VC Classes are Adversarially Robustly Learnable, but Only Improperly

Authors: Omar Montasser, Steve Hanneke, Nathan Srebro

Review by Anurag Singh, 03743384

August 4, 2021

1 Paper summary

The paper aims to investigate the learning of adversarially robust predictor. The claim made by the authors is that for any hypothesis class $\mathcal{H}$ which has a finite VC dimension, it is robustly PAC - learnable with an improper learning rule. The authors define the requirement of improper learning necessary, as they demonstrate by giving examples of hypothesis classes $\mathcal{H}$ with finite VC dimension that are not robustly PAC learnable with any proper learning rule.

1.1 Problem setup and Preliminaries

For an instance space $\mathcal{X}$ the label space $\mathcal{Y} = \{\pm 1\}$. Consider there exists an adversary $\mathcal{U} : \mathcal{X} \to 2^{\mathcal{X}}$ to protect against. Let $\mathcal{U}(x) \subseteq \mathcal{X}$ is the set of adversarial examples that can be chosen by the adversary at test time. For example, $\mathcal{U}(x)$ could be perturbations of distance at most γ w.r.t. some metric ρ: $\mathcal{U}(x) = \{z \in \mathcal{X} : ||x - z||_\rho \leq \gamma\}$. For a distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, observe m i.i.d. samples $S \sim \mathcal{D}^m$, the objective is to learn a predictor $\hat{h} : \mathcal{X} \to \mathcal{Y}$ having small robust risk defined as,

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

The common approach to adversarially robust learning is to pick a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ and learn through robust empirical risk minimization:

$$\hat{h} \in RERM_{\mathcal{H}}(S) := argmin_{h\in\mathcal{H}} \hat{R}_{\mathcal{U}}(h; S) \tag{2}$$

Where $\hat{R}_{\mathcal{U}}(\hat{h}; S) := \frac{1}{m} \Sigma_{(x,y)\in\mathcal{S}} \sup_{z\in\mathcal{U}(x)} \mathbf{1}[\hat{h}(z) \neq y]$ as studied in (7). Given a hypothesis class $H \subseteq \mathcal{Y}^X$, goal is to design a learning rule $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{Y}^{\mathcal{X}}$ such that for any distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, the rule $\mathcal{A}$ will find a predictor that competes with the best predictor $h^* \in \mathcal{H}$ in terms of the robust risk using a number of samples that is independent of the distribution $\mathcal{D}$. The following definitions formalize the notion of robust PAC learning in the realizable and agnostic settings as defined in (2).

Definition 1. Agnostic Robust PAC learning: For any $\epsilon, \delta \in (0, 1)$, the sample complexity of agnostic robust PAC learning of $\mathcal{H}$ with respect to adversary $\mathcal{U}$, is defined as the smallest $m \in \mathbb{N} \cup \{0\}$ for which there exists a learning rule $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{Y}^{\mathcal{X}}$ such that, for every data distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$, with probability at least $1 - \delta$ over $S \sim \mathcal{D}^m$,

$$R_{\mathcal{U}}(\mathcal{A}(S) = \mathcal{D}) \inf_{h\in\mathcal{H}} R_{\mathcal{U}}(h; D) + \epsilon \tag{3}$$

If no such m exists, sample complexity is infinite. We say that H is robustly PAC learnable in the agnostic setting with respect to adversary U if smallest possible m i.e. sample complexity is finite.

Definition 2. Realizable Robust PAC Learnability: For any $\epsilon, \delta \in (0,1)$, the sample complexity of realizable robust PAC learning of $\mathcal{H}$ with respect to adversary $\mathcal{U}$ is defined as the smallest $m \in \mathbb{N} \cup \{0\}$ for which there exists a learning rule $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{Y}^{\mathcal{X}}$ such that, for every data distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ where there exists a predictor $h \in \mathcal{H}$ with zero robust risk, $R_{\mathcal{U}}(h, \mathcal{D}) = 0$, with probability at least $1 - \delta$ over $S \sim \mathcal{D}^m$ then, $R_{\mathcal{U}}(\mathcal{A}(S), \mathcal{D}) \leq \epsilon$. If no such m exists, then sample complexity is infinite. We say that $\mathcal{H}$ is robustly PAC learnable in the realizable setting with respect to adversary $\mathcal{U}$ if sample complexity is finite.

Definition 3. Proper Learnability: $\mathcal{H}$ is ***properly*** robustly PAC learnable (in the agnostic or realizable setting) if it can be learned as in Definitions 1 or 2 using a learning rule $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{H}$ that always outputs a predictor in H. Learning using any learning rule $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{Y}^{\mathcal{X}}$, as in the definitions above is improper learning.

2 Main Proof Ideas

Theorem 1: There exists a hypothesis class $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ with $vc(\mathcal{H}) \leq 1$ and an adversary $\mathcal{U}$ such that $\mathcal{H}$ is not properly robustly PAC learnable with respect to $\mathcal{U}$ in the realizable setting.
The proof of above theorem requires two main lemmas in its ideas,

Lemma 2: Let $m \in \mathbb{N}$. Then, there exists $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ such that $vc(\mathcal{H}) \leq 1$ but $vc(\mathcal{L}^{\mathcal{U}}_{\mathcal{H}}) \geq m$

The prove begins by carefully constructing a hypothesis class by choosing $\{x_1, \ldots, x_m\}$ as points which have mutually disjoint perturbation sets, i.e. $\mathcal{U}(x_i) \cap \mathcal{U}(x_j) = \phi$. They start by building a set of points Z from which perturbations should not be picked and initialize it to $\{x_1, \ldots, x_m\}$. Now for each bit string $b \in \{0,1\}^m$, Z_b is made of perturbations of x_i s.t. $b_i = 1$. At the end $Z = Z \cup Z_b$ so that for next bit-string perturbations don't repeat. Then $h_b : \mathcal{X} \to \mathcal{Y}$ is defined as:

$$h_b = \begin{cases} +1 & x \notin Z_b \\ -1 & x \in Z_b \end{cases}$$

We can think of each mapping h_b as being characterized by a unique signature Z_b that indicates the points that it labels with -1. The hypothesis class is $\mathcal{H} = h_b : b \in \{0,1\}^m$. Now the proof that $vc(\mathcal{H}) \leq 1$ follows by taking $z_1, z_2 \in \mathcal{X}$ and considering cases that both belong to $\mathcal{X} - Z$, only one belongs and none of them do. It can be shown that in each of the three cases any classifier will not be able to produce (-1,-1) when none of them are in Z, (-1,+1) if only $z_2 \in Z$ and either (-1,-1) or (+1,+1) when both $z_1, z_2 \in Z$.
For proving $vc(\mathcal{L}^{\mathcal{U}}_{\mathcal{H}}) \geq m$ one can consider a set $\{(x_1, +), \ldots, (x_m, +)\}$ and show it can be shattered. Now we if pick any labeling $y \in \{0,1\}^m$ by construction of $\mathcal{H}$ we can find a h_b made by bit-string $b = y$. Then, for each $i \in [m]$, $\sup_{z \in \mathcal{U}(xi)} \mathbf{1}[h_b(z) \neq +1] = b_i = y_i$ and hence the set is shattered.

Lemma 3: Let $m \in \mathbb{N}$. Then, there exists $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ with $vc(\mathcal{H}) \leq 1$ such that for any proper learning rule $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{H}$,

- A distribution $\mathcal{D}$ over $\mathcal{X} \times \mathcal{Y}$ and a predictor $h^* \in \mathcal{H}$ where $R_{\mathcal{U}}(h^*; \mathcal{D}) = 0$.
- With probability at least $1/7$ over $S \sim \mathcal{D}^m$, $R_{\mathcal{U}}(\mathcal{A}(S); \mathcal{D}) > 1/8$.

This proof follows standard lower bound techniques that use the probabilistic method from Chapter 5 of (1). According to Lemma 2, for $\{x_1 \dots x_{3m}\}$ construct $\mathcal{H}_0$. By construction then $\mathcal{L}^{\mathcal{U}}_{\mathcal{H}_0}$ can shatter $C = \{(x_1, +) \dots (x_{3m}, +)\}$ The idea is to construct a family of distributions that are supported on 2m points of C only and keeping only $\mathcal{H} \subseteq \mathcal{H}_0$ that has classifiers robustly correct on $2m$ examples. This would make rule $\mathcal{A}$ to choose which points it can afford to be not correctly robust on. If rule $\mathcal{A}$ observes only m points, it can't do anything better than guessing which of the remaining $2m$ points of C are actually included in the support of the distribution.

For proof for theorem 1 we can construct sequences of subsets of $3m$ distinct points from $\mathcal{X}$ as X_m with no intersection in perturbation sets. We ensure that predictors in $\mathcal{H}_m$ are non-robust on the points in $X_m^{'}$ for all $m^{'} \neq m$ as $h_b \in \mathcal{H}_m$,

$$h_b = \begin{cases} -1 & x \in Z_b \text{ or } x \in X_{m'},\, m' \neq m \\ +1 & otherwise \end{cases}$$

Then, $\mathcal{H} = \bigcup_{m=1}^{\inf} \mathcal{H}_m$ and then using lemma 2 we show VC dimension of $VC(\mathcal{H}) \leq 1$. Then we can apply lemma 3 over a distribution $\mathcal{D}$ of $X_m \times \mathcal{Y}$. The robust risk for a $h^* \in \mathcal{H}_m$ is 0. This works because classifiers from classes $\mathcal{H}_{m'}$ $m' \neq m$ are non robust on X_m. Thus, rule $\mathcal{A}$ will do worse if it picks predictors from these classes. Which shows that the sample complexity to learn proper robust PAC learnable $\mathcal{H}$ is infinite.

As opposed to previous theorem that shows that finite VC dimension is not sufficient for robust PAC learning, the rest of theorems discussed after this in the paper try to show that finite VC dimension is sufficient for robust PAC learning both in realizable PAC learning setting and in agnostic PAC learning setting. We do this by providing a bound on their sample complexity in each case. We shall discuss the realizable setting and the agnostic setting follows very similar main ideas.

Theorem 4: For any $\mathcal{H}$ and $\mathcal{U}$,$\forall \epsilon, \delta \in (0, 1/2)$,

$$\mathcal{M}_{RE}(, \delta, \mathcal{H}, \mathcal{H}) = \mathcal{O}\left(vc(\mathcal{H})vc^*(\mathcal{H})\frac{1}{\epsilon}log(\frac{vc(\mathcal{H})vc^*(\mathcal{H})}{\epsilon}) + \frac{1}{\epsilon}log(\frac{1}{\delta}) \right)$$

Where $vc^*(\mathcal{H})$ is the dual VC dimension. Based on result $vc^*(\mathcal{H}) < 2^{vc(\mathcal{H})+1}$ (4) Corollary 5 immediately follows. The proof makes use of sample compression arguments taking inspiration from work in (3). Modifications made in this proof forces the compression scheme to also have zero empirical robust loss. Fix a deterministic function $RERM_{\mathcal{H}}$ mapping any labeled data set to a classifier in $\mathcal{H}$ robustly consistent with the labels in the data set, if a robustly consistent classifier exists. For a training sample set S, which is sampled iid from a robust realizable distribution, $R^{\mathcal{U}}(RERM_{\mathcal{H}}(S); S) = 0$. Then they inflate the training set S to potentially infinite set $S_{\mathcal{U}}$ containing all the possible perturbations. Then this set is discretized to denote by $\hat{S}_{\mathcal{U}} \subset S_{\mathcal{U}}$ which includes exactly one $(x, y) \in S_{\mathcal{U}}$ for each distinct classification $\{g_{(x,y)}(h)\}_{h \in \hat{\mathcal{H}}}$ of $\hat{\mathcal{H}}$ realized by functions $g_{(x,y)} \in \mathcal{G}$. Where $\hat{\mathcal{H}}$ is set of classifiers selected by robust empirical risk minimization of n sample subset of S. In other words $\hat{\mathcal{H}} = \{RERM_{\mathcal{H}}(L), L \subseteq S \text{ st. } |L| = n\}$ and $\mathcal{G}$ is the dual space of set of functions $g_{(x,y)} : mathcalH \to \{0, 1\}$ defined as $g_{(x,y)}(h) = 1[h(x) = y]$, for each $h \in \mathcal{H}$ and each $(x, y) \in S_{\mathcal{U}}$. By application of Sauer's lemma we can bound the size of $|\hat{S}_{\mathcal{U}}|$ by $(e^2 m/vc(\mathcal{H}))^{vc(\mathcal{H})vc^*(\mathcal{H})}$ for $m > 2vc(\mathcal{H})$. By this construction the majority vote of any subset of classifiers in $\hat{\mathcal{H}}$ for each point $(x, y) \in S^{\mathcal{U}}$ is greater than 1/2. In other words, $\sum_{t=1}^{T} 1[h_t(x) = y] > 1/2$. Then same will hold try for each $(x, y) \in \hat{S}^{\mathcal{U}}$ which means $\hat{R}_{\mathcal{U}}(Majority(h_1, \dots, h_T); S) = 0$.Which leaves us with the task of finding such a set of h_t functions. By the choice of n and construction of $hatS^{\mathcal{U}}$ we can find for any distribution D over $hatS^{\mathcal{U}}$, there exists $h_D \in \hat{\mathcal{H}}$ with

$\hat{er}(h_D, D) < 1/3$. Now we can run modified version of α boost on $\hat{S}^{\mathcal{U}}$ with $RERM_{\mathcal{H}}$ as a weak learner i.e.h_D as a weak hypothesis in a boosting algorithm. Using proof in (5), for an appropriate a-priori choice of in the α-Boost algorithm, and running the algorithm for rounds to give hypotheses $\hat{h}_1\hat{h}_T \in \hat{\mathcal{H}}$ s.t.

$$\forall(x, y) \in \hat{S}_{\mathcal{U}}; \frac{1}{T}\sum_{i=1}^{T} 1[h_i(x) = y] \geq 5/9$$

Using the above observation we can say for $\hat{h} = Majority(\hat{h}_1, \ldots, \hat{h}_T)$ satisfies $\hat{R}_{\mathcal{U}}(\hat{h}, S) = 0$. And thinking $\hat{h}$ as a order-dependent reconstruction function we can say following about its compression size,$nT = \mathcal{O}(vc(H)log(|\mathcal{S}_{\mathcal{U}}|) = O(vc(\mathcal{H})^2 vc(\mathcal{H})log(m/vc(\mathcal{H})))$. Using Lemma 11 and taking care of condition on m we can rewrite above as, with probability at least $1 - \delta$,

$$RU(h;P) \leq \mathcal{O}(vc(\mathcal{H})^2 vc^*(\mathcal{H})\frac{1}{m}log\left(\frac{m}{vc(\mathcal{H})}\right)log(m) + \frac{1}{m}log(1/\delta))$$

With further application of technique from (6) the bound can be further reduced.

3 Review

Novelty: This paper provides two theoretical analyses of generalization for robust PAC learning. The results are very significant in my knowledge since they try to understand the generalization for adversarially robust learning objective which has wide applications (7). More precisely the contributions of the paper are that the authors show that there exists an adversary $\mathcal{U}$ and a hypothesis class $\mathcal{H}$ with finite VC dimension that cannot be robustly PAC learned with any proper learning rule (including RERM). They also show that for any VC class $\mathcal{H}$ and any adversary $\mathcal{U}$, using an improper learning rule, $\mathcal{H}$ is agnostically robustly PAC learnable.

Significance: Their results indicate that are for some hypothesis classes there are large gaps between what can be done with proper vs. improper PAC learning rules. This means that when studying a particular class, such as classes corresponding to neural networks, one should consider the possibility that there might be such a gap and that improper learning might be necessary. However, it is still an open question to study and establish if such gaps actually exist for specific interesting neural net classes (e.g., functions represened by a specific architecture, like resnet). Assuming that such gaps exist, one of the main takeaways of the paper is that for the task of adversarially robust learning, improper pac learning rules should be considered it would be interesting to see how improper pac learning is incorporated in neural network training/optimization framework.

Clarity: The paper is technically sound and well written with proofs mostly easy to follow. There are some parts where proves could be more elaborate in the Lemmas, some comments on them are made in minor comments.

Comments: The paper assumes certain aspects about the adversarial robust learning framework, that the $\mathcal{U}$ must be in the same instance space, which may not be the case for all attacks(8). Also, it may not be necessary that the perturbations do exist for all the inputs in the set with distance γ and it can be empty or possibly finite, which would mean that construction of such special hypothesis classes for proves would not be possible. As for some minor comments, I believe there few statements in the proof of Lemma 3 are hard to follow, particularly how being robustly correct on $2m$ examples leads to given set expression of $\mathcal{H}$ i.e. the following statement, *We will only keep a subset $\mathcal{H}$ of $\mathcal{H}_0$ that includes classifiers that are robustly correct only on subsets of size $2m$, i.e. $\mathcal{H} = \{h_b \in \mathcal{H}_0 : \sum_{i=1}^{3m} b_i = m\}$.*

References

[1] Shalev-Shwartz, Shai, and Shai Ben-David. *Understanding machine learning: From theory to algorithms.* Cambridge university press, 2014.

[2] Montasser, Omar, Steve Hanneke, and Nathan Srebro. *VC classes are adversarially robustly learnable, but only improperly.* Conference on Learning Theory. PMLR, 2019.

[3] S. Hanneke, A. Kontorovich, and M. Sadigurschi. *Sample compression for real-valued learners.* In Proceedings of the 30th International Conference on Algorithmic Learning Theory, 2019

[4] P. Assouad. *Densite et dimension.* Annales de l'Institut Fourier (Grenoble), 33(3):233–282, 1983

[5] R. E. Schapire and Y. Freund. *Boosting.* Adaptive Computation and Machine Learning. MIT Press, Cambridge, MA, 2012

[6] S. Moran and A. Yehudayoff. *Sample compression schemes for VC classes.* Journal of the ACM, 63(3):21:1–21:10, 2016.

[7] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. *Explaining and harnessing adversarial examples.* arXiv preprint arXiv:1412.6572, 2014.

[8] Zeng, Xiaohui, et al textitAdversarial Attacks Beyond the Image Space CVPR 2019