NEURAL INFORMATION PROCESSING SYSTEMS

Adversarial Examples Might be Avoidable: The Role of Data Concentration in Adversarial Robustness



Ambar Pal   Jeremias Sulam   René Vidal

JOHNS HOPKINS UNIVERSITY

Penn UNIVERSITY of PENNSYLVANIA

Adversarial Examples

- Small, targeted *adversarial* perturbations mislead modern classifiers

+

=

Dog

Cat

Impossibility Results

- Small, targeted *adversarial* perturbations mislead modern classifiers


+

=


Dog

Cat

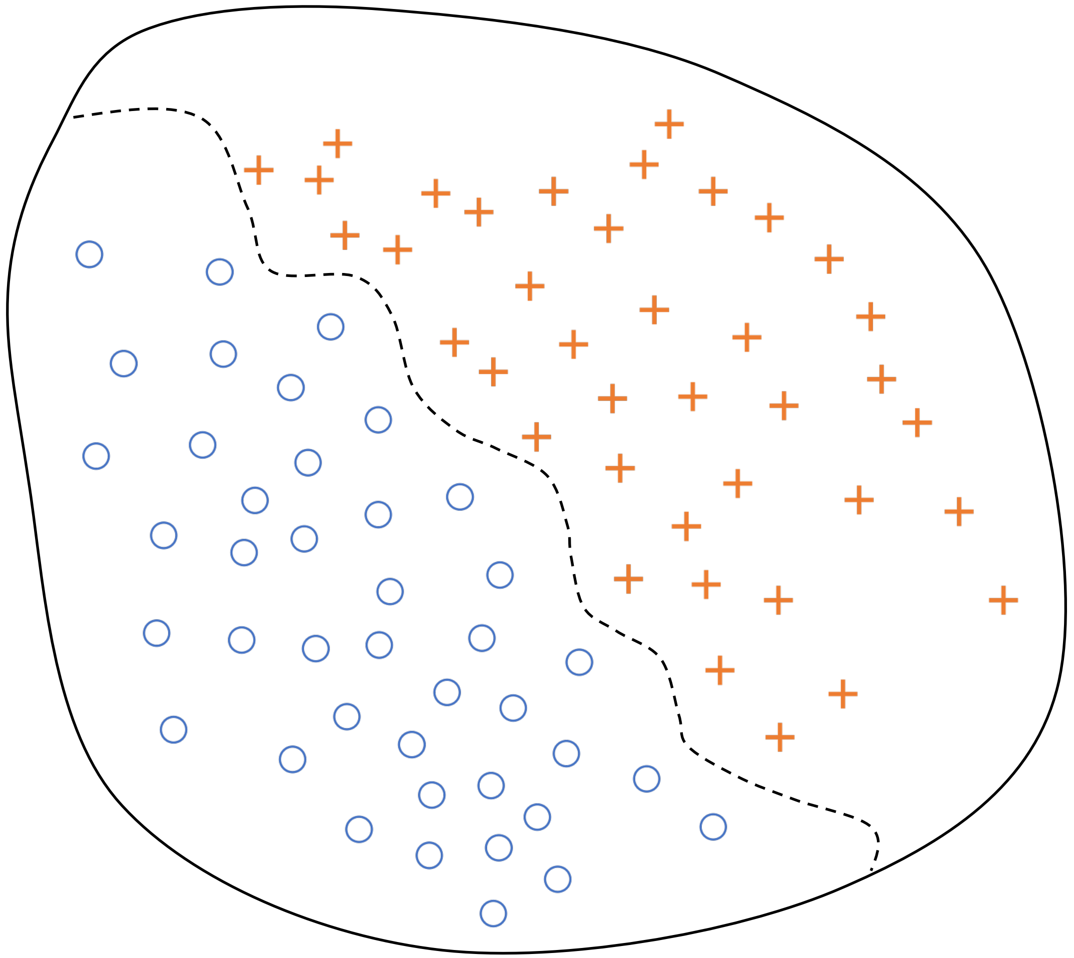
- Adversarial examples exist for *any* classifier

“any classifier admits ϵ-adversarial examples for the minority class with probability $1 - C_p \exp\left(-n\epsilon^2\right)$”

- What is going on?

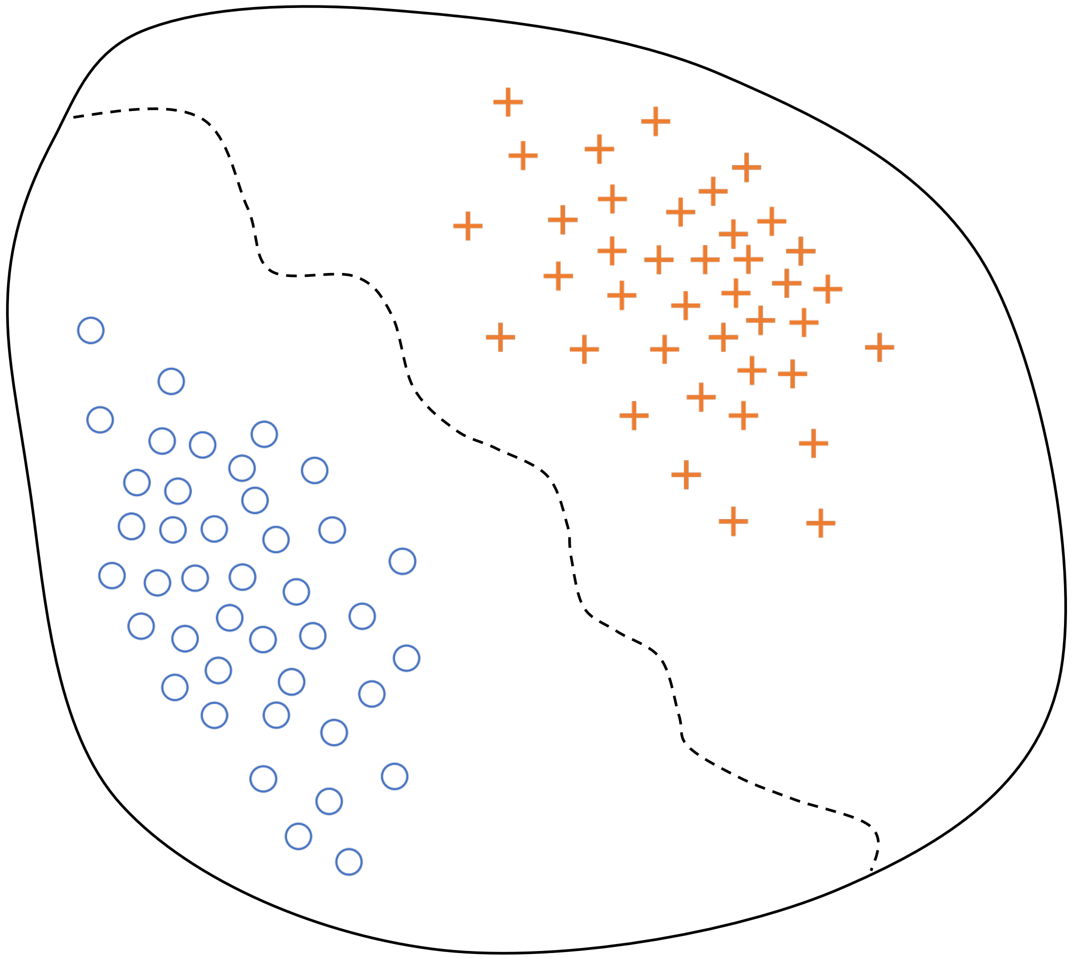
Adversarial vulnerability for any classifier, Fawzi+18. Are adversarial examples inevitable? Shafahi+18. Generalized no free lunch theorem for adversarial robustness, Dohmatob19.

Data Concentration



Uniform

Data Concentration

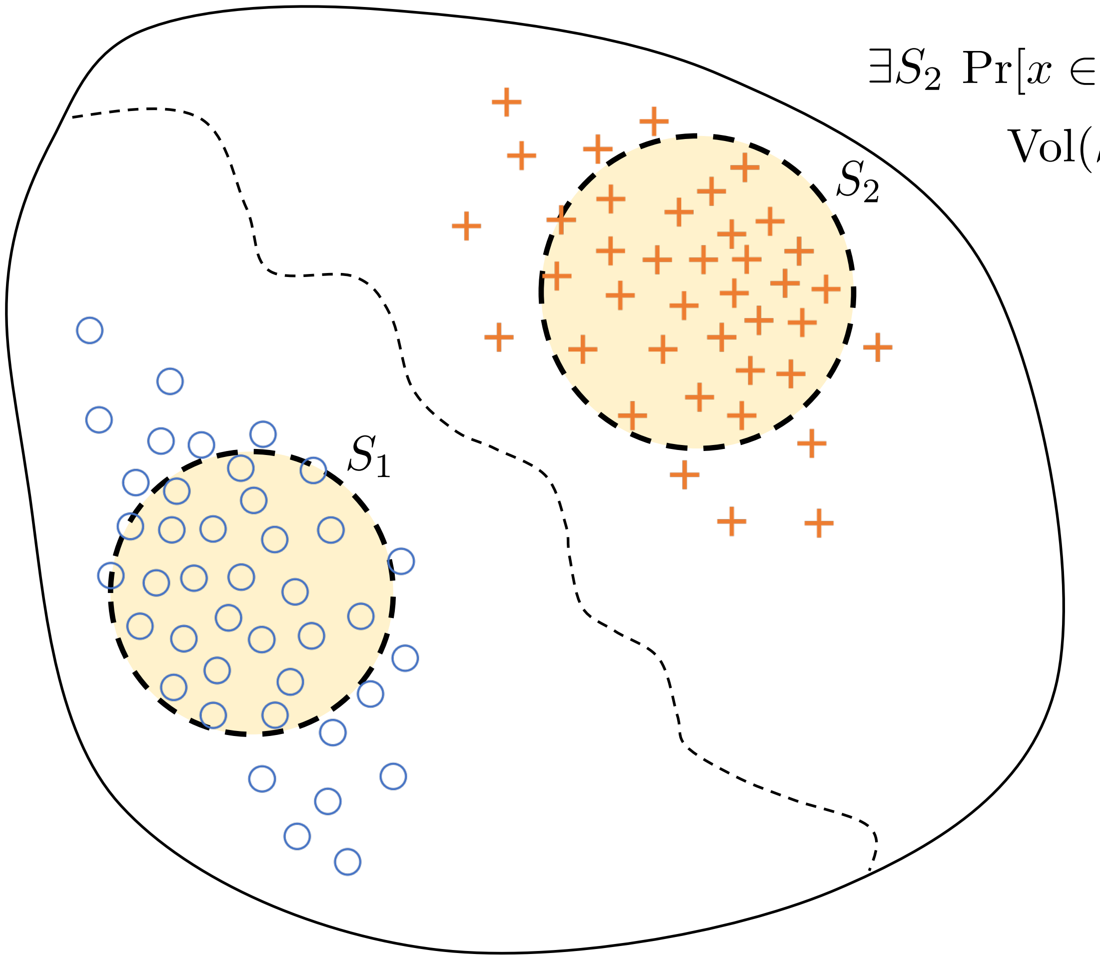


Concentrated

(ϵ, δ) - concentration

$\exists S_2 \ \Pr[x \in S_2 | y = 2] \geq 1 - \delta$

$\mathrm{Vol}(S_2) \leq \exp(-n\epsilon)$

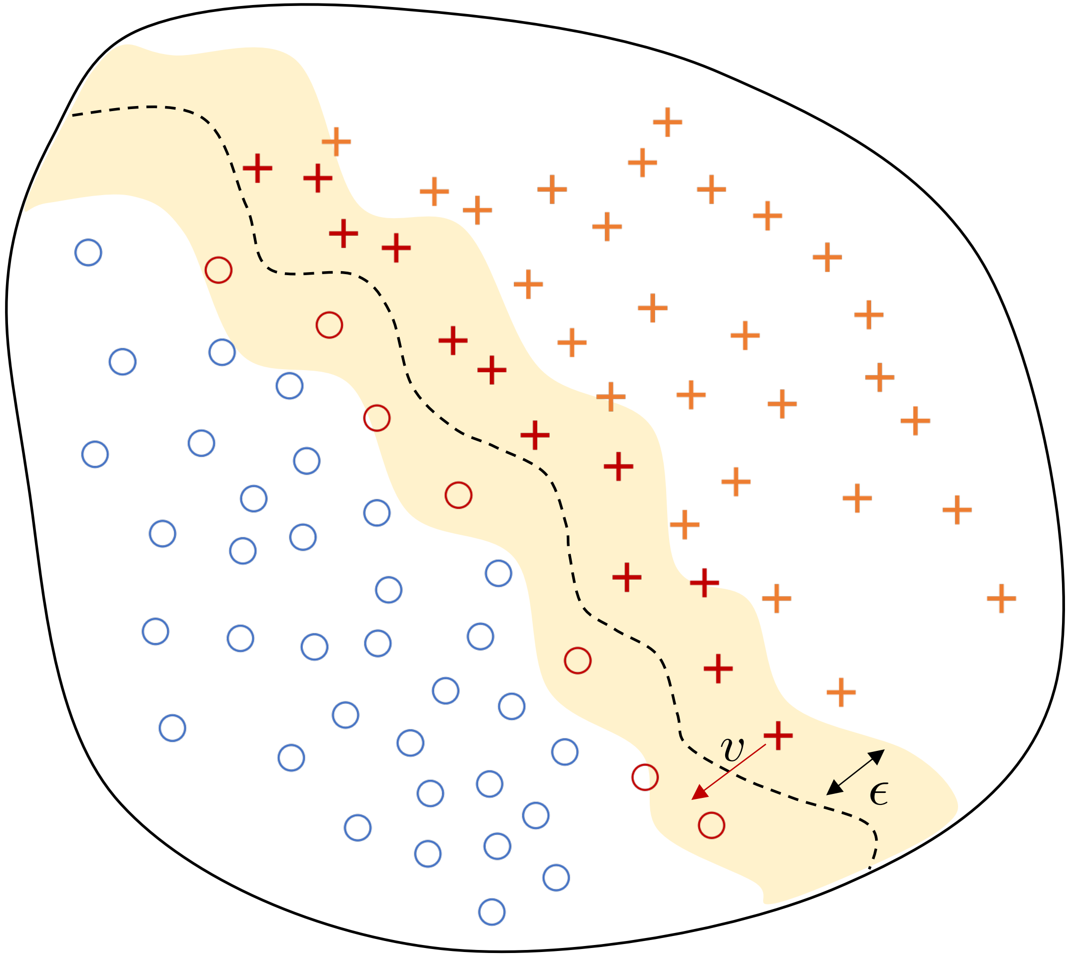


$\exists S_1 \ \Pr[x \in S_1 | y = 1] \geq 1 - \delta$

$\mathrm{Vol}(S_1) \leq \exp(-n\epsilon)$

(ϵ, δ) - robust classifier

$$\mathbb{P}(\exists v \text{ such that } \|v\| \leq \epsilon, f(x+v) \neq y) \leq \delta$$



Geometric Characterization of Robustness

Theorem 1

$\exists f$ such that f is (ϵ, δ)-robust for p

$\Downarrow$

p is (ϵ, δ)-concentrated

necessary

Theorem 2

p is strongly-$(\epsilon, \delta, \gamma)$-concentrated

$\Downarrow$

$\exists f$ such that f is $(\epsilon, \delta + \gamma)$-robust for p

sufficient

Application I

Wide class of distributions where adversarial examples do *not* exist with high probability

"Adversarial Impossibility results are vacuous for natural data-distributions" $1 - C_p \exp\left(-n\epsilon^2\right)$

Application II

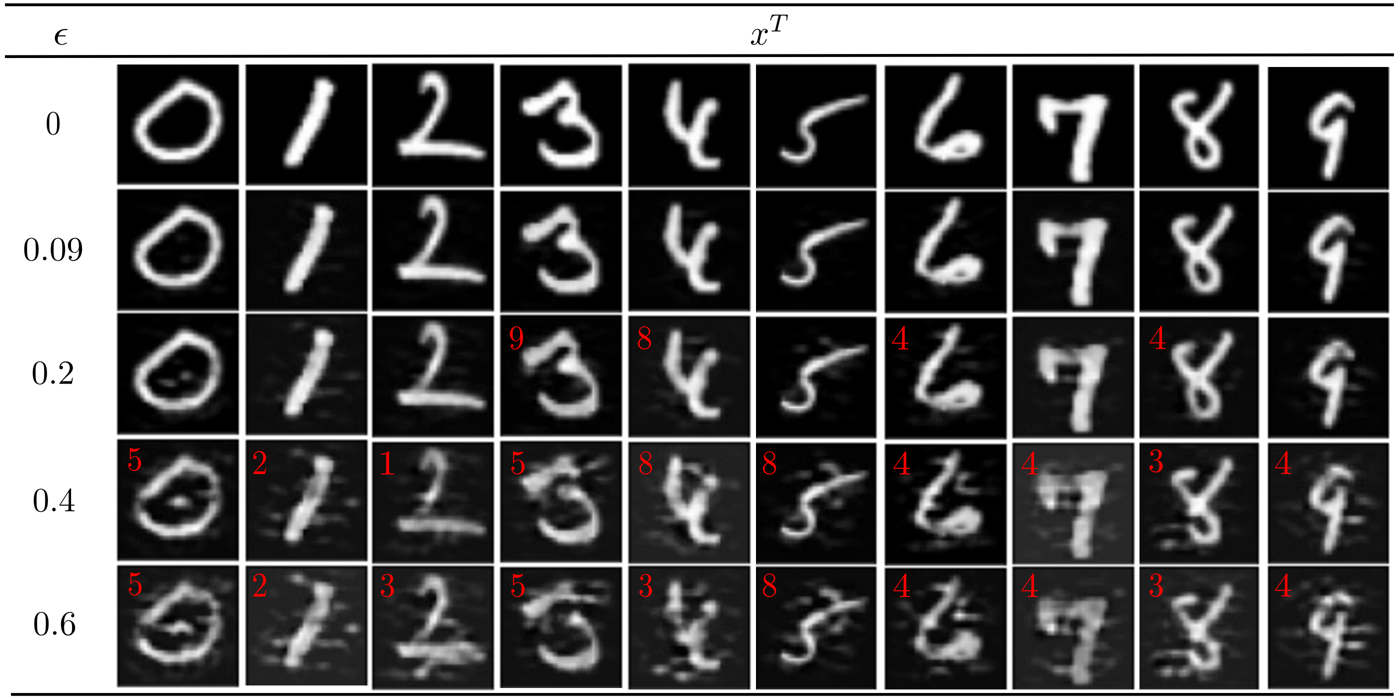
Construction of a robust classifier for distributions lying near linear subspaces (e.g., MNIST)

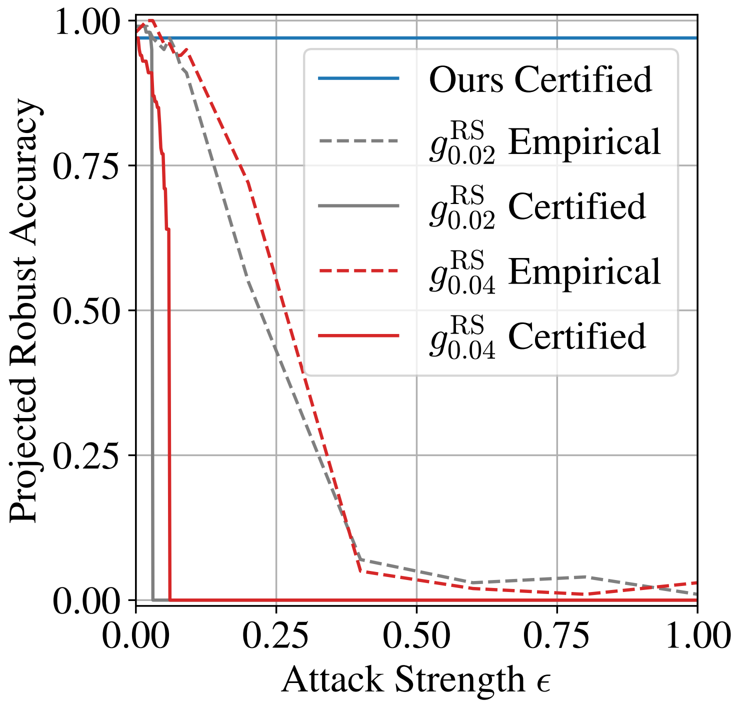
Certifying large-ℓ_p perturbations

Application II

Construction of a robust classifier for distributions lying near linear subspaces (e.g., MNIST)

- The certificate is not limited to spherical balls





Geometric Characterization of Robustness

Theorem 1

$\exists f$ such that f is (ϵ, δ)-robust for p

$$\Downarrow$$

p is (ϵ, δ)-concentrated

necessary

Theorem 2

p is strongly-$(\epsilon, \delta, \gamma)$-concentrated

$$\Downarrow$$

$\exists f$ such that f is $(\epsilon, \delta + \gamma)$-robust for p

sufficient

Application I

Wide class of distributions where adversarial examples do *not* exist with high probability

Application II

Construction of a robust classifier for distributions lying near linear subspaces (e.g., MNIST)

Thank You

Location: Great Hall & Hall B1+B2
Poster #724

Time: Wed 13 Dec
0845 - 1045 PT