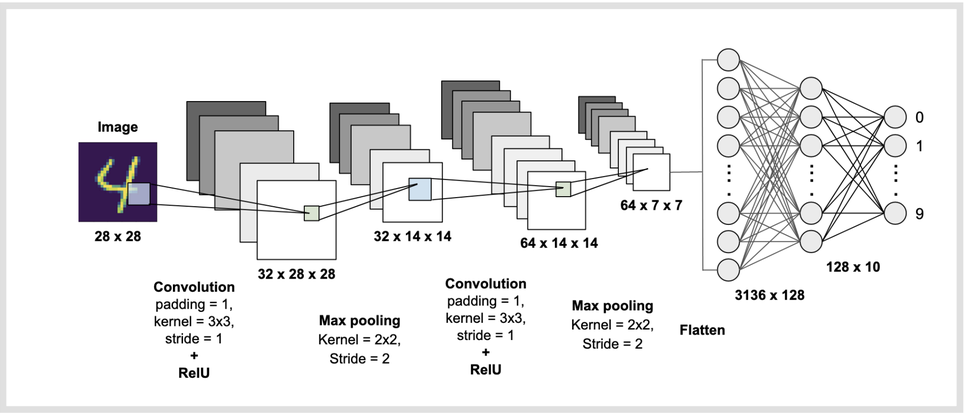
Overparametrization and Robustness

Hamed Hassani
University of Pennsylvania

NSF HDR PI meeting, October 2022

Adversarial examples: a brief introduction

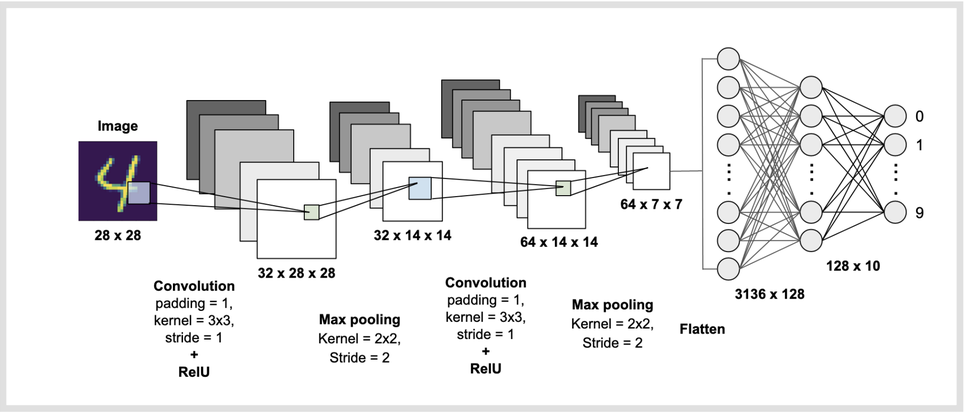
Model (predictor)



[Biggio et al 2014] [Szegedy et al 2014]

Adversarial examples: a brief introduction

Model (predictor)

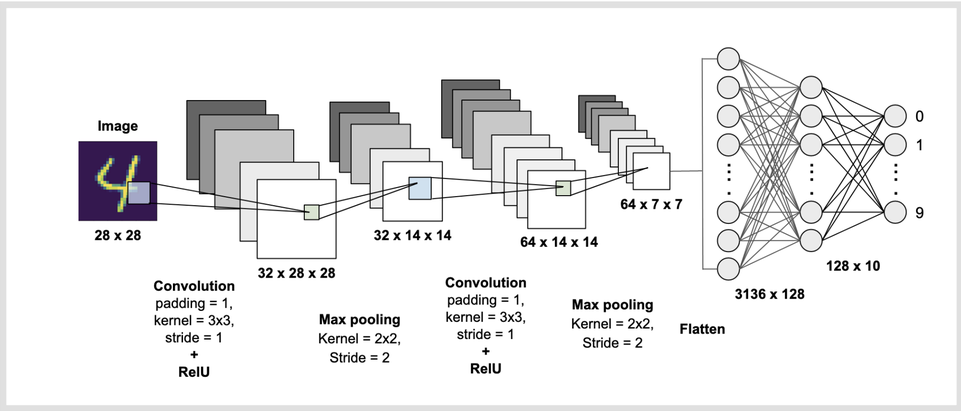


image

[Biggio et al 2014] [Szegedy et al 2014]

Adversarial examples: a brief introduction

Model (predictor)



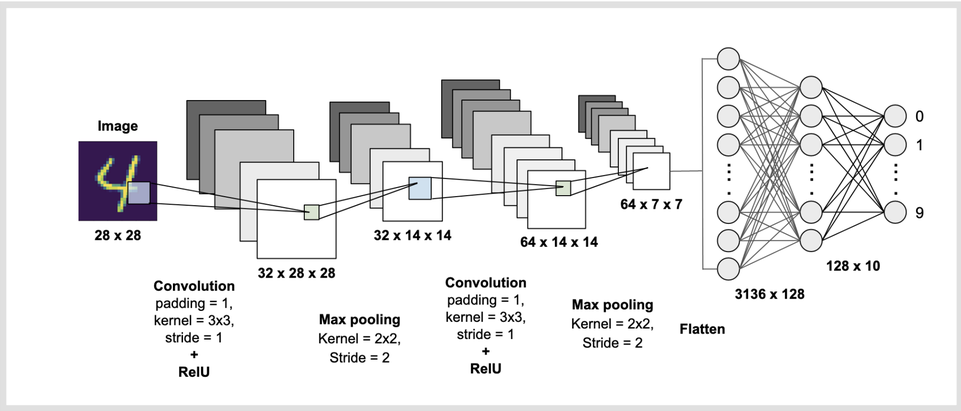
image

Model → Panda

[Biggio et al 2014] [Szegedy et al 2014]

Adversarial examples: a brief introduction

Model (predictor)



image

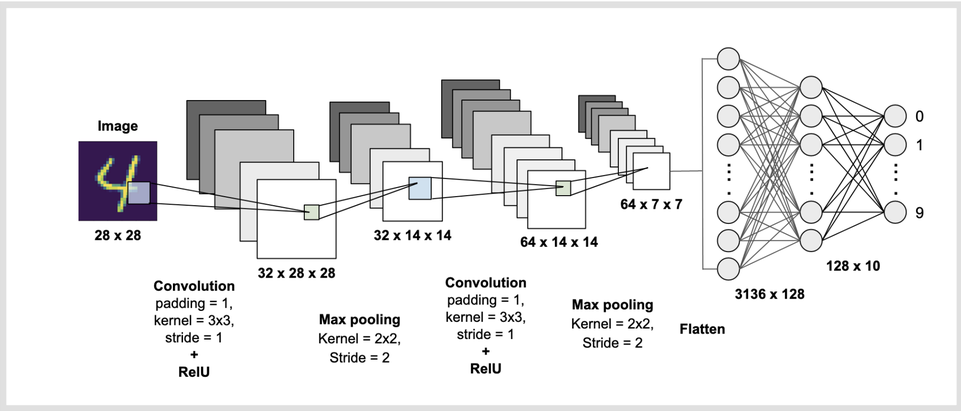
Model → Panda

image

[Biggio et al 2014] [Szegedy et al 2014]

Adversarial examples: a brief introduction

Model (predictor)



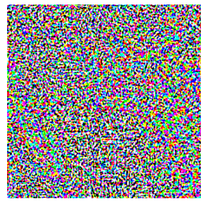
image

Model

Panda

image

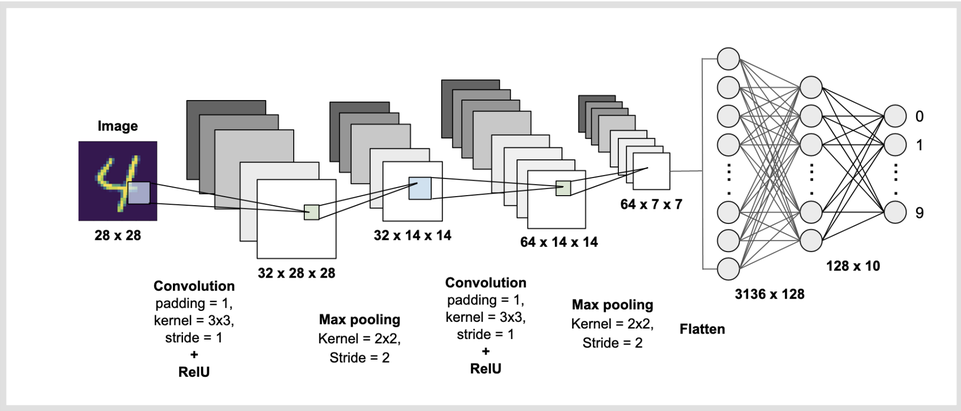
small
noise

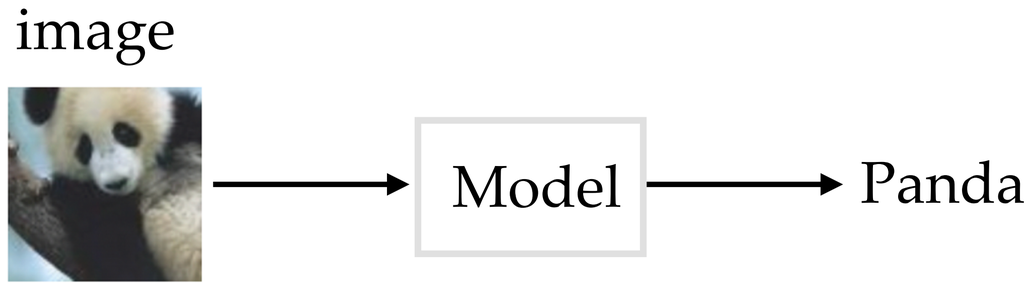


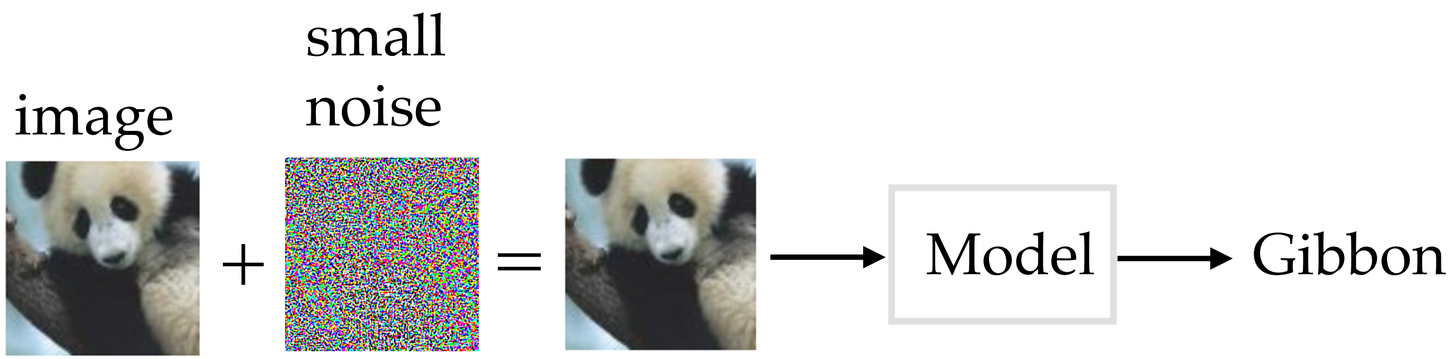
[Biggio et al 2014] [Szegedy et al 2014]

Adversarial examples: a brief introduction

Model (predictor)



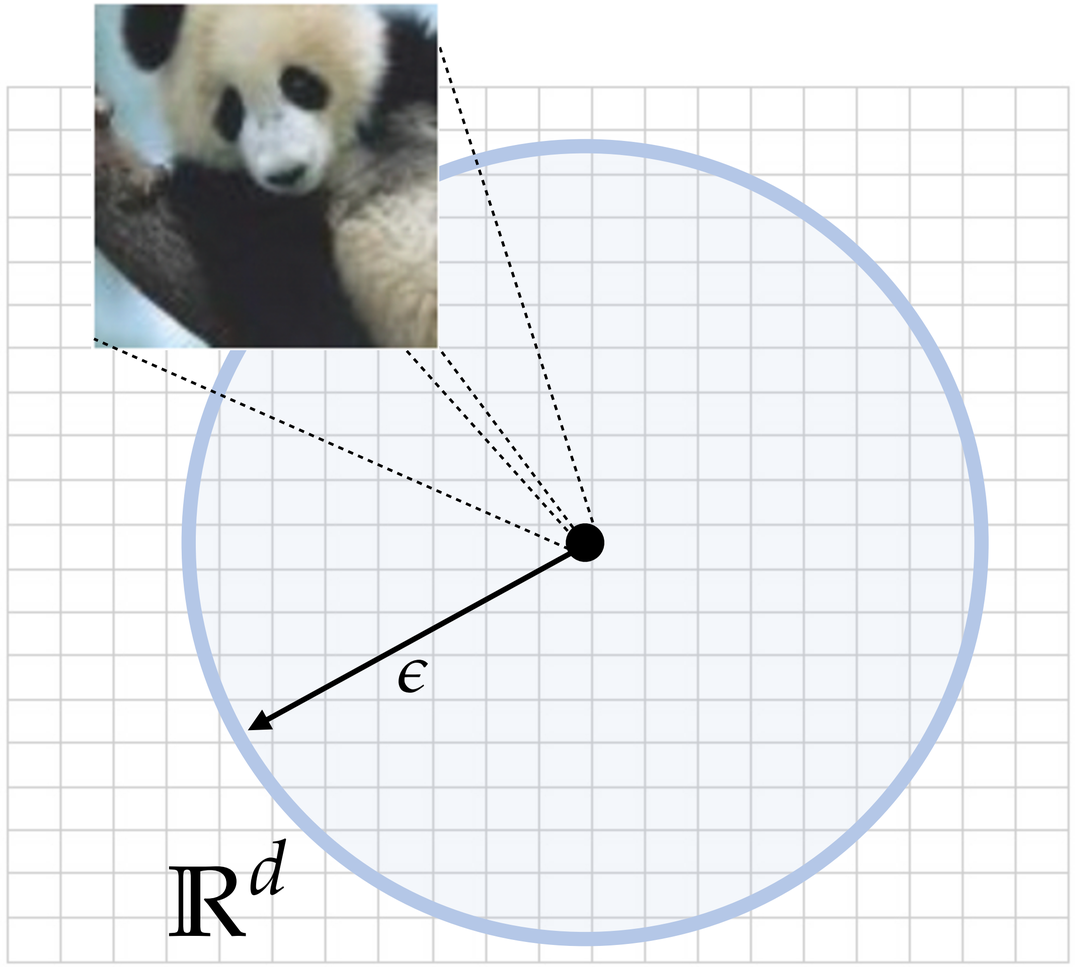




[Biggio et al 2014] [Szegedy et al 2014]

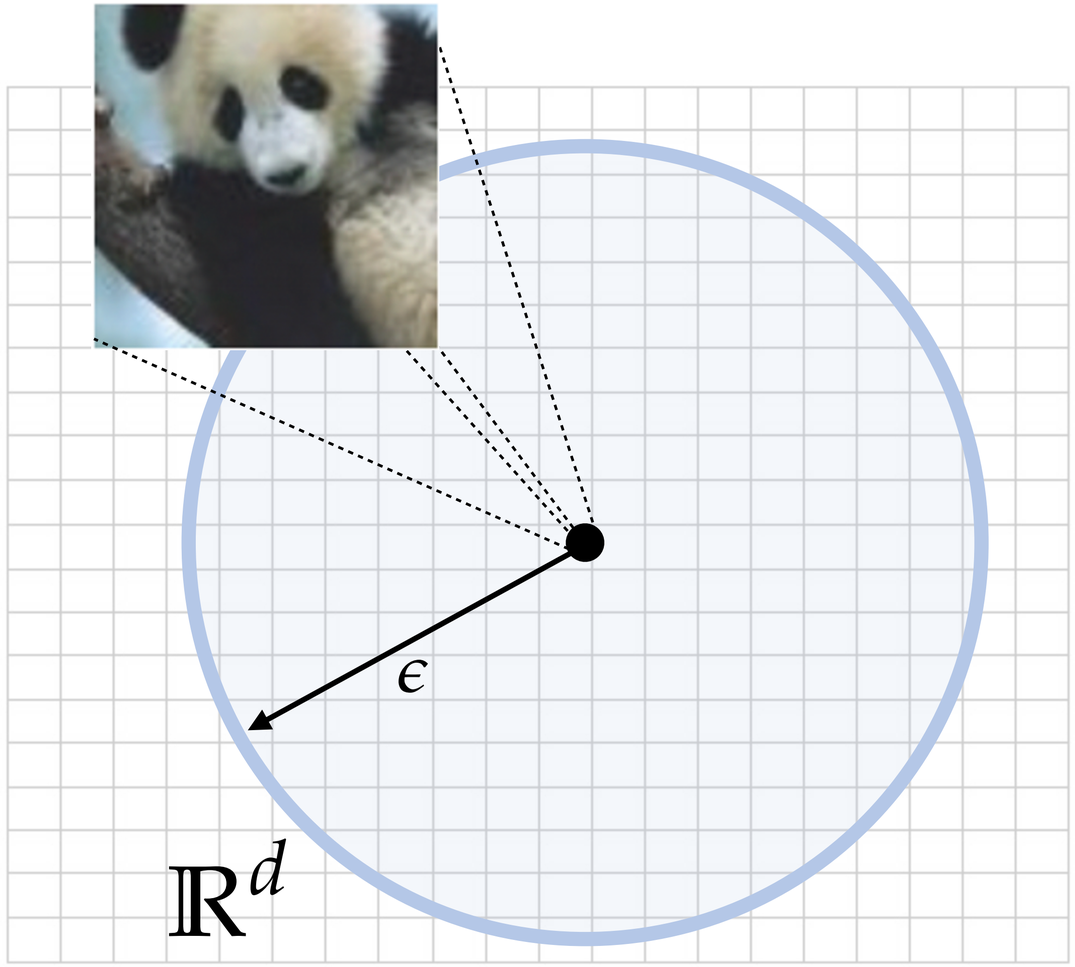
Adversarial examples: a brief introduction

Adversarial examples: a brief introduction



[Goodfellow et al. 2014]

Adversarial examples: a brief introduction

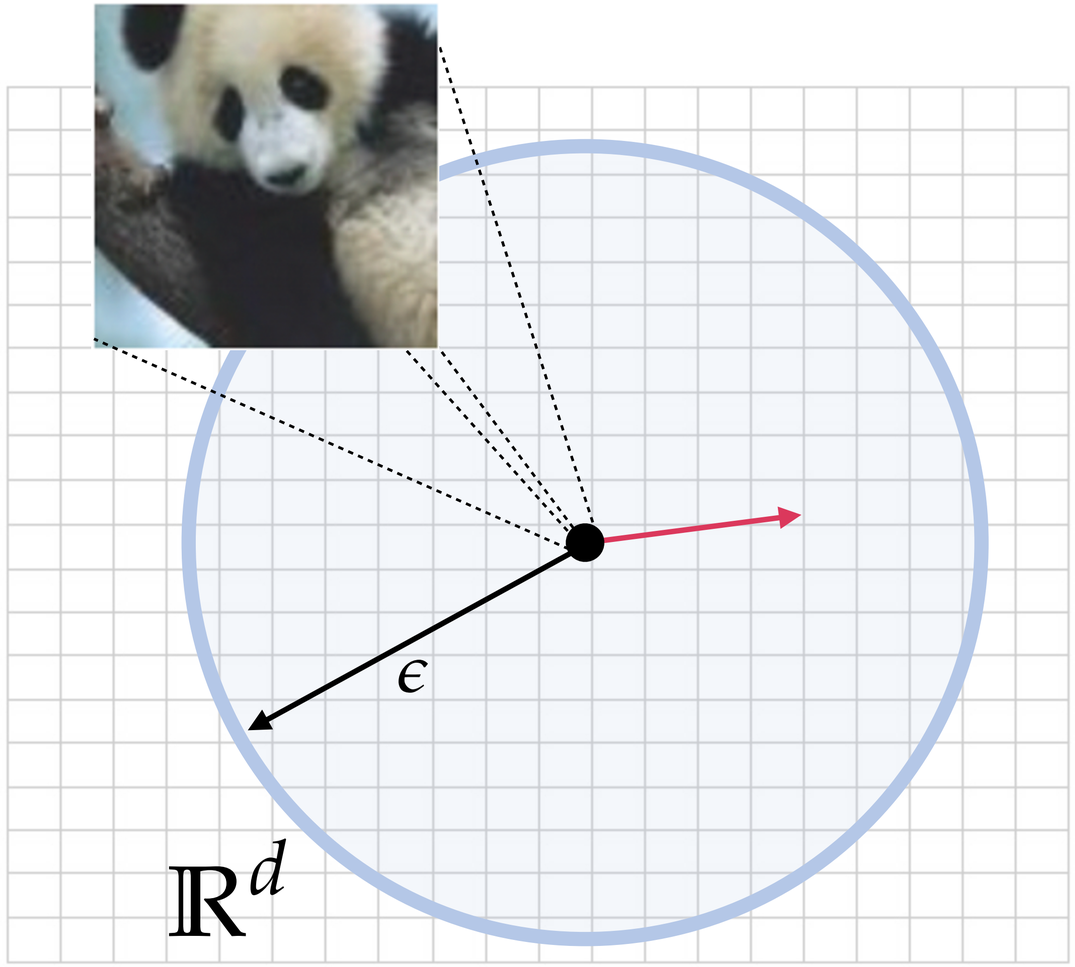


Image

x

[Goodfellow et al. 2014]

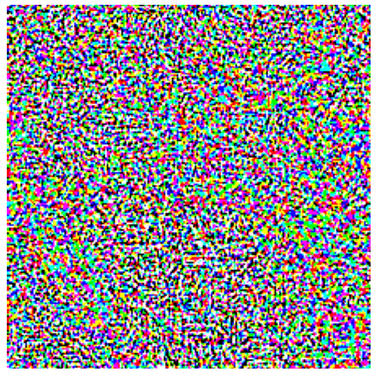
Adversarial examples: a brief introduction



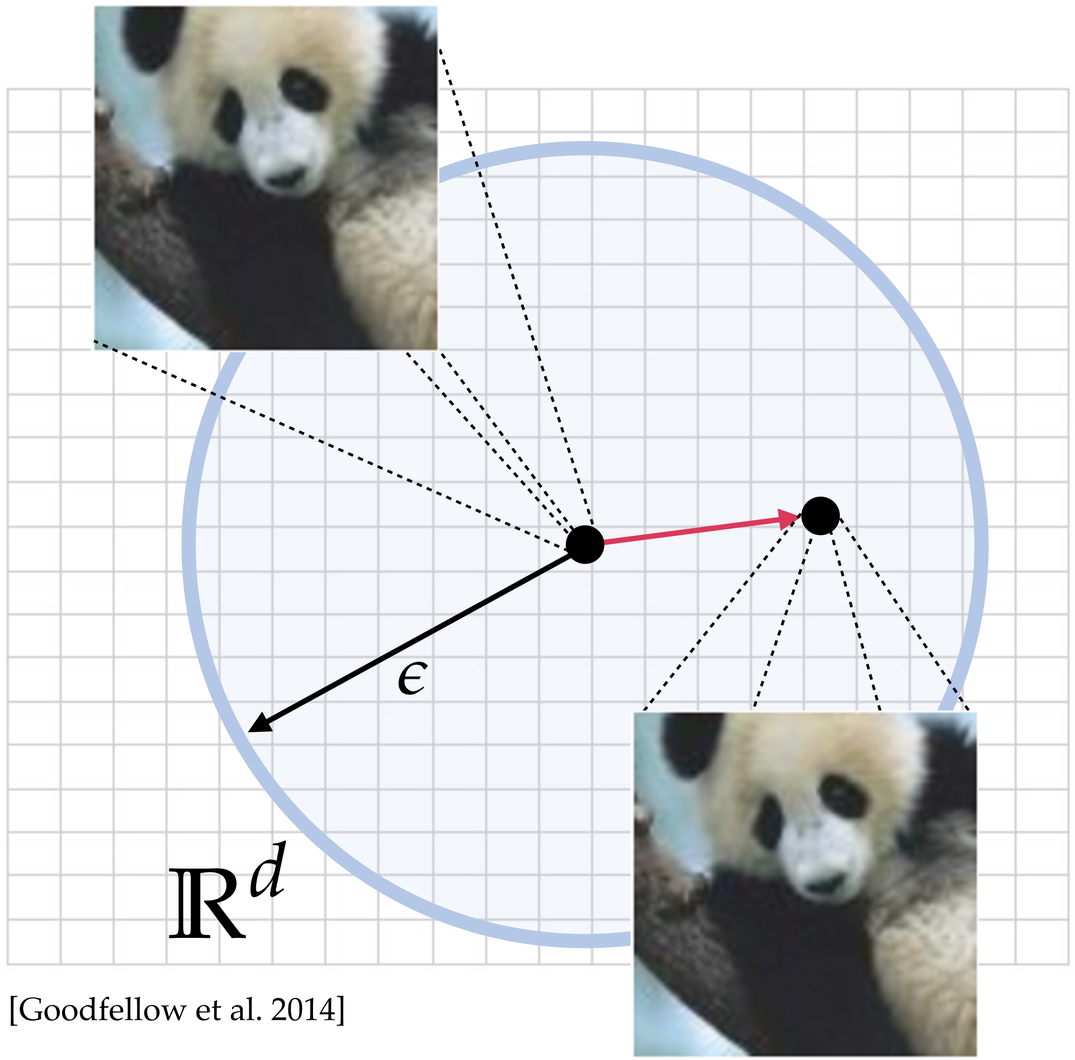
[Goodfellow et al. 2014]

Image Noise

$x + \delta$



Adversarial examples: a brief introduction

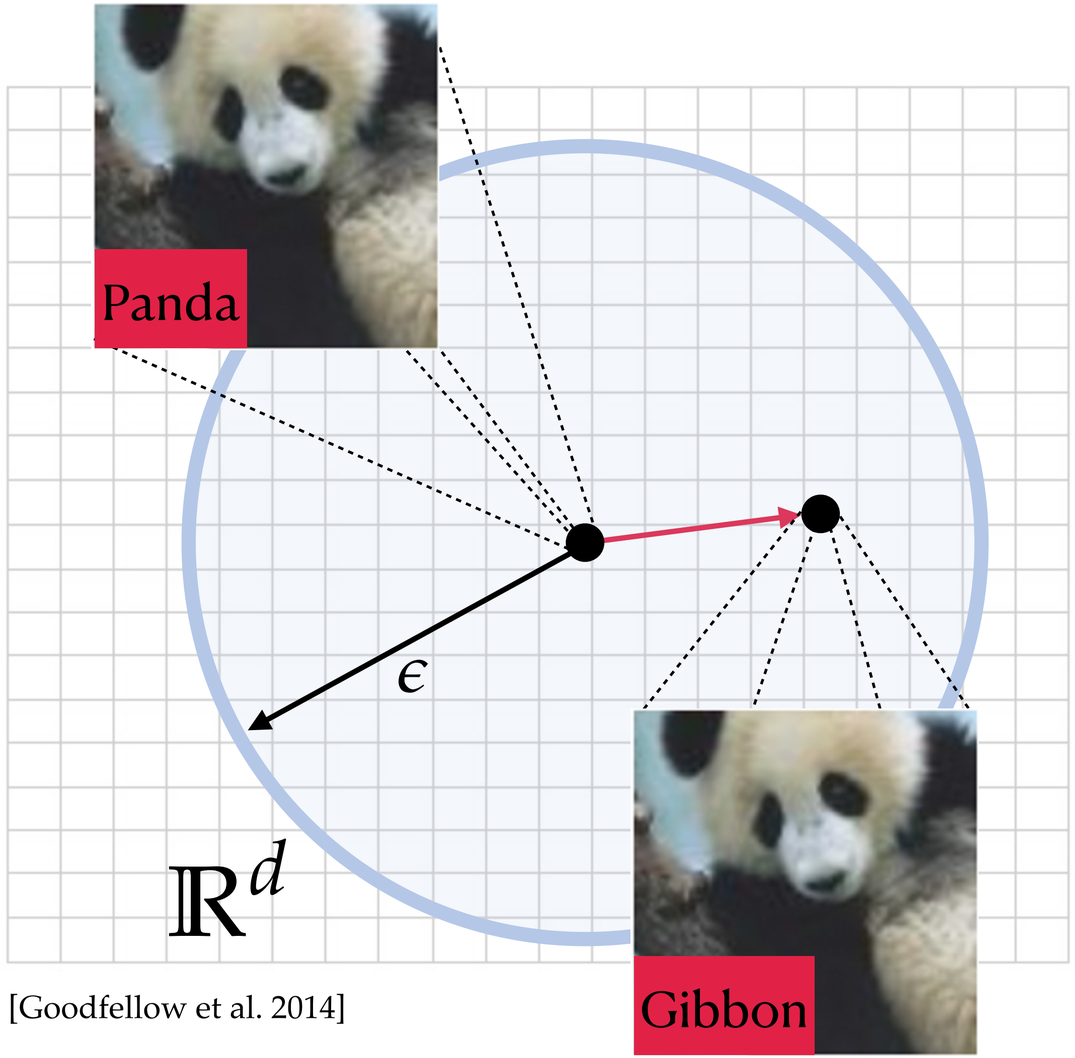


[Goodfellow et al. 2014]

Image + Noise = Adversarial example

$$x + \delta = x'$$

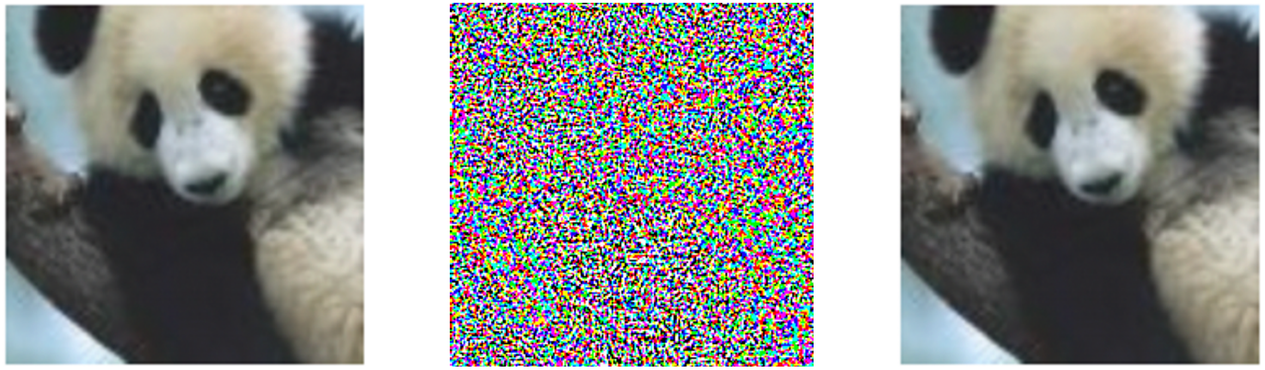
Adversarial examples: a brief introduction



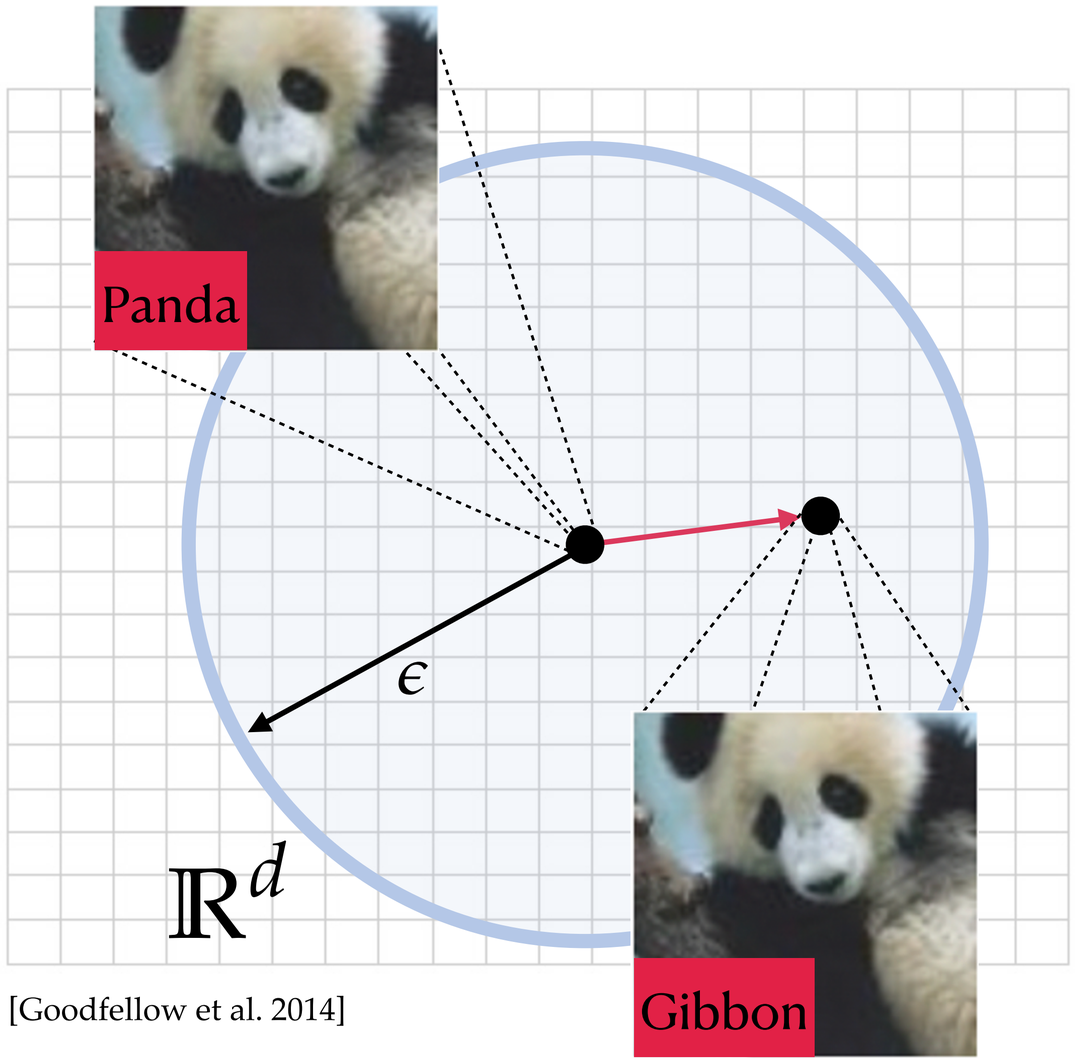
[Goodfellow et al. 2014]

Image + Noise = Adversarial example

$$x + \delta = x'$$



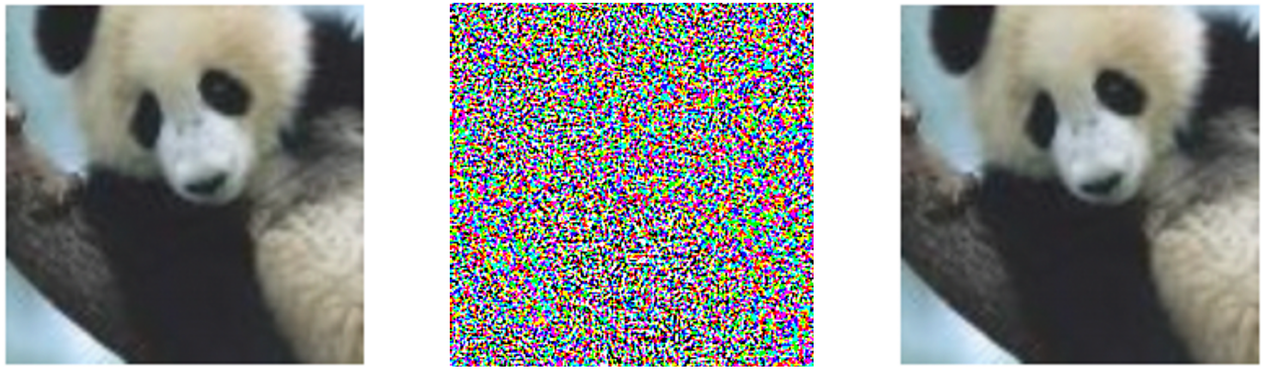
Adversarial examples: a brief introduction



[Goodfellow et al. 2014]

Image + Noise = Adversarial example

$$x + \delta = x'$$



Adding small amounts of noise can cause misclassification

Adversarial examples: problem setting

Adversarial examples: problem setting

Supervised Learning:

data: $(x, y) \sim \mathcal{D}$

problem: $\theta^* \in \arg\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\ell(x, y; \theta) \right]$

Adversarial examples: problem setting

Supervised Learning:

data: $(x, y) \sim \mathcal{D}$

problem: $\theta^* \in \arg\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\ell(x, y; \theta) \right]$

training data:

$(x_1, y_1), \cdots, (x_n, y_n) \sim \mathcal{D}$

ERM:

$$\hat{\theta} = \arg\min_\theta \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, y_i; \theta)$$

Adversarial examples: problem setting

Supervised Learning:

data: $(x, y) \sim \mathcal{D}$

problem: $\theta^* \in \arg\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\ell(x, y; \theta) \right]$

training data:

$(x_1, y_1), \cdots, (x_n, y_n) \sim \mathcal{D}$

ERM:

$$\hat{\theta} = \arg\min_\theta \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, y_i; \theta)$$

$\hat{\theta}$ works well on test data $(x, y) \sim \mathcal{D}$

Adversarial examples: problem setting

Supervised Learning:

data: $(x, y) \sim \mathcal{D}$

problem: $\theta^* \in \arg\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\ell(x, y; \theta) \right]$

training data:

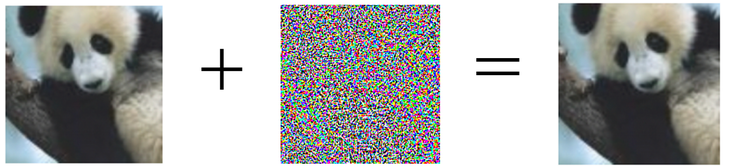
$(x_1, y_1), \cdots, (x_n, y_n) \sim \mathcal{D}$

ERM:

$$\hat{\theta} = \arg\min_\theta \frac{1}{n} \sum_{i=1}^{n} \ell(x_i, y_i; \theta)$$

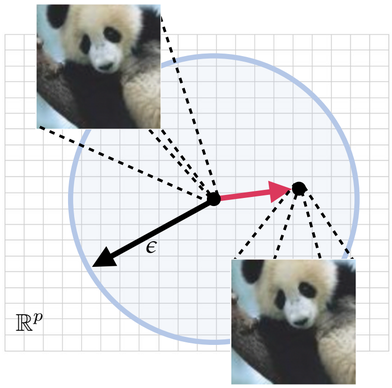
$\hat{\theta}$ works well on test data $(x, y) \sim \mathcal{D}$

but fails badly on adversarial examples



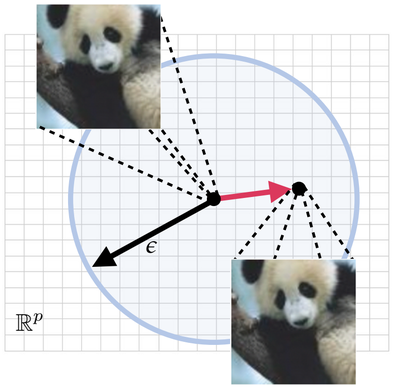
Adversarial examples: problem setting

Adversarial Learning:



Adversarial examples: problem setting

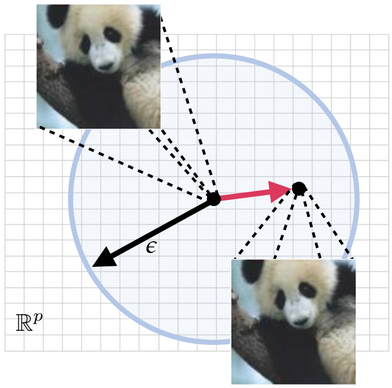
Adversarial Learning:



data: $(x, y) \sim \mathcal{D}$

Adversarial examples: problem setting

Adversarial Learning:

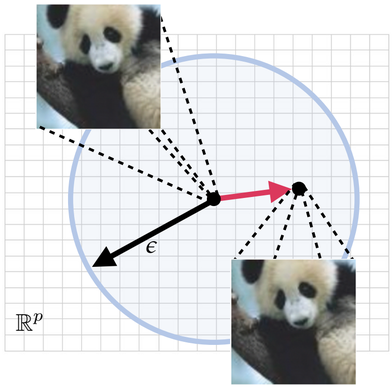


data: $(x, y) \sim \mathcal{D}$

problem: $\theta^*_{\text{adv}} \in \arg\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{||\delta|| \leq \epsilon} \ell(x + \delta, y; \theta) \right]$

Adversarial examples: problem setting

Adversarial Learning:



data: $(x, y) \sim \mathcal{D}$

problem: $\theta^*_{\text{adv}} \in \arg\min_\theta \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\max_{||\delta|| \leq \epsilon} \ell(x + \delta, y; \theta) \right]$

training data:

$(x_1, y_1), \cdots, (x_n, y_n) \sim \mathcal{D}$

Robust-ERM:

$$\hat{\theta}^\epsilon \in \arg\min_\theta \frac{1}{n} \sum_{i=1}^{n} \max_{||\delta_i|| \leq \epsilon} \ell(x_i + \delta_i, y_i; \theta)$$

[Madry et al. 2017, Tsipras et al. 2018]

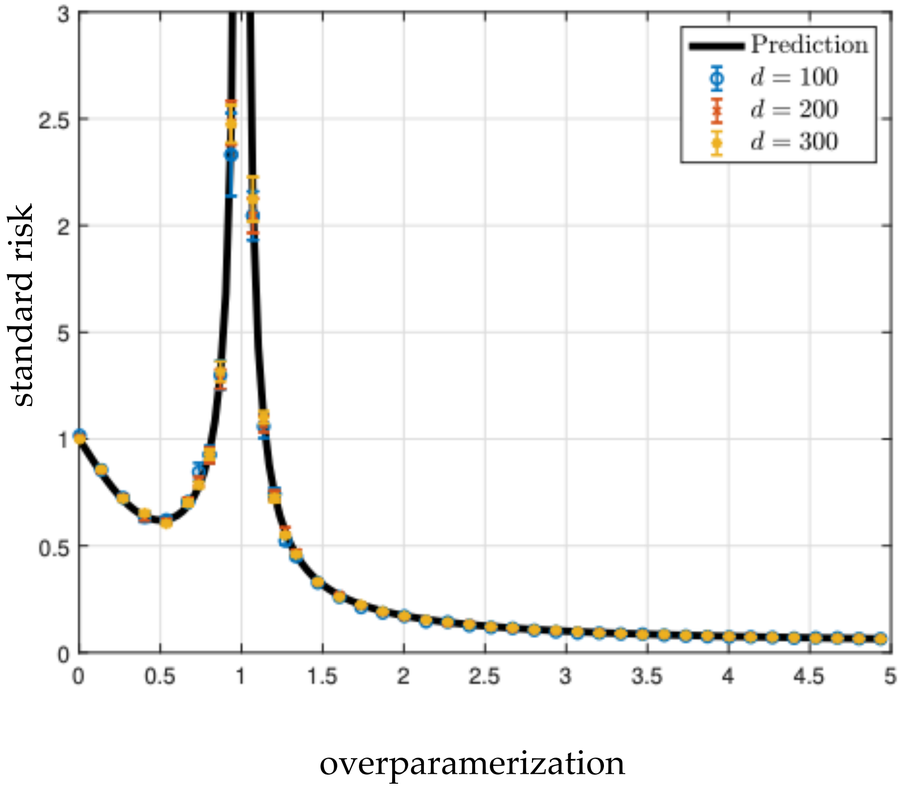
How Does Overparametrization Affect Robustness?

How Does Overparametrization Affect Robustness?

ERM (standard error, no adversary)

How Does Overparametrization Affect Robustness?

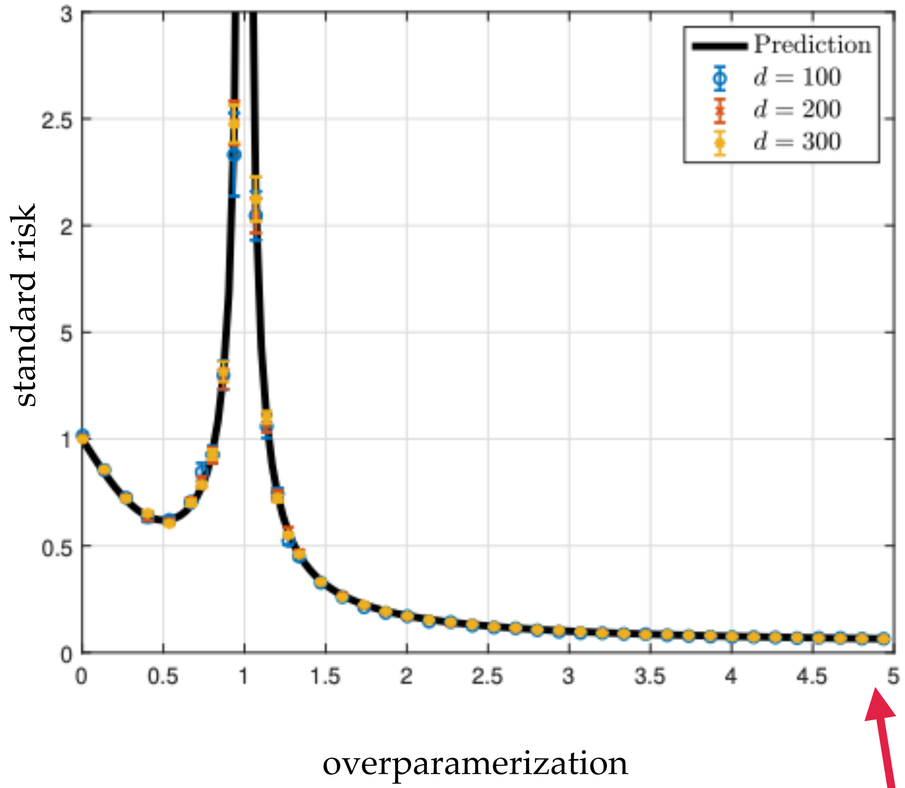
ERM (standard error, no adversary)



[Mei, Montanari '19]

How Does Overparametrization Affect Robustness?

ERM (standard error, no adversary)



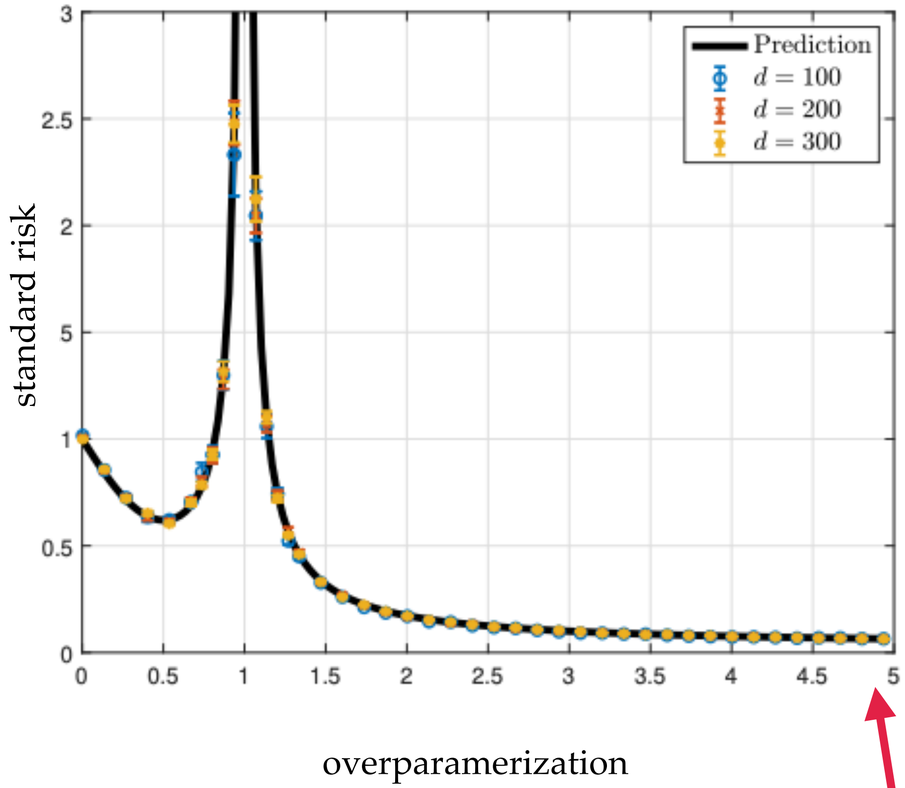
[Mei, Montanari '19]

global minimum

(infinite overparam)

How Does Overparametrization Affect Robustness?

ERM (standard error, no adversary)



[Mei, Montanari '19]

global minimum

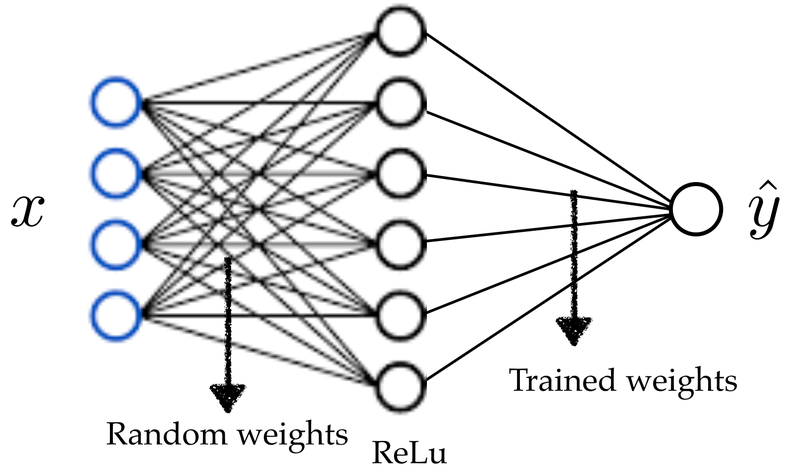
(infinite overparam)

Robust-ERM (with adversary)

?

Random Features Models

- Two-layer Neural Networks:

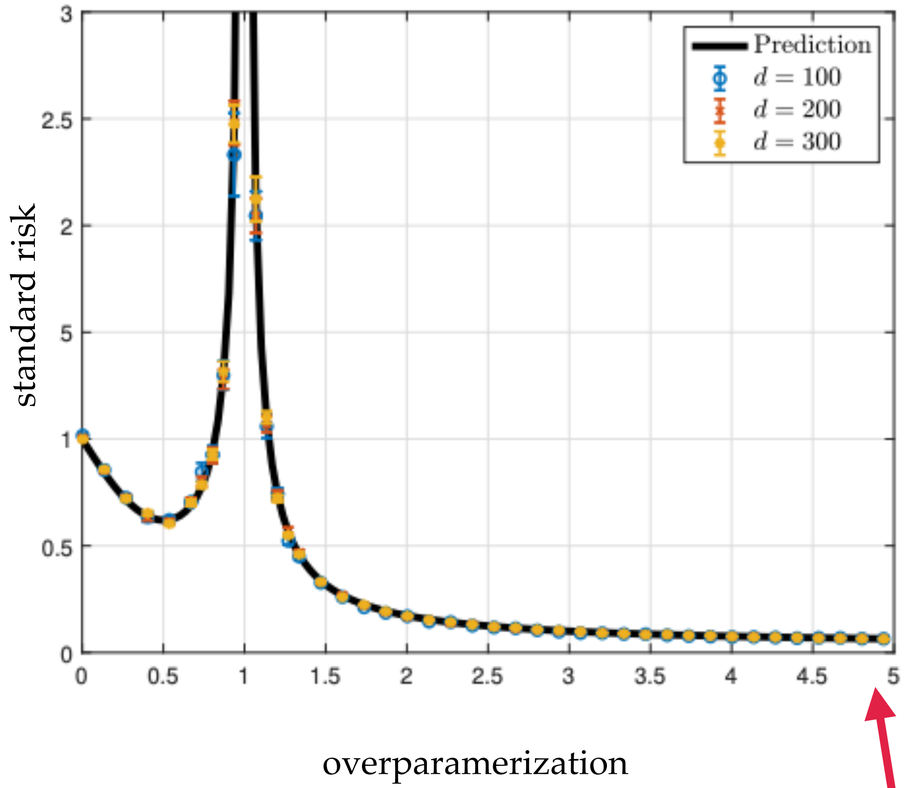


- Same setting as before: gaussian data, ℓ_2 adversarial perturbations

- The model is trained with robust-ERM

How Does Overparametrization Affect Robustness?

ERM (standard error, no adversary)



[Mei, Montanari '19]

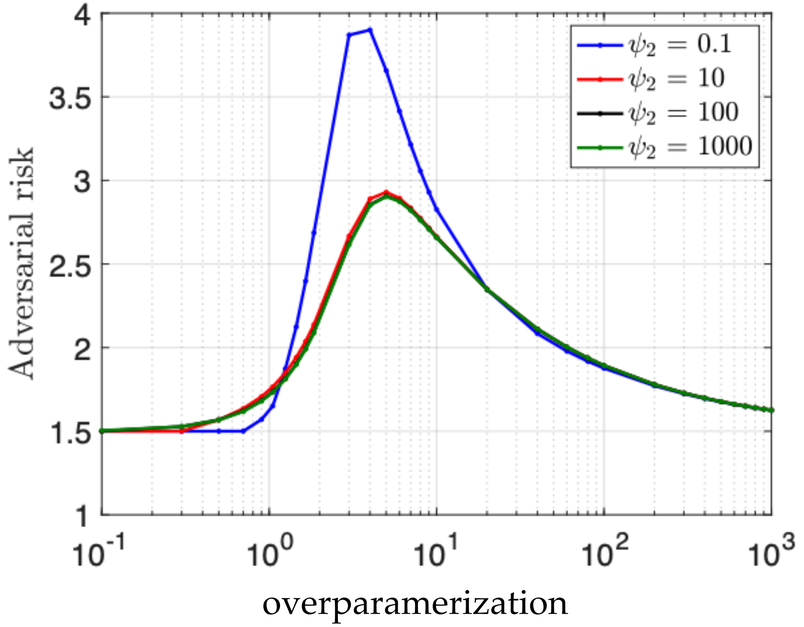
global minimum

(infinite overparam)

Robust-ERM (with adversary)

?

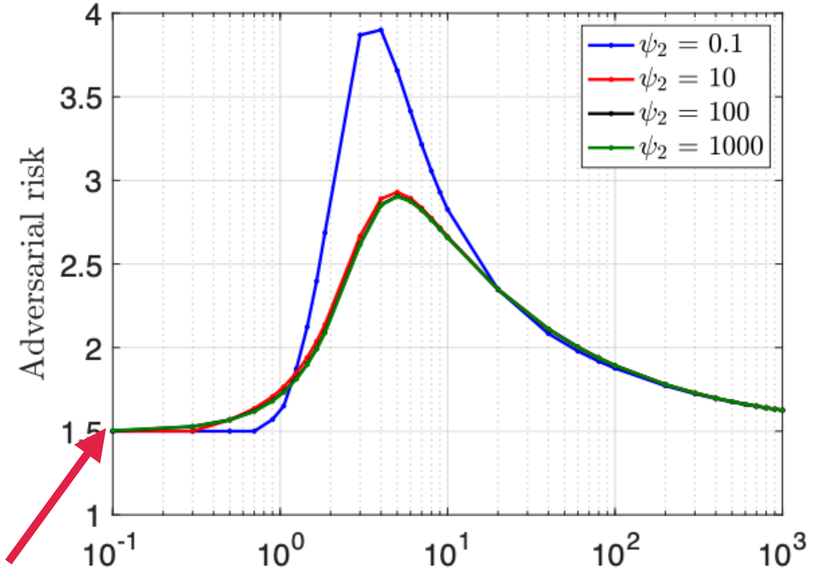
Overparametrization Can Hurt!



$\epsilon = 1$

[Hassani, Javanmard '22]

Overparametrization Can Hurt!



global minimum

(zero overparam)

$\epsilon = 1$

[Hassani, Javanmard '22]