Empirically Measuring Concentration: Fundamental Limits on Intrinsic Robustness

NeurIPS 2019

Saeed Mahloujifar*   Xiao Zhang*   Mohammad Mahmoody   David Evans

University of Virginia

Impossibility Results for Robust Learning

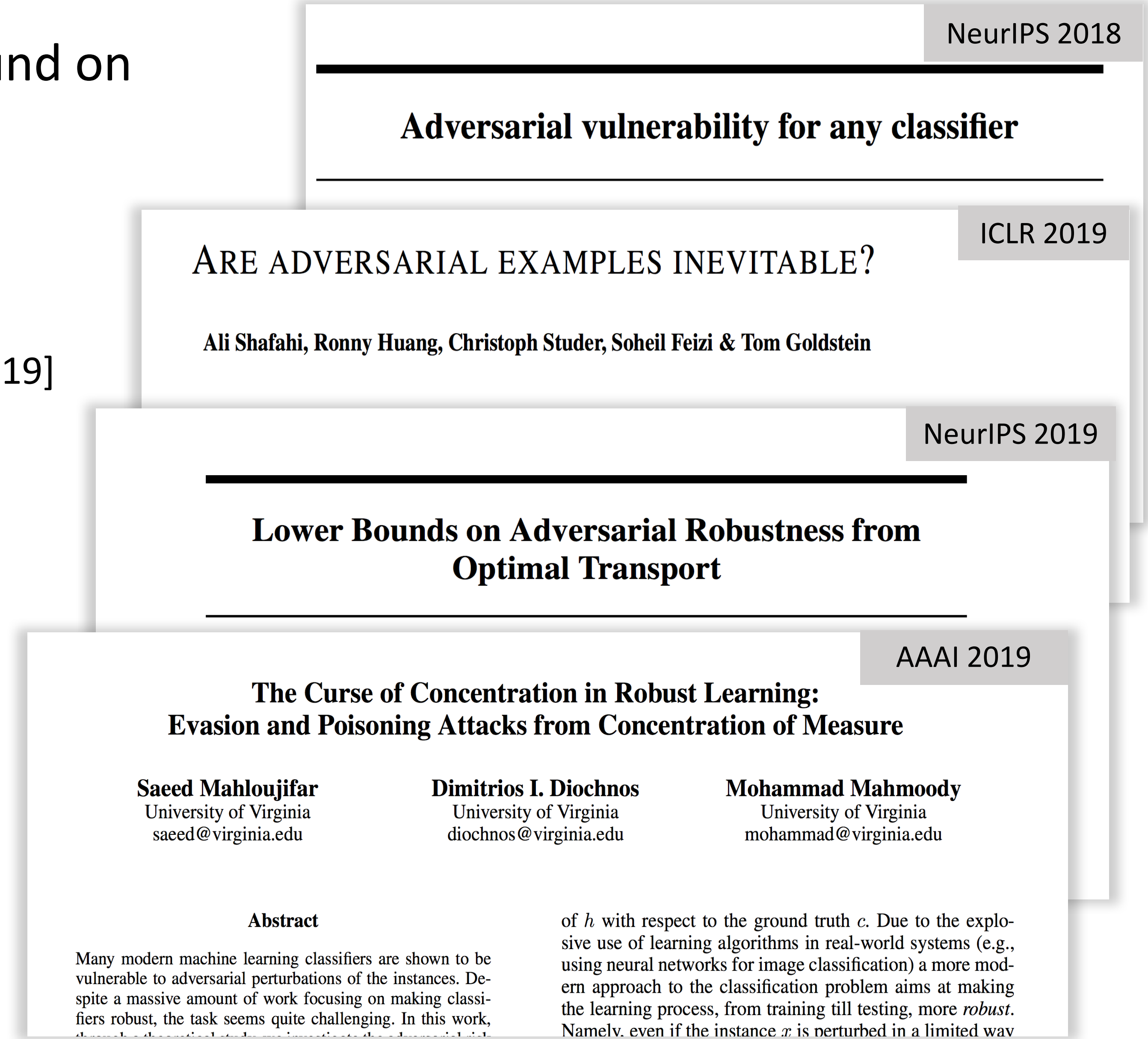
Concentration of measure gives lower bound on adversarial risk for ‘nice’ spaces:

Specific distributions:

[Gilmer+ 2018], [Fawzi+, 2018], [Diochnos+, 2018], [Shafahi+, 2019], [Bhagoji+, 2019], [Dohmatob+, 2019]

Concentrated metric probability space:

[Mahloujifar+, 2019]


NeurIPS 2018

Adversarial vulnerability for any classifier

ICLR 2019

ARE ADVERSARIAL EXAMPLES INEVITABLE?

Ali Shafahi, Ronny Huang, Christoph Studer, Soheil Feizi & Tom Goldstein

NeurIPS 2019

Lower Bounds on Adversarial Robustness from Optimal Transport

AAAI 2019

The Curse of Concentration in Robust Learning: Evasion and Poisoning Attacks from Concentration of Measure

Saeed Mahloujifar
University of Virginia
saeed@virginia.edu

Dimitrios I. Diochnos
University of Virginia
diochnos@virginia.edu

Mohammad Mahmoody
University of Virginia
mohammad@virginia.edu

Abstract

Many modern machine learning classifiers are shown to be vulnerable to adversarial perturbations of the instances. Despite a massive amount of work focusing on making classifiers robust, the task seems quite challenging. In this work,

of h with respect to the ground truth c. Due to the explosive use of learning algorithms in real-world systems (e.g., using neural networks for image classification) a more modern approach to the classification problem aims at making the learning process, from training till testing, more *robust*. Namely, even if the instance x is perturbed in a limited way

What about image distributions?

Concentration of measure gives lower bound on adversarial risk for ‘nice’ spaces:

Specific distributions:

[Gilmer+ 2018], [Fawzi+, 2018], [Diochnos+, 2018], [Shafahi+, 2019], [Bhagoji+, 2019], [Dohmatob+, 2019]

Concentrated metric probability space:

[Mahloujifar+, 2019]

NeurIPS 2018

Adversarial vulnerability for any classifier

ICLR 2019

ARE ADVERSARIAL EXAMPLES INEVITABLE?

Ali Shafahi, Ronny Huang, Christoph Studer, Soheil Feizi & Tom Goldstein

NeurIPS 2019

Lower Bounds on Adversarial Robustness from Optimal Transport

Do these results hold for real distributions like images?

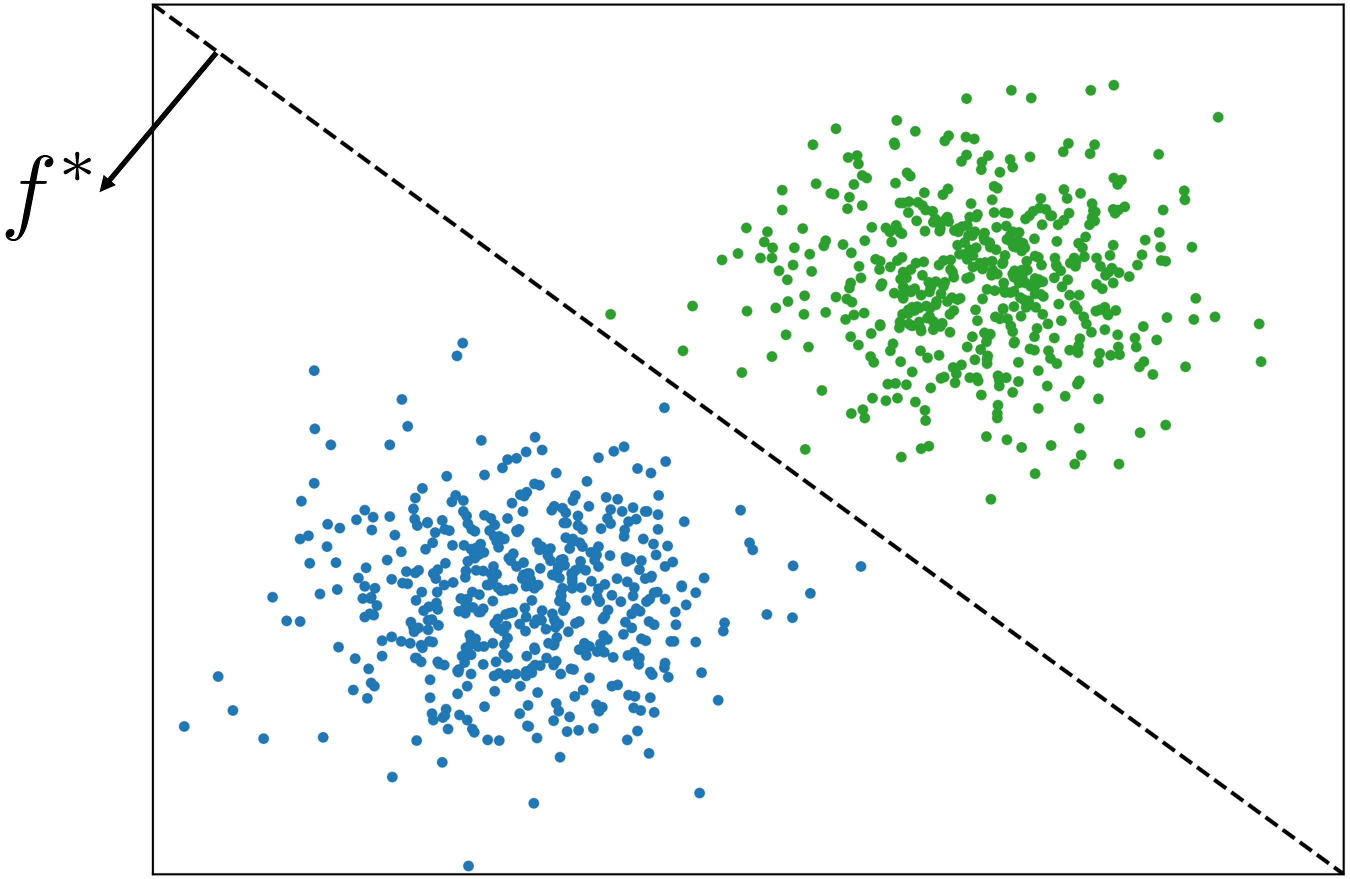
1. Provide a way to measure concentration using i.i.d. samples
2. Show these impossibility results do not simply apply to image benchmarks

Abstract

Many modern machine learning classifiers are shown to be vulnerable to adversarial perturbations of the instances. Despite a massive amount of work focusing on making classifiers robust, the task seems quite challenging. In this work,

of h with respect to the ground truth c. Due to the explosive use of learning algorithms in real-world systems (e.g., using neural networks for image classification) a more modern approach to the classification problem aims at making the learning process, from training till testing, more *robust*. Namely, even if the instance x is perturbed in a limited way

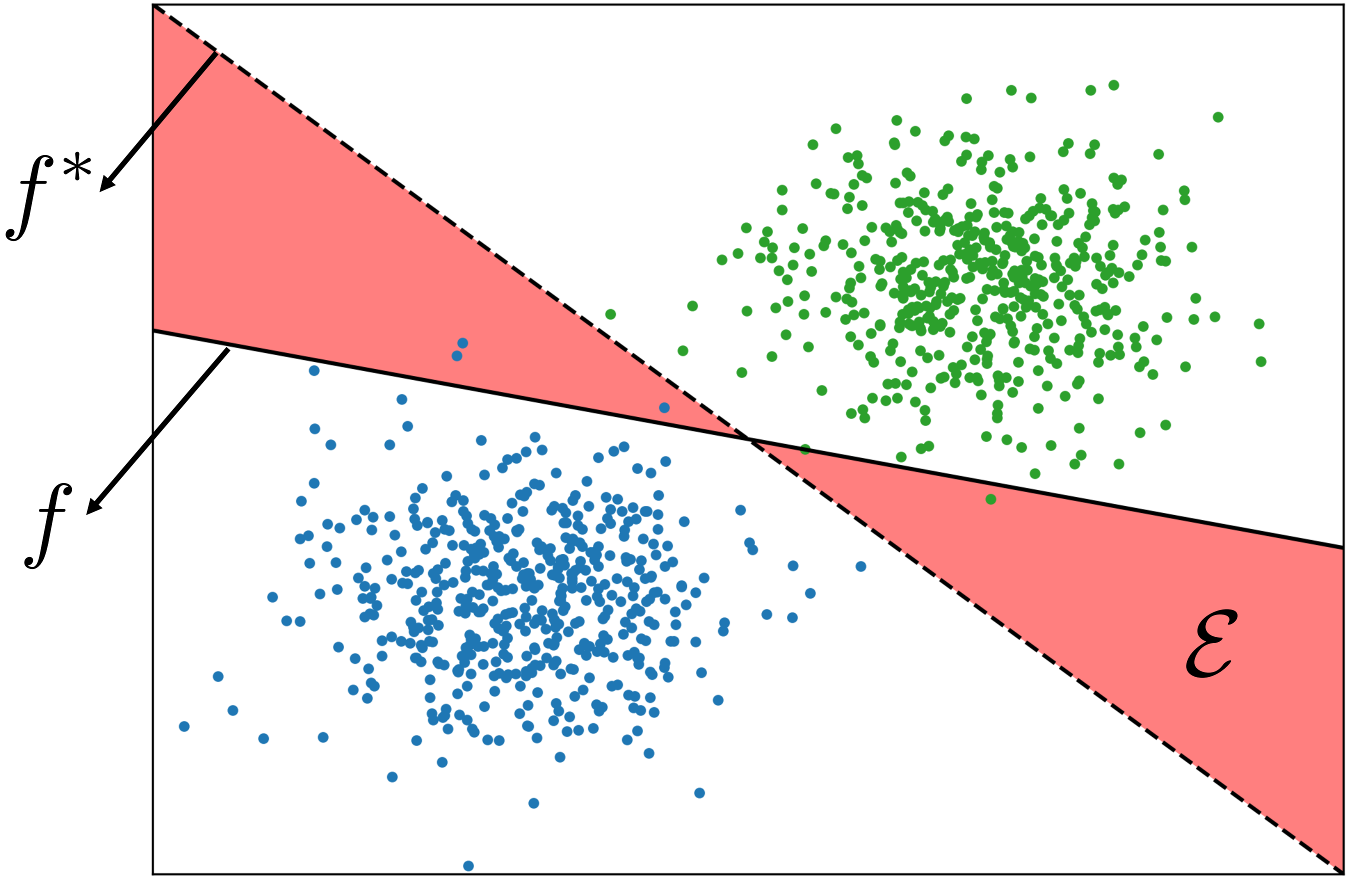
Connecting Concentration and Robust Learning



μ : underlying data distribution

f^* : ground-truth classifier

Risk and Error Region



μ : underlying data distribution

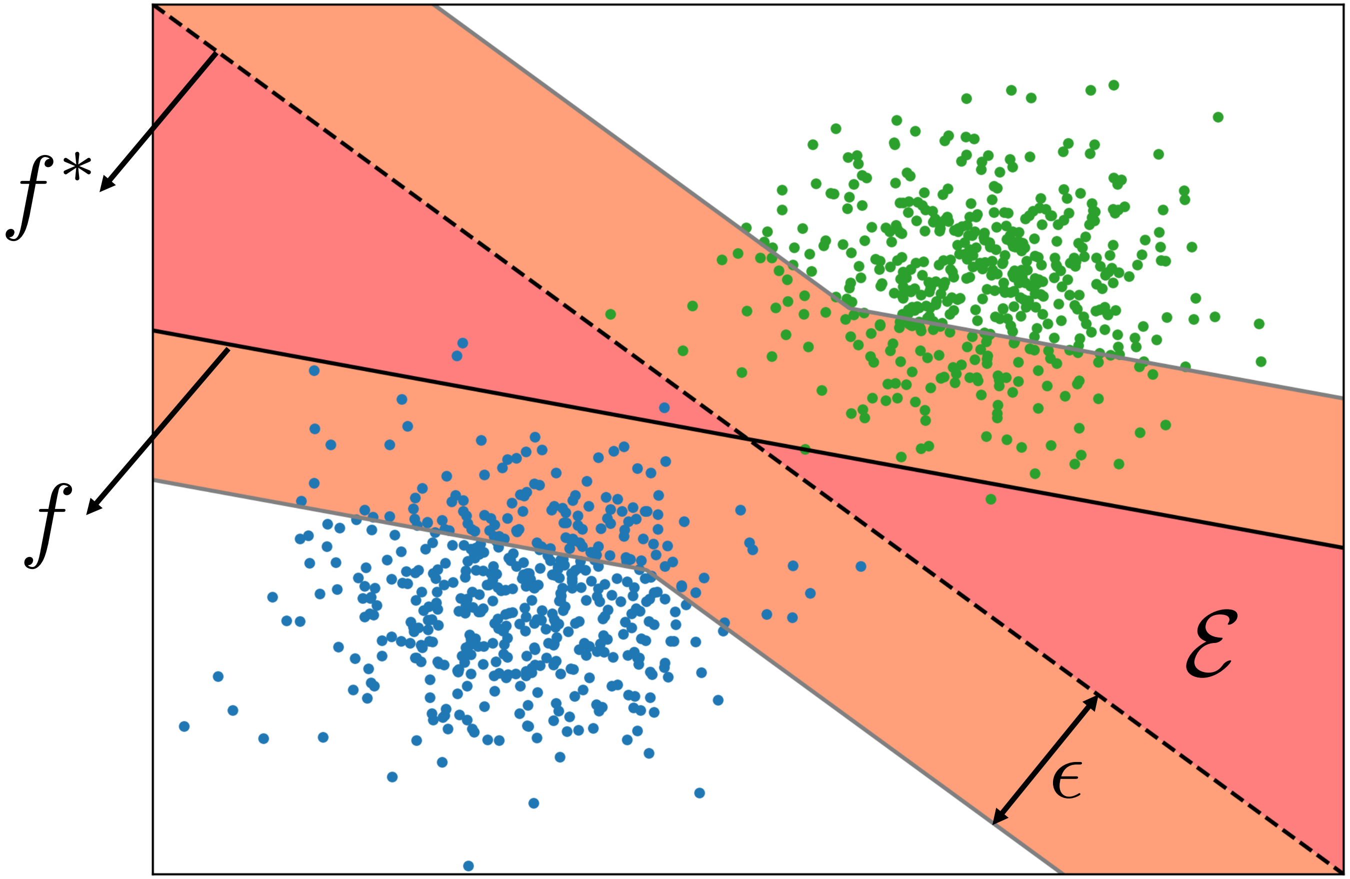
f^* : ground-truth classifier

f : any classifier

$\mathcal{E}$: error region between f and f^*

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

Adversarial Risk and Expanded Error Region



μ : underlying data distribution

f^* : ground-truth classifier

f : any classifier

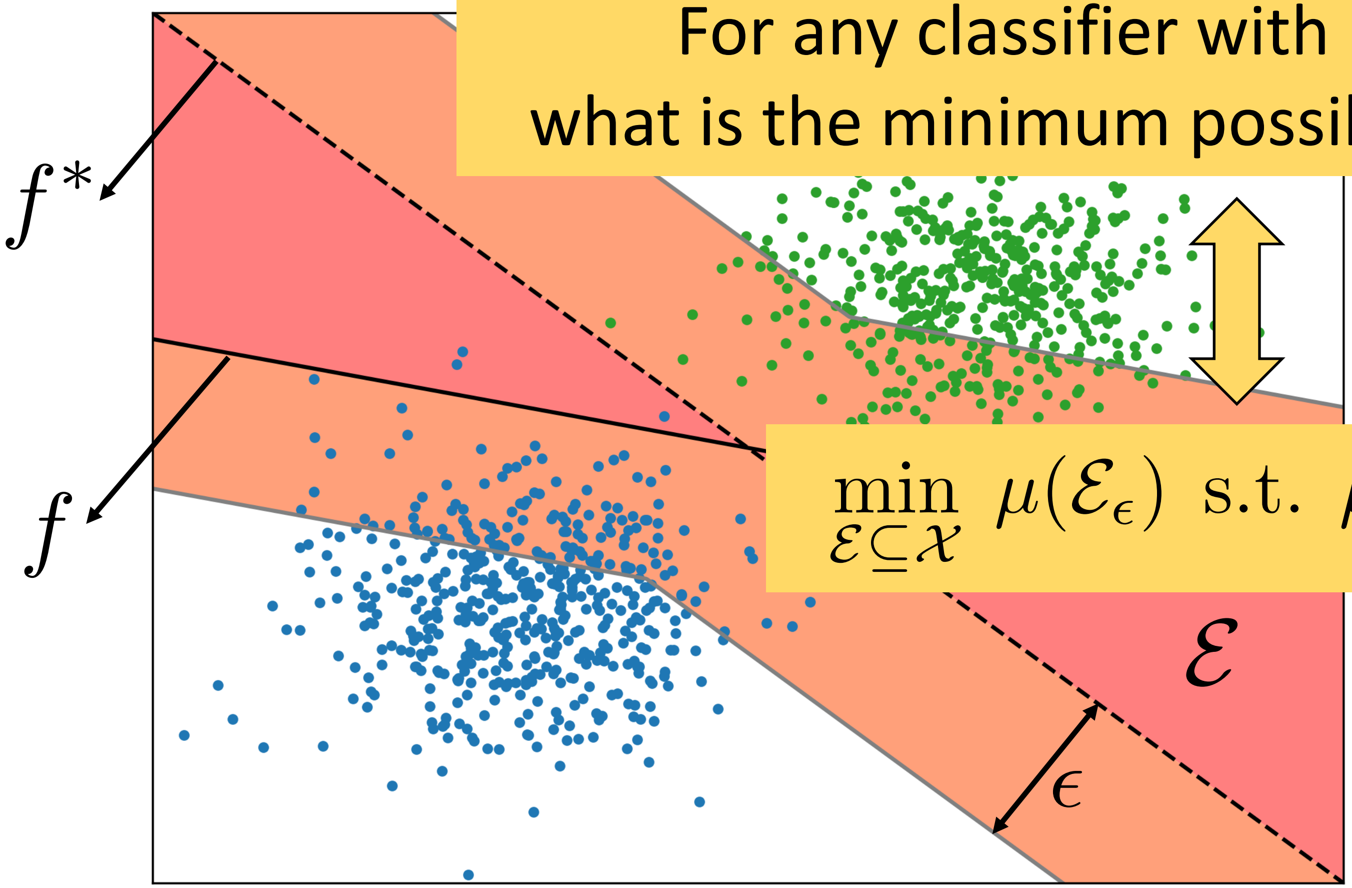
$\mathcal{E}$: error region between f and f^*

ϵ : adversarial strength

$\mathcal{E}_\epsilon$: ϵ-expansion of $\mathcal{E}$

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

Concentration of Measure



For any classifier with risk at least α,
what is the minimum possible adversarial risk?

$$\min_{\mathcal{E}\subseteq\mathcal{X}} \ \mu(\mathcal{E}_\epsilon) \ \text{ s.t. } \ \mu(\mathcal{E}) \geq \alpha$$

distribution

f^* : ground-truth classifier

f : any given classifier

r region between f and f^*

ϵ : adversarial strength

$\mathcal{E}_\epsilon$: ϵ-expansion of $\mathcal{E}$

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

Empirical Concentration Problem

Actual concentration problem:

$$\min_{\mathcal{E}\subseteq\mathcal{X}} \mu(\mathcal{E}_\epsilon) \text{ s.t. } \mu(\mathcal{E}) \geq \alpha$$

only have access to data samples

Empirical concentration problem:

$$\min_{\mathcal{E}\in\mathcal{G}} \widehat{\mu}(\mathcal{E}_\epsilon) \text{ s.t. } \widehat{\mu}(\mathcal{E}) \geq \alpha$$

$\widehat{\mu}$: empirical measure based on samples

$\mathcal{G}$: some special collection of subsets (w.r.t. perturbation metric)

Main Theoretical Result

Actual concentration problem:

$$\min_{\mathcal{E}\subseteq\mathcal{X}} \ \mu(\mathcal{E}_\epsilon) \ \text{ s.t. } \ \mu(\mathcal{E}) \geq \alpha$$

only have access to data samples

asymptotic convergence

Empirical concentration problem:

$$\min_{\mathcal{E}\in\mathcal{G}} \ \widehat{\mu}(\mathcal{E}_\epsilon) \ \text{ s.t. } \ \widehat{\mu}(\mathcal{E}) \geq \alpha$$

Key idea: increase both the sample size and the complexity of $\mathcal{G}$ in a careful way

$\widehat{\mu}$: empirical measure based on samples

$\mathcal{G}$: some special collection of subsets (w.r.t. perturbation metric)

Empirically Measuring Concentration

To solve: $\min_{\mathcal{E} \in \mathcal{G}} \widehat{\mu}(\mathcal{E}_\epsilon)$ s.t. $\widehat{\mu}(\mathcal{E}) \geq \alpha$ (ℓ_∞ metric)

$\mathcal{G}$: complement of union of rectangles

Algorithmic idea: **avoid the dense regions**

Empirically Measuring Concentration

To solve: $\min_{\mathcal{E}\in\mathcal{G}} \widehat{\mu}(\mathcal{E}_\epsilon)$ s.t. $\widehat{\mu}(\mathcal{E}) \geq \alpha$ (ℓ_∞ metric)

$\mathcal{G}$: complement of union of rectangles

Algorithmic idea: avoid the dense regions

- Select dense data points using k-nearest neighbor
- Place rectangles to capture the dense area using k-means

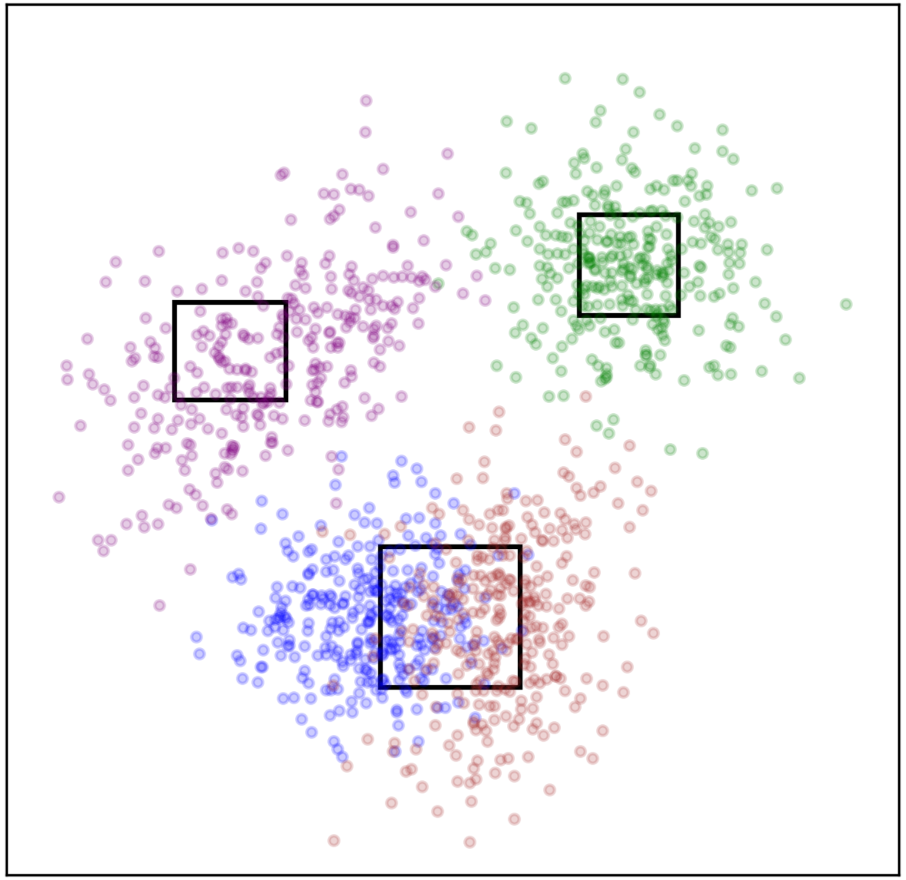


Illustration of our algorithm ($\alpha = 0.01, \epsilon = 1.0$)

Empirically Measuring Concentration

To solve: $\min_{\mathcal{E}\in\mathcal{G}} \widehat{\mu}(\mathcal{E}_\epsilon)$ s.t. $\widehat{\mu}(\mathcal{E}) \geq \alpha$ (ℓ_∞ metric)

$\mathcal{G}$: complement of union of rectangles

Algorithmic idea: avoid the dense regions

- Select dense data points using k-nearest neighbor
- Place rectangles to capture the dense area using k-means
- Expand the rectangles and treat the complement of their union as the error region
- Tune parameters (e.g. the number of rectangles) for the best results

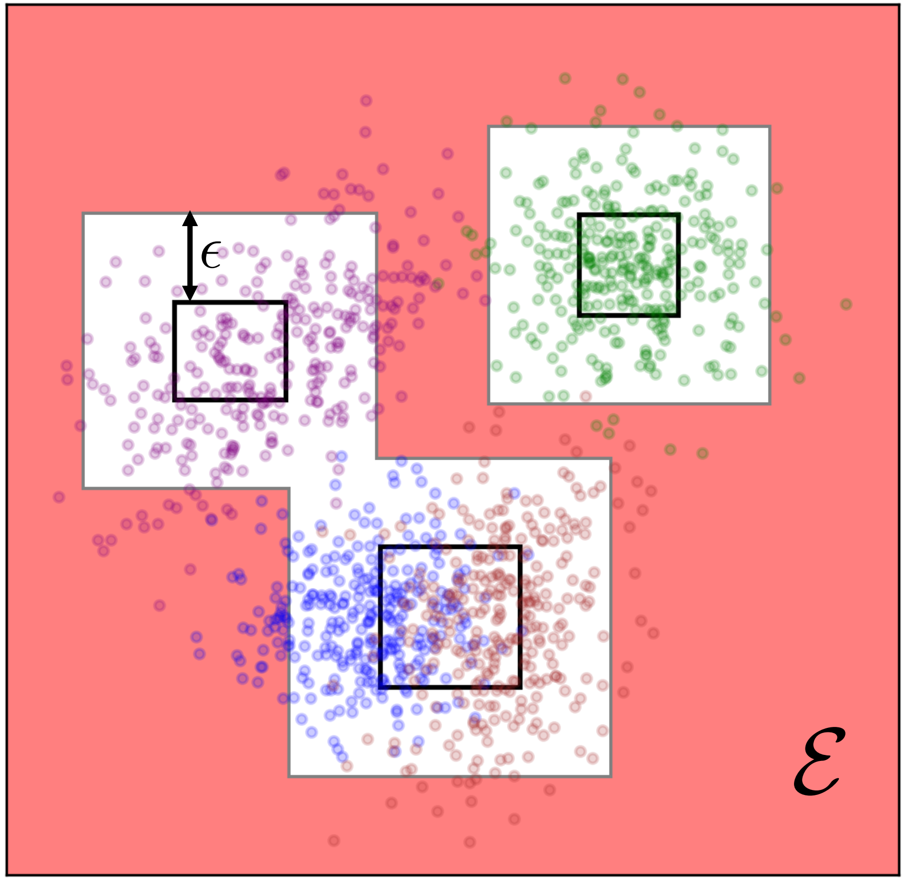


Illustration of our algorithm ($\alpha = 0.01, \epsilon = 1.0$)

$$\mu(\mathcal{E}) = 0.01 \rightarrow \mu(\mathcal{E}_\epsilon) = 0.24$$

Empirical Results on Benchmark Datasets

Datasets	Risk Constraint (α)	Max Perturbation	Lower Bound on Adversarial Risk
MNIST	0.01	$\ell_\infty \leq 0.3$	7.2%
MNIST	0.01	$\ell_2 \leq 1.5$	2.1%
CIFAR-10	0.05	$\ell_\infty \leq 8/255$	18.1%

For benchmark image datasets, there exists rather robust error regions

Compare with State-of-the-art Defenses

Datasets	Risk Constraint (α)	Max Perturbation	Lower Bound on Adversarial Risk	Attack Success Rate for State-of-the-art Defenses
MNIST	0.01	$\ell_\infty \leq 0.3$	7.2%	10.7% [Madry+, 2018]
MNIST	0.01	$\ell_2 \leq 1.5$	2.1%	20.0% [Schott+, 2019]
CIFAR-10	0.05	$\ell_\infty \leq 8/255$	18.1%	52.9% [Madry+, 2019]

a small gap

a large gap

For benchmark image datasets, there exists rather robust error regions

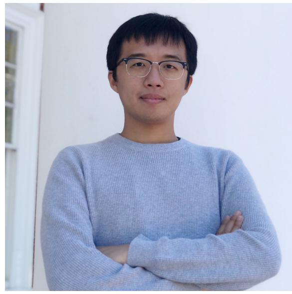
Suggest concentration is *not* the sole reason behind adversarial vulnerability

Conclusion: concentration of measure cannot explain all: either exist more robust classifiers or some other reasons explaining why

Poster: 10:45 AM -- 12:45 PM @ East Exhibition Hall B + C #10

Empirically Measuring Concentration: Fundamental Limits on Intrinsic Robustness


Saeed Mahloujifar


Xiao Zhang


Mohammad Mahmoody


David Evans

URL: https://evademl.org/concentration/