A law of adversarial risk, interpolation, and label noise

Daniel Paleka*[1], Amartya Sanyal*[1,2]
*Equal contribution; [1]Department of Computer Science, ETH Zürich; [2]ETH AI Center

ETH zürich

Overview

We study **adversarial robustness** in interpolating classifiers in presence of **label noise**.

Label noise is ubiquitous in real world datasets e.g. CIFAR-10.

Q: Does fitting label noise hurt adversarial accuracy?

Our Contribution Improve upon existing work [1]: *Give a sharper characterisation of how interpolating label noise causes large adversarial risk for sufficient sample size.*

Mathematical notation and setting

Data Distribution μ on $\mathbb{R}^d$ with norm $\|\cdot\|$.
Ground truth binary classifier $f^* : \mathbb{R}^d \to \{0,1\}$.
Adversarial risk of a classifier f with regards to balls of radius ρ

$$\mathcal{R}_{\mathrm{Adv},\rho}(f,\mu) = \mathbb{P}_{\mathbf{x}\sim\mu}\left[\exists \mathbf{z} \in B_\rho(\mathbf{x}),\ f^*(\mathbf{x}) \neq f(\mathbf{z})\right].$$

where $z \in B_\rho(x)$ means $\|\mathrm{z} - \mathrm{x}\| \leq \rho$.

Setting: Adversarial risk in interpolation regime under uniform label noise

Dataset of size m sampled uniformly from μ.
Label the dataset with f and flip each label with probability η.
Classifier f obtains zero training error on this dataset.

Q: Can we lower bound $\mathcal{R}_{\mathrm{Adv},\rho}(f,\mu)$?

References and QR

[1] Amartya Sanyal, Puneet K. Dokania, Varun Kanade, and Philip Torr. *How benign is benign overfitting?* ICLR (2021)

Main Result

Theorem (Informal): *Any classifier that interpolates training data with uniform label noise, has large adversarial risk when the training set size m is large.*
Formally, with label noise η, we have

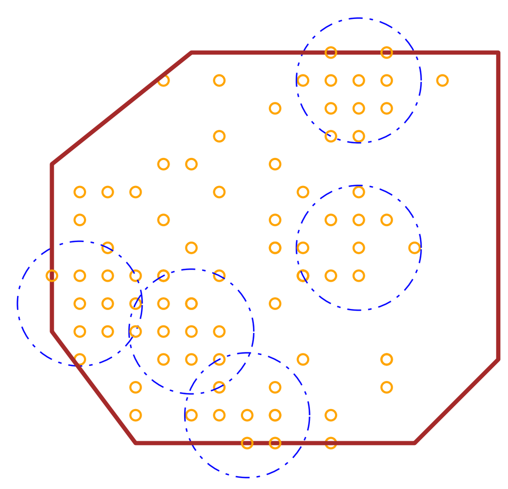
$$\mathcal{R}_{\mathrm{Adv},\rho}(f,\mu) \geq \mathrm{const.} > 0$$

for f trained on $\mathbf{x}_1, \ldots, \mathbf{x}_m \sim \mu$.

Let N be the covering number of $\mathrm{Support}\,(\mu)$.
Our result holds for dataset set size $m \gtrsim \frac{N \log N}{\eta}$.

Proof Sketch:
Observation 1: If a $\|\cdot\|$-ball of radius $\frac{\rho}{2}$ contains a noisy point, the entire ball is vulnerable.
Observation 2: The expected number of noisy training points is ηm; however a priori those could be anywhere in $\mathrm{Support}\,(\mu)$.

Key lemma: Can always find a set of $\|\cdot\|$-balls of radius ρ covering a significant portion of μ, with each of the balls having a large enough density of μ.



When m is large enough, with high probability, each chosen $\|\cdot\|$-ball will contain noisy labels, resulting in adversarial risk.

Tightness in sample size

Theorem (Informal) *For arbitrary interpolators f on arbitrary distributions μ on $\mathbb{R}^d$, no guarantees possible unless m is exponential in d.*

Proof Sketch: Let f^* be the threshold classifier $\mathbb{I}\left\{x_1 > \frac{1}{2}\right\}$. Sample $\mathbf{z}_1, \ldots, \mathbf{z}_m$ from the unit sphere $\mathbb{S}^{d-1} \subseteq \mathbb{R}^d$ with label noise η.
For $m \leq \lfloor 1.01^d \rfloor$, there is an interpolator with adversarial risk $o(1)$.

Q: What about empirical results regarding required sample size ?

Empirically, much smaller sample size required for large adversarial risk.

Inductive bias affects adversarial risk

Theorem (Informal) For any ρ, there exists model classes $\mathcal{H}, \mathcal{C}$ such that for $m = \Theta\left(\frac{1}{\eta}\right)$

All interpolators $h \in \mathcal{H}$ suffer constant $\mathcal{R}_{\mathrm{Adv},\rho}(h,\mu)$.
Exists interpolators $c \in \mathcal{C}$ with vanishing $\mathcal{R}_{\mathrm{Adv},\rho}(c,\mu)$.

Illustration of $\mathcal{C}$ below.