Projet Sciences des données
Adversarial examples

Geovani Rizk

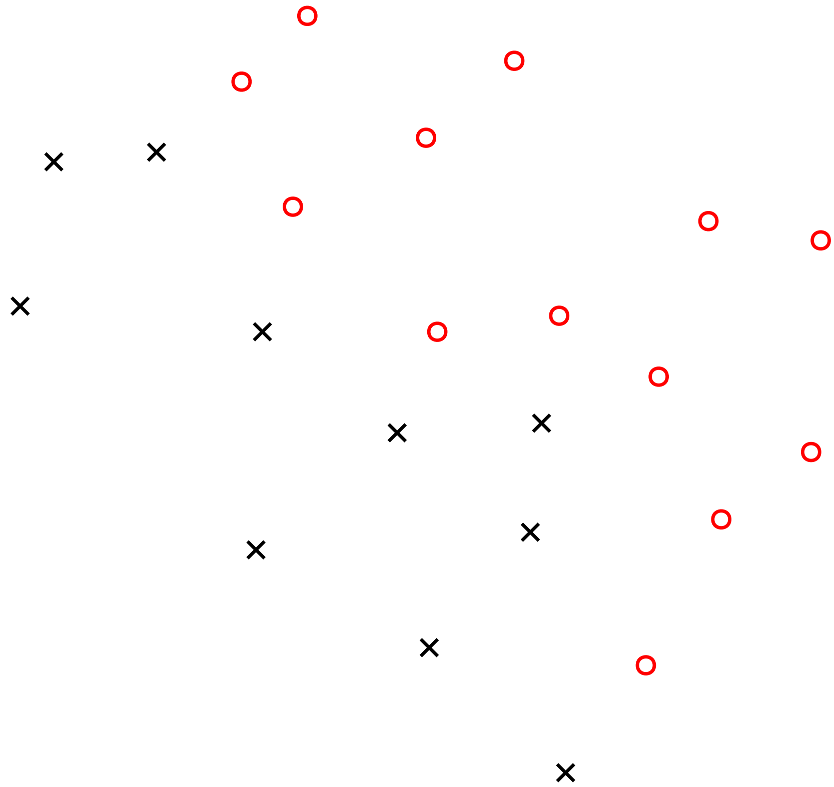
Université Paris Dauphine - PSL

September 23, 2020

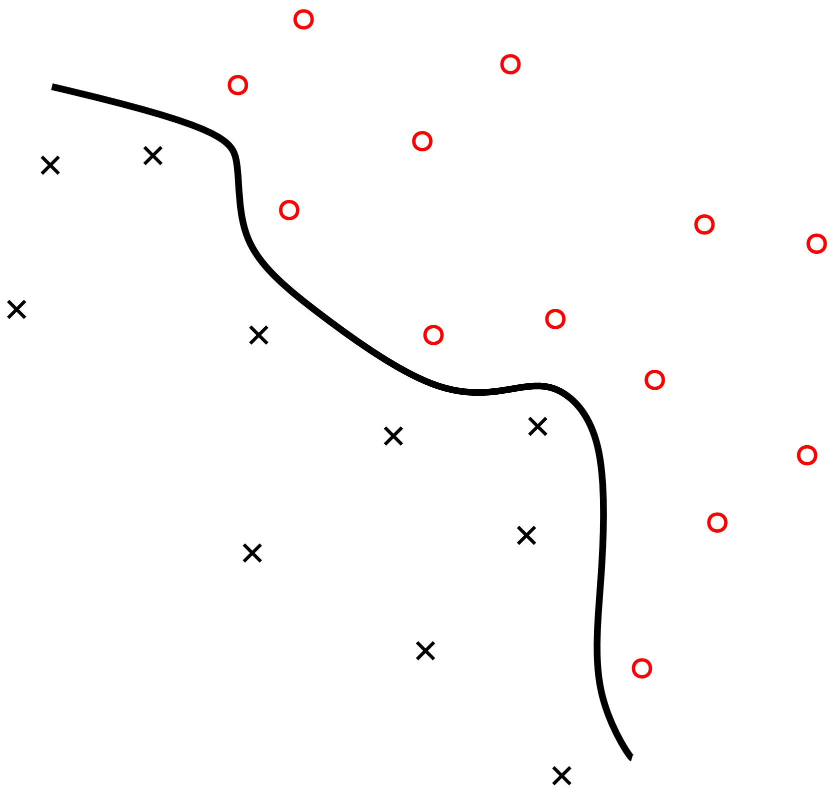
Table of contents

Table of contents

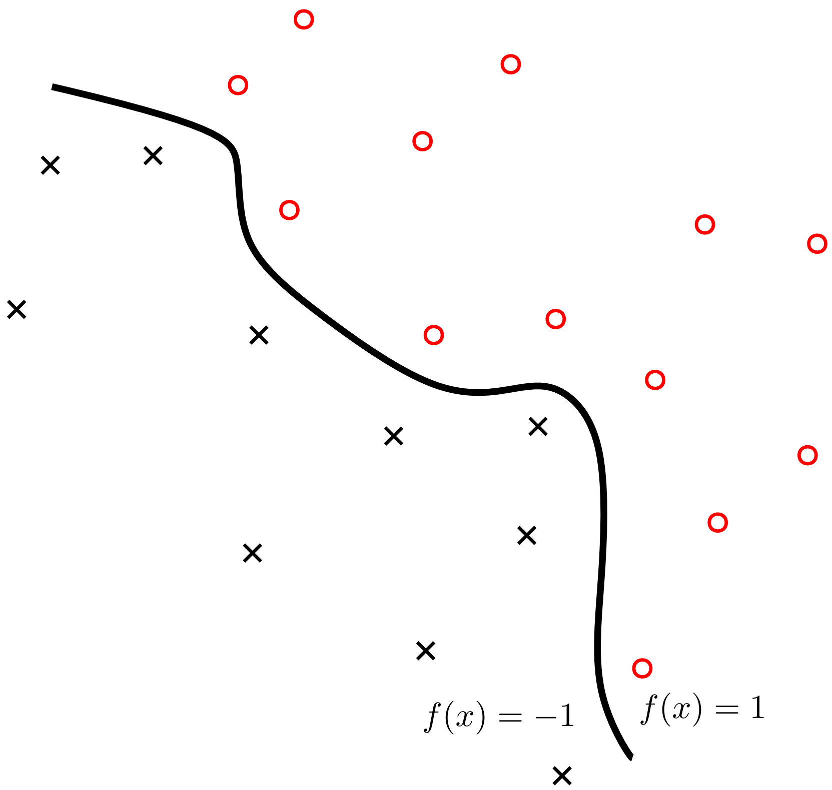
Adversarial Attacks



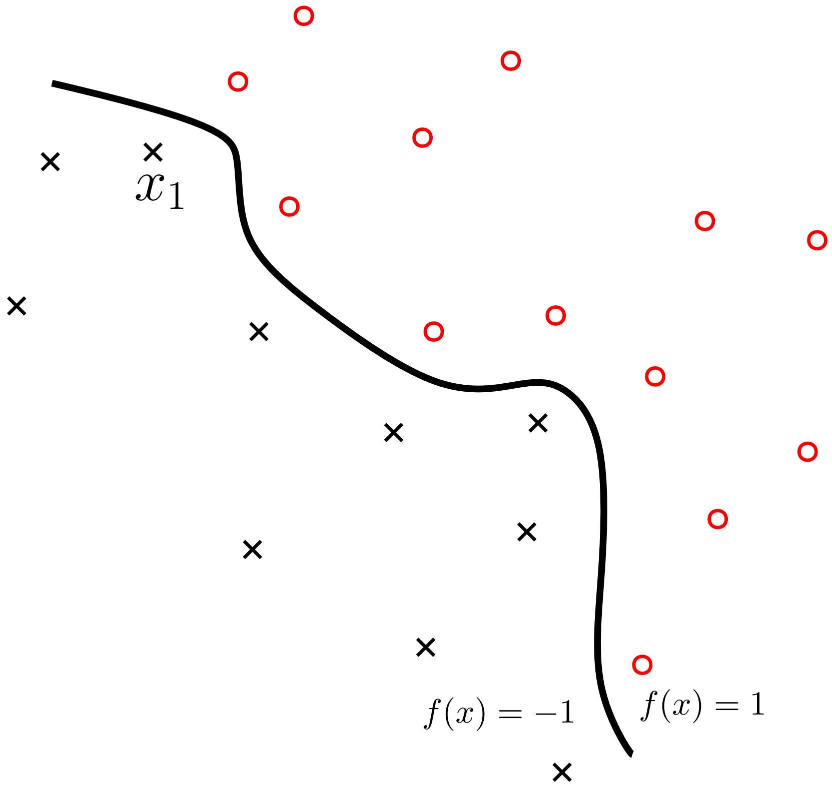
Adversarial Attacks



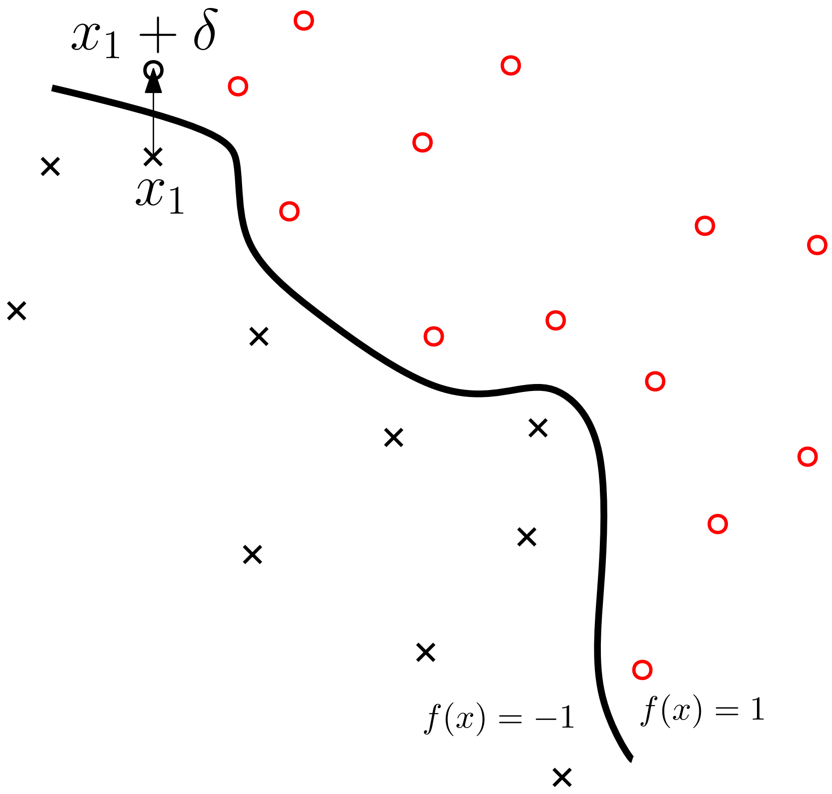
Adversarial Attacks



Adversarial Attacks



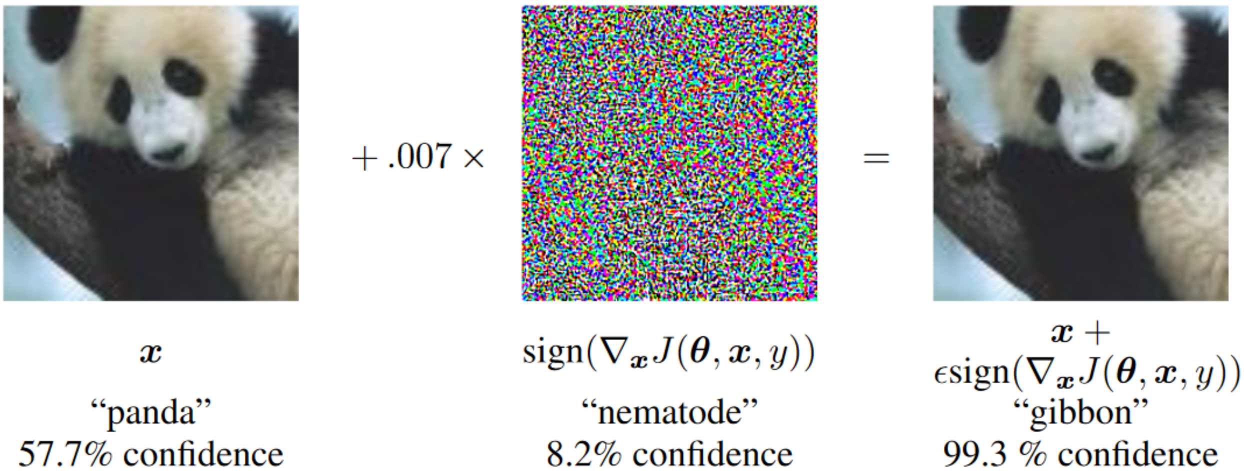
Adversarial Attacks



Adversarial Attacks

What if δ is imperceptible ?

Adversarial Attacks in Image recognition



Source : *Explaining and Harnessing Adversarial Examples*, Goodfellow et al, ICLR 2015.

Adversarial Attacks in Image recognition

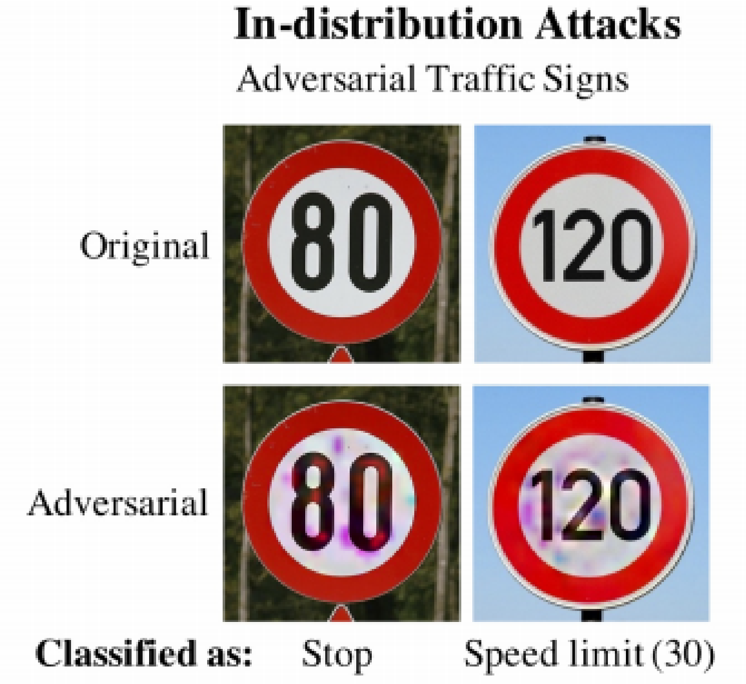


Figure 1: Adversarial traffic signs (Sitawarin, Bhagoji et al., 2018)

Adversarial Attacks

To be imperceptible, the norm of the perturbation is bounded

We define an $\epsilon \in \mathbb{R}$ such that $\|\delta\|_p \leq \epsilon$.
In practice, we use ℓ_2 and ℓ_∞ norm to bound the perturbation.

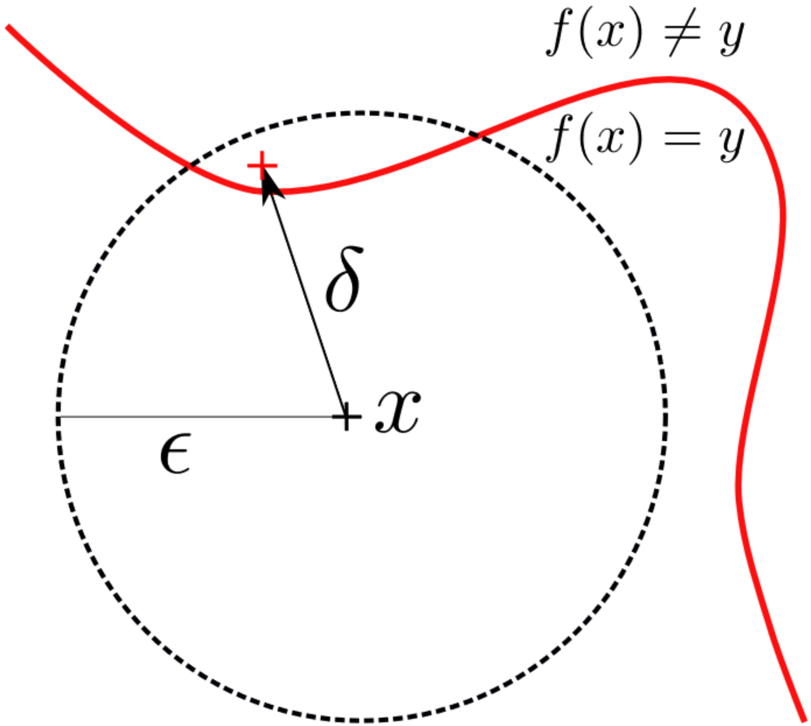
Generating a adversarial example

Let $f : \mathbb{R}^d \rightarrow \mathcal{Y}$ be a classifier. Given an example $x \in \mathcal{X} \subset \mathbb{R}^d$ and its true label $y \in \mathcal{Y}$, the goal is to find $\delta \in \mathbb{R}^d$ such that :

Untargeted attacks
$\|\delta\|_p \leq \epsilon$ and $f(x + \delta) \neq y$

Targeted attacks
$\|\delta\|_p \leq \epsilon$ and $f(x + \delta) = t$ with $t \neq y$

Generating an adversarial example with ℓ_2-norm



Generating an adversarial example with ℓ_2-norm

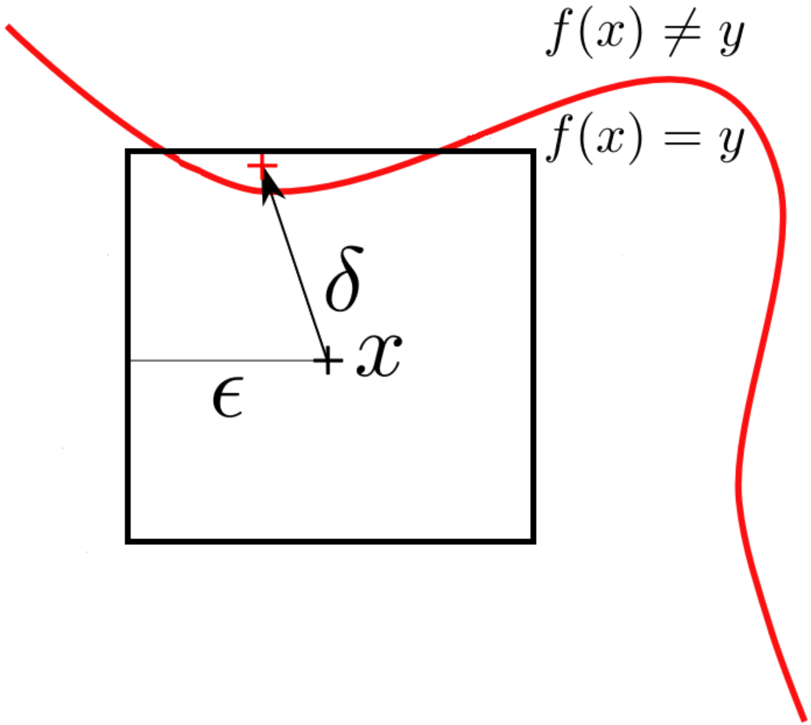


Table of contents

ℓ_∞-PGD Attack

ℓ_∞-PGD is an iterative method that constructs the perturbed data as follows :

1. $x_0 \leftarrow x$
2. repeat n times :

$$x_{t+1} = \Pi_{B_\infty(x,\epsilon)}\left(x_t + \eta sign(\nabla_x L_\theta(x_t, y))\right)$$

Paper :

[3] Towards Deep Learning Models Resistant to Adversarial Attacks, Madry et. al, ICLR 2018.

ℓ_2-Carlini & Wagner

For a given example $x \in \mathcal{X}$ of the class $y \in \mathcal{Y}$, the ℓ_2 Carlini & Wagner attack (C&W) aims to resolve the following optimization problem :

$$\min_{x+\delta} c\|\delta\|_2 + g(x+\delta) \tag{1}$$

where $g(x+\delta) \leq 0$ iff $f(x+\delta) \neq y$. You can find the different functions g in the paper :

[1] Towards Evaluating the Robustness of Neural Networks, Carlini and Wagner, IEEE 2017.

Table of contents

Adversarial Training

Adversarial training is a method that aims to optimize (Goodfellow, 2015) :

$$\min_{\theta} \mathbb{E}_{(x,y)} \left(\max_{\|\delta\|_p \leq \epsilon} L_{\theta} \left(x + \delta, y \right) \right) \tag{2}$$

To solve the inner maximization problem, we use in practice PGD attack. ([3] Madry et al. 2017)

Randomized Networks

An other defense is to inject noise into the input data during the training and inference phases(Cohen, 2019; Pinot et al., 2019). It is shown that predicting $\mathbb{E}_{\eta}\left(f(x+\eta)\right)$, where η is the injected noise, brings more robustness.

Papers :

[2] Certified adversarial robustness via randomized smoothing, Cohen et. al, ICML 2019.
[4] Theoretical evidence for adversarial robustness through randomization, Pinot et. al, NeurIPS 2019.
[5] Randomization matters. How to defend against strong adversarial attacks, Pinot et. al, ICML 2020.

Projects

- Datasets : (MNIST,) CIFAR10
- First part (common) :
  - Code ℓ_∞-PGD attack & Observe robustness of neural networks
  - Code Adversarial training & Observe robustness of neural networks
- Second part (choose one) :
  - Adversarial robustness competition: open project! The goal is to build the most robust classifier against classical attacks.
  - Adversarial Attacks competition: open project! The goal is to build the strongest attack.

Evaluation for this projet

You are evaluated in groups of 2 or 3 or 4 students.

- A latex document explaining what you did in the **first and second part**.
- A **10 min** presentation on the **second part** of the project.
- Your code.

References I

[1] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. *arXiv preprint arXiv:1608.04644*, 2017.

[2] J. M. Cohen, E. Rosenfeld, and J. Z. Kolter. Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv:1902.02918*, 2019.

[3] A. Madry, A. Makelov, L. Schmidt, D. Tsipras, and A. Vladu. Towards deep learning models resistant to adversarial attacks. *arXiv preprint arXiv:1706.06083*, 2017.

[4] R. Pinot, L. Meunier, A. Araujo, H. Kashima, F. Yger, C. Gouy-Pailler, and J. Atif. Theoretical evidence for adversarial robustness through randomization. In *Advances in Neural Information Processing Systems*, pages 11838–11848, 2019.

[5] R. Pinot, R. Ettedgui, G. Rizk, Y. Chevaleyre, and J. Atif. Randomization matters. how to defend against strong adversarial attacks. *arXiv preprint arXiv:2002.11565*, 2020.