VC Classes are Adversarially Robustly Learnable, but Only Improperly

Authors: Omar Montasser and Steve Hanneke and Nathan Srebro

Anurag Singh

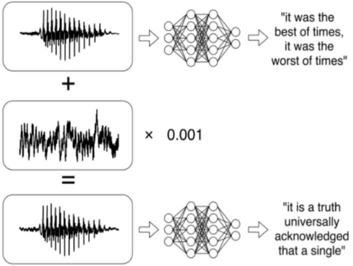
Technical University of Munich

Adversarial Attacks in Visual Computing

Figure: Attacks on Visual Computing systems for multiple tasks. [1] [2]

[1] Brown et al. "Adversarial patch." arXiv (2017).
[2] Goodfellow et al. Explaining and harnessing adversarial examples." arXiv (2014).

Adversarial Attacks in Speech & NLP



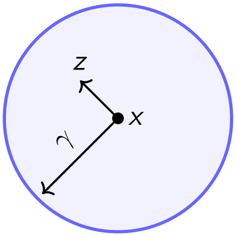
Original Input	Connoisseurs of Chinese film will be pleased to discover that Tian's meticulous talent has not withered during his enforced hiatus.	Prediction: <u>Positive (77%)</u>
Adversarial example [Visually similar]	<u>Aonnoisseurs</u> of Chinese film will be pleased to discover that Tian's meticulous talent has not withered during his enforced hiatus.	Prediction: <u>Negative (52%)</u>
Adversarial example [Semantically similar]	Connoisseurs of Chinese <u>footage</u> will be pleased to discover that Tian's meticulous talent has not withered during his enforced hiatus.	Prediction: <u>Negative (54%)</u>

Figure: Attacks on Speech and NLP tasks. [3] [4]

[3] Carlini et al. Audio Adversarial Examples: Targeted Attacks on Speech-to-Text
[4] Jin et al. Is BERT Really Robust? A Strong Baseline for Natural Language Attack on Text Classification

Problem setup

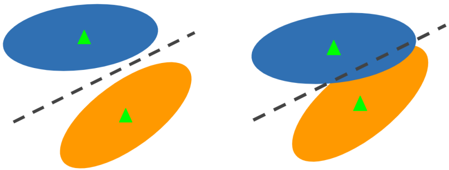
- Instance space $\mathcal{X}$ and label space $\mathcal{Y} \in \{0, 1\}$.
- An adversary $\mathcal{U} : \mathcal{X} \to 2^{\mathcal{X}}$
- There is following conditions on the adversary $\mathcal{U}$ that perturbations can be distance at most γ w.r.t metric ρ.



$$\mathcal{U} = \{z \in \mathcal{X} : \ ||x - z||_\rho \leq \gamma\}$$

Robust Risk

- The robust risk is defined as $R_{\mathcal{U}}(h, \mathcal{D}) = \mathbb{P}_{(x,y)\sim\mathcal{D}}[\exists z \in \mathcal{U}(x) \text{ s.t. } h(z) \neq y]$

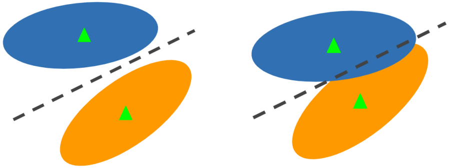


(a) Robust (b) Non Robust

Figure: Classification boundaries for robust and non robust classifiers

Robust Risk

- The robust risk is defined as $R_{\mathcal{U}}(h, \mathcal{D}) = \mathbb{P}_{(x,y)\sim\mathcal{D}}[\exists z \in \mathcal{U}(x) \text{ s.t. } h(z) \neq y]$
- Equivalently, $R_{\mathcal{U}}(h, \mathcal{D}) = \mathbb{E}_{(x,y)\sim\mathcal{D}}\left[\sup_{z\sim\mathcal{U}(x)} 1[h(z) \neq y]\right]$



(a) Robust (b) Non Robust

Figure: Classification boundaries for robust and non robust classifiers

Robust PAC Learning - Realizable setting

- We can say that $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is *Realizable* **Robust-PAC Learnable** with respect to $\mathcal{U}$ if there exists a predictor $h^* \in \mathcal{H}$ with zero robust risk i.e. $R_{\mathcal{U}}(h^*, \mathcal{D}) = 0$ and $\forall\ \epsilon, \delta \in (0,1)\ \exists\ m(\epsilon, \delta)$ and a learning rule $\mathcal{A}$ for all distributions $\mathcal{D}$ st. st. following holds with probability $1 - \delta$

$$\mathbb{E}_{S \sim \mathcal{D}^m}[R_{\mathcal{U}}(\mathcal{A}(S), \mathcal{D})] \leq \epsilon$$

Robust PAC Learning - Agnostic setting

- We can say that $\mathcal{H} \subseteq \mathcal{Y}^{\mathcal{X}}$ is *Agnostically* **Robust-PAC Learnable** with respect to $\mathcal{U}$ if $\forall\ \epsilon, \delta \in (0,1)\ \exists\ m(\epsilon, \delta)$ and a learning rule $\mathcal{A}$ for all distributions $\mathcal{D}$ over $(\mathcal{X} \times \mathcal{Y})$ st. following holds with probability $1-\delta$

$$\mathbb{E}_{S\sim\mathcal{D}^m}[R_{\mathcal{U}}(\mathcal{A}(S), \mathcal{D})] \leq \inf_{h\in\mathcal{H}} R_{\mathcal{U}}(h, \mathcal{D}) + \epsilon$$

Proper and Improper Learning

- We can say that $\mathcal{H}$ is **properly** robustly PAC learnable (in the agnostic or realizable setting) if it can be learned using a learning rule $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{H}$ that always outputs a predictor in $\mathcal{H}$. Learning using any learning rule $\mathcal{A} : (\mathcal{X} \times \mathcal{Y})^* \to \mathcal{Y}^{\mathcal{X}}$, is improper learning.

Population Risk Estimation

How can we ensure that we have a small population risk $R_{\mathcal{U}}(h, \mathcal{D})$?

$$\hat{h} \in RERM_{\mathcal{H}}(S) := argmin_{h \in \mathcal{H}} \hat{R}_{\mathcal{U}}(h; S) \quad (1)$$

Where $\hat{R}_{\mathcal{U}}(\hat{h}; S) := \frac{1}{m} \Sigma_{(x,y) \in \mathcal{S}} \sup_{z \in \mathcal{U}(x)} 1[\hat{h}(z) \neq y]$

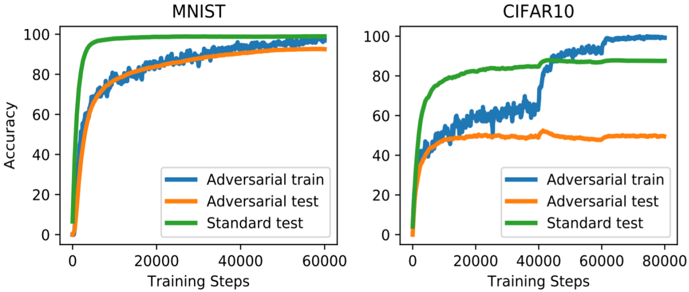


Figure: Classification Accuracy on MNIST and CIFAR10 [5]

[5] Schmidt, Ludwig, et al. Adversarially robust generalization requires more data.

Sometimes there are no proper robust learners

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.

Sometimes there are no proper robust learners

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.

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 $vc(\mathcal{L}_{\mathcal{H}}^{\mathcal{U}}) \geq m$

Sometimes there are no proper robust learners

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.

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 $vc(\mathcal{L}_{\mathcal{H}}^{\mathcal{U}}) \geq m$.

$$\mathcal{L}_{\mathcal{H}}^{\mathcal{U}} = \left\{ (x, y) \to \sup_{z \sim \mathcal{U}(x)} 1[\hat{h}(z) \neq y] \;:\; h \in \mathcal{H} \right\}$$

If $vc(\mathcal{L}_{\mathcal{H}}^{\mathcal{U}}) < \infty$ then $\mathcal{H}$ is robustly PAC learnable.

Sometimes there are no proper robust learners

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.

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 $S \sim \mathcal{D}^m$, $R_{\mathcal{U}}(\mathcal{A}(S); \mathcal{D}) > 1/8$.

Intuition for Proof of Theorem 1

We aim to show that even for hypothesis classes with finite VC dimension, indeed even if $vc(H) = 1$, robust PAC learning might not be possible using *any* proper learning rule.

- Create infinite sequence of sets $(X_m)_{m \in \mathbb{N}}$ from $\mathcal{X}$.
- Construct hypothesis class $\mathcal{H}_m$ st. $\mathcal{H}_m$ are non-robust on the points in $X_{m'}$ for all $m' \neq m$
- Consider $\mathcal{H} = \bigcup_{m=1}^{\infty} \mathcal{H}_m$
- Show $vc(\mathcal{H}) \leq 1$ using Lemma 2
- Use Lemma 3 to show $\mathcal{H}$ is not robust PAC learnable.

Improper Robust PAC Learning is possible

Finite VC Dimension is Sufficient for (Improper) Robust PAC Learning.

- if H is learnable, it is also robustly learnable
- improper learning is necessary for some hypothesis classes.

Improper Robust PAC Learning is possible

Finite VC Dimension is Sufficient for (Improper) Robust PAC Learning.

- if H is learnable, it is also robustly learnable
- improper learning is necessary for some hypothesis classes.

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. Can be further simplified using $vc^*(\mathcal{H}) < 2^{vc(\mathcal{H})+1}$.

Intuition for Proof of Theorem 4

In the realizable setting

- Inflate the training set to a (possibly infinite) set $S_{\mathcal{U}}$ that includes all permutations.
- Discretize the set $S_{\mathcal{U}}$ to $\hat{S}_{\mathcal{U}}$
- Run a modified version of α-Boost on $\hat{S}_{\mathcal{U}}$ with $RERM_{\mathcal{H}}$ as a weak leaner.
- Use robust generalization guarantee through sample compression[6]
- Extend to agnostic case via[7]

[6] Sample compression for real-valued learners. In COLT 2019

[7] Supervised learning through the lens of compression. NIPS 2016

Implications

- There exists an adversary $\mathcal{U}$ and hypothesis class $\mathcal{H}$ with $vc(\mathcal{H}) = 1$ s.t.
  - RERM cannot robustly PAC learn $\mathcal{H}$ even for realizable case.
  - No proper learning rule can robustly PAC learn $\mathcal{H}$ even in realizable case.
- For any hypothesis class $\mathcal{H}$ and any adversary $\mathcal{U}$, $\mathcal{H}$ is agnostically robustly PAC learnable with an *improper* learning rule.

Open Questions

ERM	RERM
Proper Learning always possible	Improper learning is sometimes needed
Finite VC dim is necessary and sufficient	Finite VC dim is sufficient but not necessary.
Sample complexity $O(\frac{vc(\mathcal{H})}{\epsilon^2}))$	Sample complexity $O(\frac{2^{vc(\mathcal{H})}}{\epsilon})$

Table: Differences in standard loss and robust empirical risk

- What are necessary and sufficient conditions for robust PAC learning ?
- What is optimal sample complexity for robust PAC learning?

Takeaway

Start considering improper learning for adversarially robust learning !

Review

Pros:

- Tackles an interesting area with real applications.
- Results are significant and novel
- Direct applications of results in training of models.

Cons:

- Claims may generalize to adversarial attacks in NLP.
- No empirical analysis
- Bound on sample complexity is non optimal.

Thank You for your attention