Empirically Measuring Concentration: Fundamental Limits on Intrinsic Robustness

University of Virginia

Saeed Mahloujifar* Xiao Zhang* Mohammad Mahmoody David Evans

NeurIPS 2019

Adversarial Risk

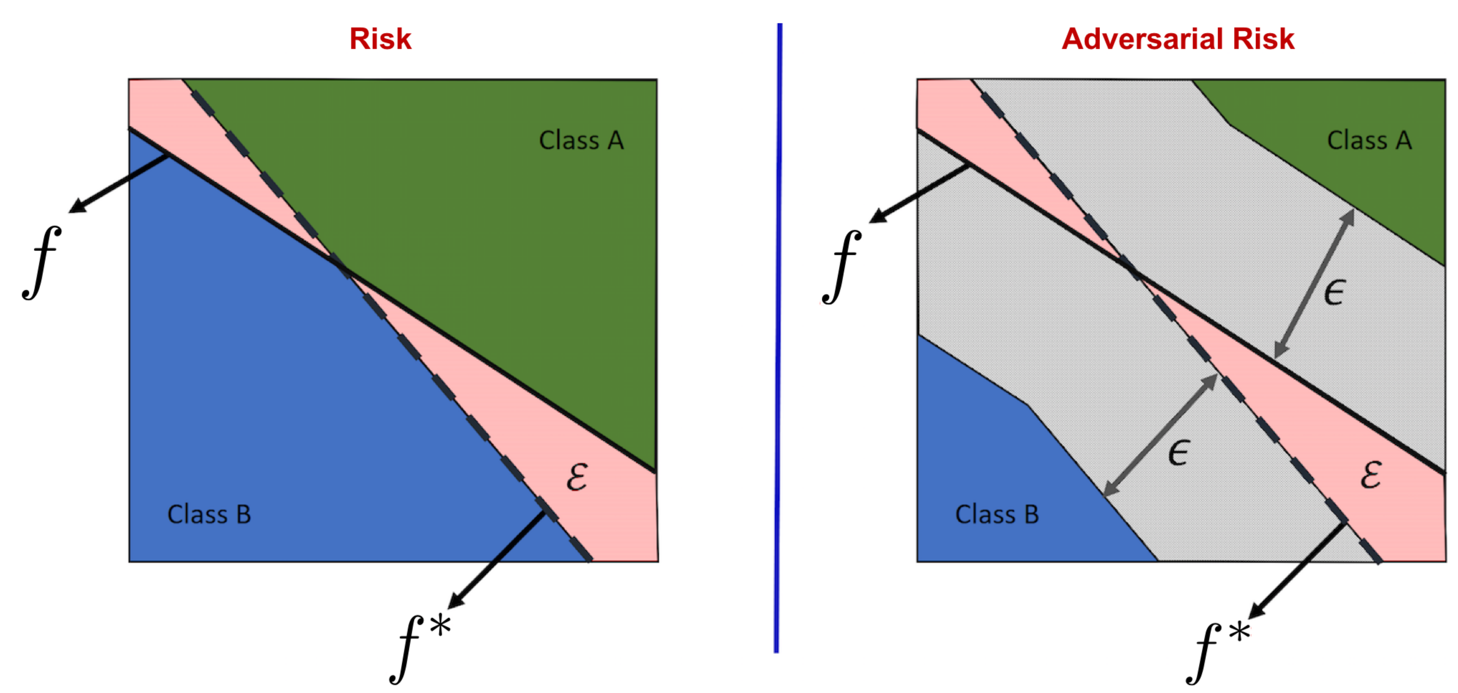
Given distribution μ, ground-truth classifier f^* and some classifier f.

Define **risk** as

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

Define **adversarial risk** w.r.t. ϵ perturbation as

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



Concentration for Real Distributions?

What is the minimum possible adversarial risk given 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 gives lower bound for **nice distributions**:

- Uniform distribution over spheres under ℓ_2 (Gilmer et al., 2018)
- Gaussian distribution under ℓ_2 (Fawzi et al., 2018)
- Any product distribution under ℓ_0 (Mahloujifar et al., 2018)
- Uniform distribution over hypercube under ℓ_2 (Shafahi et al., 2019)

Can we estimate concentration of measure for **real distributions?**

Our Empirical Framework

Challenge: do not know the PDF of the distribution.

Solution: replace μ with empirical distribution $\widehat{\mu}$ using samples $\mathcal{S}$.

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

Challenge: cannot search through all the possible subsets.

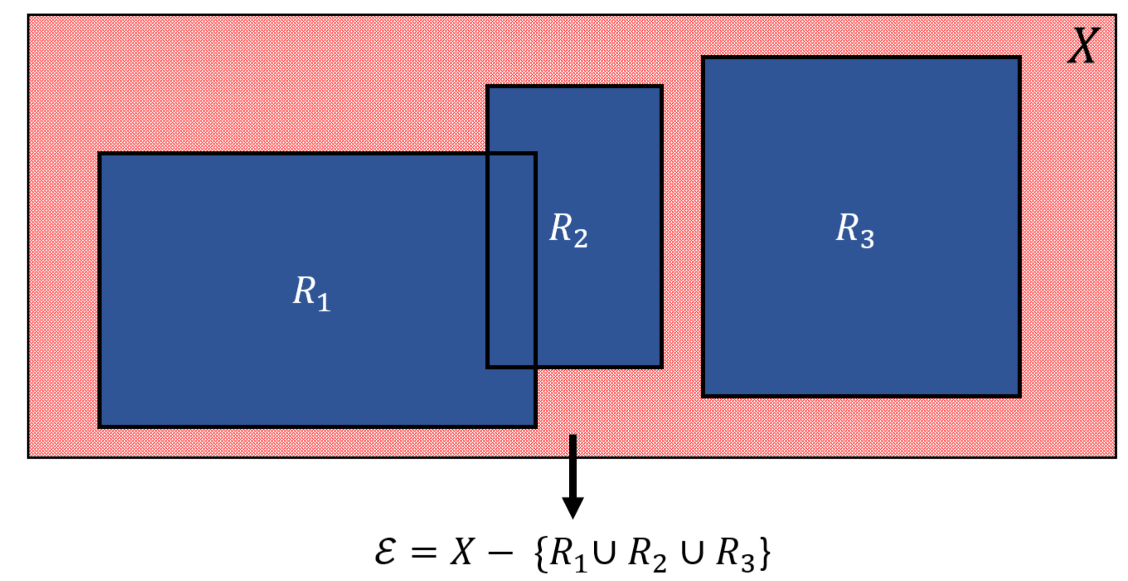
Solution: limit the search space to a special collection of subsets $\mathcal{G}$.

Remaining task: solve the following empirical problem:

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

Theoretical Results for ℓ_∞

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



Main Theorem: Define

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

Let $\widehat{\mu}_T$ be the empirical distribution with sample size T^4. Define

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

With probability 1 over the randomness of training data, we have

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

Finding Robust Error Region for ℓ_∞

1. Sort the dataset using ℓ_1 distance to the k-th nearest neighbor.
2. Perform kmeans clustering on the top-q densest images.
3. Obtain T rectangular image clusters and expand them by ϵ in ℓ_∞.
4. Treat the complement of these hyperrectangles as our error region.

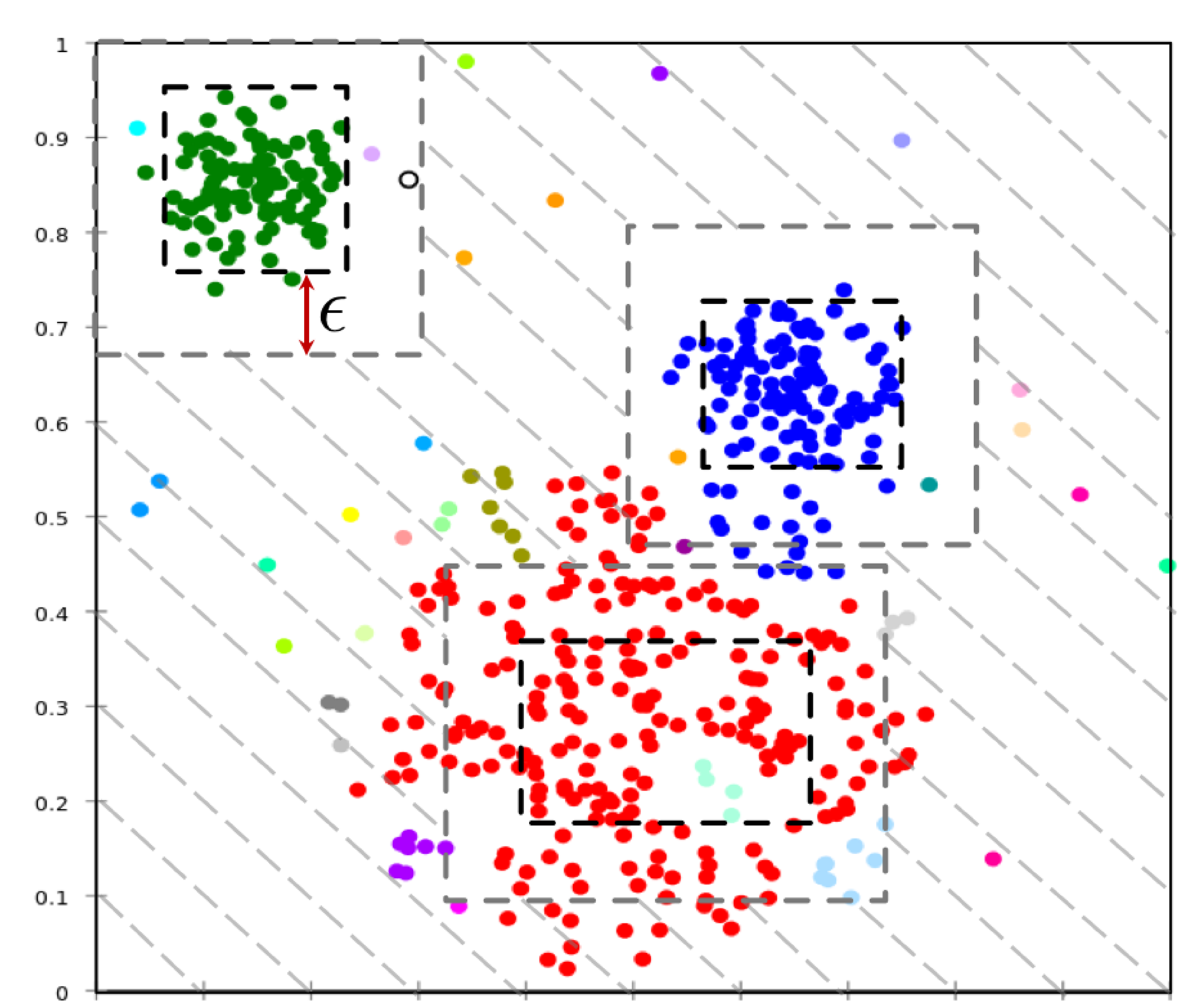


Figure: Illustration of the proposed algorithm using generated data

Conclusions and Future Work

- Impossibility results, such as Gilmer et al. (2018), **should not** make the community hopeless in finding more robust classifiers.
- Concentration of measure is **not the sole reason** behind the vulnerability of existing classifiers to adversarial examples.
- Study the error regions of practical machine learning classifers would be an interesting future direction.

Main Experimental Results

Table: Experiments for ℓ_∞ (Complement of Union of Hyperrectangles)

Dataset	α	ϵ	Risk	Adversarial Risk
MNIST	0.01	0.1	$1.23\% \pm 0.12\%$	$3.64\% \pm 0.30\%$
		0.3	$1.15\% \pm 0.13\%$	$7.24\% \pm 0.38\%$
CIFAR-10	0.05	2/255	$5.72\% \pm 0.25\%$	$8.13\% \pm 0.26\%$
		8/255	**$5.94\% \pm 0.34\%$**	**$18.13\% \pm 0.30\%$**

Table: Experiments for ℓ_2 (Union of Balls)

Dataset	α	ϵ_2	Risk	Adversarial Risk
MNIST	0.01	3.16	1.02%	4.15%
		4.74	1.07%	10.09%
CIFAR-10	0.05	0.4905	5.14%	5.83%
		0.9810	**5.12%**	**6.56%**

Table: Comparisons with state-of-the-art robust classifiers

Dataset	Strength	Method	Risk	Adversarial Risk
MNIST	$\epsilon_\infty = 0.3$	Madry et al. (2017)	1.20%	10.70%
		Our Bound	1.35%	8.28%
MNIST	$\epsilon_2 = 1.5$	Schott et al. (2018)	1.00%	20.00%
		Our Bound	1.08%	2.12%
CIFAR-10	$\epsilon_\infty = 8/255$	Madry et al. (2017)	12.70%	52.96%
		Our Bound	14.22%	29.21%

NSF Center for Trustworthy Machine Learning

PennState Stanford Berkeley UC San Diego Wisconsin University of Virginia

https://ctml.psu.edu/

shawn@virginia.edu