A law of Adversarial Risk, Interpolation, and Label Noise

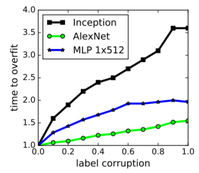
Daniel Paleka, Amartya Sanyal

ETH Zürich, Max-Planck Institute for Intelligent Systems, Tübingen

Interpolation & Adversarial Robustness

Two well known properties of neural networks

- **Interpolation**: Neural networks **interpolate label noise**, if trained long enough.[1]



- **Adversarial Risk**: Neural Networks known to suffer from high Adversarial Risk.

Adversarial Risk$(h; \epsilon, \|\cdot\|) = \mathbb{P}_{x,y\sim D}\left[\text{exists } \delta \text{ s.t } h(x+\delta) \neq y \text{ and } \|\delta\| \leq \epsilon\right]$

Goal: Understanding the impact of Interpolating label noise on adversarial robustness.

Theorem 1: Uniform Label Noise & Robustness

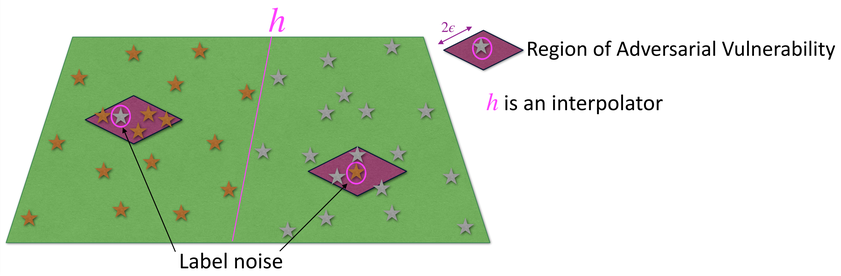
For

- Any distribution μ on $\mathbb{R}^d$, **and any compact** $C \subset \mathbb{R}^d$
- An i.i.d. dataset of "sufficiently large" size $\gtrsim \dfrac{N(C, \epsilon, \|\cdot\|)}{\mu(C)\eta}$
- With uniform label noise,

We prove a **lower bound** on adversarial risk for **ANY** interpolator

$$\text{Adversarial Risk}(h) = \Omega(1) \gtrsim \mu(C)$$

Illustrative Proof Sketch



Remark: Required dataset size may be exponential in dimension!

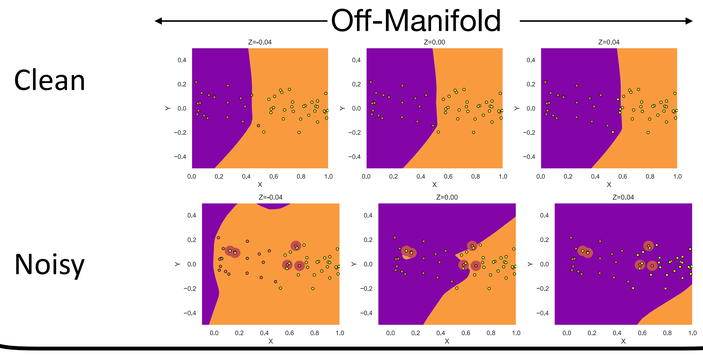
Theorem 2: Tightness of result

Question: Is the required dataset size in Theorem 1 tight ?

Without assumptions on interpolator, Theorem 1 is tight.

Inductive biases of interpolators

Interpolating label noise in Neural Networks increases vulnerability to off-manifold adversarial attacks



Theorem 3: Poison & Uniform Noise Noise

Uniform noise **randomly** flips k point(s).

Poisoner **chooses** ℓ point(s) to flip.

For $k = \mathcal{O}(\ell \log \ell)$, adversarial error for uniform label noise and poisoner are nearly equal.

Proof Idea: A poison is only harmful if it lies in a region of high density, where uniform noise will sample from anyway.

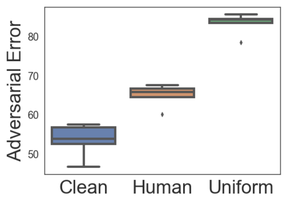
Main Takeaways

1. Interpolating Label Noise **provably increases Adversarial Risk** even when it does not increase test error (Benign Overfitting)
2. Uniform label noise is **nearly as bad as Poisoning** for increasing adversarial error.
3. Label noise from human annotation concentrates in the tail of data and is more benign.

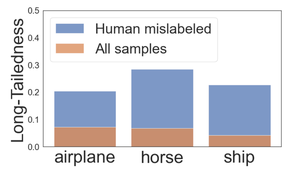
Label Noise in Human Annotation

Experiment using CIFAR10-N dataset

Observation: Interpolating human label noise is more benign than uniform noise.



Reason: Human label noise concentrates on the long-tail of data.



References & QR code for full paper

[1] Zhang, Chiyuan, et al. "Understanding deep learning (still) requires rethinking generalization." *Communications of the ACM* 64.3 (2021): 107-115

[2] Sanyal, Amartya, et al. "How benign is benign overfitting?." International Conference on Learning Representations (2021)