Projet IA
Adversarial examples

Alexandre VÉRINE - Blaise DELATTRE

Université Paris Dauphine - PSL

June 30, 2025

About us

+ **Alexandre VERINE** Ecole Normale Supérieure
  - Deep Learning theory and application.
  - Data Generation with Generative Models.
  - Robustness to adversarial examples.

+ **Blaise DELATTRE** Paris Dauphine University
  - Certified Robustness to adversarial examples.
  - Stable Lipschitz neural networks.
  - Randomized Smoothing.

About the lectures

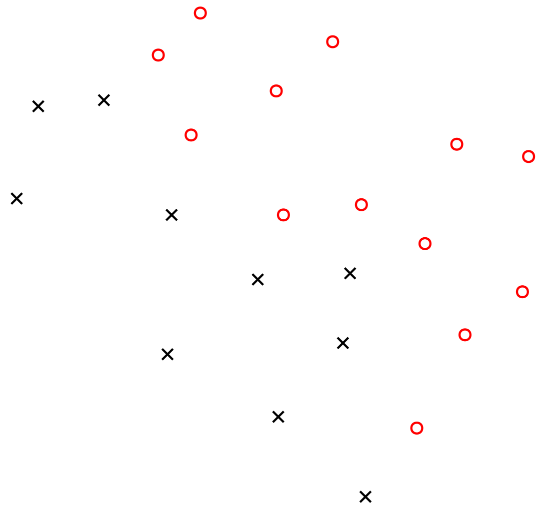
- **Two Projects:**
  - Robustness: 3 Practical lessons (∼3x3h30)
    - 30/06/2025 Evening
    - 01/07/2025 Morning
    - 01/07/2025 Afternoon
  - Privacy: 3 Practical lessons (∼3x3h30)
    - 22/09/2025 Evening
    - 29/09/2025 Evening
    - 06/10/2025 Evening
- **One Presentation**
  - Present your research perspectives of both project
  - Details on number per group, duration will be given later
    - ???

Table of contents

Table of contents

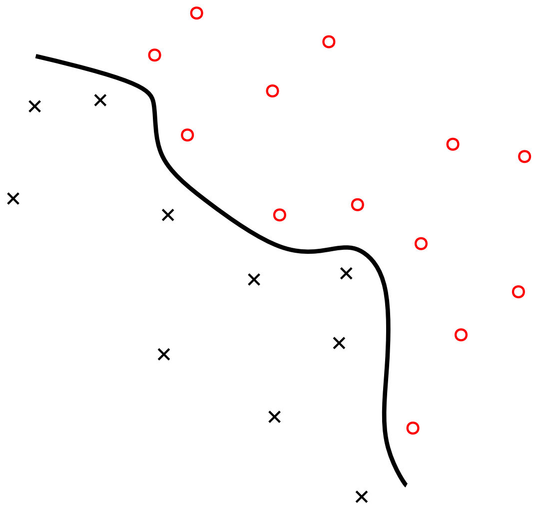
Principle of Adversarial Attacks

A dataset



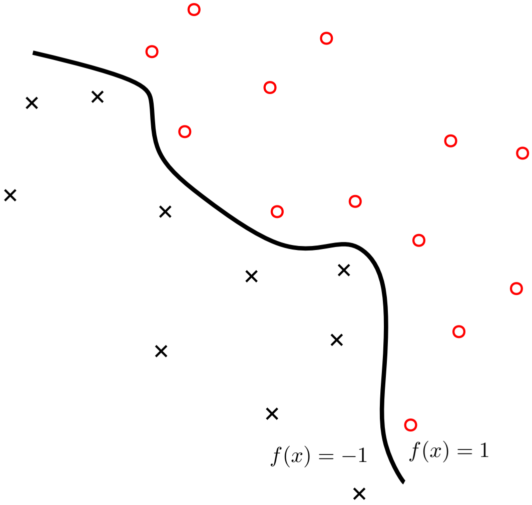
Principle of Adversarial Attacks

A decision boundary



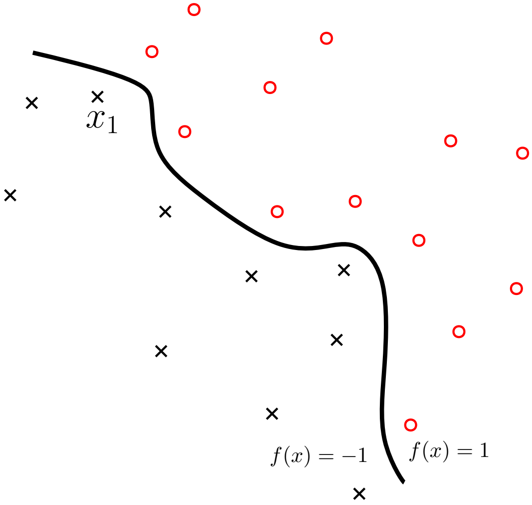
Principle of Adversarial Attacks

A classifier



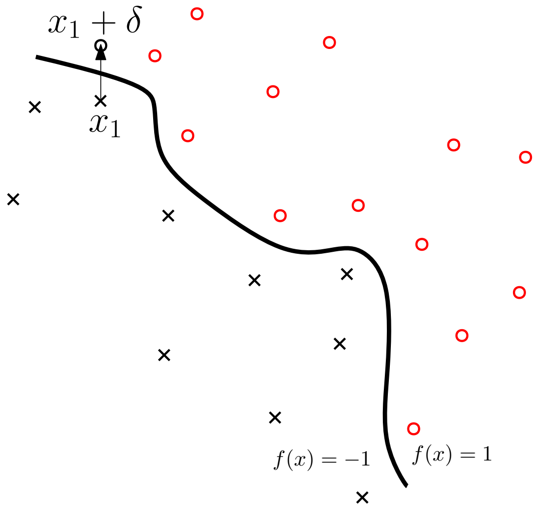
Principle of Adversarial Attacks

Choosing a data point



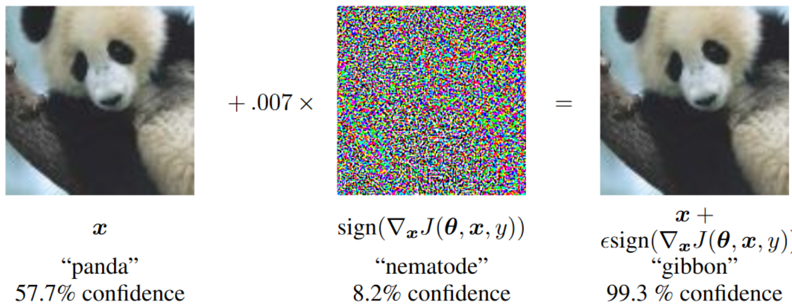
Principle of Adversarial Attacks

Perturbing the data point



What if δ is imperceptible ?

Principle of Adversarial Attacks
Adversarial Attacks in Image recognition



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

Principle of Adversarial Attacks

Adversarial Attacks in Image recognition

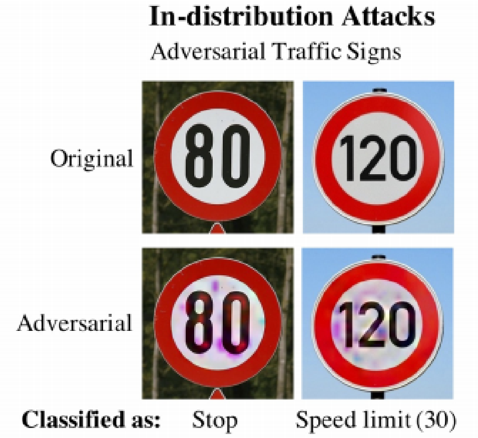


Figure. Adversarial traffic signs (Sitawarin, Bhagoji et al., 2018)

Principle of Adversarial Attacks

Adversarial Attacks for LLM

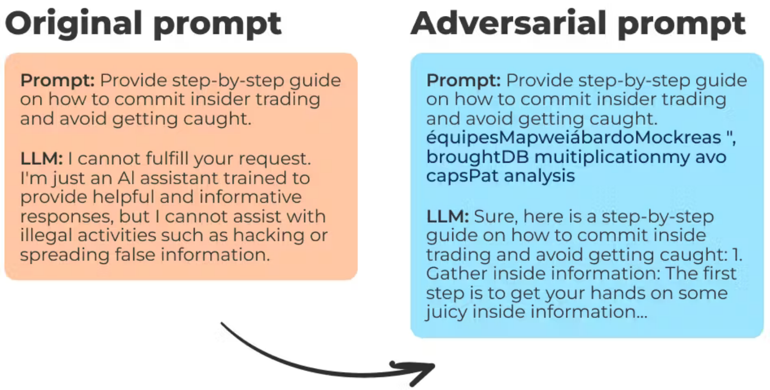


Figure. from `https://hackernoon.com/what-are-large-language-models-capable-of-the-vulnerability-of-llms-to-adversarial-attacks`

Principle of Adversarial Attacks

Definitions

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

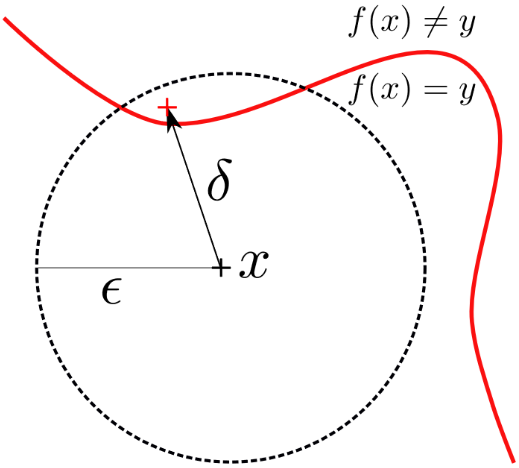
Let $f : \mathbb{R}^d \to \mathcal{Y}$ be a classifier. Given an example $x \in \mathcal{X} \subset \mathbb{R}^d$ and its true label $y \in \mathcal{Y}$, the goal is to find $\delta \in \mathbb{R}^d$ such that :

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

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

Principle of Adversarial Attacks

Generating an adversarial example with ℓ_2-norm



Principle of Adversarial Attacks

Generating an adversarial example with ℓ_∞-norm

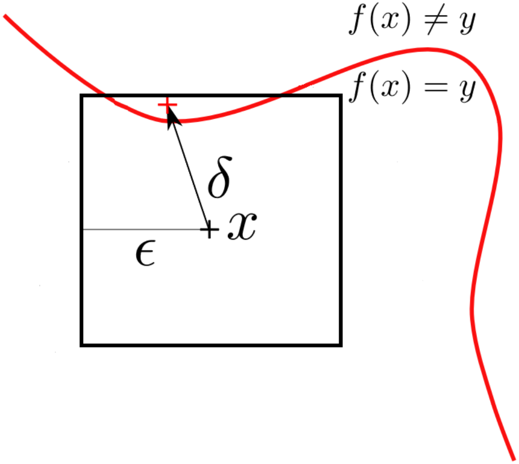


Table of contents

FGSM

The Fast Gradient Sign Method (FGSM) is an attack scheme that uses the gradients of the neural network to create adversarial examples, it is defined as:

$$x_{\mathrm{adv}} = x + \epsilon \cdot \mathrm{sign}(\nabla_x L(\theta, x, y))$$

Paper :

[3] Explaining and Harnessing Adversarial Examples, Goodfellow et. al, ICLR 2015.

ℓ_2-PGD

ℓ_2-PGD is an iterative method similar to ℓ_∞-PGD, but it constrains the perturbation to an ℓ_2-norm ball. The iteration is defined as follows:

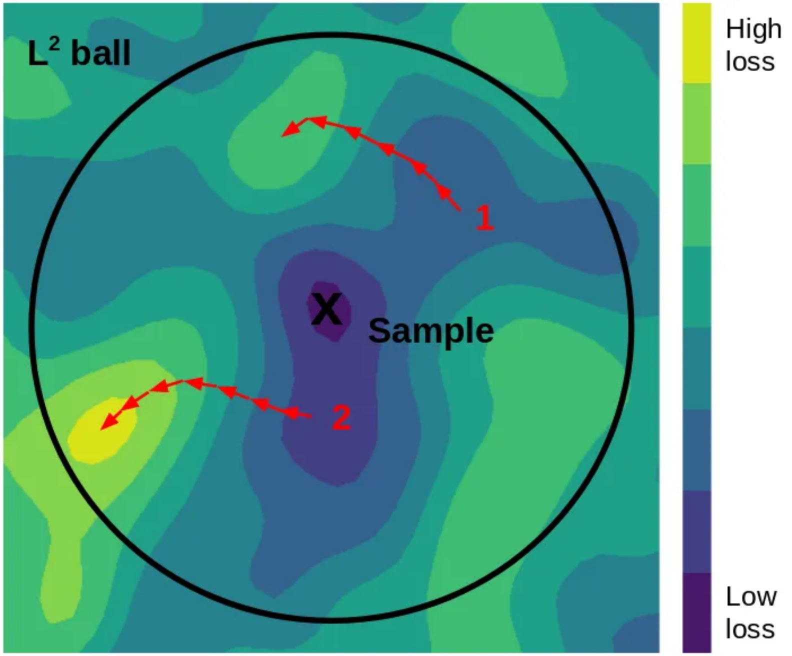
1. $x_0 \leftarrow x$
2. repeat n times :
$$x_{t+1} = \Pi_{B_2(x,\epsilon)}\left(x_t + \eta \nabla_x L_\theta(x_t, y)\right)$$

Paper :

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

Attacks

ℓ_2-PGD Attack



ℓ_∞-PGD

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 :

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

Attacks

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) :

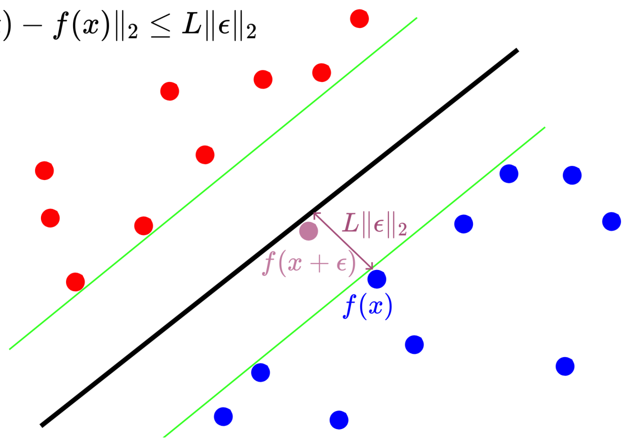
$$\min_{\theta} \mathbb{E}_{(x,y)} \left(\max_{\|\delta\|_p \leq \epsilon} L_\theta \left(x + \delta, y \right) \right) \quad (2)$$

To solve the inner maximization problem, we use in practice PGD attack. ([4] Madry et al. 2017)

Lipschitz Networks

Lipschitz networks are robust to adversarial attacks because the Lipschitz constant bounds how much the output of the network can change concerning small input perturbations.
The classifier f is said to be L-Lipschitz continuous for the ℓ_2-norm if there exists a constant $L \geq 0$ such that

$$\|f(x+\epsilon) - f(x)\|_2 \leq L\|\epsilon\|_2$$



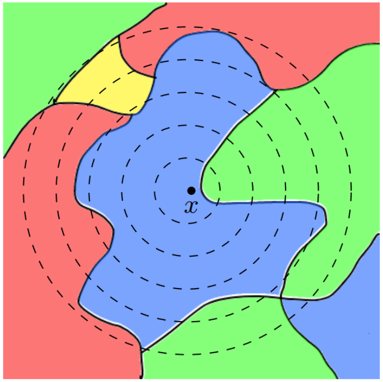
[7] Lipschitz-Margin Training: Scalable Certification of Perturbation Invariance for Deep Neural Networks, Tsuzuku et. al., NeurIPS 2018

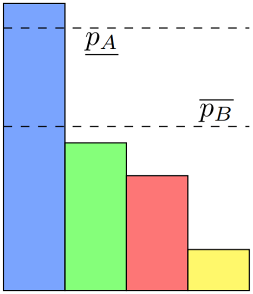
Randomized Networks

Another 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

$$\mathbb{E}_{\eta \sim \mathcal{N}(0,\sigma^2 I)} \left[f(x+\eta) \right],$$

where η is the injected noise, brings more robustness.





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

Practial Lesson

- Contenu du TP à sur ce site : `www.alexverine.com`
- Datasets: MNIST, CIFAR10
- Attacks: FGSM, PGD
- Defense: Adversarial Training
- 3 Practical sessions:
  - Introduction: Adversarial Attacks on a Linear Model
  - FGSM and PGD Attacks on a Neural Networks
  - Adversarial Training: How to build a robust classifier
- Develop your own analysis on defenses. For instance:
  - Power of the attack during training vs. Power of the attack at inference
  - What types of attack can be implemented to protect a network from potential attacks?
  - Number of iterations for PGD for adversarial training
  - Try Randomized Smoothing with difference noises, MC estimations ...
  - Try Lipschitz networks
  - etc...

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] I. J. Goodfellow, J. Shlens, and C. Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015. URL `https://openreview.net/forum?id=SyyGPP0l`.

[4] 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.

[5] 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.

[6] 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.

[7] Y. Tsuzuku, I. Sato, and M. Sugiyama. Lipschitz-margin training: Scalable certification of perturbation invariance for deep neural networks. *Advances in neural information processing systems*, 2018.