Empirically Measuring Concentration: Fundamental Limits on Intrinsic Robustness

UNIVERSITY of VIRGINIA

Saeed Mahloujifar* Xiao Zhang* Mohammad Mahmoody David Evans

Adversarial Examples

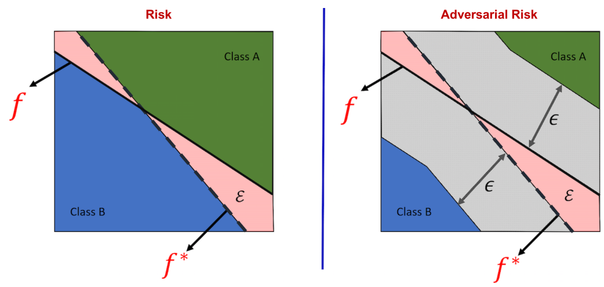
Let $(\mathcal{X}, \mu)$ be the underlying probability space and f^* be the ground-truth classifier.

- **Risk:**

$$\mathrm{Risk}(f, f^*) = \Pr_{\boldsymbol{x} \sim \mu}\left[f(\boldsymbol{x}) \neq f^*(\boldsymbol{x})\right].$$

- **Adversarial Risk w.r.t ϵ perturbations:**

$$\mathrm{AdvRisk}_\epsilon(f, f^*) = \Pr_{\boldsymbol{x} \sim \mu}\left[\exists\, \boldsymbol{x}' \in \mathrm{Ball}(\boldsymbol{x}, \epsilon) \text{ s.t. } f(\boldsymbol{x}') \neq f^*(\boldsymbol{x}')\right].$$



Main Question

What is the minimum possible Adversarial Risk, given that Risk is at least α?

$$\min_{\mathcal{E} \subseteq \mathcal{X}} \; \mu(\mathcal{E}_\epsilon) \quad \text{such that } \mu(\mathcal{E}) \geq \alpha.$$

Concentration of measure of nice distributions gives lower bound on adversarial risk:

- Spheres under ℓ_2 (Gilmer et al., 2018)
- Gaussian under ℓ_2 (Fawzi et al., 2018)
- Any product distribution under ℓ_0 (Mahloujifar et al., 2018)

Can we estimate concentration of measure for real world distributions (e.g. MNIST)?

Empirical Framework to Measure Concentration

- **Challenge 1:** We do not know the PDF of the distribution.
  Our solution: replace the actual distribution μ with empirical distribution $\widehat{\mu}$ based on a set of i.i.d. samples $\mathcal{S}$

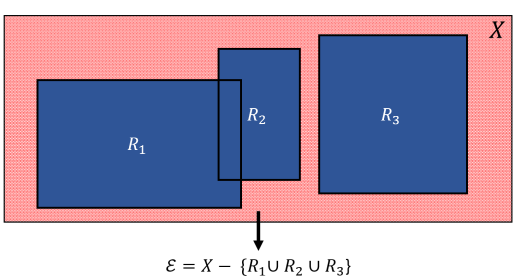
$$\widehat{\mu}(\mathcal{A}) \equiv \sum_{\boldsymbol{x} \in \mathcal{S}} \mathbb{1}_{\mathcal{A}}(\boldsymbol{x}) / |\mathcal{S}|.$$

- **Challenge 2:** We cannot search through all possible subsets.
  Our solution: limit the search space to a carefully chosen collection of subsets $\mathcal{G}$.
- **Remaining task:** solve the following optimization problem:

$$\underset{\mathcal{E} \subseteq \mathcal{G}}{\text{minimize}} \;\; \widehat{\mu}(\mathcal{E}_\epsilon) \quad \text{such that } \widehat{\mu}(\mathcal{E}) \geq \alpha.$$

Theoretical Results for ℓ_∞

Let $\mathcal{G}_T$ be the collection of subsets specified by complement of union of T hyperrectangles.



Let $\widehat{\mu}_T$ be the empirical distribution based on a i.i.d. dataset of size T^4. Define

$$c = \min_{\mathcal{E} \subseteq \mathcal{X}} \; \mu(\mathcal{E}_\epsilon) \quad \text{such that } \mu(\mathcal{E}) \geq \alpha.$$

Also define

$$c_T = \min_{\mathcal{E} \in \mathcal{G}_T} \; \widehat{\mu}_T(\mathcal{E}_\epsilon) \quad \text{such that } \widehat{\mu}_T(\mathcal{E}) \geq \alpha.$$

Main Theorem: With probability 1 over the randomness of training data we have

$$\lim_{T \to \infty} c_T = c.$$

Finding Robust Error Region for ℓ_∞

- Sort all the training data using ℓ_1 distance to the k-th nearest neighbors.
- Perform kmeans clustering on the top-q densest images.
- Obtain T hyperrectangular image clusters and expand each of them by ϵ in ℓ_∞.
- Treat the complement of union of these hyperrectangles as our error region.

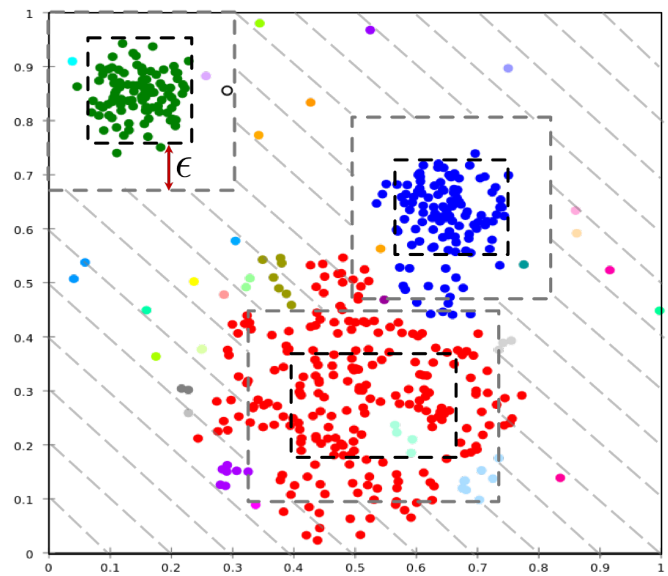
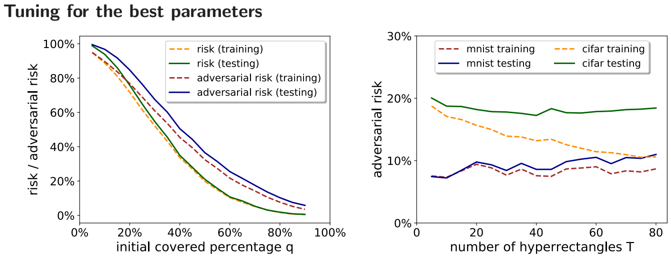


Illustration of the proposed method for finding robust error region under ℓ_∞

Experimental for ℓ_∞

- **Tuning for the best parameters**



Tuning results of our method on MNIST and CIFAR-10 datasets

- **Main experimental results**

Table: Summary of the main results using our empirical method under ℓ_∞ perturbations

Dataset	α	ϵ	T	q	Risk	Adversarial Risk
MNIST	0.01	0.1	5	0.662	$1.23\% \pm 0.12\%$	$3.64\% \pm 0.30\%$
		0.2	10	0.660	$1.11\% \pm 0.10\%$	$5.89\% \pm 0.44\%$
		0.3	10	0.629	$1.15\% \pm 0.13\%$	$7.24\% \pm 0.38\%$
		0.4	10	0.598	$1.21\% \pm 0.09\%$	$9.92\% \pm 0.60\%$
CIFAR-10	0.05	2/255	10	0.680	$5.72\% \pm 0.25\%$	$8.13\% \pm 0.26\%$
		4/255	20	0.688	$6.05\% \pm 0.40\%$	$13.66\% \pm 0.33\%$
		8/255	40	0.734	$5.94\% \pm 0.34\%$	$18.13\% \pm 0.30\%$
		16/255	75	0.719	$5.28\% \pm 0.23\%$	$28.83\% \pm 0.46\%$
SVHN	0.05	0.01	10	0.812	$8.83\% \pm 0.30\%$	$10.17\% \pm 0.29\%$
		0.02	10	0.773	$8.86\% \pm 0.20\%$	$12.46\% \pm 0.15\%$
		0.03	10	0.750	$8.55\% \pm 0.22\%$	$13.82\% \pm 0.25\%$

- **Implications of our experiments**
  - Provide examples of rather robust error regions for real image datasets.
  - Suggest the concentration of measure phenomenon is not the sole reason behind vulnerability of the existing classifiers to adversarial examples.
  - Suggest the impossibility results, such as Gilmer et al. (2018) and Mahloujifar et al. (2018), should not make the community hopeless in finding more robust image classifiers.

{saeed, xiao, mohammad, evans}@virginia.edu