning

In Tolerantly Robust PAC Learning [Bhattacharjee et al., 2022, Ashtiani et al., 2022], the learner's adversarially robust risk under a perturbation set $\mathcal{G}$ is compared with the best achievable adversarially robust risk for a larger perturbation set $\mathcal{G}' \supset \mathcal{G}$. Ashtiani et al. [2022] study the setting where both $\mathcal{G}$ and $\mathcal{G}'$ induce ℓ_p balls with radius r and $(1+\gamma)r$ respectively. In the work of Bhattacharjee et al. [2022], $\mathcal{G}$ is arbitrary, but $\mathcal{G}'$ is constructed such that it induces perturbation sets that are the union of balls with radius γ that cover $\mathcal{G}$. Critically, Bhattacharjee et al. [2022] show that, under certain assumptions, running RERM over a larger perturbation set $\mathcal{G}'$ is sufficient for Tolerantly Robust PAC learning. In this section, we take a slightly different approach to Tolerantly Robust PAC learning. Instead of having the learner compete against the best possible risk for a larger perturbation set, we have the learner still compete against the best possible adversarially robust risk over $\mathcal{G}$, but evaluate the learner's adversarially robust risk using a *smaller* perturbation set $\mathcal{G}' \subset \mathcal{G}$.

For what $\mathcal{G}' \subset \mathcal{G}$ is Tolerantly Robust PAC learning via RERM possible? As an immediate result of Lemma 5.1 and Vapnik's "General Learning", finite VC dimension of the loss class $\mathcal{L}^{\mathcal{H}}_{\mathcal{G}'} = \{(x,y) \mapsto \ell_{\mathcal{G}'}(h,(x,y)) : h \in \mathcal{H}\}$ is sufficient. Note that finite VC dimension of $\mathcal{L}^{\mathcal{H}}_{\mathcal{G}'}$ implies that the loss function $\ell_{\mathcal{G}'}(h,(x,y))$ enjoys the uniform convergence property with sample complexity $O\left(\frac{\mathrm{VC}(\mathcal{L}^{\mathcal{G}'}_{\mathcal{H}})+\ln(\frac{1}{\delta})}{\epsilon^2}\right)$. Thus, taking $\ell_1(h,(x,y)) = \tilde{\ell}(h,(x,y)) = \ell_{\mathcal{G}'}(h,(x,y))$ and $\ell_2(h,(x,y)) = \ell_{\mathcal{G}}(h,(x,y))$ in Lemma 5.1, we have that if there exists a $\mathcal{G}' \subset \mathcal{G}$ such that $\mathrm{VC}(\mathcal{L}^{\mathcal{G}'}_{\mathcal{H}}) < \infty$, then with probability $1-\delta$ over a sample $S \sim \mathcal{D}^n$ of size $n = O\left(\frac{\mathrm{VC}(\mathcal{L}^{\mathcal{H}}_{\mathcal{G}'})+\ln(\frac{1}{\delta})}{\epsilon^2}\right)$,

$$R_{\mathcal{G}'}(\mathcal{A}(S);\mathcal{D}) \leq \inf_{h \in \mathcal{H}} R_{\mathcal{G}}(h;\mathcal{D}) + \epsilon,$$

where $\mathcal{A}(S) = \mathrm{RERM}(S;\mathcal{G})$.

Alternatively, if $\mathcal{G}' \subset \mathcal{G}$ such that there exists a *finite* subset $\tilde{\mathcal{G}} \subset \mathcal{G}$ where $\ell_{\mathcal{G}'}(h,(x,y)) \leq \ell_{\tilde{\mathcal{G}}}(h,(x,y))$, then Tolerantly Robust PAC learning via RERM is possible with sample complexity that scales according to $O\left(\frac{\mathrm{VC}(\mathcal{H})\log(|\tilde{\mathcal{G}}|)+\ln(\frac{1}{\delta})}{\epsilon^2}\right)$. This result essentially comes from the fact that the VC dimension of the loss class for any finite perturbation set $\tilde{\mathcal{G}}$ incurs only a $\log(|\tilde{\mathcal{G}}|)$ blow-up from the VC dimension of $\mathcal{H}$ (see Lemma 1.1 in Attias et al. [2021]). Thus, finite VC dimension of

$\mathcal{H}$ implies finite VC dimension of the loss class $\mathcal{L}_{\mathcal{H}}^{\tilde{\mathcal{G}}}$ which implies uniform convergence of the loss $\ell_{\tilde{\mathcal{G}}}(h, (x, y))$, as needed for Lemma 5.1 to hold.

We now give an example where such a finite approximation of $\mathcal{G}'$ is possible. In order to do so, we will need to consider a *metric space* of perturbation functions $(\mathcal{G}, d)$ and define a notion of "nice" perturbation sets, similar to"regular" hypothesis classes from Bhattacharjee et al. [2022].

Definition 4 (r-Nice Perturbation Set)**.** *Let $\mathcal{H}$ be a hypothesis class and $(\mathcal{G}, d)$ a metric space of perturbation functions. Let $B_r(g) := \{g' \in \mathcal{G} : d(g, g') \leq r\}$ denote a closed ball of radius r centered around $g \in \mathcal{G}$. We say that $\mathcal{G}' \subset \mathcal{G}$ is r-*Nice *with respect to $\mathcal{H}$, if for all $x \in \mathcal{X}$, $h \in \mathcal{H}$, and $g \in \mathcal{G}'$, there exists a $g^* \in \mathcal{G}$, such that $g \in B_r(g^*)$ and $h(g(x)) = h(g'(x))$ for all $g' \in B_r(g^*)$.*

Definition 4 prevents a situation where a hypothesis $h \in \mathcal{H}$ is non-robust to an isolated perturbation function $g \in \mathcal{G}'$ for any given labelled example $(x, y) \in \mathcal{X} \times \mathcal{Y}$. If a hypothesis h is non-robust to a perturbation $g \in \mathcal{G}'$, then Definition 4 asserts that there must exist a small ball of perturbation functions in $\mathcal{G}$ over which h is also non-robust. Next, we define the covering number.

Definition 5 (Covering Number)**.** *Let $(\mathcal{M}, d)$ be a metric space, let $\mathcal{K} \subset \mathcal{M}$ be a subset, and $r > 0$. Let $B_r(x) = \{x' \in \mathcal{M} : d(x, x') \leq r\}$ denote the ball of radius r centered around $x \in \mathcal{M}$. A subset $\mathcal{C} \subset \mathcal{M}$ is an r-covering of $\mathcal{K}$ if $\mathcal{K} \subset \bigcup_{c \in \mathcal{C}} B_r(c)$. The *covering number *of $\mathcal{K}$, denoted $\mathcal{N}_r(\mathcal{K}, d)$, is the smallest cardinality of any r-covering of $\mathcal{K}$.*

Finally, let $\mathcal{G}'_{2r} = \bigcup_{g \in \mathcal{G}'} B_{2r}(g)$ denote the union over all balls of radius $2r$ with centers in $\mathcal{G}'$. Theorem 5.4 then states that if there exists a set $\mathcal{G}' \subset \mathcal{G}$ that is r-Nice with respect to $\mathcal{H}$, then Tolerantly Robust PAC learning is possible via RERM with sample complexity that scales logarithmically with $\mathcal{N}_r(\mathcal{G}'_{2r}, d)$.

Theorem 5.4 (Tolerantly Robust PAC learning under Nice Perturbations)**.** *Let $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ be a hypothesis class and $(\mathcal{G}, d)$ be a metric space of perturbation functions. Given a subset $\mathcal{G}' \subset \mathcal{G}$ such that $\mathcal{G}'$ is r-*Nice *with respect to $\mathcal{H}$, then the proper learning rule $\mathcal{A}(S) = \text{RERM}(S; \mathcal{G})$, for any distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, achieves, with probability at least $1 - \delta$ over a sample $S \sim \mathcal{D}^n$ of size $n \geq O\left(\frac{\text{VC}(\mathcal{H}) \log(\mathcal{N}_r(\mathcal{G}'_{2r}, d)) + \ln(\frac{1}{\delta})}{\epsilon^2}\right)$, the guarantee*

$$R_{\mathcal{G}'}(\mathcal{A}(S); \mathcal{D}) \leq \inf_{h \in \mathcal{H}} R_{\mathcal{G}}(h; \mathcal{D}) + \epsilon.$$

In Appendix D, we give a full proof and show that *ℓ_p balls are r-Nice perturbation sets for robustly learning halfspaces*. Note that Theorem 5.4 does not require apriori knowledge of the r-Nice perturbation set $\mathcal{G}'$, but just *its existence*. Therefore, Theorem 5.4 applies to the largest possible r-Nice perturbation subset of $\mathcal{G}$. This is important from a practical standpoint as computing an r-Nice perturbation set might not be computationally tractable. Accordingly, while RERM may not be a proper adversarially robust PAC learner [Montasser et al., 2019], Theorem 5.4 shows that RERM is a proper tolerantly robust PAC learner.

6 Discussion

In this work, we show that there exists natural relaxations of the adversarially robust loss for which finite VC dimension is still not sufficient for proper learning. On the other hand, we identify a large set of Lipschitz robust loss relaxations for which finite VC dimension is sufficient for proper learning. In addition, we give new generalization guarantees for RERM. As future work, we are interested in understanding whether our robust loss relaxations can be used to mitigate the tradeoff between achieving adversarial robustness and maintaining high nominal performance. In addition, it is also interesting to find a combinatorial characterization of proper probabilistically robust PAC learning with respect to $\ell^{\rho}_{\mathcal{G},\mu}$.

Acknowledgements

We acknowledge the support of NSF via grant IIS-2007055. We also acknowledge Idan Attias for introducing us to Tolerantly Robust PAC Learning. VR acknowledges the support of the NSF Graduate Research Fellowship.

References

Hassan Ashtiani, Vinayak Pathak, and Ruth Urner. Black-box certification and learning under adversarial perturbations. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 388–398. PMLR, 13–18 Jul 2020.

Hassan Ashtiani, Vinayak Pathak, and Ruth Urner. Adversarially robust learning with tolerance. *arXiv preprint arXiv:2203.00849*, 2022.

Idan Attias and Steve Hanneke. Adversarially robust pac learnability of real-valued functions. 2023.

Idan Attias, Aryeh Kontorovich, and Yishay Mansour. Improved generalization bounds for adversarially robust learning. 2021.

Idan Attias, Steve Hanneke, and Yishay Mansour. A characterization of semi-supervised adversarially robust pac learnability. *Advances in Neural Information Processing Systems*, 35:23646–23659, 2022.

Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. H-consistency bounds for surrogate loss minimizers. In *International Conference on Machine Learning*, pages 1117–1174. PMLR, 2022a.

Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. Multi-class h-consistency bounds. *Advances in neural information processing systems*, 35:782–795, 2022b.

Pranjal Awasthi, Anqi Mao, Mehryar Mohri, and Yutao Zhong. Theoretically grounded loss functions and algorithms for adversarial robustness. In *International Conference on Artificial Intelligence and Statistics*, pages 10077–10094. PMLR, 2023.

Han Bao, Clay Scott, and Masashi Sugiyama. Calibrated surrogate losses for adversarially robust classification. In *Conference on Learning Theory*, pages 408–451. PMLR, 2020.

Peter Bartlett. Lecture notes in theoretical statistics, 2013.

Peter L Bartlett and Shahar Mendelson. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3(Nov):463–482, 2002.

Robi Bhattacharjee, Max Hopkins, Akash Kumar, Hantao Yu, and Kamalika Chaudhuri. Robust empirical risk minimization with tolerance. *arXiv preprint arXiv:2210.00635*, 2022.

Olivier Chapelle, Jason Weston, Léon Bottou, and Vladimir Vapnik. Vicinal risk minimization. *Advances in neural information processing systems*, 13, 2000.

Daniel Cullina, Arjun Nitin Bhagoji, and Prateek Mittal. Pac-learning in the presence of adversaries. *Advances in Neural Information Processing Systems*, 31, 2018.

Edgar Dobriban, Hamed Hassani, David Hong, and Alexander Robey. Provable tradeoffs in adversarially robust classification. *arXiv preprint arXiv:2006.05161*, 2020.

Cassidy Laidlaw and Soheil Feizi. Functional adversarial attacks. *Advances in neural information processing systems*, 32, 2019.

Michel Ledoux and Michel Talagrand. *Probability in Banach Spaces: isoperimetry and processes*, volume 23. Springer Science & Business Media, 1991.

Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. Tilted empirical risk minimization. *arXiv preprint arXiv:2007.01162*, 2020.

Tian Li, Ahmad Beirami, Maziar Sanjabi, and Virginia Smith. On tilted losses in machine learning: Theory and applications. *arXiv preprint arXiv:2109.06141*, 2021.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

A Mao, M Mohri, and Y Zhong. Cross-entropy loss functions: Theoretical analysis and applications. arxiv. *arXiv preprint arXiv:2304.07288*, 2023.

Mehryar Mohri, Afshin Rostamizadeh, and Ameet Talwalkar. *Foundations of machine learning*. MIT press, 2018.

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

Aditi Raghunathan, Jacob Steinhardt, and Percy Liang. Certified defenses against adversarial examples. *arXiv preprint arXiv:1801.09344*, 2018.

Aditi Raghunathan, Sang Michael Xie, Fanny Yang, John C Duchi, and Percy Liang. Adversarial training can hurt generalization. *arXiv preprint arXiv:1906.06032*, 2019.

Leslie Rice, Anna Bair, Huan Zhang, and J Zico Kolter. Robustness between the worst and average case. *Advances in Neural Information Processing Systems*, 34:27840–27851, 2021.

Alexander Robey, Luiz FO Chamon, George J Pappas, and Hamed Hassani. Probabilistically robust learning: Balancing average-and worst-case performance. *arXiv preprint arXiv:2202.01136*, 2022.

Ludwig Schmidt, Shibani Santurkar, Dimitris Tsipras, Kunal Talwar, and Aleksander Madry. Adversarially robust generalization requires more data. *Advances in neural information processing systems*, 31, 2018.

Dong Su, Huan Zhang, Hongge Chen, Jinfeng Yi, Pin-Yu Chen, and Yupeng Gao. Is robustness the cost of accuracy?–a comprehensive study on the robustness of 18 deep image classification models. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 631–648, 2018.

Dimitris Tsipras, Shibani Santurkar, Logan Engstrom, Alexander Turner, and Aleksander Madry. Robustness may be at odds with accuracy. *arXiv preprint arXiv:1805.12152*, 2018.

Vladimir Vapnik. *Estimation of dependences based on empirical data*. Springer Science & Business Media, 2006.

Vladimir Naumovich Vapnik and Aleksei Yakovlevich Chervonenkis. On uniform convergence of the frequencies of events to their probabilities. *Teoriya Veroyatnostei i ee Primeneniya*, 16(2):264–279, 1971.

Eric Wong and Zico Kolter. Provable defenses against adversarial examples via the convex outer adversarial polytope. In *International conference on machine learning*, pages 5286–5295. PMLR, 2018.

Yao-Yuan Yang, Cyrus Rashtchian, Hongyang Zhang, Russ R Salakhutdinov, and Kamalika Chaudhuri. A closer look at accuracy vs. robustness. *Advances in neural information processing systems*, 33:8588–8601, 2020.

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International conference on machine learning*, pages 7472–7482. PMLR, 2019.

A Equivalence between Adversarial Robustness Models

We show that the perturbation set and perturbation function models are equivalent.

Theorem A.1 (Equivalence between $\mathcal{G}$ and $\mathcal{U}$)**.** *Let $\mathcal{X}$ be an arbitrary domain. There exists a perturbation set $\mathcal{U} : \mathcal{X} \to 2^{\mathcal{X}}$ if and only if there exists a set of perturbation functions $\mathcal{G}$ such that $\mathcal{G}(x) = \{g(x) : g \in \mathcal{G}\} = \mathcal{U}(x)$ for all $x \in \mathcal{X}$.*

Proof. We first show that every set of perturbation functions $\mathcal{G}$ induces a perturbation set $\mathcal{U}$. Let $\mathcal{G}$ be an arbitrary set of perturbation functions $g : \mathcal{X} \to \mathcal{X}$. Then, for each $x \in \mathcal{X}$, define $\mathcal{U}(x) := \{g(x) : g \in \mathcal{G}\}$, which completes the proof of this direction.

Now we will show the converse - every perturbation set $\mathcal{U}$ induces a point-wise equivalent set $\mathcal{G}$ of perturbation functions. Let $\mathcal{U}$ be an arbitrary perturbation set mapping points in $\mathcal{X}$ to subsets in $\mathcal{X}$. Assume that $\mathcal{U}(x)$ is not empty for all $x \in \mathcal{X}$. Let $\tilde{z}_x$ denote an arbitrary perturbation from $\mathcal{U}(x)$. For every $x \in \mathcal{X}$, and every $z \in \mathcal{U}(x)$, define the perturbation function $g_z^x(t) = z\mathbb{1}\{t = x\} + \tilde{z}_t\mathbb{1}\{t \neq x\}$ for $t \in \mathcal{X}$. Observe that $g_z^x(x) = z \in \mathcal{U}(x)$ and $g_z^x(x') = \tilde{z}_{x'} \in \mathcal{U}(x')$. Finally, let $\mathcal{G} = \bigcup_{x \in \mathcal{X}} \bigcup_{z \in \mathcal{U}(x)} \{g_z^x\}$. To verify that $\mathcal{G} = \mathcal{U}$, consider an arbitrary point $x' \in \mathcal{X}$. Then,

$$\begin{aligned}
\mathcal{G}(x') &= \bigcup_{x \in \mathcal{X}} \bigcup_{z \in \mathcal{U}(x)} \{g_z^x(x')\} \\
&= \left(\bigcup_{z \in \mathcal{U}(x')} \{g_z^{x'}(x')\} \right) \cup \left(\bigcup_{x \in \mathcal{X} \setminus x'} \bigcup_{z \in \mathcal{U}(x)} \{g_z^x(x')\} \right) \\
&= \left(\bigcup_{z \in \mathcal{U}(x')} \{z\} \right) \cup \left(\bigcup_{x \in \mathcal{X} \setminus x'} \bigcup_{z \in \mathcal{U}(x)} \{\tilde{z}_{x'}\} \right) \\
&= \mathcal{U}(x') \cup \tilde{z}_{x'} \\
&= \mathcal{U}(x').
\end{aligned}$$

as needed. □

B Proofs for Section 3

B.1 Proper ρ-Probabilistically Robust PAC Learning for finite $\mathcal{G}$

We show that if $\mathcal{G}$ is *finite* then VC classes are ρ-probabilistically robustly learnable.

Theorem B.1 (Proper ρ-Probabilistically Robust PAC Learner)**.** *For every hypothesis class $\mathcal{H}$, threshold $\rho \in [0, 1)$, perturbation set $\mathcal{G}$, and perturbation measure μ such that $|\mathcal{G}| \leq K$, there exists a proper learning rule $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^n \to \mathcal{H}$ such that for every distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, with probability at least $1 - \delta$ over $S \sim \mathcal{D}^n$, algorithm $\mathcal{A}$ achieves*

$$R^{\rho}_{\mathcal{G},\mu}(\mathcal{A}(S); \mathcal{D}) \leq \inf_{h \in \mathcal{H}} R^{\rho}_{\mathcal{G},\mu}(h; \mathcal{D}) + \epsilon$$

with

$$n(\epsilon, \delta, \rho; \mathcal{H}, \mathcal{G}, \mu) = O\left(\frac{\text{VC}(\mathcal{H}) \ln(K) + \ln(\frac{1}{\delta})}{\epsilon^2}\right)$$

samples.

Proof. Fix $\rho \in (0, 1)$. Our main strategy will be to upper bound the VC dimension of the ρ-probabilistically robust loss class by some function of the VC dimension of $\mathcal{H}$. Then, finite VC dimension of $\mathcal{H}$ implies finite VC dimension of the loss class, which ultimately implies uniform convergence over the ρ-probabilistically robust loss. Finally, uniform convergence of $\ell^{\rho}_{\mathcal{G},\mu}(h, (x, y))$ implies that ERM is sufficient for ρ-probabilistically robust PAC learning. To that end, let

$$\mathcal{L}^{\mathcal{H},\rho}_{\mathcal{G},\mu} = \{(x, y) \mapsto \mathbb{1}\{\mathbb{P}_{g \sim \mu}\left(h(g(x)) \neq y\right) > \rho\} : h \in \mathcal{H}\}$$

be the ρ-probabilistically robust loss class of $\mathcal{H}$. Let $S = \{(x_1, y_1), ..., (x_n, y_n)\} \in (\mathcal{X} \times \mathcal{Y})^n$ be an arbitrary labeled sample of size n. Inflate S to $S_\mathcal{G}$ by adding for each labelled example $(x, y) \in S$ all possible perturbed examples $(g(x), y)$ for $g \in \mathcal{G}$. That is, $S_\mathcal{G} = \bigcup_{(x,y)\in S}\{(g(x), y) : g \in \mathcal{G}\}$. Note that $|S_\mathcal{G}| \leq nK$. Let $\mathcal{L}^{\mathcal{H},\rho}_{\mathcal{G},\mu}(S)$ denote the set of all possible behaviors of functions in $\mathcal{L}^{\mathcal{H},\rho}_{\mathcal{G},\mu}$ on S. Likewise, let $\mathcal{H}(S_\mathcal{G})$ denote the set of all possible behaviors of functions in $\mathcal{H}$ on the inflated set $S_\mathcal{G}$. Note that each behavior in $\mathcal{L}^{\mathcal{H},\rho}_{\mathcal{G},\mu}(S)$ maps to at least 1 behavior in $\mathcal{H}$. Therefore $|\mathcal{L}^{\mathcal{H},\rho}_{\mathcal{G},\mu}(S)| \leq |\mathcal{H}(S_\mathcal{G})|$. By Sauer-Shelah's lemma, $|\mathcal{H}(S_\mathcal{G})| \leq (nK)^{\text{VC}(\mathcal{H})}$. Solving for n such that $(nK)^{\text{VC}(\mathcal{H})} < 2^n$ gives that $n = O(\text{VC}(\mathcal{H}) \ln(K))$, ultimately implying that $\text{VC}(\mathcal{L}^{\mathcal{H},\rho}_{\mathcal{G},\mu}) \leq O(\text{VC}(\mathcal{H}) \ln(K))$ (see Lemma 1.1 in Attias et al. [2021]).

Since for VC classes, the VC dimension of $\mathcal{L}^{\mathcal{H},\rho}_{\mathcal{G},\mu}$ is bounded, by Vapnik's "General Learning", we have that for VC classes the loss function $\ell^{\rho}_{\mathcal{G},\mu}(h, (x, y))$ enjoys the uniform convergence property. Namely, let $\mathcal{D}$ be a distribution over $\mathcal{X} \times \mathcal{Y}$. For a sample of size $n \geq O(\frac{\text{VC}(\mathcal{H}) \ln(K)+\ln(\frac{1}{\delta})}{\epsilon^2})$, we have that with probability at least $1 - \delta$ over $S \sim \mathcal{D}^n$, for all $h \in \mathcal{H}$

$$|\mathbb{E}_\mathcal{D}\left[\ell^{\rho}_{\mathcal{G},\mu}(h, (x, y))\right] - \hat{\mathbb{E}}_\mathcal{S}\left[\ell^{\rho}_{\mathcal{G},\mu}(h, (x, y))\right]| \leq \epsilon.$$

Standard arguments yield that the proper learning rule $\mathcal{A}(S) = \arg\min_{h\in\mathcal{H}} \hat{\mathbb{E}}_S\left[\ell^{\rho}_{\mathcal{G},\mu}(h, (x, y))\right]$ is a ρ-probabilistically robust PAC learner with sample complexity $O(\frac{\text{VC}(\mathcal{H}) \ln(K)+\ln(\frac{1}{\delta})}{\epsilon^2})$. □

B.2 Proof of Lemma 3.2

Proof. Fix $\rho \in [0, 1)$ and let $m \in \mathbb{N}$. Pick m center points $c_1, ..., c_m$ in $\mathcal{X}$ such that for all $i, j \in [m]$, $\mathcal{G}(c_i) \cap \mathcal{G}(c_j) = \emptyset$. For each center c_i, consider $2^{m-1} + 1$ disjoint subsets of its perturbation set $\mathcal{G}(c_i)$ which do not contain c_i. Label 2^{m-1} of these subsets with a unique bitstring $b \in \{0, 1\}^m$ fixing $b_i = 1$. Let $\mathcal{B}^b_i$ denote the subset labeled by bitstring b and let $\mathcal{B}_i$ denote the single remaining subset that was not labeled. Furthermore, for each $i \in [m]$ and $b \in \{\{0, 1\}^m | b_i = 1\}$, pick $\mathcal{B}_i$ and $\mathcal{B}^b_i$'s such that $\mu_{c_i}(\mathcal{B}_i) = \rho$ and $0 < \mu_{c_i}(\mathcal{B}^b_i) \leq \frac{1-\rho}{2^m}$. If $b_i = 0$, let $\mathcal{B}^b_i = \emptyset$. If $\rho = 0$, let $\mathcal{B}_i = \emptyset$ for all $i \in [m]$. Finally, define $\mathcal{B} = \bigcup_{i=1}^m \bigcup_{b\in\{0,1\}^m} \mathcal{B}^b_i \cup \mathcal{B}_i$ as the union of all the subsets. Crucially, observe that for all $i \in [m]$, $\mu_{c_i}\left(\mathcal{B}_i \cup \left(\bigcup_b \mathcal{B}^b_i\right)\right) \leq \frac{1+\rho}{2} < 1$.

For bitstring $b \in \{0, 1\}^m$, define the hypothesis h_b as

$$h_b(z) = \begin{cases} -1 & \text{if } z \in \bigcup_{i=1}^m \mathcal{B}^b_i \cup \mathcal{B}_i \\ 1 & \text{otherwise} \end{cases}$$

and consider the hypothesis class $\mathcal{H} = \{h_b | b \in \{0, 1\}^m\}$ which consists of all 2^m hypothesis, one for each bitstring. We first show that $\mathcal{H}$ has VC dimension at most 1. Consider two points $x_1, x_2 \in \mathcal{X}$. We will show case by case that every possible pair of points cannot be shattered by $\mathcal{H}$. First, consider the case where, wlog, $x_1 \notin \mathcal{B}$. Then, $\forall h \in \mathcal{H}$, $h(x_1) = 1$, and thus shattering is not possible. Now, consider the case where both $x_1 \in \mathcal{B}$ and $x_2 \in \mathcal{B}$. If either x_1 or x_2 is in $\bigcup_{i=1}^m \mathcal{B}_i$, then every hypothesis $h \in \mathcal{H}$ will label it as -1, and thus these two points cannot be shattered. If $x_1 \in \mathcal{B}^b_i$ and $x_2 \in \mathcal{B}^b_j$ for $i \neq j$, then $h_b(x_1) = h_b(x_2) = -1$, but $\forall h \in \mathcal{H}$ such that $h \neq h_b, h(x_1) = h(x_2) = 1$. If $x_1 \in \mathcal{B}^{b_1}_i$ and $x_2 \in \mathcal{B}^{b_2}_j$ for $b_1 \neq b_2$, then there exists no hypothesis in $\mathcal{H}$ that can label (x_1, x_2) as $(-1, -1)$. Thus, overall, no two points $x_1, x_2 \in \mathcal{X}$ can be shattered by $\mathcal{H}$.

Now we are ready to show that the VC dimension of the loss class is at least m. Specifically, given the sample of labelled points $S = \{(c_1, 1), ..., (c_m, 1)\}$, we will show that the loss behavior corresponding to hypothesis h_b on the sample S is exactly b. Since $\mathcal{H}$ contains all the hypothesis corresponding to every single bitstring $b \in \{0, 1\}^m$, the loss class of $\mathcal{H}$ will shatter S. In order to prove that the loss behavior of h_b on the sample S is exactly b, it suffices to show that the probabilistic loss of h_b on example $(c_i, 1)$ is b_i, where b_i denotes the ith bit of b. By definition,

$$\begin{aligned}
\ell^{\rho}_{\mathcal{G},\mu}(h_b, (c_i, 1)) &= \mathbb{1}\{\mathbb{P}_{g\sim\mu}\left(h_b(g(c_i)) \neq 1\right) > \rho\} \\
&= \mathbb{1}\{\mathbb{P}_{z\sim\mu_{c_i}}\left(h_b(z) = 0\right) > \rho\} \\
&= \mathbb{1}\{\mathbb{P}_{z\sim\mu_{c_i}}\left(z \in \mathcal{B}^b_i \cup \mathcal{B}_i\right) > \rho\} \\
&= \mathbb{1}\{\mu_{c_i}(\mathcal{B}^b_i \cup \mathcal{B}_i) > \rho\} \\
&= b_i.
\end{aligned}$$

Thus, the loss behavior of h_b on S is b, and the total number of distinct loss behaviors over each hypothesis in $\mathcal{H}$ on S is 2^m, implying that the VC dimension of the loss class is at least m. This completes the construction and proof of the claim. □

B.3 Proof of Lemma 3.3

Proof. (of Lemma 3.3) This proof closely follows Lemma 3 from Montasser et al. [2019]. In fact, the only difference is in the construction of the hypothesis class, which we will describe below.

Fix $\rho \in [0, 1)$. Let $m \in \mathbb{N}$. Construct a hypothesis class $\mathcal{H}_0$ as in Lemma 3.2 on $3m$ centers $c_1, ..., c_{3m}$ based on ρ. By the construction in Lemma 3.2, we know that $\mathcal{L}^{\mathcal{H},\rho}_{\mathcal{G},\mu}$ shatters the sample $C = \{(c_1, 1), ..., (c_{3m}, 1)\}$. Instead of keeping all of $\mathcal{H}_0$, we will only keep a subset $\mathcal{H}$ of $\mathcal{H}_0$, namely those classifiers that are probabilistically robustly correct on subsets of size $2m$ of C. More specifically, recall from the construction in Lemma 3.2, that each hypothesis $h_b \in \mathcal{H}_0$ is parameterized by a bitstring $b \in \{0, 1\}^{3m}$ where if $b_i = 1$, then h_b is not robust to example $(c_i, 1)$. Therefore, $\mathcal{H} = \{h_b \in \mathcal{H}_0 : \sum_{i=1}^{3m} b_i = m\}$. Now, let $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{H}$ be an arbitrary proper learning rule. Consider a set of distributions $\mathcal{D}_1, ..., \mathcal{D}_L$ where $L = \binom{3m}{2m}$. Each distribution $\mathcal{D}_i$ is uniform over exactly $2m$ centers in C. Critically, note that by our construction of $\mathcal{H}$, every distribution $\mathcal{D}_i$ is probabilistically robustly realizable by a hypothesis in $\mathcal{H}$. That is, for all $\mathcal{D}_i$, there exists a hypothesis $h^* \in \mathcal{H}$ such that $R^{\rho}_{\mathcal{G},\mu}(h^*; \mathcal{D}_i) = 0$. Observe that this satisfies the first condition in Lemma 3.3. For the second condition, at a high-level, the idea is to use the probabilistic method to show that there exists a distribution $\mathcal{D}_i$ where $\mathbb{E}_{S\sim\mathcal{D}^m_i}\left[R^{\rho}_{\mathcal{G},\mu}(\mathcal{A}(S); \mathcal{D})\right] \geq \frac{1}{4}$ and then use a variant of Markov's inequality to show that with probability at least $1/7$ over $S \sim \mathcal{D}^m$, $R^{\rho}_{\mathcal{G},\mu}(\mathcal{A}(S); \mathcal{D}) > 1/8$.

Let $S \in C^m$ be an arbitrary set of m points. Let $\mathcal{C}$ be a uniform distribution over C. Let $\mathcal{P}$ be a uniform distribution over $\mathcal{D}_1, ..., \mathcal{D}_T$. Let E_S denote the event that $S \subset \text{supp}(\mathcal{D}_i)$ for $\mathcal{D}_i \sim \mathcal{P}$. Given the event E_S, we will lower bound the expected probabilistic robust loss of the hypothesis the proper learning rule $\mathcal{A}$ outputs,

$$\mathbb{E}_{\mathcal{D}_i\sim\mathcal{P}}\left[R^{\rho}_{\mathcal{G},\mu}(\mathcal{A}(S); \mathcal{D}_i)|E_S\right] = \mathbb{E}_{\mathcal{D}_i\sim\mathcal{P}}\left[\mathbb{E}_{(x,y)\sim\mathcal{D}_i}\left[\mathbb{1}\{\mathbb{P}_{g\sim\mu}\left(\mathcal{A}(S)(g(x)) \neq y\right) > \rho\}\right]|E_S\right].$$

Conditioning on the event that $(x, y) \notin S$, denoted, $E_{(x,y)\notin S}$,

$$\begin{aligned}
\mathbb{E}_{(x,y)\sim\mathcal{D}_i}\left[\mathbb{1}\{\mathbb{P}_{g\sim\mu}\left(\mathcal{A}(S)(g(x)) \neq y\right) > \rho\}\right] \geq \mathbb{P}_{(x,y)\sim\mathcal{D}_i}&\left[E_{(x,y)\notin S}\right] \\
\times\, \mathbb{E}_{(x,y)\sim\mathcal{D}_i}&\left[\mathbb{1}\{\mathbb{P}_{g\sim\mu}\left(\mathcal{A}(S)(g(x)) \neq y\right) > \rho\}|E_{(x,y)\notin S}\right]
\end{aligned}$$

Since $\mathcal{D}_i$ is supported over $2m$ points and $|S| = m$, $\mathbb{P}_{(x,y)\sim\mathcal{D}_i}\left[E_{(x,y)\notin S}\right] \geq \frac{1}{2}$ since in the worst-case $S \subset \text{supp}(\mathcal{D}_i)$. Thus, we obtain the lower bound,

$$\mathbb{E}_{\mathcal{D}_i\sim\mathcal{P}}\left[R^{\rho}_{\mathcal{G},\mu}(\mathcal{A}(S); \mathcal{D}_i)|E_S\right] \geq \frac{1}{2}\mathbb{E}_{\mathcal{D}_i\sim\mathcal{P}}\left[\mathbb{E}_{(x,y)\sim\mathcal{D}_i}\left[\mathbb{1}\{\mathbb{P}_{g\sim\mu}\left(\mathcal{A}(S)(g(x)) \neq y\right) > \rho\}|E_{(x,y)\notin S}\right]|E_S\right].$$

Unravelling the expectation over the draw from $\mathcal{D}_i$ given the event E_S, we have,

$$\mathbb{E}_{(x,y)\sim\mathcal{D}_i}\left[\mathbb{1}\{\mathbb{P}_{g\sim\mu}\left(\mathcal{A}(S)(g(x))\neq y\right)>\rho\}|E_{(x,y)\notin S}\right]\geq\frac{1}{m}\sum_{(x,y)\in\text{supp}(\mathcal{D}_i)\setminus S}\mathbb{1}\{\mathbb{P}_{g\sim\mu}\left(\mathcal{A}(S)(g(x))\neq y\right)>\rho\}$$

Observing that $\mathbb{E}_{\mathcal{D}_i\sim\mathcal{P}}\left[\mathbb{1}\{(x,y)\in\text{supp}(\mathcal{D}_i)\}|E_S\right]\geq\frac{1}{2}$ yields,

$$\mathbb{E}_{\mathcal{D}_i\sim\mathcal{P}}\left[\mathbb{E}_{(x,y)\sim\mathcal{D}_i}\left[\mathbb{1}\{\mathbb{P}_{g\sim\mu}\left(\mathcal{A}(S)(g(x))\neq y\right)>\rho\}|E_{(x,y)\notin S}\right]|E_S\right]\geq\frac{1}{2m}\sum_{(x,y)\notin S}\mathbb{1}\{\mathbb{P}_{g\sim\mu}\left(\mathcal{A}(S)(g(x))\neq y\right)>\rho\}.$$

Since $\mathcal{A}(S)\in\mathcal{H}$, by construction of $\mathcal{H}$, there are at least m points in C where $\mathcal{A}$ is not probabilistically robustly correct. Therefore,

$$\frac{1}{2m}\sum_{(x,y)\notin S}\mathbb{1}\{\mathbb{P}_{g\sim\mu}\left(\mathcal{A}(S)(g(x))\neq y\right)>\rho\}\geq\frac{1}{2},$$

from which we have that, $\mathbb{E}_{\mathcal{D}_i\sim\mathcal{P}}\left[R^{\rho}_{\mathcal{G},\mu}(\mathcal{A}(S);\mathcal{D}_i)|E_S\right]\geq\frac{1}{4}$. By the law of total expectation, we have that

$$\begin{aligned}\mathbb{E}_{\mathcal{D}_i\sim\mathcal{P}}\left[\mathbb{E}_{S\sim\mathcal{D}_i^m}\left[R^{\rho}_{\mathcal{G},\mu}(\mathcal{A}(S);\mathcal{D}_i)\right]\right]&=\mathbb{E}_{S\sim\mathcal{C}}\left[\mathbb{E}_{\mathcal{D}_i\sim\mathcal{P}|E_S}\left[R^{\rho}_{\mathcal{G},\mu}(\mathcal{A}(S);\mathcal{D}_i)\right]\right]\\&=\mathbb{E}_{S\sim\mathcal{C}}\left[\mathbb{E}_{\mathcal{D}_i\sim\mathcal{P}}\left[R^{\rho}_{\mathcal{G},\mu}(\mathcal{A}(S);\mathcal{D}_i)|E_S\right]\right]\\&\geq 1/4\end{aligned}$$

Since the expectation over $\mathcal{D}_1,...,\mathcal{D}_T$ is at least $1/4$, there must exist a distribution $\mathcal{D}_i$ where $\mathbb{E}_{S\sim\mathcal{D}_i^m}\left[R^{\rho}_{\mathcal{G},\mu}(\mathcal{A}(S);\mathcal{D}_i)\right]\geq 1/4$. Using a variant of Markov's inequality, gives

$$\mathbb{P}_{S\sim\mathcal{D}_i^m}\left[R^{\rho}_{\mathcal{G},\mu}(\mathcal{A}(S);\mathcal{D}_i)>1/8\right]\geq 1/7$$

which completes the proof. □

B.4 Proof of Theorem 3.1

Proof. (of Theorem 3.1) Fix $\rho\in[0,1)$. Let $(C_m)_{m\in\mathbb{N}}$ be an infinite sequence of disjoint sets such that each set C_m contains $3m$ distinct center points from $\mathcal{X}$, where for any $c_i,c_j\in\bigcup_{m=1}^{\infty}C_m$ such that $c_i\neq c_j$, we have $\mathcal{G}(c_i)\cap\mathcal{G}(c_j)=\emptyset$. For every $m\in\mathbb{N}$, construct $\mathcal{H}_m$ on C_m as in Lemma 3.2. In addition, a key part of this proof is to ensure that the hypothesis in $\mathcal{H}_m$ are non-robust to points in $C_{m'}$ for all $m'\neq m$. To do so, we will need to adjust each hypothesis $h_b\in\mathcal{H}_m$ carefully. By definition, for every $m\in\mathbb{N}$, $\mathcal{H}_m$ consists of 2^{3m} hypothesis of the form

$$h_b(z)=\begin{cases}-1 & \text{if } z\in\bigcup_{i=1}^{3m}\mathcal{B}_i^b\cup\mathcal{B}_i\\ 1 & \text{otherwise}\end{cases}$$

for each bitstring $b\in\{0,1\}^{3m}$. Note that the same set $\bigcup_{i=1}^{3m}\mathcal{B}_i$ is shared across every hypothesis $h_b\in\mathcal{H}_m$. For each $m\in\mathbb{N}$, let $\mathcal{B}^m=\bigcup_{i=1}^{3m}\mathcal{B}_i$ be exactly the union of these $3m$ sets. Next, from the construction in Lemma 3.2, for every center $c_i\in C_m$, $\mu_{c_i}\left(\mathcal{B}_i\cup\left(\bigcup_b\mathcal{B}_i^b\right)\right)\leq\frac{1+\rho}{2}<1$. Thus, there exists a set $\tilde{\mathcal{B}}_i\subset\mathcal{G}(c_i)$ such that $\mu_{c_i}(\tilde{\mathcal{B}}_i)>0$ and $\tilde{\mathcal{B}}_i\cap\left(\mathcal{B}_i\cup\left(\bigcup_b\mathcal{B}_i^b\right)\right)=\emptyset$. Consider one such subset $\tilde{\mathcal{B}}_i$ from each of the $3m$ centers in C_m and let $\tilde{\mathcal{B}}^m=\bigcup_{i=1}^{3m}\tilde{\mathcal{B}}_i$. Finally, make the following adjustment to each $h_b\in\mathcal{H}_m$,

$$h_b(z)=\begin{cases}-1 & \text{if } z\in\bigcup_{i=1}^{3m}\mathcal{B}_i^b\cup\mathcal{B}_i \text{ or } z\in\mathcal{B}^{m'}\cup\tilde{\mathcal{B}}^{m'} \text{ for } m'\neq m\\ 1 & \text{otherwise}\end{cases}$$

One can verify that every hypothesis in $\mathcal{H}_m$ has a non-robust region (i.e. $\mathcal{B}^{m'} \cup \tilde{\mathcal{B}}^{m'}$ for $m' \neq m$) with mass strictly bigger than ρ in every center in $C_{m'}$ for every $m' \neq m$. Thus, the hypotheses in $\mathcal{H}_m$ are non-robust to points in $C_{m'}$ for all $m' \neq m$. Finally, as we did in Lemma 3.3, for each m, we only keep the subset of hypothesis $\mathcal{H}'_m = \{h_b \in \mathcal{H}_m : \sum_{i=1}^{3m} b_i = m\}$. Note that for each $m \in \mathbb{N}$, the hypothesis class $\mathcal{H}'_m$ behaves exactly like the hypothesis class from Lemma 3.3 on C_m.

Let $\mathcal{H} := \bigcup_{m=1}^{\infty} \mathcal{H}'_m$ and $\mathcal{G}(C_m) := \bigcup_{i=1}^{3m} \mathcal{G}(c_i)$. Since we have modified the hypothesis class, we need to reprove that its VC dimension is still at most 1. Consider two points $x_1, x_2 \in \mathcal{X}$. If either x_1 or x_2 is not in $\bigcup_{m=1}^{\infty} \mathcal{G}(C_m)$ and not in $\bigcup_{m=1}^{\infty} \mathcal{B}^m \cup \tilde{\mathcal{B}}^m$, then all hypothesis predict x_1 or x_2 as 1. If both x_1 and x_2 are in $\mathcal{B}^m \cup \tilde{\mathcal{B}}^m$ for some $m \in \mathbb{N}$, then:

- if either x_1 or x_2 are in $\mathcal{B}^m$, every hypothesis in $\mathcal{H}$ labels either x_1 or x_2 as -1.
- if both x_1 and x_2 are in $\tilde{\mathcal{B}}^m$, we can only get the labeling $(1, 1)$ from hypotheses in $\mathcal{H}_m$ and the labeling $(-1, -1)$ from the hypotheses in $\mathcal{H}_{m'}$ for $m' \neq m$.

In the case both x_1 and x_2 are in $\mathcal{G}(C_m) \setminus (\mathcal{B}^m \cup \tilde{\mathcal{B}}^m)$, then, they cannot be shattered by Lemma 3.2. In the case $x_1 \in \mathcal{B}^m \cup \tilde{\mathcal{B}}^m$ and $x_2 \in \mathcal{G}(C_m) \setminus (\mathcal{B}^m \cup \tilde{\mathcal{B}}^m)$:

- if x_1 is in $\mathcal{B}^m$, every hypothesis in $\mathcal{H}$ labels x_1 as -1.
- if x_1 is in $\tilde{\mathcal{B}}^m$ then, we can never get the labeling $(-1, -1)$.

If $x_1 \in \mathcal{B}^i \cup \tilde{\mathcal{B}}^i$ and $x_2 \in \mathcal{B}^j \cup \tilde{\mathcal{B}}^j$ for $i \neq j$, then:

- if either x_1 or x_2 are in $\mathcal{B}^i$ or $\mathcal{B}^j$ respectively, every hypothesis in $\mathcal{H}$ labels either x_1 or x_2 as -1.
- if both x_1 and x_2 are in $\tilde{\mathcal{B}}^i$ and $\tilde{\mathcal{B}}^j$ respectively, we can never get the labeling $(1, 1)$.

In the case $x_1 \in \mathcal{B}^i \cup \tilde{\mathcal{B}}^i$ and $x_2 \in \mathcal{G}(C_j) \setminus (\mathcal{B}^j \cup \tilde{\mathcal{B}}^j)$ for $j \neq i$, then we cannot obtain the labeling $(1, -1)$. If $x_1 \in \mathcal{G}(C_i) \setminus (\mathcal{B}^i \cup \tilde{\mathcal{B}}^i)$ and $x_2 \in \mathcal{G}(C_j) \setminus (\mathcal{B}^j \cup \tilde{\mathcal{B}}^j)$ for $i \neq j$, then we cannot obtain the labeling $(-1, -1)$. Since we shown that for all possible x_1 and x_2, $\mathcal{H}$ cannot shatter them, VC$(\mathcal{H}) \leq 1$.

We now use the same reasoning in Montasser et al. [2019], to show that no proper learning rule works. By Lemma 3.3, for any proper learning rule $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{H}$ and for any $m \in \mathbb{N}$, we can construct a distribution $\mathcal{D}$ over C_m (which has $3m$ points from $\mathcal{X}$) where there exists a hypothesis $h^* \in \mathcal{H}'_m$ that achieves $R^{\rho}_{\mathcal{G},\mu}(h^*; \mathcal{D}) = 0$, but with probability at least $1/7$ over $S \sim \mathcal{D}^m$, $R^{\rho}_{\mathcal{G},\mu}(\mathcal{A}(S); \mathcal{D}) > 1/8$. Note that it suffices to only consider hypothesis in $\mathcal{H}'_m$ because, by construction, all hypothesis in $\mathcal{H}'_{m'}$ for $m' \neq m$ are not probabilistically robust on C_m, and thus always achieve loss 1 on all points in C_m. Thus, rule $\mathcal{A}$ will do worse if it picks hypotheses from these classes. This shows that the sample complexity of properly probabilistically robustly PAC learning $\mathcal{H}$ is arbitrarily large, allowing us to conclude that $\mathcal{H}$ is not properly learnable. □

C Proofs for Section 4

C.1 Proof of Theorem 4.2

Proof. (of Theorem 4.2) Let VC$(\mathcal{H}) = d$ and $S = \{(x_1, y_1), ..., (x_m, y_m)\}$ an i.i.d. sample of size m from $\mathcal{D}$. Consider the learning algorithm $\mathcal{A}(S) = \arg\min_{h \in \mathcal{H}} \hat{\mathbb{E}}_S \left[\ell_{\mathcal{G},\mu}(h, (x, y))\right]$. Note that $\mathcal{A}$ is a proper learning algorithm. Let $\hat{h} = \mathcal{A}(S)$ denote hypothesis output by $\mathcal{A}$ and $h^* = \inf_{h \in \mathcal{H}} \mathbb{E}_{\mathcal{D}} \left[\ell_{\mathcal{G},\mu}(h, (x, y))\right]$.

We now show that if the sample size $m = O\left(\frac{dL^2 \ln(\frac{L}{\epsilon}) + \ln(\frac{1}{\delta})}{\epsilon^2}\right)$, then $\hat{h}$ achieves the stated generalization bound with probability $1 - \delta$. By Lemma 4.1, if $m = O\left(\frac{dL^2 \ln(\frac{L}{\epsilon}) + \ln(\frac{1}{\delta})}{\epsilon^2}\right)$, we have that with probability $1-\delta$, for all $h \in \mathcal{H}$ simultaneously,

$$\left|\mathbb{E}_{\mathcal{D}}\left[\ell_{\mathcal{G},\mu}(h,(x,y))\right]-\hat{\mathbb{E}}_S\left[\ell_{\mathcal{G},\mu}(h,(x,y))\right]\right| \leq \frac{\epsilon}{2}.$$

This means that both $\mathbb{E}_{\mathcal{D}}\left[\ell_{\mathcal{G},\mu}(\hat{h},(x,y))\right]-\hat{\mathbb{E}}_S\left[\ell_{\mathcal{G},\mu}(\hat{h},(x,y))\right] \leq \frac{\epsilon}{2}$ and $\hat{\mathbb{E}}_S\left[\ell_{\mathcal{G},\mu}(h^*,(x,y))\right]-\mathbb{E}_{\mathcal{D}}\left[\ell_{\mathcal{G},\mu}(h^*,(x,y))\right] \leq \frac{\epsilon}{2}$. By definition of $\hat{h}$, note that $\hat{\mathbb{E}}_S\left[\ell_{\mathcal{G},\mu}(\hat{h},(x,y))\right] \leq \hat{\mathbb{E}}_S\left[\ell_{\mathcal{G},\mu}(h^*,(x,y))\right]$. Putting these observations together, we have that

$$\begin{aligned}\mathbb{E}_{\mathcal{D}}\left[\ell_{\mathcal{G},\mu}(\hat{h},(x,y))\right]-(\mathbb{E}_{\mathcal{D}}\left[\ell_{\mathcal{G},\mu}(h^*,(x,y))\right]+\frac{\epsilon}{2}) &\leq \mathbb{E}_{\mathcal{D}}\left[\ell_{\mathcal{G},\mu}(\hat{h},(x,y))\right]-\hat{\mathbb{E}}_S\left[\ell_{\mathcal{G},\mu}(h^*,(x,y))\right]\\ &\leq \mathbb{E}_{\mathcal{D}}\left[\ell_{\mathcal{G},\mu}(\hat{h},(x,y))\right]-\hat{\mathbb{E}}_S\left[\ell_{\mathcal{G},\mu}(\hat{h},(x,y))\right]\\ &\leq \frac{\epsilon}{2},\end{aligned}$$

from which we can deduce that

$$\mathbb{E}_{\mathcal{D}}\left[\ell_{\mathcal{G},\mu}(\hat{h},(x,y))\right]-\inf_{h\in\mathcal{H}}\mathbb{E}_{\mathcal{D}}\left[\ell_{\mathcal{G},\mu}(h,(x,y))\right] \leq \epsilon.$$

Thus, $\mathcal{A}$ achieves the stated generalization bound with sample complexity $m = O\left(\frac{dL^2\ln(\frac{L}{\epsilon})+\ln(\frac{1}{\delta})}{\epsilon^2}\right)$, completing the proof. □

C.2 Proof of Theorem 4.3

For the proof in this section, it will be useful to define the $(\mathcal{G},\mu)$-smoothed hypothesis class $\mathcal{H}$:

$$\mathcal{F}^{\mathcal{H}}_{\mathcal{G},\mu} := \{\mathbb{E}_{g\sim\mu}\left[h(g(x))\right] : h \in \mathcal{H}\}.$$

Proof. (of Theorem 4.3) Let $\mathcal{X}=\mathbb{R}$ and $\mathcal{H}=\{\text{sign}(\sin(\omega x)) : \omega \in \mathbb{R}\}$. Without loss of generality, assume $\text{sign}(\sin(0))=1$. For every $x \in \mathcal{X}$ and $c \in [-1,1]$, define $g_c(x)=cx$. Then, let $\mathcal{G}=\{g_c : c \in [-1,1]\}$ and μ be uniform over $\mathcal{G}$. First, $\mathit{VC}(\mathcal{H})=\infty$ as desired. Next, to show learnability, it suffices to show that the loss

$$\ell_{\mathcal{G},\mu}(h,(x,y))=\ell(y\mathbb{E}_{g\sim\mu}\left[h(g(x))\right]).$$

enjoys the uniform convergence property despite $\text{VC}(\mathcal{H})=\infty$. By Theorem 2.1 and similar to the proof of Lemma 4.1, it suffices upperbound the Rademacher complexity of the loss class $\mathcal{L}^{\mathcal{H}}_{\mathcal{G},\mu}=\{(x,y)\mapsto \ell_{\mathcal{G},\mu}(h,(x,y)) : h \in \mathcal{H}\}$. Since for every fixed y, $\ell_{\mathcal{G},\mu}(h,(x,y))$ is L-Lipschitz with respect to the real-valued function $\mathbb{E}_{g\sim\mu}\left[h(g(x))\right]$, by Ledoux-Talagrand's contraction principle $\hat{\mathfrak{R}}_m(\mathcal{L}^{\mathcal{H}}_{\mathcal{G},\mu}) \leq L\cdot\hat{\mathfrak{R}}_m(\mathcal{F}^{\mathcal{H}}_{\mathcal{G},\mu})$ where $\mathcal{F}^{\mathcal{H}}_{\mathcal{G},\mu}$ is the $(\mathcal{G},\mu)$-smoothed hypothesis classed defined previously. Thus, it suffices to upper-bound $\hat{\mathfrak{R}}_m(\mathcal{F}^{\mathcal{H}}_{\mathcal{G},\mu})$ by a sublinear function of m to show that $\ell_{\mathcal{G},\mu}(h,(x,y))$ enjoys the uniform convergence property. But for every $h_\omega \in \mathcal{H}$,

$$\mathbb{E}_{g\sim\mu}\left[h_\omega(g(x))\right]=\mathbb{E}_{c\sim\text{Unif}(-1,1)}\left[\text{sign}(\sin(\omega(cx))\right]=\frac{1}{2}\int_{-1}^{1}\text{sign}(\sin(c(\omega x)))dc.$$

Since $\sin(ax)$ is an odd function, $\text{sign}(\sin(ax))$ is also odd, from which it follows that for all $h_\omega \in \mathcal{H}$:

$$\mathbb{E}_{g\sim\mu}\left[h_\omega(g(x))\right]=\begin{cases}0 & \text{if } x \neq 0 \text{ and } \omega \neq 0\\ 1 & \text{otherwise}\end{cases}.$$

Therefore, $\mathcal{F}^{\mathcal{H}}_{\mathcal{G},\mu}=\{f_1,f_2\}$ where $f_1(x)=1$ for all $x \in \mathbb{R}$ and $f_2(x)=1$ if $x=0$ and $f_2(x)=0$ if $x \neq 0$. Since $\mathcal{F}^{\mathcal{H}}_{\mathcal{G},\mu}$ is finite, by Massart's Lemma [Mohri et al., 2018], $\hat{\mathfrak{R}}_m(\mathcal{F}^{\mathcal{H}}_{\mathcal{G},\mu})$ is upper-bounded by a sublinear function of m such that $\ell_{\mathcal{G},\mu}(h,(x,y))$ enjoys the uniform convergence property with sample complexity $O(\frac{L^2+\ln(\frac{1}{\delta})}{\epsilon^2})$. Therefore, $(\mathcal{H},\mathcal{G},\mu)$ is PAC learnable with respect to $\ell_{\mathcal{G},\mu}(h,(x,y))$ by the learning rule $\mathcal{A}(S)=\arg\min_{h\in\mathcal{H}}\hat{\mathbb{E}}_S\left[\ell_{\mathcal{G},\mu}(h,(x,y))\right]$ with sample complexity that scales according to $O(\frac{L^2+\ln(\frac{1}{\delta})}{\epsilon^2})$. □

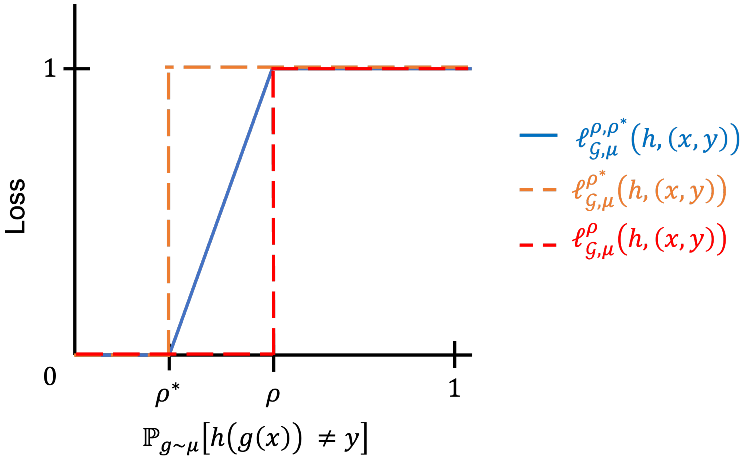

Figure 1: Comparison of probabilistic robust *ramp* loss to probabilistic robust losses of hypothesis h on example (x, y). The probabilistic robust losses at ρ and ρ^* sandwich the probabilistic robust ramp loss at ρ, ρ^*.

D Proofs for Section 5

D.1 Proof of Theorem 5.2

Proof. (of Theorem 5.2) Fix $0 \leq \rho^* < \rho < 1$ and let $\mathcal{H}$ be a hypothesis class with $\text{VC}(\mathcal{H}) = d$. Let $(\mathcal{G}, \mu)$ be an arbitrary perturbation set and measure, $\mathcal{D}$ be an arbitrary distribution over $\mathcal{X} \times \mathcal{Y}$, and $S = \{(x_1, y_1), ..., (x_m, y_m)\}$ an i.i.d. sample of size m. Let $\mathcal{A}(S) = \text{PRERM}(S; (\mathcal{G}, \mu), \rho^*)$.

By Lemma 5.1, it suffices to show that there exists a loss function $\ell(h, (x, y))$ such that $\ell_{\mathcal{G},\mu}^{\rho}(h, (x, y)) \leq \ell(h, (x, y)) \leq \ell_{\mathcal{G},\mu}^{\rho^*}(h, (x, y)))$ and $\ell(h, (x, y))$ enjoys the uniform convergence property with sample complexity $n = O\left(\frac{\frac{d}{(\rho-\rho^*)^2}\ln(\frac{1}{(\rho-\rho^*)\epsilon})+\ln(\frac{1}{\delta})}{\epsilon^2}\right)$. Consider the probabilistically robust ramp loss:

$$\ell_{\mathcal{G},\mu}^{\rho,\rho^*}(h, (x, y)) = \min(1, \max(0, \frac{\mathbb{P}_{g\sim\mu}\left[h(g(x)) \neq y\right] - \rho^*}{\rho - \rho^*})).$$

Figure 1 visually showcases how the probabilistic robust losses at ρ and ρ^* sandwich the probabilistic ramp loss at ρ, ρ^*.

Its not too hard to see that $\ell_{\mathcal{G},\mu}^{\rho}(h, (x, y)) \leq \ell_{\mathcal{G},\mu}^{\rho,\rho^*}(h, (x, y)) \leq \ell_{\mathcal{G},\mu}^{\rho^*}(h, (x, y)))$. Furthermore, since $\ell_{\mathcal{G},\mu}^{\rho,\rho^*}(h, (x, y))$ is $O(\frac{1}{\rho-\rho^*})$-Lipschitz in $y\mathbb{E}_{g\sim\mu}\left[h(g(x)) \neq y\right]$, by Lemma 4.1, we have that $\ell_{\mathcal{G},\mu}^{\rho,\rho^*}(h, (x, y))$ enjoys the uniform convergence property with sample complexity $O\left(\frac{\frac{d}{(\rho-\rho^*)^2}\ln(\frac{1}{(\rho-\rho^*)\epsilon})+\ln(\frac{1}{\delta})}{\epsilon^2}\right)$. This completes the proof, as the conditions for Lemma 5.1 have been met, and therefore the learning rule $\mathcal{A}(S) = \text{PRERM}(S; \mathcal{G}, \rho^*)$ enjoys the stated generalization guarantee with the specified sample complexity. □

D.2 Proof of Theorem 5.3

Proof. (of Theorem 5.3) Fix $0 < \rho$ and let $\mathcal{H}$ be a hypothesis class with $\text{VC}(\mathcal{H}) = d$. Let $\mathcal{G}$ be an arbitrary perturbation set, $\mathcal{D}$ be an arbitrary distribution over $\mathcal{X} \times \mathcal{Y}$, and $S = \{(x_1, y_1), ..., (x_m, y_m)\}$ an i.i.d. sample of size m. Let $\mathcal{A}(S) = \text{RERM}(S; \mathcal{G})$.

Fix a measure μ over $\mathcal{G}$. By Lemma 5.1, it suffices to show that there exists a loss function $\ell(h, (x, y))$ such that $\ell_{\mathcal{G},\mu}^{\rho}(h, (x, y)) \leq \ell(h, (x, y)) \leq \ell_{\mathcal{G}}(h, (x, y)))$ and $\ell(h, (x, y))$ enjoys the uniform convergence property with sample complexity $n = O\left(\frac{\frac{d}{\rho^2}\ln(\frac{1}{\rho\epsilon})+\ln(\frac{1}{\delta})}{\epsilon^2}\right)$. Recall the probabilistically robust ramp loss:

$$\ell_{\mathcal{G},\mu}^{\rho,\rho^*}(h,(x,y)) = \min(1, \max(0, \frac{\mathbb{P}_{g\sim\mu}\left[h(g(x)) \neq y\right] - \rho^*}{\rho - \rho^*})).$$

Letting $\rho^* = 0$, its not too hard to see that $\ell^{\rho}_{\mathcal{G},\mu}(h,(x,y)) \leq \ell^{\rho,0}_{\mathcal{G},\mu}(h,(x,y)) \leq \ell_{\mathcal{G}}(h,(x,y)))$. Furthermore, since $\ell^{\rho,0}_{\mathcal{G},\mu}(h,(x,y))$ is $O(\frac{1}{\rho})$-Lipschitz in $y\mathbb{E}_{g\sim\mu}\left[h(g(x)) \neq y\right]$, by Lemma 4.1, we have that $\ell^{\rho,0}_{\mathcal{G},\mu}(h,(x,y))$ enjoys the uniform convergence property with sample complexity $O\left(\frac{\frac{d}{\rho^2}\ln(\frac{1}{\rho\epsilon})+\ln(\frac{1}{\delta})}{\epsilon^2}\right)$. This completes the proof, as the conditions for Lemma 5.1 have been met, and therefore the learning rule $\mathcal{A}(S)$ enjoys the stated generalization guarantee with the specified sample complexity. □

D.3 Proof of Theorem 5.4

Proof. (of Theorem 5.4) Assume that there exists a subset $\mathcal{G}' \subset \mathcal{G}$, that is r-Nice with respect to $\mathcal{H}$. By Lemma 5.1, it is sufficient to find a perturbation set $\tilde{\mathcal{G}}$ such that (1) $\ell_{\mathcal{G}'}(h,(x,y)) \leq \ell_{\tilde{\mathcal{G}}}(h,(x,y)) \leq \ell_{\mathcal{G}}(h,(x,y))$ and (2) $\ell_{\tilde{\mathcal{G}}}(h,(x,y))$ enjoys the uniform convergence property with sample complexity $O\left(\frac{\text{VC}(\mathcal{H})\log(\mathcal{N}_r(\mathcal{G}'_{2r},d))\ln(\frac{1}{\epsilon})\ +\ln(\frac{1}{\delta})}{\epsilon^2}\right)$. Let $\tilde{\mathcal{G}} \subset \mathcal{G}$ be the minimal r-cover of $\mathcal{G}'_{2r}$ with cardinality $\mathcal{N}_r(\mathcal{G}'_{2r}, d)$. By Lemma 1.1 of Attias et al. [2021], the loss class $\mathcal{L}^{\tilde{\mathcal{G}}}_{\mathcal{H}}$ has VC dimension at most $O(\text{VC}(\mathcal{H})\log(|\tilde{\mathcal{G}}|)) = O(\text{VC}(\mathcal{H})\log(\mathcal{N}_r(\mathcal{G}'_{2r})))$, implying that $\ell_{\tilde{\mathcal{G}}}(h,(x,y))$ enjoys the uniform convergence property with the previously stated sample complexity $O\left(\frac{\text{VC}(\mathcal{H})\log(\mathcal{N}_r(\mathcal{G}'_{2r},d))\ln(\frac{1}{\epsilon})\ +\ln(\frac{1}{\delta})}{\epsilon^2}\right)$. Now, it remains to show that for our choice of $\tilde{\mathcal{G}}$, we have $\ell_{\mathcal{G}'}(h,(x,y)) \leq \ell_{\tilde{\mathcal{G}}}(h,(x,y)) \leq \ell_{\mathcal{G}}(h,(x,y))$. Since, $\tilde{\mathcal{G}} \subset \mathcal{G}$,the upperbound is trivial. Thus, we only focus on proving the lowerbound, $\ell_{\mathcal{G}'}(h,(x,y)) \leq \ell_{\tilde{\mathcal{G}}}(h,(x,y))$ for all $h \in \mathcal{H}$ and $(x,y) \in \mathcal{X} \times \mathcal{Y}$. Fix $h \in \mathcal{H}$ and $(x,y) \in \mathcal{X} \times \mathcal{Y}$. If $\ell_{\mathcal{G}'}(h,(x,y)) = 1$, then there exists a $g \in \mathcal{G}'$ such that $h(g(x)) \neq y$. Let g denote one such perturbation function. By the r-Niceness property of $\mathcal{G}'$ with respect to $\mathcal{H}$, there must exist $B_r(g^*)$ centered at some $g^* \in \mathcal{G}$ such that $g \in B_r(g^*)$ and $h(g(x)) = h(g'(x))$ for all $g' \in B_r(g^*)$. This implies that $h(g'(x)) \neq y$ for all $g' \in B_r(g^*)$. Furthermore, since $B_{2r}(g)$ is the union of all balls of radius r that contain g, we have that $B_r(g^*) \subset B_{2r}(g)$. From here, its not too hard to see that $B_r(g^*) \subset \mathcal{G}'_{2r}$ by definition. Finally, since $\tilde{\mathcal{G}}$ is an r-cover of $\mathcal{G}'_{2r}$, it must contain at least one function from $B_r(g^*)$. This completes the proof as we have shown that there exists a perturbation function $\hat{g} \in \tilde{\mathcal{G}}$ such that $h(\hat{g}(x)) \neq y$. □

D.4 ℓ_p balls are r-Nice perturbation sets for linear classifiers

In this section, we give a concrete example of a hypothesis class $\mathcal{H}$ and metric space of perturbation functions $(\mathcal{G}, d)$ for which there exists an r-nice perturbation subset $\mathcal{G}' \subset \mathcal{G}$. Let $\mathcal{X} = \mathbb{R}^q$ and fix $r \in \mathbb{R}_{\geq 0}$. For the hypothesis class, consider the set of homogeneous halfspaces, $\mathcal{H} = \{h_w | w \in \mathbb{R}^q\}$, where $h_w(x) = w^T x$. Let $\hat{\mathcal{G}} = \{g_\delta : \delta \in \mathbb{R}^q, ||\delta||_p \leq 3r\}$ where $g_\delta(x) = x + \delta$ for all $x \in \mathcal{X}$ and consider *any* perturbation set $\mathcal{G}$ such that $\mathcal{G} \supset \hat{\mathcal{G}}$. That is, $\hat{\mathcal{G}}(x) = \{g(x) : g \in \hat{\mathcal{G}}\}$ induces a ℓ_p ball of radius $3r$ around x. We will accordingly consider the distance metric $d(g_{\delta_1}, g_{\delta_2}) = \sup_{x\in\mathcal{X}} ||g_{\delta_1}(x) - g_{\delta_2}(x)||_p$. Restricted to the set $\hat{\mathcal{G}}$, this distance metric reduces to $d(g_{\delta_1}, g_{\delta_2}) = ||\delta_1 - \delta_2||_p = \ell_p(\delta_1, \delta_2)$ for $g_{\delta_1}, g_{\delta_2} \in \hat{\mathcal{G}}$. Finally, consider $\mathcal{G}' = \{g_\tau : \tau \in \mathbb{R}^q, ||\tau||_p \leq r\} \subset \hat{\mathcal{G}} \subset \mathcal{G}$ which induces an ℓ_p ball of radius r around x.

We will now show that $\mathcal{G}'$ is r-nice perturbation set with respect to $\mathcal{H}$. Let $x \in \mathcal{X}$, $h_w \in \mathcal{H}$, and $g_\tau \in \mathcal{G}'$. Let $c = h(g_\tau(x)) \in \{\pm 1\}$. Consider the function $g_{\tau + \frac{crw}{||w||_p}}$. By definition, we have that $g_\tau \in B_r(g_{\tau+\frac{crw}{||w||_p}}) \subset \hat{\mathcal{G}} \subset \mathcal{G}$. To see this, observe that $||\tau + \frac{crw}{||w||_p}||_p \leq 2r$ by the triangle inequality. Finally, it remains to show that for every $g' \in B_r(g_{\tau+\frac{crw}{||w||_p}}) = \{g_{\tau+\frac{crw}{||w||_p}+\kappa} | \kappa \in \mathbb{R}^d, ||\kappa||_p \leq r\}$, $h_w(g'(x)) = h_w(g_\tau(x)) = c$. Let $c = +1$ and consider the function $g'_{\tau+\frac{rw}{||w||_p}+\kappa} \in B_r(g_{\tau+\frac{rw}{||w||_p}})$.

Note that $w^T(x+\tau+\frac{rw}{||w||_p}+\kappa) = w^T(x+\tau)+r||w||_p+w^T\kappa$. By Cauchy-Schwartz, we can lower bound $w^T\kappa \geq -||w||_p||\kappa||_p \geq -r||w||_p$. Therefore, we have that $w^T(x+\tau+\frac{rw}{||w||_p}+\kappa) \geq w^T(x+\tau) > 0$, where the last inequality comes from the fact that $+1 = c = h_w(g_\tau) = \text{sign}(w^T(x+\tau))$. Therefore, $h(g'_{\tau+\frac{rw}{||w||_p}+\kappa}(x)) = \text{sign}(w^T(x+\tau+\frac{rw}{||w||_p}+\kappa)) = \text{sign}(w^T(x+\tau)) = h(g_\tau(x))$ as desired. A similar proof holds when $c = -1$. Therefore, we have shown that $\mathcal{G}'$ is a r-nice perturbation set with respect to $\mathcal{H}$.

We now can use Theorem 5.4 to provide sample complexity guarantees on Tolerantly Robust PAC Learning with $\mathcal{G}'$ and $\mathcal{G}$. The main quantity of interest is $\log(\mathcal{N}_r(\mathcal{G}'_{2r}, d))$. However, note that $\mathcal{G}'_{2r} = \hat{\mathcal{G}}$. Therefore, we just need to compute $\log(\mathcal{N}_r(\hat{\mathcal{G}}, d)) = \log(\mathcal{N}_r(\{g_\delta : \delta \in \mathbb{R}^q, ||\delta||_p \leq 3r\}, d))$. However, this is equal to $\log(\mathcal{N}_r(\{\delta \in \mathbb{R}^q : ||\delta||_p \leq 3r\}, \ell_p))$ using the ℓ_p distance metric since g_δ maps one-to-one to δ. Using standard arguments, $\log(\mathcal{N}_r(\{\delta \in \mathbb{R}^q : ||\delta||_p \leq 3r\}, \ell_p)) = \log(\mathcal{N}_{\frac{1}{3}}(\{\delta \in \mathbb{R}^q : ||\delta||_p \leq 1\}, \ell_p)) = O(q)$ (Bartlett [2013]). Thus, overall, $\mathcal{H}$ is tolerantly PAC learnable with respect to $(\mathcal{G}, \mathcal{G}')$ with sample complexity close to what one would require in the standard PAC setting.