On the impact of randomization on robustness in Machine Learning

A probabilistic point of view on adversarial examples

Candidate: Rafael Pinot

Université Paris-Dauphine PSL & Institut CEA LIST

Machine Intelligence and Learning Systems

Date: 02/12/2020

PhD advisors: Jamal Atif
Cédric Gouy-Pailler
Florian Yger

Jury members: Stéphane Canu (Reviewer)
Panayotis Mertikopoulos (Reviewer)
Francis Bach (Examiner)
Sébastien Bubeck (Examiner)
Cordelia Schmid (Examiner)
Michèle Sebag (Examiner)

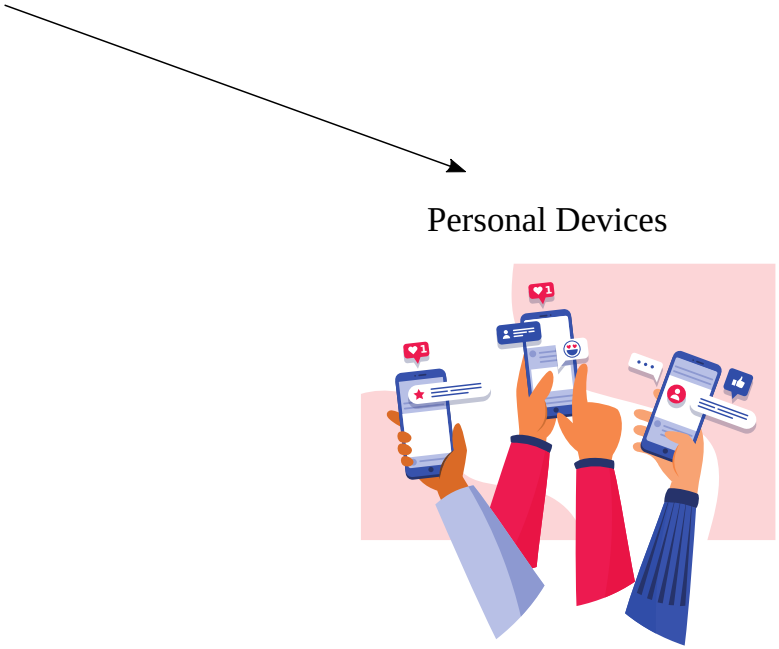
Context & motivations

Machine learning models are everywhere

- Machine learning models recently gave **outstanding results** (*e.g.* vision, NLP)
- Industries and governments are starting to use them in **critical applications**

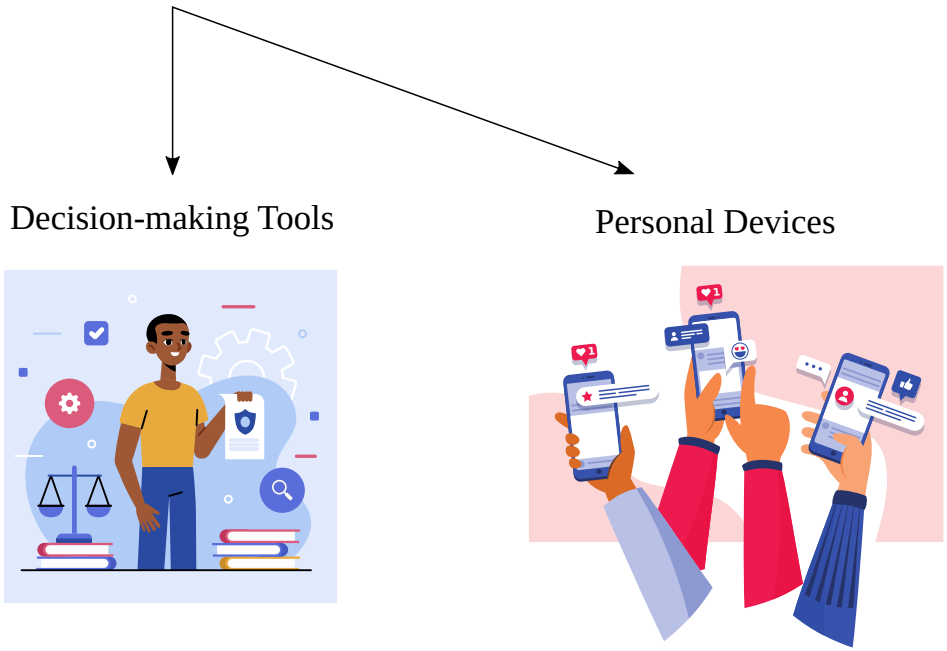
Machine learning models are everywhere

- Machine learning models recently gave **outstanding results** (*e.g.* vision, NLP)
- Industries and governments are starting to use them in **critical applications**



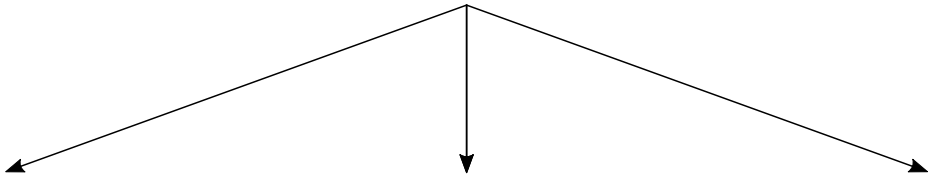
Machine learning models are everywhere

- Machine learning models recently gave **outstanding results** (*e.g.* vision, NLP)
- Industries and governments are starting to use them in **critical applications**

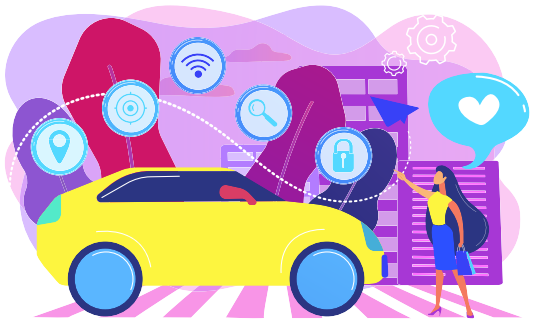


Machine learning models are everywhere

- Machine learning models recently gave **outstanding results** (*e.g.* vision, NLP)
- Industries and governments are starting to use them in **critical applications**



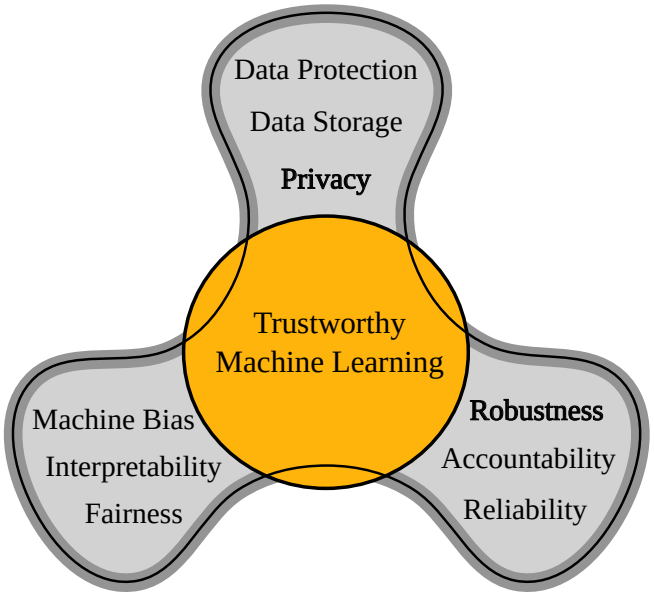
AI-driven Technologies



Decision-making Tools

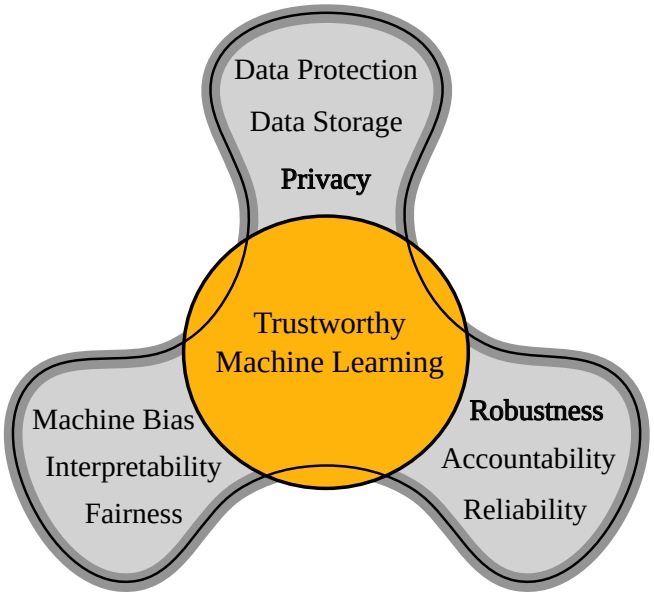
Personal Devices

Trustworthy Machine Learning



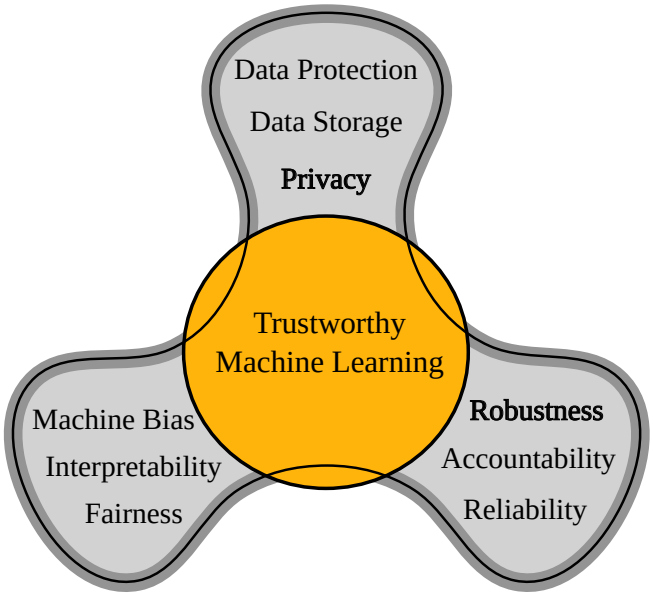
- Massive use of machine learning algorithms raises major issues (*e.g.* privacy, fairness)

Trustworthy Machine Learning



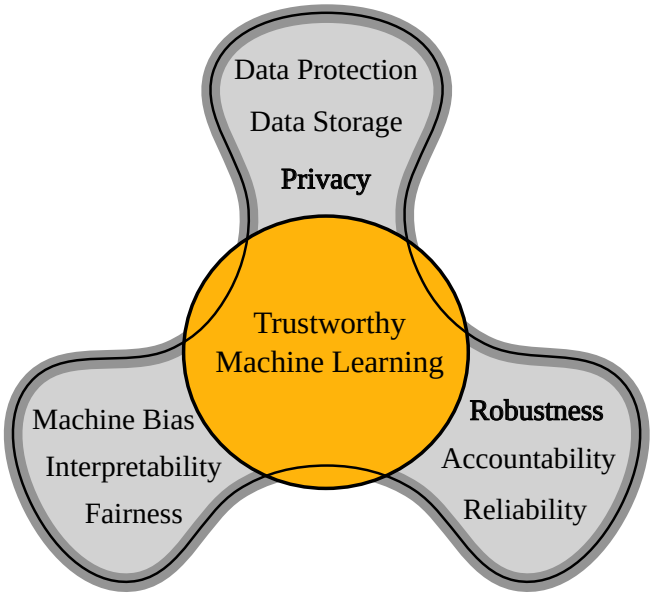
- Massive use of machine learning algorithms raises major issues (*e.g.* privacy, fairness)
- Industries and governments **have to** comply to new regulations (*e.g.* GDPR 2018)

Trustworthy Machine Learning



- Massive use of machine learning algorithms raises major issues (*e.g.* privacy, fairness)
- Industries and governments **have to** comply to new regulations (*e.g.* GDPR 2018)
- We focused on **Privacy** and **Robustness**

Trustworthy Machine Learning

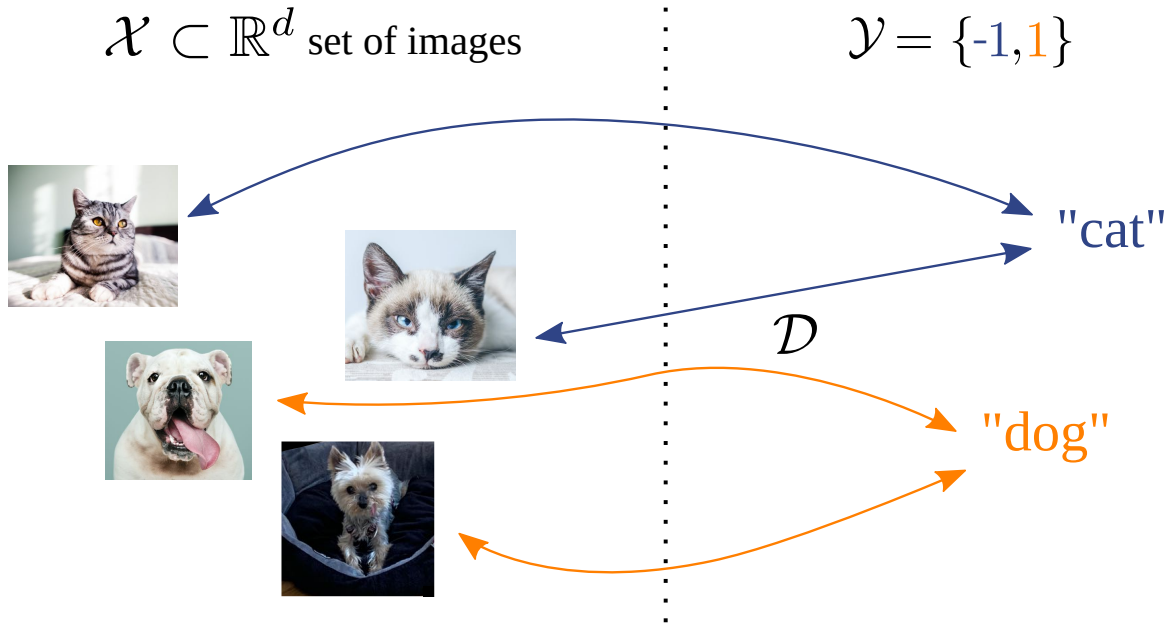


- Massive use of machine learning algorithms raises major issues (*e.g.* privacy, fairness)
- Industries and governments **have to** comply to new regulations (*e.g.* GDPR 2018)
- We focused on **Privacy** and **Robustness**

In this talk: How can we better understand the problem? Can we build machine learning models that are more robust to input manipulation?

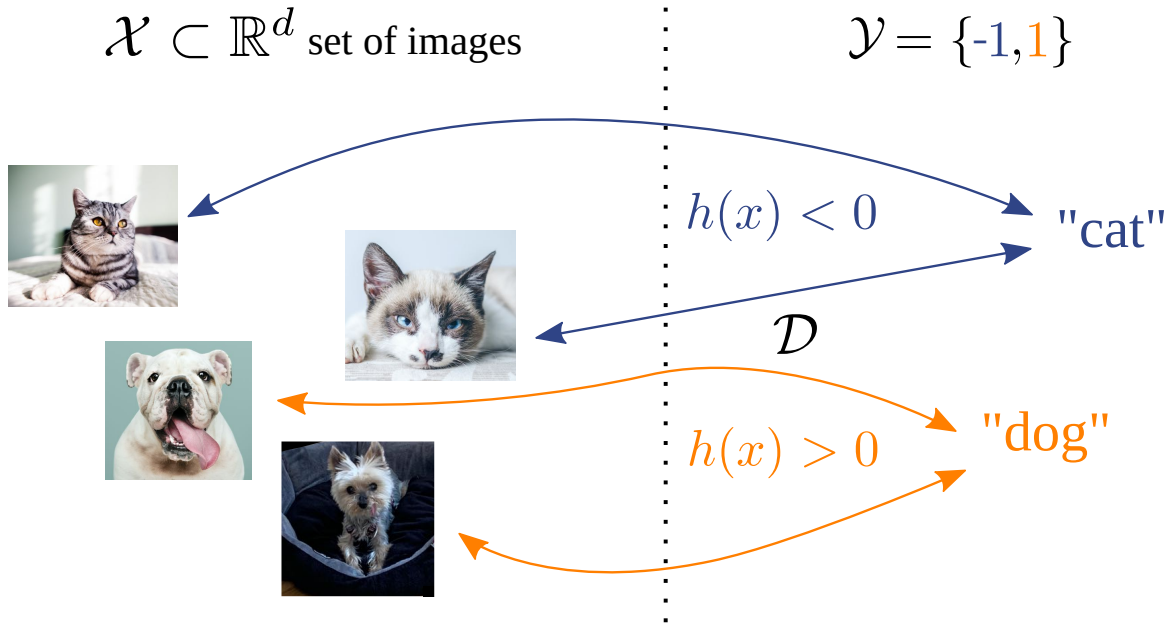
Image classification and adversarial examples

Image classification



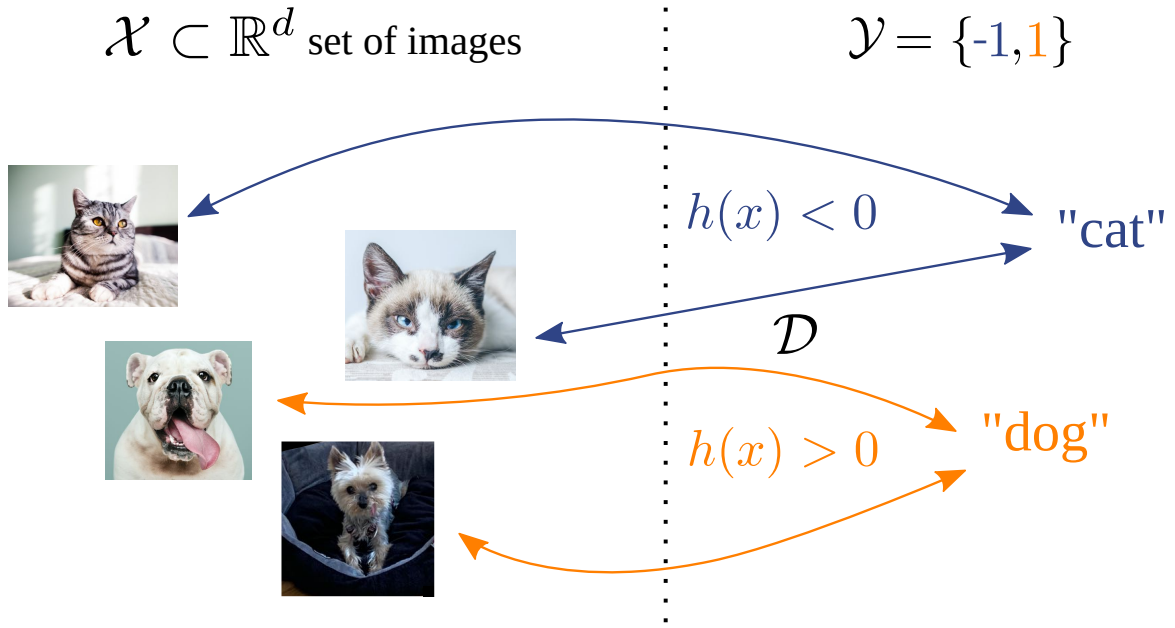
- **Assumption:** A ground-truth distribution $\mathcal{D}$ explains the link between $\mathcal{X}$ and $\mathcal{Y}$

Image classification



- **Assumption:** A ground-truth distribution $\mathcal{D}$ explains the link between $\mathcal{X}$ and $\mathcal{Y}$
- **Goal:** Use $\mathcal{D}$ to design a mapping $h : \mathcal{X} \to \mathbb{R}$ matching images $\mathcal{X}$ to labels $\mathcal{Y}$

Image classification



- **Assumption:** A ground-truth distribution $\mathcal{D}$ explains the link between $\mathcal{X}$ and $\mathcal{Y}$
- **Goal:** Use $\mathcal{D}$ to design a mapping $h : \mathcal{X} \to \mathbb{R}$ matching images $\mathcal{X}$ to labels $\mathcal{Y}$

  1) Define a loss function $\ell : \mathbb{R} \times \mathcal{Y} \to \mathbb{R}^+$ and a hypothesis class $\mathcal{H}$

  2) Find $h \in \mathcal{H}$ to minimize the risk $\mathcal{R}(h) := \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\ell\left(h(x), y\right)\right]$

Image classification

In theory:

We have access to the distribution $\mathcal{D} = \mathbb{P}(y = \text{-}1)\,\mathcal{D}_{\text{-}1} + \mathbb{P}(y = 1)\,\mathcal{D}_1$

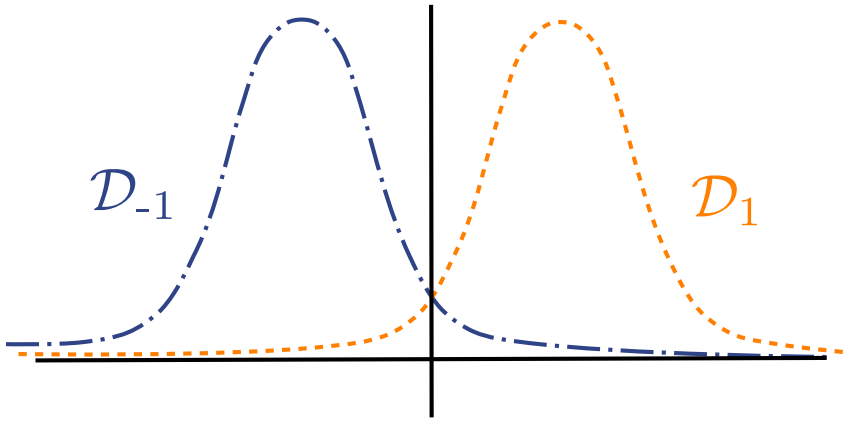
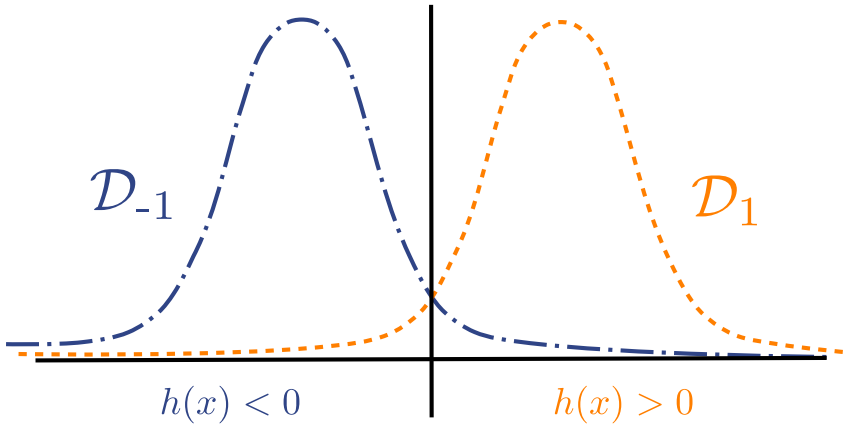


Image classification

In theory:

We have access to the distribution $\mathcal{D} = \mathbb{P}(y = \text{-}1)\,\mathcal{D}_{\text{-}1} + \mathbb{P}(y = 1)\,\mathcal{D}_1$



In theory:

We have access to the distribution $\mathcal{D} = \mathbb{P}(y = \text{-}1)\,\mathcal{D}_{\text{-}1} + \mathbb{P}(y = 1)\,\mathcal{D}_1$

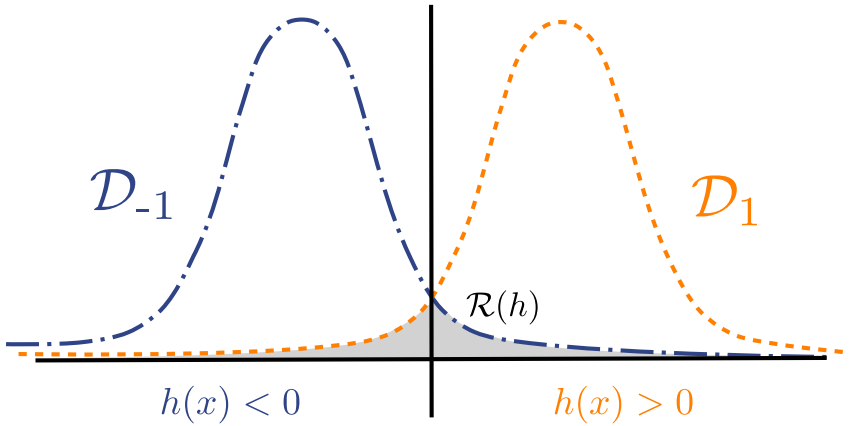
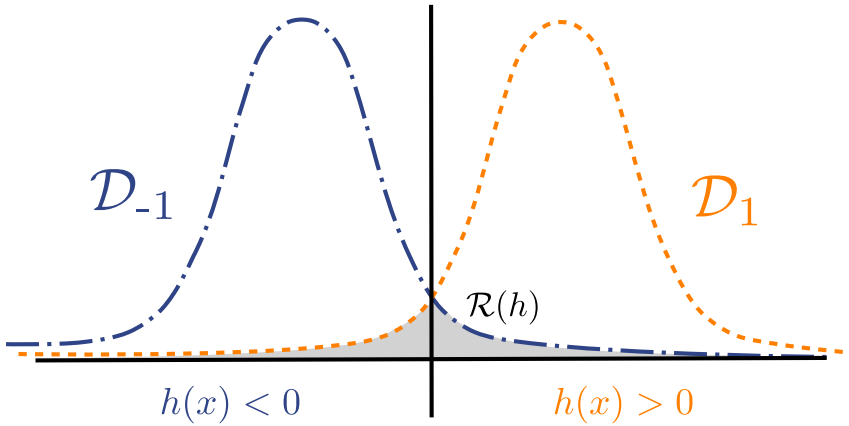


Image classification

In theory:

We have access to the distribution $\mathcal{D} = \mathbb{P}(y = \text{-}1)\,\mathcal{D}_{\text{-}1} + \mathbb{P}(y = 1)\,\mathcal{D}_1$

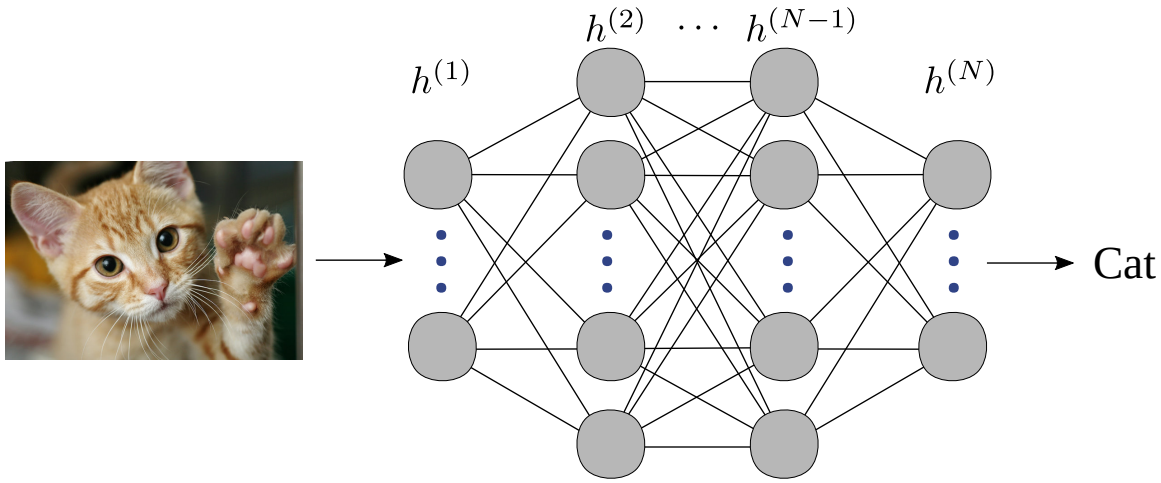


In practice:

- We only have access to a **training sample** $\{(x_1, y_1), ..., (x_n, y_n)\} \sim \mathcal{D}$
- Find $h \in \mathcal{H}$ the **empirical risk** instead $\mathcal{R}_{\text{n}}(h) := \frac{1}{n}\sum_{i=1}^{n} \ell(h(x_i), y_i)$
- The **more data**, the best $\mathcal{R}_{\text{n}}(h) \underset{n\to\infty}{\to} \mathcal{R}(h)$ (depending on $\mathcal{H}$)

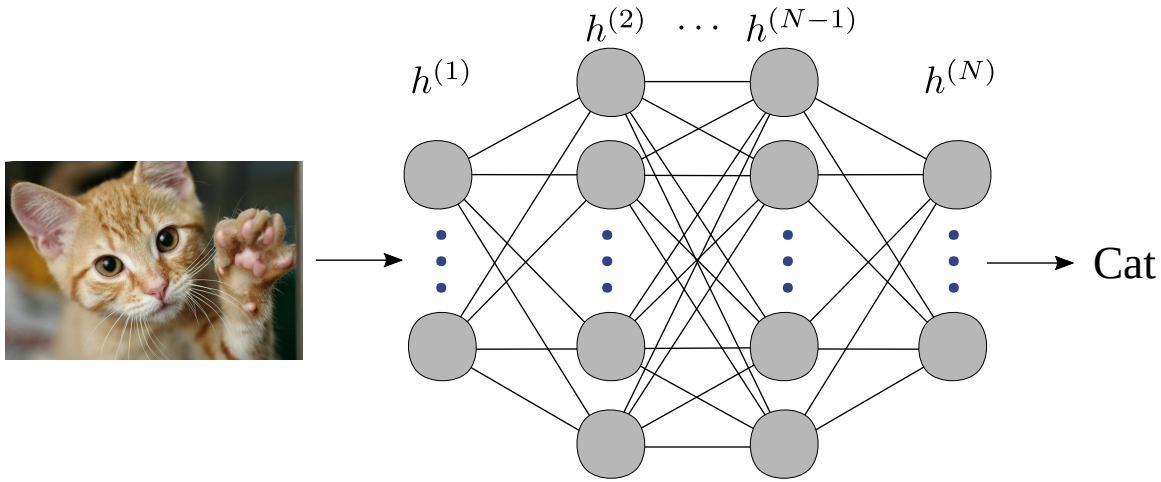
Neural networks: the state-of-the-art for image classification

Hypothesis class of neural networks $\mathcal{H} := \left\{ x \mapsto h^{(N)} \circ h^{(N-1)} \circ \cdots \circ h^{(1)}(x) \right\}$



Neural networks: the state-of-the-art for image classification

Hypothesis class of neural networks $\mathcal{H} := \left\{ x \mapsto h^{(N)} \circ h^{(N-1)} \circ \cdots \circ h^{(1)}(x) \right\}$

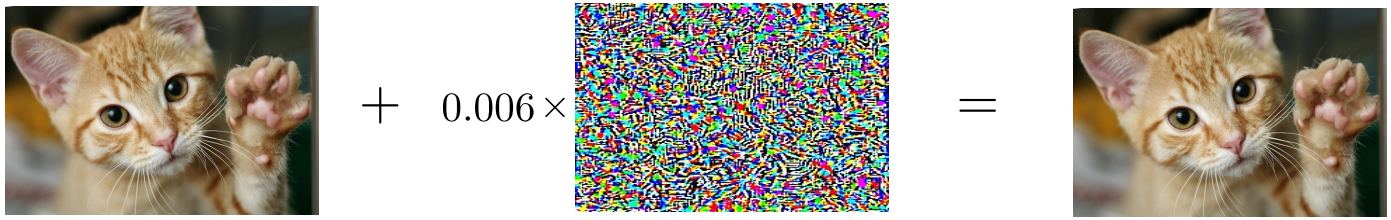


- **Deep convolutional neural networks** offer strong performances
- Still **lack theoretical guarantees** but widely used in **real-life applications**
- **Very vulnerable** to malicious data manipulations (*a.k.a.* adversarial attacks)

Adversarial examples for image classification

Adversarial attack (Biggio et al., 2013; Szegedy et al., 2014): small, imperceptible change of an image maliciously designed to fool the model

Clean example x | Perturbation τ | Adversarial example $x + \tau$

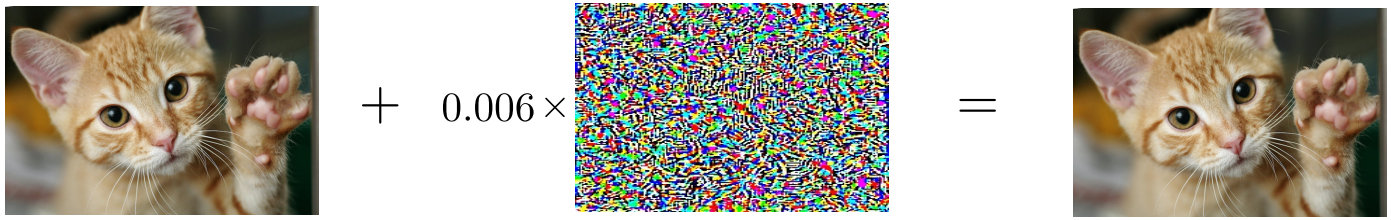


Label: "cat" | Label: "dog"

Adversarial examples for image classification

Adversarial attack (Biggio et al., 2013; Szegedy et al., 2014): small, imperceptible change of an image maliciously designed to fool the model

Clean example x | Perturbation τ | Adversarial example $x + \tau$

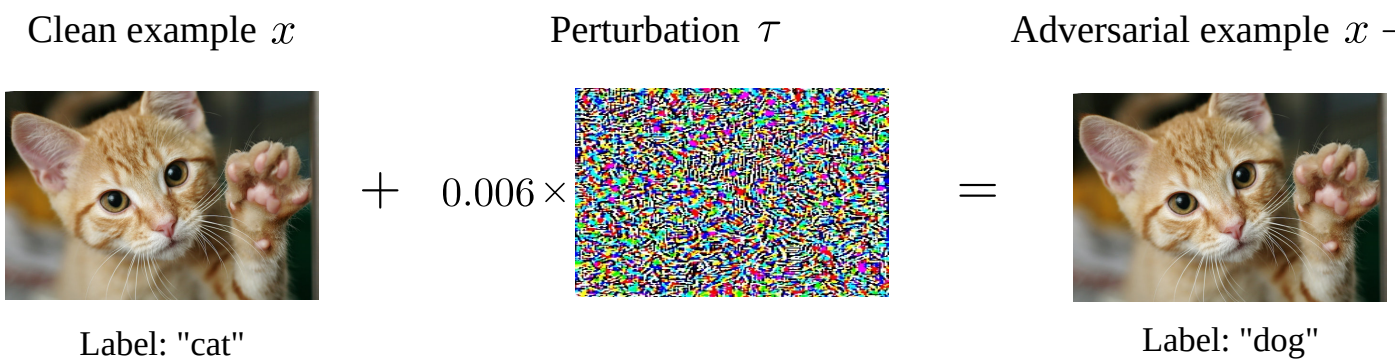


Label: "cat" | Label: "dog"

- Surrogate notion of imperceptibility is modeled with $||\tau||_p \leq \alpha^*$

Adversarial examples for image classification

Adversarial attack (Biggio et al., 2013; Szegedy et al., 2014): small, imperceptible change of an image maliciously designed to fool the model



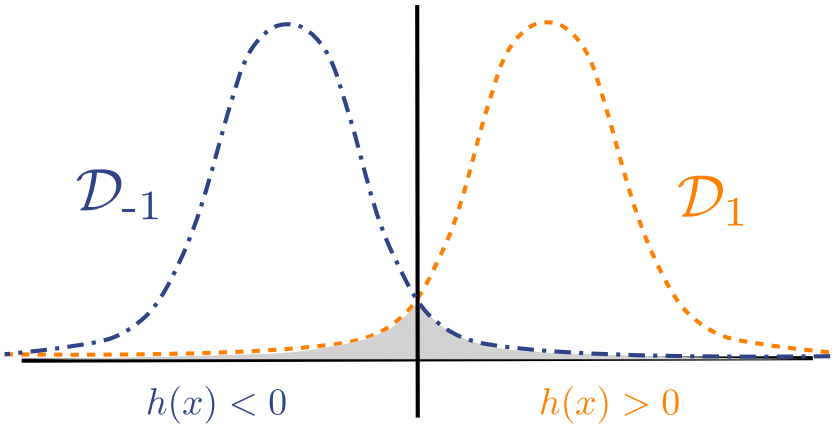
- Surrogate notion of imperceptibility is modeled with $||\tau||_p \leq \alpha^*$
- Numerous attack methods exist: attacks are **easy** to design for **most models**
- **Genuine security issues:** Face recognition in smartphones or autonomous cars

Adversary's point of view

Adversary's goal: find $\tau \in B(\alpha^*)$ to maximize $\ell(h(x+\tau), y)$

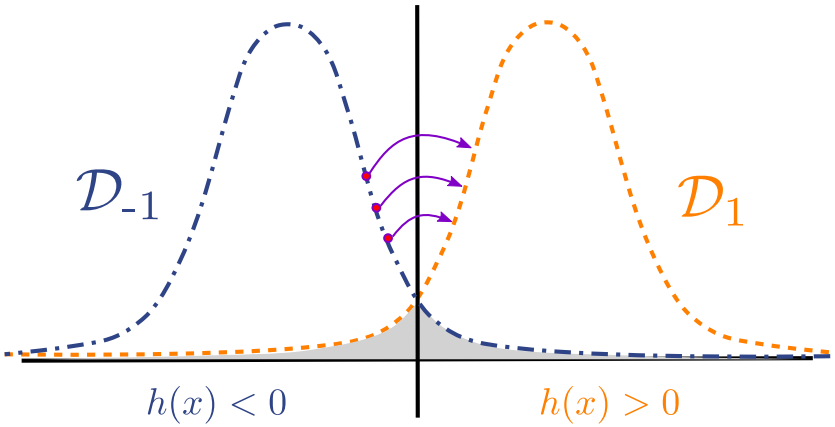
Adversary's point of view

Adversary's goal: find $\tau \in B(\alpha^*)$ to maximize $\ell(h(x+\tau), y)$



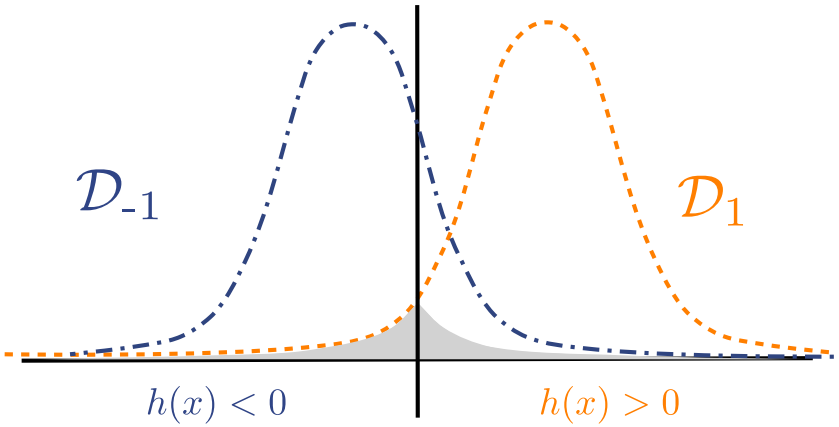
Adversary's point of view

Adversary's goal: find $\tau \in B(\alpha^*)$ to maximize $\ell(h(x+\tau), y)$



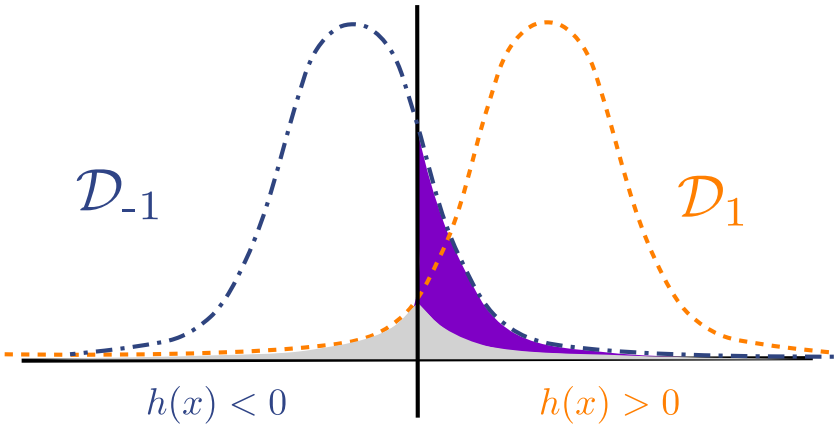
Adversary's point of view

Adversary's goal: find $\tau \in B(\alpha^*)$ to maximize $\ell(h(x+\tau), y)$



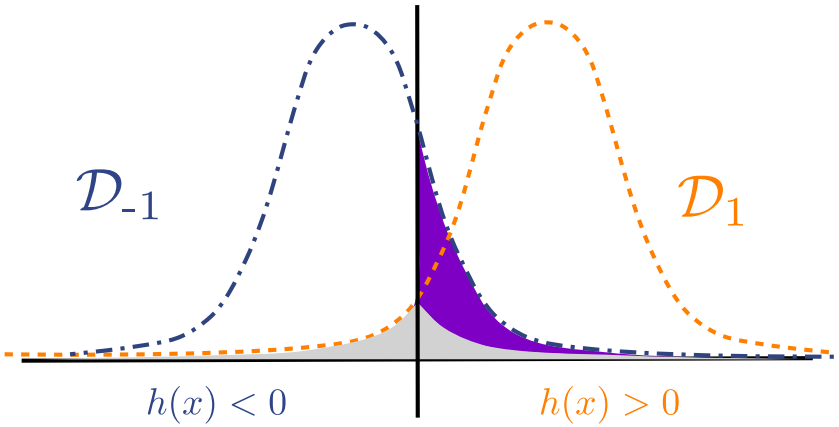
Adversary's point of view

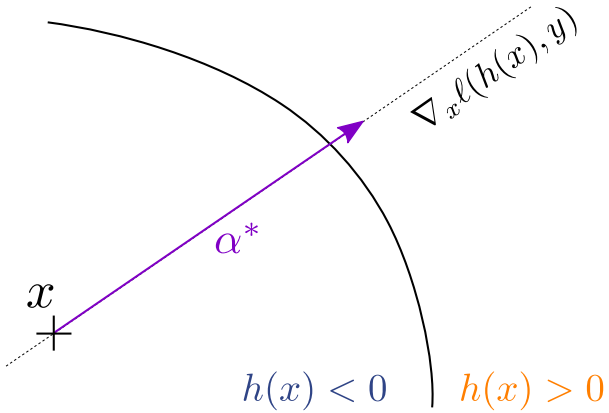
Adversary's goal: find $\tau \in B(\alpha^*)$ to maximize $\ell(h(x+\tau), y)$



Adversary's point of view

Adversary's goal: find $\tau \in B(\alpha^*)$ to maximize $\ell(h(x+\tau), y)$

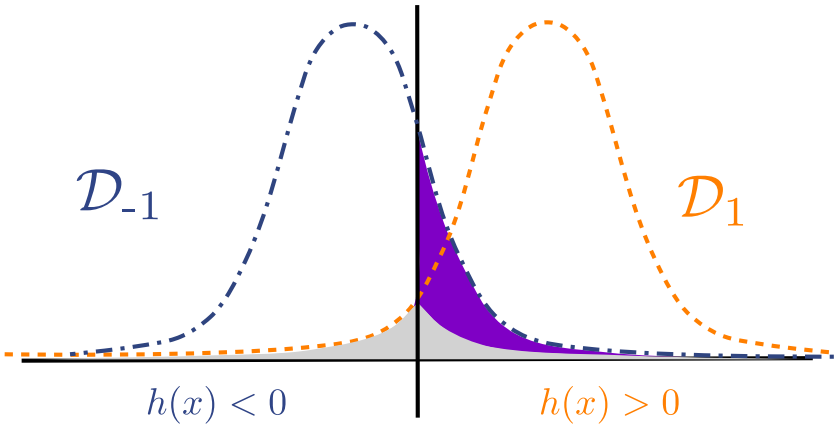


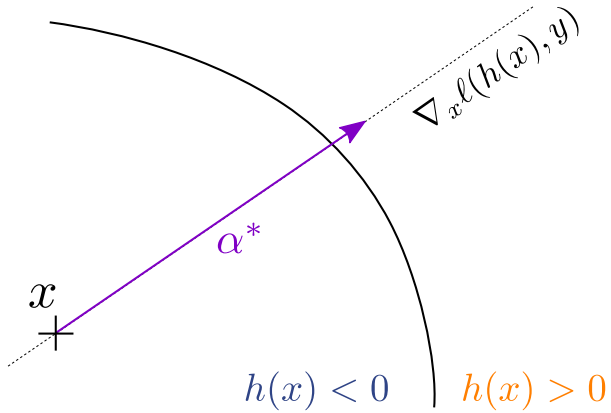


- In practice, attacks use the gradient of the loss **at x** $\nabla_x \ell(h(x), y)$

Adversary's point of view

Adversary's goal: find $\tau \in B(\alpha^*)$ to maximize $\ell(h(x+\tau), y)$

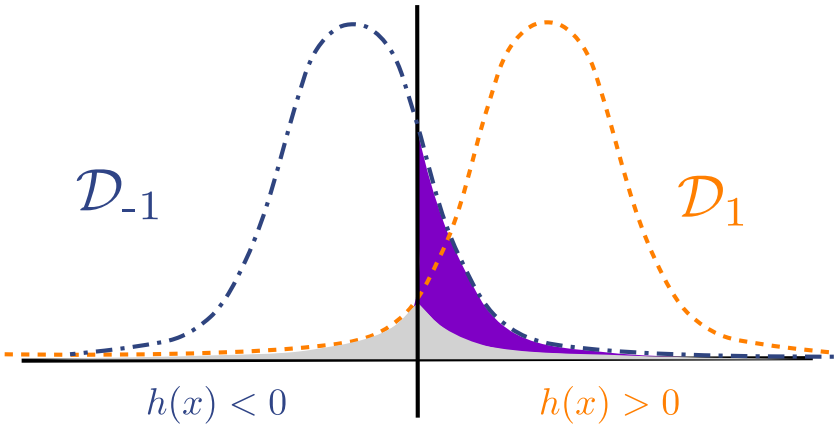


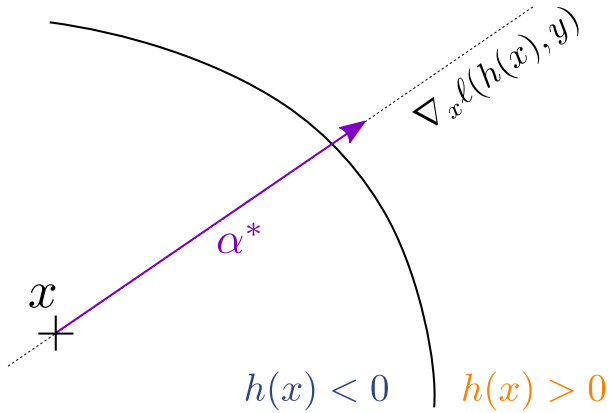


- In practice, attacks use the gradient of the loss **at x** $\nabla_x \ell(h(x), y)$
  - → **F**ast **G**radient **M**ethod (Goodfellow et al., 2015)
  - → **P**rojected **G**radient **D**escent (Kurakin et al., 2016; Madry et al., 2018)

Adversary's point of view

Adversary's goal: find $\tau \in B(\alpha^*)$ to maximize $\ell(h(x+\tau), y)$





- In practice, attacks use the gradient of the loss **at x** $\nabla_x \ell(h(x), y)$
  - → **F**ast **G**radient **M**ethod (Goodfellow et al., 2015)
  - → **P**rojected **G**radient **D**escent (Kurakin et al., 2016; Madry et al., 2018)
- Attacking every point in the dataset makes the accuracy **drop to 0%**

State-of-the-art defense strategies

New goal: find $h \in \mathcal{H}$ minimizing the adversarial risk

$$\mathcal{R}_{\text{adv}}(h) := \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\sup_{\tau \in B(\alpha *)} \ell(h(x+\tau), y) \right]$$

State-of-the-art defense strategies

New goal: find $h \in \mathcal{H}$ minimizing the adversarial risk

$$\mathcal{R}_{\text{adv}}(h) := \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\sup_{\tau \in B(\alpha *)} \ell(h(x+\tau), y) \right]$$

- **Adversarial training:** during the learning procedure, replace x with x_{adv}
  State-of-the-art **empirical** defense for ℓ_∞ attacks (Goodfellow et al., 2015)

State-of-the-art defense strategies

New goal: find $h \in \mathcal{H}$ minimizing the adversarial risk

$$\mathcal{R}_{\text{adv}}(h) := \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\sup_{\tau \in B(\alpha *)} \ell(h(x+\tau), y) \right]$$

- **Adversarial training:** during the learning procedure, replace x with x_{adv}
  State-of-the-art **empirical** defense for ℓ_∞ attacks (Goodfellow et al., 2015)

- **Randomized smoothing:** point-wise certification of the classification
  State-of-the-art **certified** defense for ℓ_2 attacks (Cohen et al.; Salman et al.,2019)

$$h_{\text{RS}}(x) = \mathop{\mathbb{E}}_{\delta\sim\mathcal{N}(0,\sigma^2\mathbb{I})} [h(x+\delta)]$$

On the hardness of adversarial defense

- Adversarial examples are **hard to mitigate** ($\approx 50\%$ on CIFAR10)

On the hardness of adversarial defense

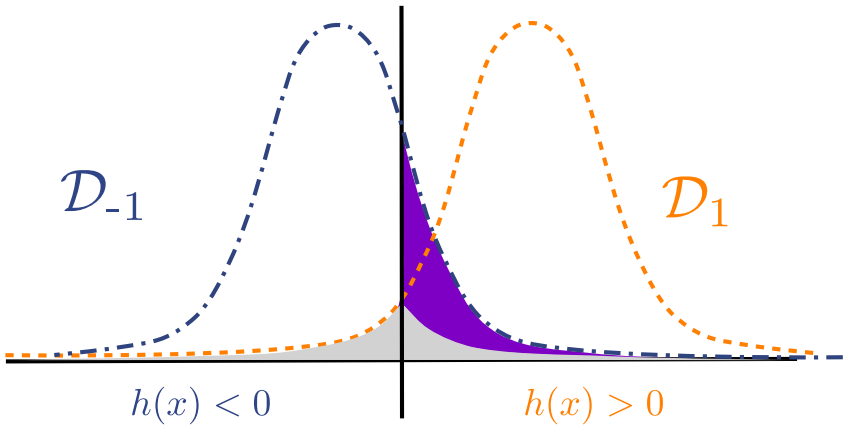
- Adversarial examples are **hard to mitigate** ($\approx 50\%$ on CIFAR10)
- Several works evaluated this hardness **more formally**
  - $\rightarrow$ Lower bound on the adversarial risk (Bhagoji et al.;Pydi and Jog, 2019)
  - $\rightarrow$ Sample complexity (Yin et al., 2019; Awasthi et al., 2020)
  - $\rightarrow$ Computational constraints (Bubeck et al., 2019)

On the hardness of adversarial defense

- Adversarial examples are **hard to mitigate** ($\approx 50\%$ on CIFAR10)
- Several works evaluated this hardness **more formally**
  - → Lower bound on the adversarial risk (Bhagoji et al.;Pydi and Jog, 2019)
  - → Sample complexity (Yin et al., 2019; Awasthi et al., 2020)
  - → Computational constraints (Bubeck et al., 2019)
- Evidence for a **trade-off** between accuracy and robustness (Zhang et al., 2019)

On the hardness of adversarial defense

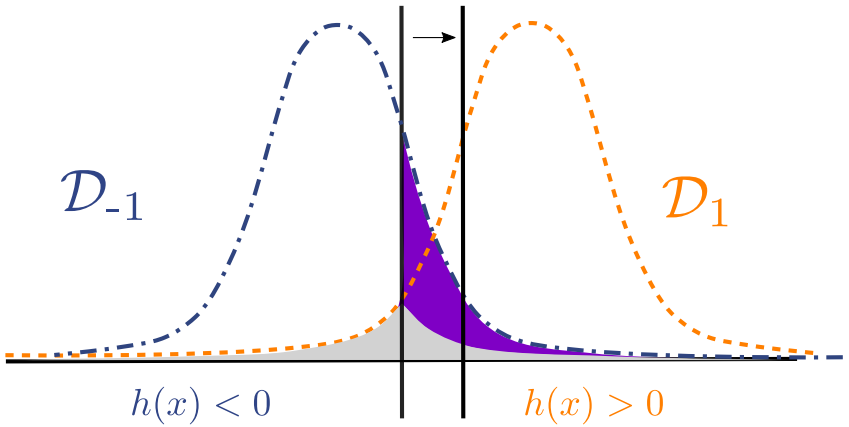
- Adversarial examples are **hard to mitigate** ($\approx 50\%$ on CIFAR10)
- Several works evaluated this hardness **more formally**
  - $\rightarrow$ Lower bound on the adversarial risk (Bhagoji et al.;Pydi and Jog, 2019)
  - $\rightarrow$ Sample complexity (Yin et al., 2019; Awasthi et al., 2020)
  - $\rightarrow$ Computational constraints (Bubeck et al., 2019)
- Evidence for a **trade-off** between accuracy and robustness (Zhang et al., 2019)



- $\mathcal{R}_{\text{adv}}(h) = \mathcal{R}(h) + \mathcal{R}_{\text{adv}>0}(h)$

On the hardness of adversarial defense

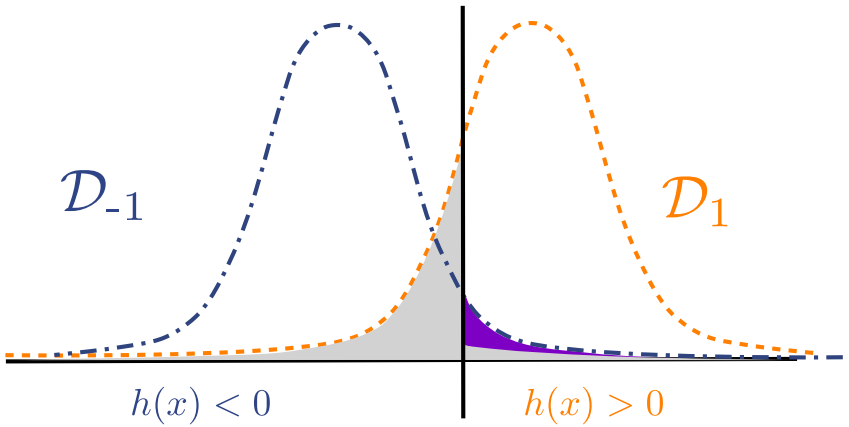
- Adversarial examples are **hard to mitigate** ($\approx 50\%$ on CIFAR10)
- Several works evaluated this hardness **more formally**
  - $\rightarrow$ Lower bound on the adversarial risk (Bhagoji et al.;Pydi and Jog, 2019)
  - $\rightarrow$ Sample complexity (Yin et al., 2019; Awasthi et al., 2020)
  - $\rightarrow$ Computational constraints (Bubeck et al., 2019)
- Evidence for a **trade-off** between accuracy and robustness (Zhang et al., 2019)



- $\mathcal{R}_{\text{adv}}(h) = \mathcal{R}(h) + \mathcal{R}_{\text{adv}>0}(h)$
- $\mathcal{R}_{\text{adv}>0}(h) \searrow$

On the hardness of adversarial defense

- Adversarial examples are **hard to mitigate** ($\approx 50\%$ on CIFAR10)
- Several works evaluated this hardness **more formally**
  - $\rightarrow$ Lower bound on the adversarial risk (Bhagoji et al.;Pydi and Jog, 2019)
  - $\rightarrow$ Sample complexity (Yin et al., 2019; Awasthi et al., 2020)
  - $\rightarrow$ Computational constraints (Bubeck et al., 2019)
- Evidence for a **trade-off** between accuracy and robustness (Zhang et al., 2019)



- $\mathcal{R}_{\text{adv}}(h) = \mathcal{R}(h) + \mathcal{R}_{\text{adv}>0}(h)$
- $\mathcal{R}_{\text{adv}>0}(h) \searrow$ but $\mathcal{R}(h) \nearrow$

Presented contributions

Main Question: Is there a class of hypotheses that could ensure both accuracy and robustness against adversarial attacks? $\rightarrow$ randomized classifiers

Presented contributions

Main Question: Is there a class of hypotheses that could ensure both accuracy and robustness against adversarial attacks? $\rightarrow$ randomized classifiers

Part 1: Justifying the use of randomized classifiers

- **Formalization:** cast the problem as a (regularized) zero-sum game
- **Result 1:** non-existence of a Pure Nash Equilibrium in this game
- **Result 2:** randomized classifiers are more robust than deterministic ones

Presented contributions

Main Question: Is there a class of hypotheses that could ensure both accuracy and robustness against adversarial attacks? $\rightarrow$ randomized classifiers

Part 1: Justifying the use of randomized classifiers

- **Formalization:** cast the problem as a (regularized) zero-sum game
- **Result 1:** non-existence of a Pure Nash Equilibrium in this game
- **Result 2:** randomized classifiers are more robust than deterministic ones

Part 2: Demonstrating the robustness of randomized classifiers

- **Formalization:** define adversarial robustness for randomized classifier
- **Result 1:** bound the accuracy vs robustness trade-off for these classifiers
- **Result 2:** design classes of robust (randomized) neural networks

Part 1:
Justifying the use of randomized classifiers

Recall on Zero-sum games

Generic optimization problem:

$$\inf_{h} \sup_{\phi} \text{Score}(h, \phi)$$

Recall on Zero-sum games

Generic optimization problem:

$$\inf_{h} \sup_{\phi} \text{Score}(h, \phi)$$

First player $h \in \mathcal{H}$

Recall on Zero-sum games

Generic optimization problem:

$$\inf_{h} \sup_{\phi} \text{Score}(h, \phi)$$

First player $h \in \mathcal{H}$

Second player $\phi \in \mathcal{F}$

Recall on Zero-sum games

Generic optimization problem:

$$\inf_{h} \sup_{\phi} \text{Score}(h, \phi)$$

First player $h \in \mathcal{H}$

$$\mathfrak{BR}(\phi) = \operatorname*{argmin}_{h} \text{Score}(h, \phi)$$

Second player $\phi \in \mathcal{F}$

Recall on Zero-sum games

Generic optimization problem:

$$\inf_{h} \sup_{\phi} \text{Score}(h, \phi)$$

First player $h \in \mathcal{H}$

$$\mathfrak{BR}(\phi) = \operatorname*{argmin}_{h} \ \text{Score}(h, \phi)$$

Second player $\phi \in \mathcal{F}$

$$\mathfrak{BR}(h) = \operatorname*{argmax}_{\phi} \ \text{Score}(h, \phi)$$

Recall on Zero-sum games

Generic optimization problem:

$$\inf_{h} \sup_{\phi} \text{Score}(h, \phi)$$

First player $h \in \mathcal{H}$

$$\mathfrak{BR}(\phi) = \operatorname*{argmin}_{h} \ \text{Score}(h, \phi)$$

Second player $\phi \in \mathcal{F}$

$$\mathfrak{BR}(h) = \operatorname*{argmax}_{\phi} \ \text{Score}(h, \phi)$$

Nash Equilibrium: stable state (h_{NE}, ϕ_{NE}) s.t. $\begin{cases} h_{NE} \in \mathfrak{BR}(\phi_{NE}) \\ \phi_{NE} \in \mathfrak{BR}(h_{NE}) \end{cases}$

A closer look at adversarial risk

In the following we set:

$$\mathcal{H} := \{h : \mathcal{X} \mapsto \mathbb{R} \text{ continuous}\} \quad \text{and} \quad \ell_{0/1}(h(x), y) = \mathbb{1}\{\operatorname{sgn}(h(x)) \neq y\}$$

Decomposition of the adversarial risk:

$$\mathcal{R}_{\text{adv}}(h) := \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\sup_{\tau \in B(\alpha^*)} \ell_{0/1}(h(x+\tau), y) \right]$$

A closer look at adversarial risk

In the following we set:

$$\mathcal{H} := \{h : \mathcal{X} \mapsto \mathbb{R} \text{ continuous}\} \quad \text{and} \quad \ell_{0/1}(h(x), y) = \mathbb{1}\{\operatorname{sgn}(h(x)) \neq y\}$$

Decomposition of the adversarial risk:

$$\begin{aligned} \mathcal{R}_{\text{adv}}(h) &:= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\sup_{\tau \in B(\alpha^*)} \ell_{0/1}(h(x+\tau), y) \right] \\ &= \mathop{\mathbb{E}}_{y\sim\rho} \left[\mathop{\mathbb{E}}_{x\sim\mathcal{D}_y} \left[\sup_{\tau \in B(\alpha^*)} \ell_{0/1}\left(h(x+\tau), y\right) \right] \right] \end{aligned}$$

A closer look at adversarial risk

In the following we set:

$$\mathcal{H} := \{h : \mathcal{X} \mapsto \mathbb{R} \text{ continuous}\} \quad \text{and} \quad \ell_{0/1}(h(x), y) = \mathbb{1}\{\operatorname{sgn}(h(x)) \neq y\}$$

Decomposition of the adversarial risk:

$$\begin{aligned}
\mathcal{R}_{\text{adv}}(h) &:= \mathop{\mathbb{E}}_{(x,y)\sim\mathcal{D}} \left[\sup_{\tau \in B(\alpha^*)} \ell_{0/1}(h(x+\tau), y) \right] \\
&= \mathop{\mathbb{E}}_{y\sim\rho} \left[\mathop{\mathbb{E}}_{x\sim\mathcal{D}_y} \left[\sup_{\tau \in B(\alpha^*)} \ell_{0/1}\left(h(x+\tau), y\right) \right] \right] \\
&= \mathop{\mathbb{E}}_{y\sim\rho} \left[\sup_{\phi \in \mathcal{F}} \mathop{\mathbb{E}}_{x\sim\phi\#\mathcal{D}_y} \left[\ell_{0/1}\left(h(x), y\right) \right] \right]
\end{aligned}$$

Where $\mathcal{F} := \left\{ \phi : \mathcal{X} \to \mathcal{X} \mid \phi \text{ measurable, and } \sup_{x\in\mathcal{X}} \|\phi(x) - x\|_p \leq \alpha^* \right\}$

Adversarial examples game and trivial equilibrium

- Everyone has perfect information (distribution, other player's strategy)

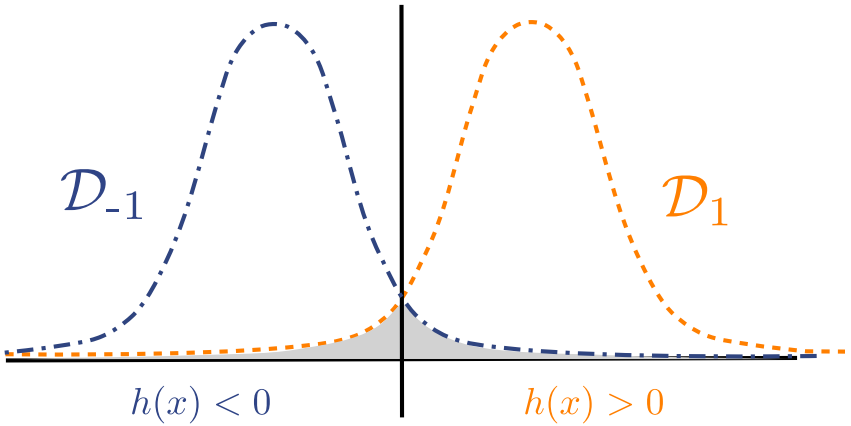
Adversarial examples game and trivial equilibrium

- Everyone has perfect information (distribution, other player's strategy)
- The adversary has no limitation (computational, perceptual)

Adversarial examples game and trivial equilibrium

- Everyone has perfect information (distribution, other player's strategy)
- The adversary has no limitation (computational, perceptual)

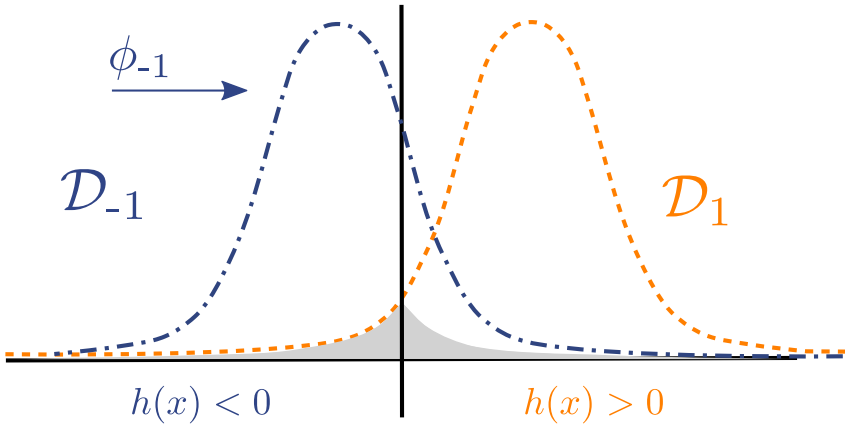
$$\inf_{h\in\mathcal{H}} \sup_{\phi\in\mathcal{F}^2} \underbrace{\mathop{\mathbb{E}}_{y\sim\rho}\left[\mathop{\mathbb{E}}_{x\sim\phi_y\#\mathcal{D}_y}\left[\ell_{0/1}(h(x),y)\right]\right]}_{\mathrm{Score}(h,\phi)}$$



Adversarial examples game and trivial equilibrium

- Everyone has perfect information (distribution, other player's strategy)
- The adversary has no limitation (computational, perceptual)

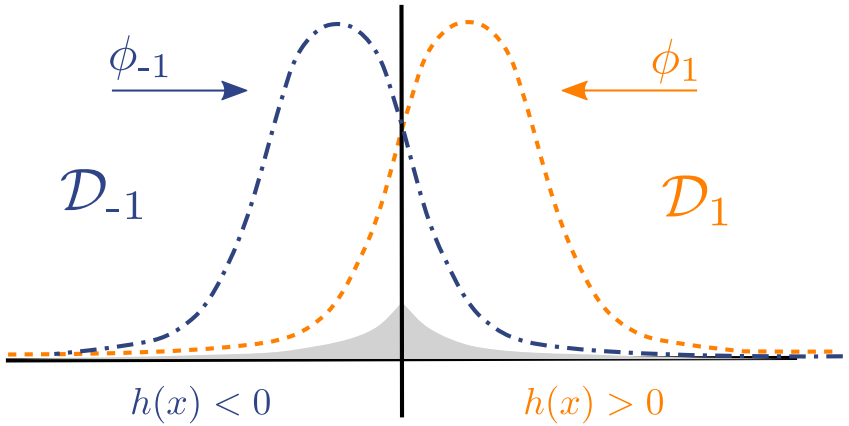
$$\inf_{h\in\mathcal{H}} \sup_{\phi\in\mathcal{F}^2} \underbrace{\mathop{\mathbb{E}}_{y\sim\rho}\left[\mathop{\mathbb{E}}_{x\sim\phi_y\#\mathcal{D}_y}\left[\ell_{0/1}(h(x),y)\right]\right]}_{\text{Score}(h,\phi)}$$



Adversarial examples game and trivial equilibrium

- Everyone has perfect information (distribution, other player's strategy)
- The adversary has no limitation (computational, perceptual)

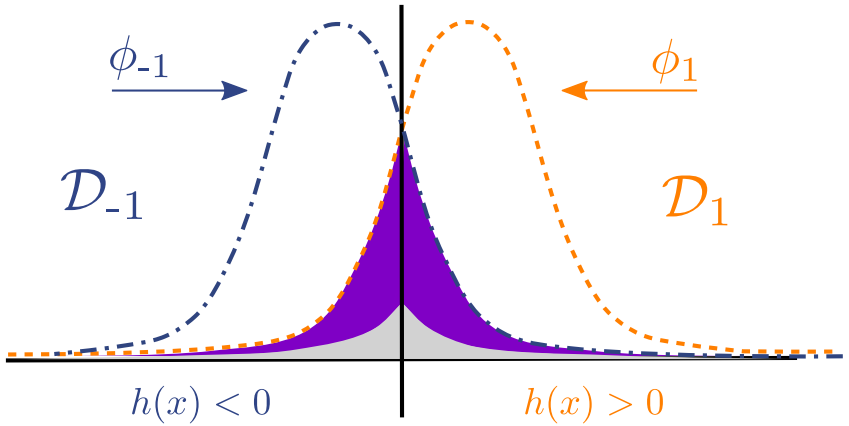
$$\inf_{h\in\mathcal{H}} \sup_{\phi\in\mathcal{F}^2} \underbrace{\mathop{\mathbb{E}}_{y\sim\rho}\left[\mathop{\mathbb{E}}_{x\sim\phi_y\#\mathcal{D}_y}\left[\ell_{0/1}(h(x),y)\right]\right]}_{\text{Score}(h,\phi)}$$



Adversarial examples game and trivial equilibrium

- Everyone has perfect information (distribution, other player's strategy)
- The adversary has no limitation (computational, perceptual)

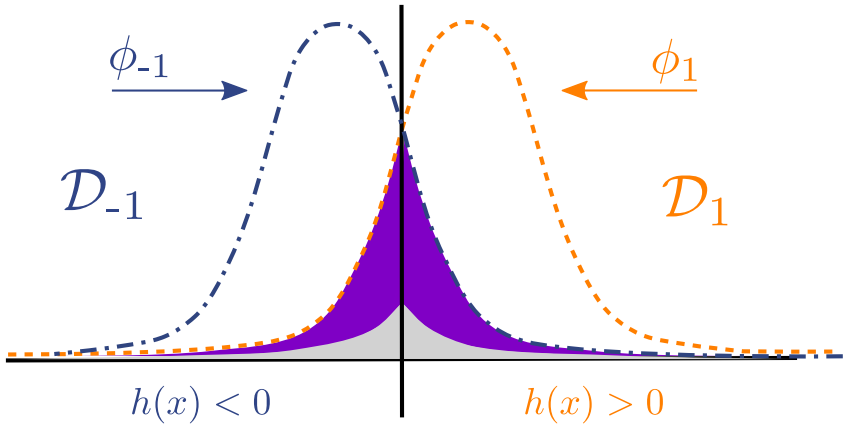
$$\inf_{h \in \mathcal{H}} \sup_{\phi \in \mathcal{F}^2} \underbrace{\mathop{\mathbb{E}}_{y \sim \rho} \left[\mathop{\mathbb{E}}_{x \sim \phi_y \# \mathcal{D}_y} \left[\ell_{0/1}(h(x), y) \right] \right]}_{\text{Score}(h, \phi)}$$



Adversarial examples game and trivial equilibrium

- Everyone has perfect information (distribution, other player's strategy)
- The adversary has no limitation (computational, perceptual)

$$\inf_{h \in \mathcal{H}} \sup_{\phi \in \mathcal{F}^2} \underbrace{\mathop{\mathbb{E}}_{y \sim \rho} \left[\mathop{\mathbb{E}}_{x \sim \phi_y \# \mathcal{D}_y} \left[\ell_{0/1}(h(x), y) \right] \right]}_{\text{Score}(h, \phi)}$$

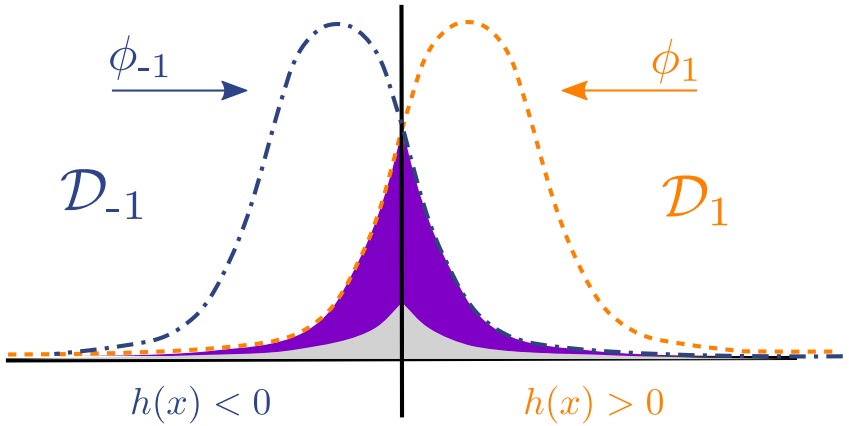


- Trivial Nash equilibrium

Adversarial examples game and trivial equilibrium

- Everyone has perfect information (distribution, other player's strategy)
- The adversary has no limitation (computational, perceptual)

$$\inf_{h\in\mathcal{H}} \sup_{\phi\in\mathcal{F}^2} \underbrace{\mathop{\mathbb{E}}_{y\sim\rho}\left[\mathop{\mathbb{E}}_{x\sim\phi_y\#\mathcal{D}_y}\left[\ell_{0/1}(h(x),y)\right]\right]}_{\mathrm{Score}(h,\phi)}$$



- Trivial Nash equilibrium
- Inevitably high adversarial risk

Unrealistic behavior of the adversary and regularization

Behavior of the adversary:

- The adversary attacks all points (no budget constraint)

Unrealistic behavior of the adversary and regularization

Behavior of the adversary:

- The adversary attacks all points (no budget constraint)
- Autonomous car example: unrealistic to attack every image processed

Unrealistic behavior of the adversary and regularization

Behavior of the adversary:

- The adversary attacks all points (no budget constraint)
- Autonomous car example: unrealistic to attack every image processed

→ Regularization to account for the adversary's budget

Unrealistic behavior of the adversary and regularization

Behavior of the adversary:

- The adversary attacks all points (no budget constraint)
- Autonomous car example: unrealistic to attack every image processed

$\rightarrow$ Regularization to account for the adversary's budget

Regularized Score:

$$\inf_{h \in \mathcal{H}} \sup_{\phi \in \mathcal{F}^2} \underbrace{\text{Score}(h, \phi) - \lambda \, \Omega \, (\phi)}_{\text{Score}_\Omega(h, \phi)}$$

Unrealistic behavior of the adversary and regularization

Behavior of the adversary:

- The adversary attacks all points (no budget constraint)
- Autonomous car example: unrealistic to attack every image processed

$\rightarrow$ Regularization to account for the adversary's budget

Regularized Score:

$$\inf_{h \in \mathcal{H}} \sup_{\phi \in \mathcal{F}^2} \underbrace{\text{Score}(h, \phi) - \lambda \, \Omega \, (\phi)}_{\text{Score}_\Omega(h, \phi)}$$

Typical penalty: penalize the expected number of queries

$$\Omega(\phi) := \mathop{\mathbb{E}}_{(x,y) \sim \mathcal{D}} \left[\mathbb{1} \left[x \neq \phi_y(x) \right] \right]$$

Non-existence of a Pure Nash Equilibrium

Theorem (Non-existence of PNE)

Let us consider the above zero-sum game with penalty Ω and $\lambda > 0$.

If $\mathcal{D}_1$ and $\mathcal{D}_{\text{-}1}$ have full support, then the game has no Pure Nash Equilibrium.

Non-existence of a Pure Nash Equilibrium

Theorem (Non-existence of PNE)

Let us consider the above zero-sum game with penalty Ω and $\lambda > 0$.

If $\mathcal{D}_1$ and $\mathcal{D}_{\text{-}1}$ have full support, then the game has no Pure Nash Equilibrium.

Idea of the proof: Let $h \in \mathcal{H}$, and $\phi \in \mathfrak{BR}(h)$. Then $h \notin \mathfrak{BR}(\phi)$.

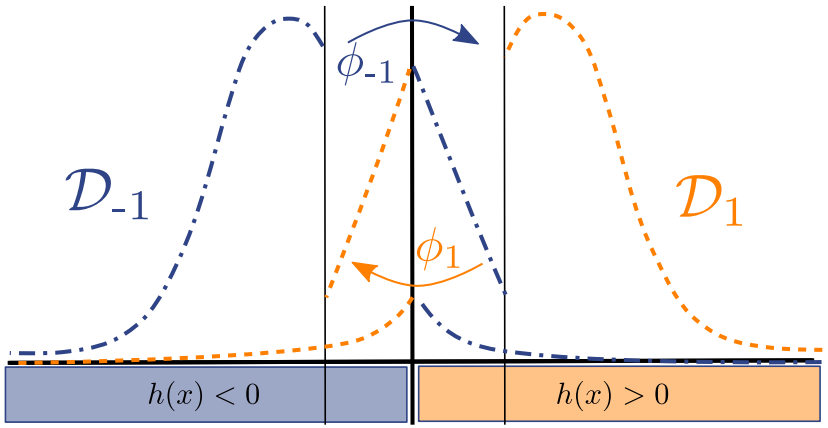
Non-existence of a Pure Nash Equilibrium

Theorem (Non-existence of PNE)

Let us consider the above zero-sum game with penalty Ω and $\lambda > 0$.

If $\mathcal{D}_1$ and $\mathcal{D}_{\text{-}1}$ have full support, then the game has no Pure Nash Equilibrium.

Idea of the proof: Let $h \in \mathcal{H}$, and $\phi \in \mathfrak{BR}(h)$. Then $h \notin \mathfrak{BR}(\phi)$.



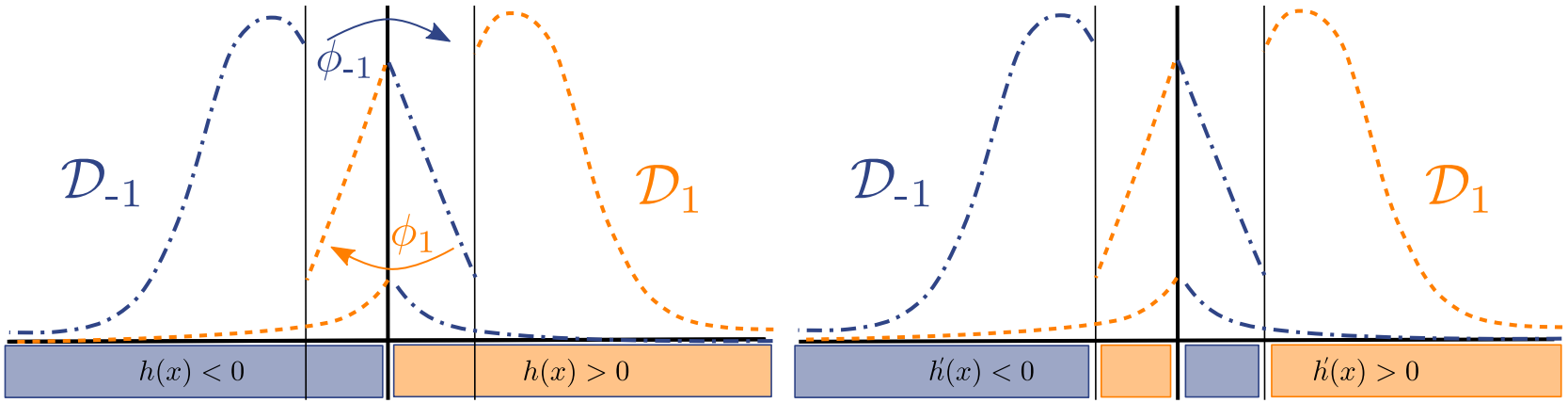
Non-existence of a Pure Nash Equilibrium

Theorem (Non-existence of PNE)

Let us consider the above zero-sum game with penalty Ω and $\lambda > 0$.

If $\mathcal{D}_1$ and $\mathcal{D}_{\text{-}1}$ have full support, then the game has no Pure Nash Equilibrium.

Idea of the proof: Let $h \in \mathcal{H}$, and $\phi \in \mathfrak{BR}(h)$. Then $h \notin \mathfrak{BR}(\phi)$.



Consequences and next step

Consequences:

- The race between the attacks and defenses is still on!

Consequences and next step

Consequences:

- The race between the attacks and defenses is still on!
- **No free lunch for transferable attacks** (duality gap)

$$\sup_{\phi \in \mathcal{F}^2} \inf_{h \in \mathcal{H}} \text{Score}_\Omega(h, \phi) < \inf_{h \in \mathcal{H}} \sup_{\phi \in \mathcal{F}^2} \text{Score}_\Omega(h, \phi)$$

Consequences and next step

Consequences:

- The race between the attacks and defenses is still on!
- **No free lunch for transferable attacks** (duality gap)

$$\sup_{\phi \in \mathcal{F}^2} \inf_{h \in \mathcal{H}} \text{Score}_\Omega(h, \phi) < \inf_{h \in \mathcal{H}} \sup_{\phi \in \mathcal{F}^2} \text{Score}_\Omega(h, \phi)$$

Next step:

- What happens if we use **mixed strategies** (Mixed Nash Equilibrium)

Consequences and next step

Consequences:

- The race between the attacks and defenses is still on!
- **No free lunch for transferable attacks** (duality gap)

$$\sup_{\phi \in \mathcal{F}^2} \inf_{h \in \mathcal{H}} \text{Score}_\Omega(h, \phi) < \inf_{h \in \mathcal{H}} \sup_{\phi \in \mathcal{F}^2} \text{Score}_\Omega(h, \phi)$$

Next step:

- What happens if we use **mixed strategies** (Mixed Nash Equilibrium)

→ *Study what happens if **only the defender** uses randomized strategies*

Formalize randomized strategies

Randomized classifier:

- A randomized classifier $\mathrm{m} : x \longmapsto \mathrm{m}(x)$ outputs a measure on $\mathbb{R}$

Formalize randomized strategies

Randomized classifier:

- A randomized classifier $\mathrm{m} : x \longmapsto \mathrm{m}(x)$ outputs a measure on $\mathbb{R}$
- To get a numerical out of $\mathrm{m}(x)$, sample $z \sim \mathrm{m}(x)$

Formalize randomized strategies

Randomized classifier:

- A randomized classifier $\mathrm{m} : x \longmapsto \mathrm{m}(x)$ outputs a measure on $\mathbb{R}$
- To get a numerical out of $\mathrm{m}(x)$, sample $z \sim \mathrm{m}(x)$
- We denote Δ the set of all randomized classifiers

Formalize randomized strategies

Randomized classifier:

- A randomized classifier $\mathrm{m} : x \longmapsto \mathrm{m}(x)$ outputs a measure on $\mathbb{R}$
- To get a numerical out of $\mathrm{m}(x)$, sample $z \sim \mathrm{m}(x)$
- We denote Δ the set of all randomized classifiers

Adaptation of the score:

$$\mathrm{Score}_{\Omega}(\mathrm{m}, \phi) := \mathop{\mathbb{E}}_{y \sim \rho} \left[\mathop{\mathbb{E}}_{x \sim \phi_y \# \mathcal{D}_y} \left[\mathop{\mathbb{E}}_{z \sim \mathrm{m}(x)} \left[\ell_{0/1}\left(z, y\right) \right] \right] \right] - \lambda\, \Omega\left(\phi\right)$$

Randomization matters

Theorem (Randomization matters)

Let us consider $h \in \mathcal{H}$ and $\lambda \in (0, 1)$. There exists $\mathrm{m} \in \Delta$ such that

$$\sup_{\phi \in \mathcal{F}^2} \mathrm{Score}_\Omega(\mathrm{m}, \phi) < \sup_{\phi \in \mathcal{F}^2} \mathrm{Score}_\Omega(h, \phi).$$

Randomization matters

Theorem (Randomization matters)

Let us consider $h \in \mathcal{H}$ and $\lambda \in (0, 1)$. There exists $\mathrm{m} \in \Delta$ such that

$$\sup_{\phi \in \mathcal{F}^2} \mathrm{Score}_\Omega(\mathrm{m}, \phi) < \sup_{\phi \in \mathcal{F}^2} \mathrm{Score}_\Omega(h, \phi).$$

Idea of the proof:

- Let $\phi \in \mathfrak{BR}(h)$, and $h' \in \mathfrak{BR}(\phi)$

Randomization matters

Theorem (Randomization matters)

Let us consider $h \in \mathcal{H}$ and $\lambda \in (0, 1)$. There exists $\mathrm{m} \in \Delta$ such that

$$\sup_{\phi \in \mathcal{F}^2} \mathrm{Score}_\Omega(\mathrm{m}, \phi) < \sup_{\phi \in \mathcal{F}^2} \mathrm{Score}_\Omega(h, \phi).$$

Idea of the proof:

- Let $\phi \in \mathfrak{BR}(h)$, and $h' \in \mathfrak{BR}(\phi)$
- $\mathrm{m}(x)$ outputs $h(x)$ with probability q and $h'(x)$ otherwise

Randomization matters

Theorem (Randomization matters)

Let us consider $h \in \mathcal{H}$ and $\lambda \in (0, 1)$. There exists $\mathrm{m} \in \Delta$ such that

$$\sup_{\phi \in \mathcal{F}^2} \mathrm{Score}_\Omega(\mathrm{m}, \phi) < \sup_{\phi \in \mathcal{F}^2} \mathrm{Score}_\Omega(h, \phi).$$

Idea of the proof:

- Let $\phi \in \mathfrak{BR}(h)$, and $h' \in \mathfrak{BR}(\phi)$
- $\mathrm{m}(x)$ outputs $h(x)$ with probability q and $h'(x)$ otherwise
- If $1 > q > \max(\lambda, 1 - \lambda)$ the result holds

From theorem to algorithm

Algorithm: Boosted adversarial training

Inputs: D the training set and $q \in (0, 1)$

> Train h on D with adversarial training

> Generate an adversarial data set D' against h

> Train h' on D' with standard training

Return: $h \sim q$ and $h' \sim 1 - q$

From theorem to algorithm

Algorithm: Boosted adversarial training

Inputs: D the training set and $q \in (0, 1)$

> Train h on D with adversarial training

> Generate an adversarial data set D' against h

> Train h' on D' with standard training

Return: $h \sim q$ and $h' \sim 1 - q$

$\mathcal{D}$

From theorem to algorithm

Algorithm: Boosted adversarial training

Inputs: D the training set and $q \in (0, 1)$

> Train h on D with adversarial training

> Generate an adversarial data set D' against h

> Train h' on D' with standard training

Return: $h \sim q$ and $h' \sim 1 - q$

From theorem to algorithm

Algorithm: Boosted adversarial training

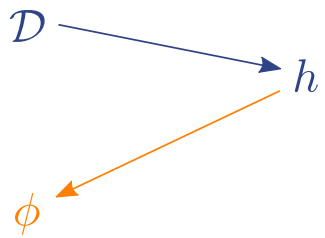
Inputs: D the training set and $q \in (0, 1)$

> Train h on D with adversarial training

> Generate an adversarial data set D' against h

> Train h' on D' with standard training

Return: $h \sim q$ and $h' \sim 1 - q$



From theorem to algorithm

Algorithm: Boosted adversarial training

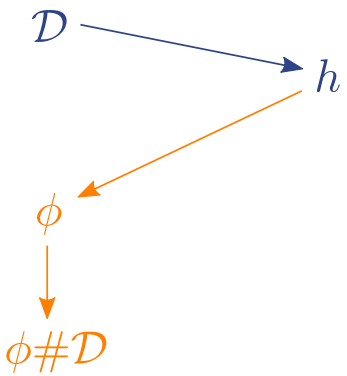
Inputs: D the training set and $q \in (0, 1)$

> Train h on D with adversarial training

> Generate an adversarial data set D' against h

> Train h' on D' with standard training

Return: $h \sim q$ and $h' \sim 1 - q$



From theorem to algorithm

Algorithm: Boosted adversarial training

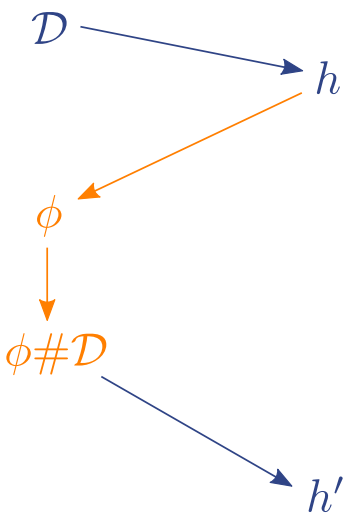
Inputs: D the training set and $q \in (0, 1)$

> Train h on D with adversarial training

> Generate an adversarial data set D' against h

> Train h' on D' with standard training

Return: $h \sim q$ and $h' \sim 1 - q$



From theorem to algorithm

Algorithm: Boosted adversarial training

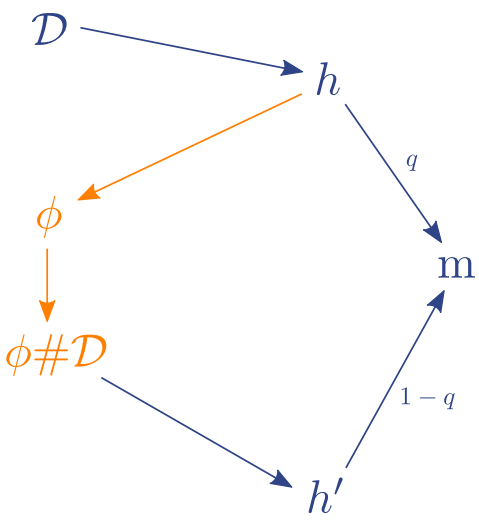
Inputs: D the training set and $q \in (0, 1)$

> Train h on D with adversarial training

> Generate an adversarial data set D' against h

> Train h' on D' with standard training

Return: $h \sim q$ and $h' \sim 1 - q$



From theorem to algorithm

Algorithm: Boosted adversarial training

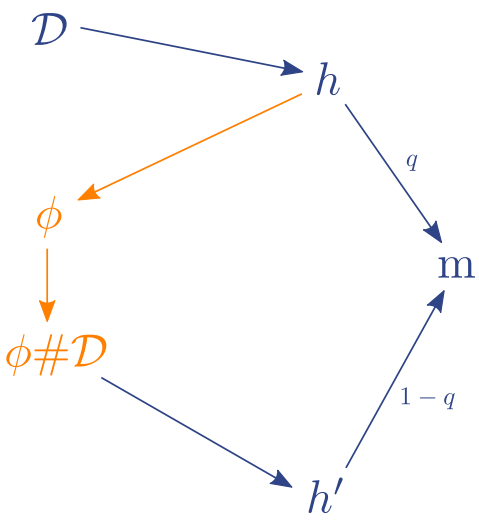
Inputs: D the training set and $q \in (0, 1)$

> Train h on D with adversarial training

> Generate an adversarial data set D' against h

> Train h' on D' with standard training

Return: $h \sim q$ and $h' \sim 1 - q$



Dataset	Method	Accuracy	l_∞-PGD
CIFAR10	Undefended	0.88	0.00
	Adversarial training	0.83	0.42
	Boosted adversarial training	0.80	**0.55**

Summary and next step

Summary:

- These results give a **formal motivation** for studying randomized classifiers

Summary and next step

Summary:

- These results give a **formal motivation** for studying randomized classifiers
- But the algorithm we just presented still lacks certification guarantees

Summary and next step

Summary:

- These results give a **formal motivation** for studying randomized classifiers
- But the algorithm we just presented still lacks certification guarantees

Next step

- We want to be able to bound gap the between the risks

Summary and next step

Summary:

- These results give a **formal motivation** for studying randomized classifiers
- But the algorithm we just presented still lacks certification guarantees

Next step

- We want to be able to bound gap the between the risks

→ To do so, we focus on another class of randomized classifiers

Summary and next step

Summary:

- These results give a **formal motivation** for studying randomized classifiers
- But the algorithm we just presented still lacks certification guarantees

Next step

- We want to be able to bound gap the between the risks

→ To do so, we focus on another class of randomized classifiers

$$\Delta_\Sigma := \{\mathrm{m} : x \mapsto h(x+z) \mid h : \mathcal{X} \to \mathbb{R} \text{ continuous, and } z \sim \mathcal{N}(0, \Sigma)\}$$

Summary and next step

Summary:

- These results give a **formal motivation** for studying randomized classifiers
- But the algorithm we just presented still lacks certification guarantees

Next step

- We want to be able to bound gap the between the risks

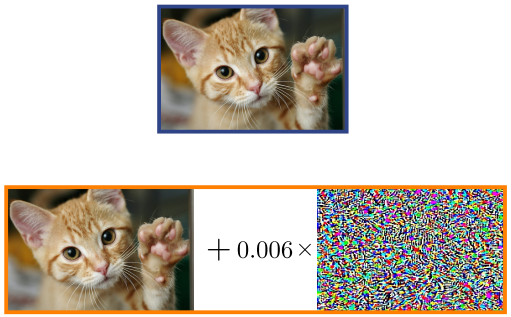
$\rightarrow$ To do so, we focus on another class of randomized classifiers

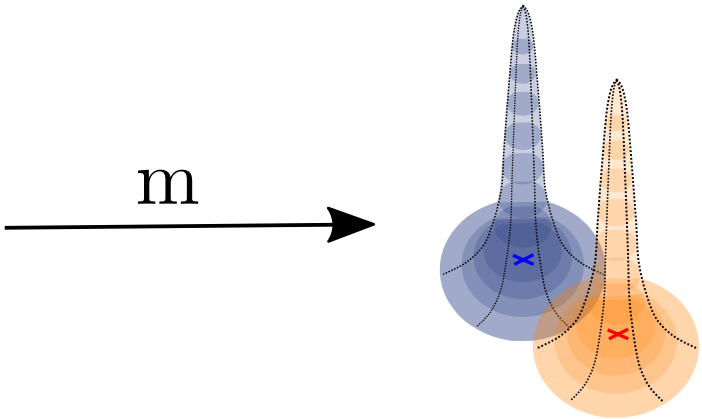
$$\Delta_\Sigma := \{\mathrm{m} : x \mapsto h(x+z) \mid h : \mathcal{X} \to \mathbb{R} \text{ continuous, and } z \sim \mathcal{N}(0, \Sigma)\}$$

Remark: Gaussian noise injection has been proved to regularize the training phase
Here we use it at **test time** to get robustness to data manipulation

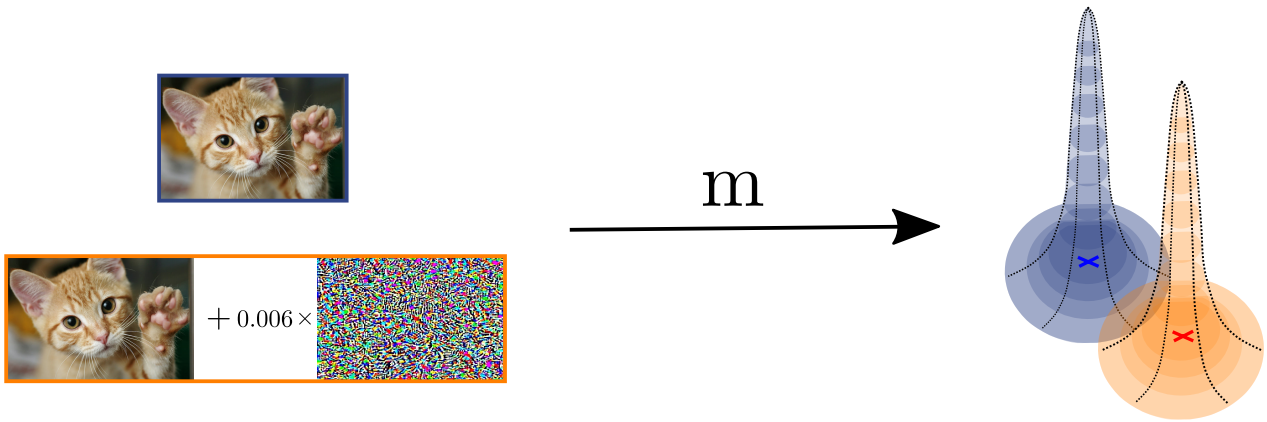
Part 2:
Demonstrating the robustness of $\Delta_{\sum}$

Randomized robustness: a Lipschitz condition





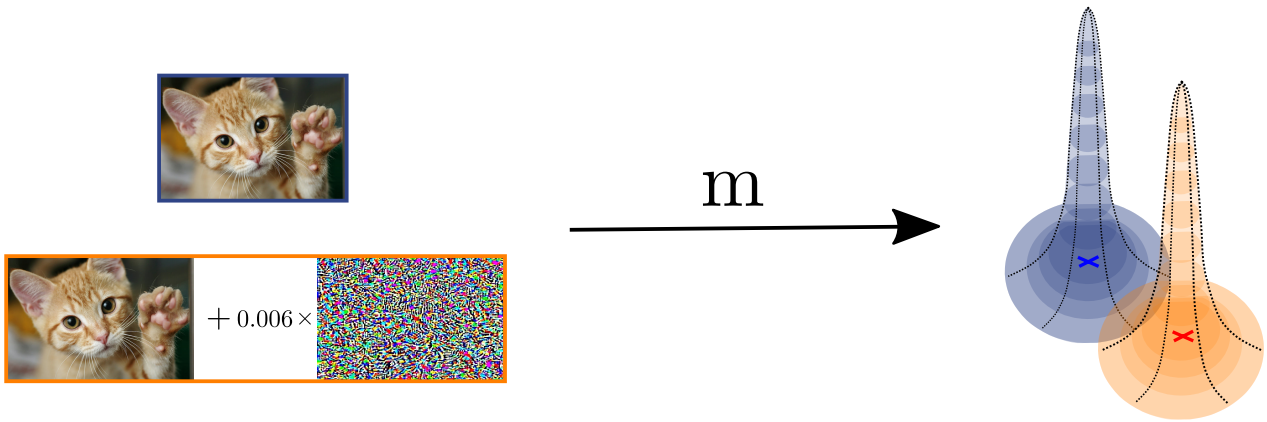
Randomized robustness: a Lipschitz condition



We define robustness for randomized classifiers with a **stability condition**

$$\forall \tau \in B(\alpha^*), \quad D\left(\mathrm{m}(x+\tau), \mathrm{m}(x)\right) \leq \epsilon$$

Randomized robustness: a Lipschitz condition

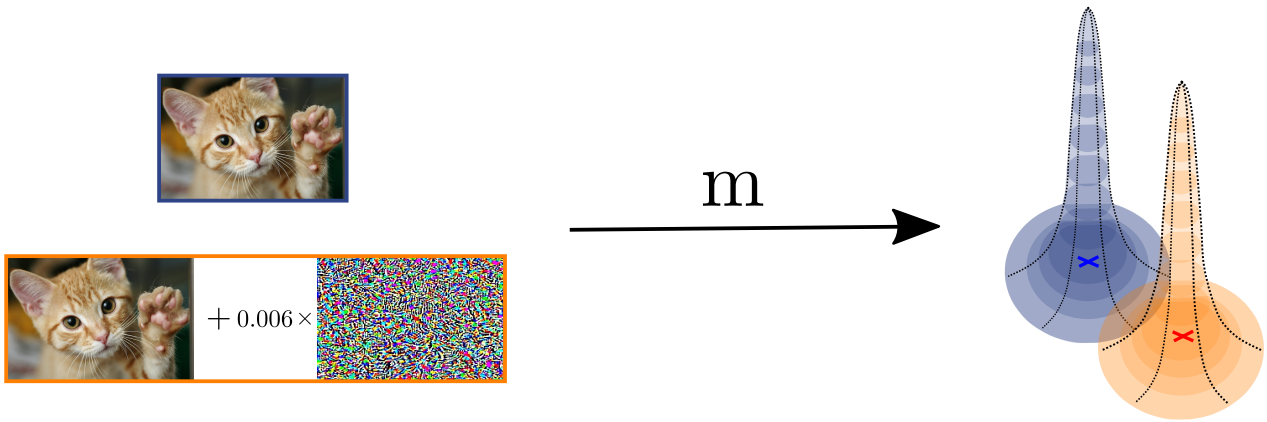


We define robustness for randomized classifiers with a **stability condition**

$$\forall \tau \in B(\alpha^*), \quad D\left(\mathrm{m}(x+\tau), \mathrm{m}(x)\right) \leq \epsilon$$

$$D_{TV}\left(P, Q\right) := \sup_{Z \in \mathcal{B}(\mathbb{R})} \mid P(Z) - Q(Z) \mid$$

Randomized robustness: a Lipschitz condition



We define robustness for randomized classifiers with a **stability condition**

$$\forall \tau \in B(\alpha^*), \quad D\left(\mathrm{m}(x+\tau), \mathrm{m}(x)\right) \leq \epsilon$$

$$D_{TV}\left(P, Q\right) := \sup_{Z \in \mathcal{B}(\mathbb{R})} \mid P(Z) - Q(Z) \mid$$

$$D_{\beta}\left(P, Q\right) := \frac{1}{\lambda - 1} \log\left(\mathbb{E}_{z \sim Q}\left[\frac{p^{\alpha}}{q^{\alpha}}(z)\right]\right)$$

Noise injection offers stability

Proposition (local Lipschitzness)

Let $\mathrm{m} \in \Delta_\Sigma$, then for any $x \in \mathcal{X}$, the following holds

$$\|\tau\|_{\Sigma^{-1}} \leq \alpha^* \implies D_\beta\left(\mathrm{m}(x), \mathrm{m}(x+\tau)\right) \leq \frac{(\alpha^*)^2 \beta}{2}.$$

Noise injection offers stability

Proposition (local Lipschitzness)

Let $\mathrm{m} \in \Delta_\Sigma$, then for any $x \in \mathcal{X}$, the following holds

$$\|\tau\|_{\Sigma^{-1}} \leq \alpha^* \implies D_\beta\left(\mathrm{m}(x), \mathrm{m}(x+\tau)\right) \leq \frac{(\alpha^*)^2 \beta}{2}.$$

Idea of the proof:

- Simple calculus $D_\beta\left(\mathcal{N}(x, \Sigma), \mathcal{N}(x+\tau, \Sigma)\right) \leq \frac{\beta}{2}\|\tau\|^2_{\Sigma^{-1}}$

Noise injection offers stability

Proposition (local Lipschitzness)

Let $\mathrm{m} \in \Delta_\Sigma$, then for any $x \in \mathcal{X}$, the following holds

$$\|\tau\|_{\Sigma^{-1}} \leq \alpha^* \implies D_\beta\left(\mathrm{m}(x), \mathrm{m}(x+\tau)\right) \leq \frac{(\alpha^*)^2 \beta}{2}.$$

Idea of the proof:

- Simple calculus $D_\beta\left(\mathcal{N}(x,\Sigma), \mathcal{N}(x+\tau,\Sigma)\right) \leq \frac{\beta}{2}\|\tau\|^2_{\Sigma^{-1}}$
- Data-processing inequality (Beaudry and Renner, 2012; Dwork and Roth, 2013)

$$D_\beta\left(h\#\mathcal{N}(x,\Sigma), h\#\mathcal{N}(x+\tau,\Sigma)\right) \leq D_\beta\left(\mathcal{N}(x,\Sigma), \mathcal{N}(x+\tau,\Sigma)\right)$$

Bounding the adversarial gap

Theorem (Bounding the accuracy vs robustness trade-off)

Let $\Sigma = \sigma^2 \times \mathbb{I}$ and $\mathrm{m} \in \Delta_\Sigma$. Then the following holds

$$\mathcal{R}_{\mathrm{adv}}(\mathrm{m}) - \mathcal{R}(\mathrm{m}) \leq 1 - e^{-\frac{(\alpha^*)^2}{2\sigma}} \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{|\mathcal{X}}} \left[e^{-H(\mathrm{m}(x))} \right].$$

Where H is the Shannon entropy $H(P) = -\mathbb{E}_{z \sim P} \left[\log(p(z)) \right]$.

Bounding the adversarial gap

Theorem (Bounding the accuracy vs robustness trade-off)

Let $\Sigma = \sigma^2 \times \mathbb{I}$ and $\mathrm{m} \in \Delta_\Sigma$. Then the following holds

$$\mathcal{R}_{\mathrm{adv}}(\mathrm{m}) - \mathcal{R}(\mathrm{m}) \leq 1 - e^{-\frac{(\alpha^*)^2}{2\sigma}} \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{|\mathcal{X}}} \left[e^{-H(\mathrm{m}(x))} \right].$$

Where H is the Shannon entropy $H(P) = -\mathbb{E}_{z \sim P} \left[\log(p(z)) \right]$.

Interpretation:

- σ is **large** (the entropy may be large too). We get **robustness over accuracy**

Bounding the adversarial gap

Theorem (Bounding the accuracy vs robustness trade-off)

Let $\Sigma = \sigma^2 \times \mathbb{I}$ and $\mathrm{m} \in \Delta_\Sigma$. Then the following holds

$$\mathcal{R}_{\mathrm{adv}}(\mathrm{m}) - \mathcal{R}(\mathrm{m}) \leq 1 - e^{-\frac{(\alpha^*)^2}{2\sigma}} \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{|\mathcal{X}}} \left[e^{-H(\mathrm{m}(x))} \right].$$

Where H is the Shannon entropy $H(P) = -\mathbb{E}_{z \sim P} \left[\log(p(z)) \right]$.

Interpretation:

- σ is **large** (the entropy may be large too). We get **robustness over accuracy**
- σ is **small** (the entropy will be too). We get **accuracy over robustness**

Bounding the adversarial gap

Theorem (Bounding the accuracy vs robustness trade-off)

Let $\Sigma = \sigma^2 \times \mathbb{I}$ and $\mathrm{m} \in \Delta_\Sigma$. Then the following holds

$$\mathcal{R}_{\mathrm{adv}}(\mathrm{m}) - \mathcal{R}(\mathrm{m}) \leq 1 - e^{-\frac{(\alpha^*)^2}{2\sigma}} \mathop{\mathbb{E}}_{x \sim \mathcal{D}_{|\mathcal{X}}} \left[e^{-H(\mathrm{m}(x))} \right].$$

Where H is the Shannon entropy $H(P) = -\mathbb{E}_{z \sim P} \left[\log(p(z)) \right]$.

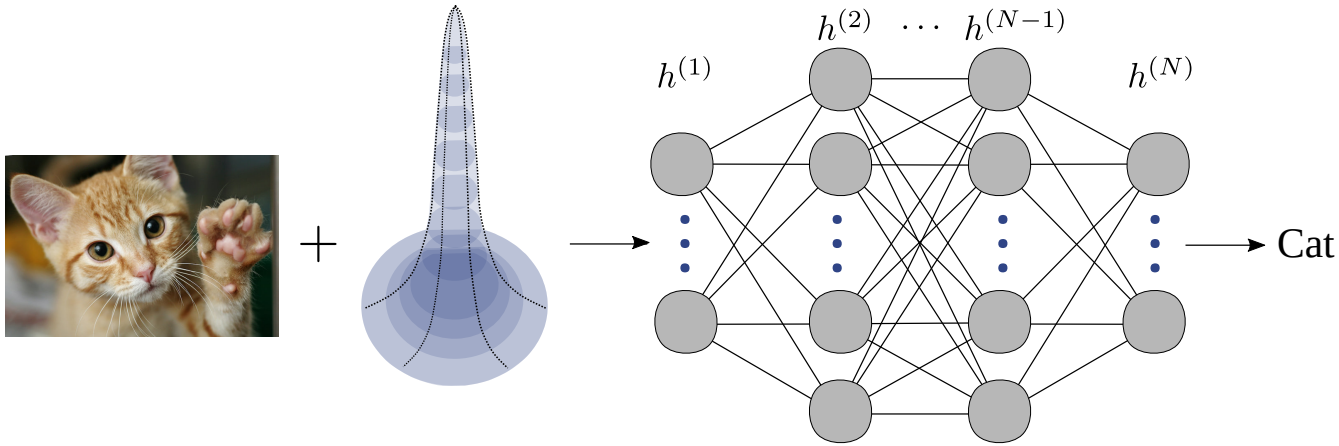
Interpretation:

- σ is **large** (the entropy may be large too). We get **robustness over accuracy**
- σ is **small** (the entropy will be too). We get **accuracy over robustness**

Remark: we get similar results with $\mathcal{R}_{\mathrm{adv}}(\mathrm{m}) - \mathcal{R}(\mathrm{m}) \leq 2F_{\mathcal{N}(0,1)}\left(\frac{\alpha^*}{2\sigma}\right) - 1$

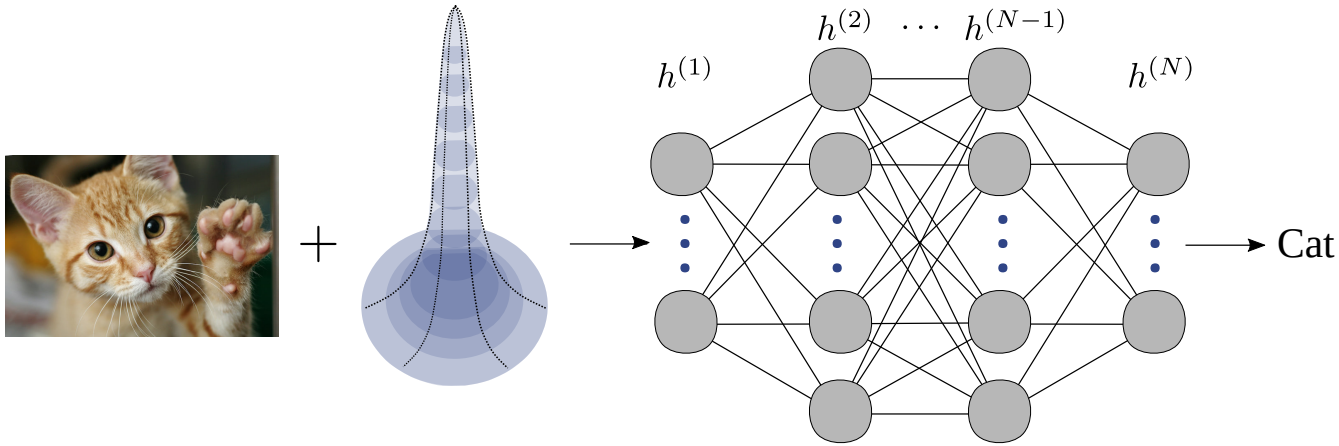
Consequences and applications

Simple noise injection methods to build robust randomized neural networks



Consequences and applications

Simple noise injection methods to build robust randomized neural networks

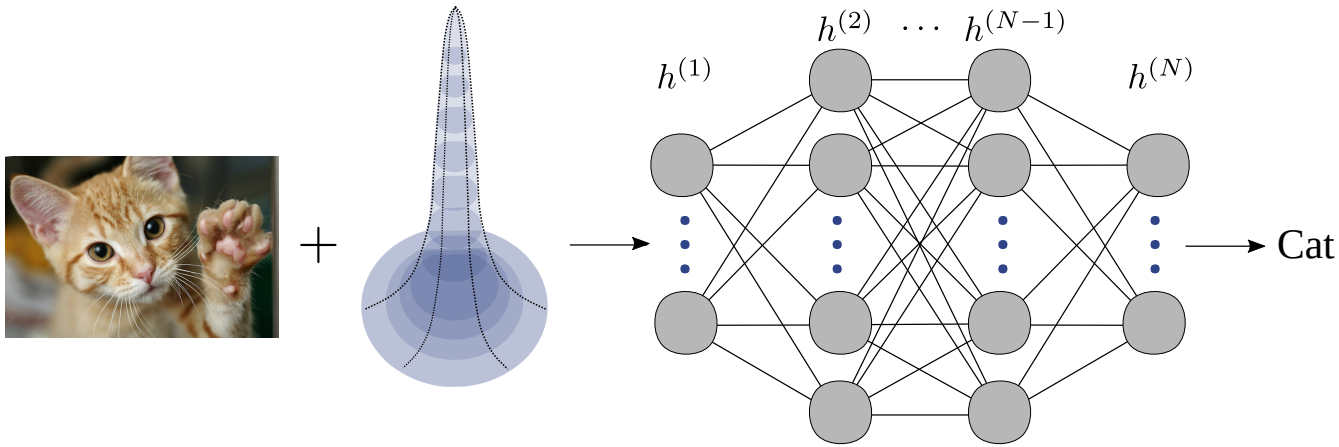


Remarks:

- In theory, we can inject noise **anywhere** (data-processing inequality)

Consequences and applications

Simple noise injection methods to build robust randomized neural networks

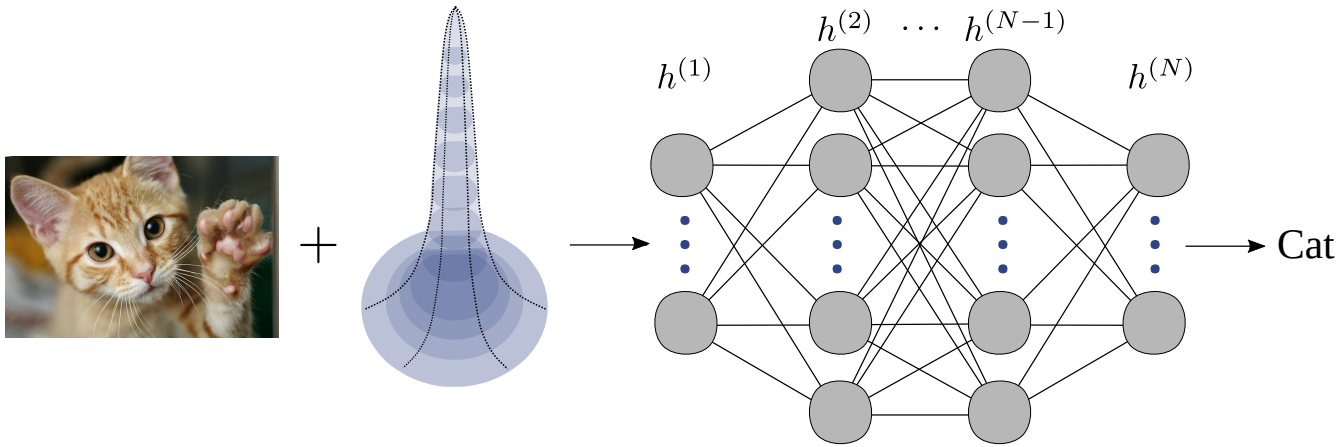


Remarks:

- In theory, we can inject noise **anywhere** (data-processing inequality)
- Similar results for exponential family noises (Laplace, Weibull, etc.)

Consequences and applications

Simple noise injection methods to build robust randomized neural networks



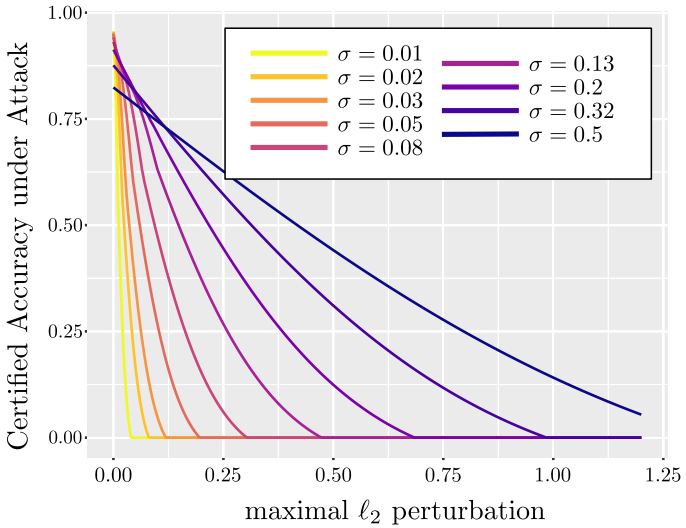
Remarks:

- In theory, we can inject noise **anywhere** (data-processing inequality)
- Similar results for exponential family noises (Laplace, Weibull, etc.)
- Results work for K-class classification (application to benchmark datasets)

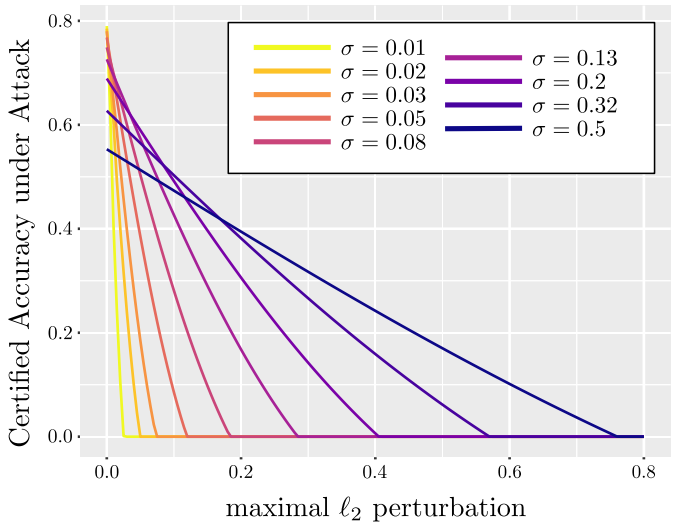
Application to benchmark datasets

Some results for WideResNets 28 (state-of-the-art model + Gaussian noise)

CIFAR-10



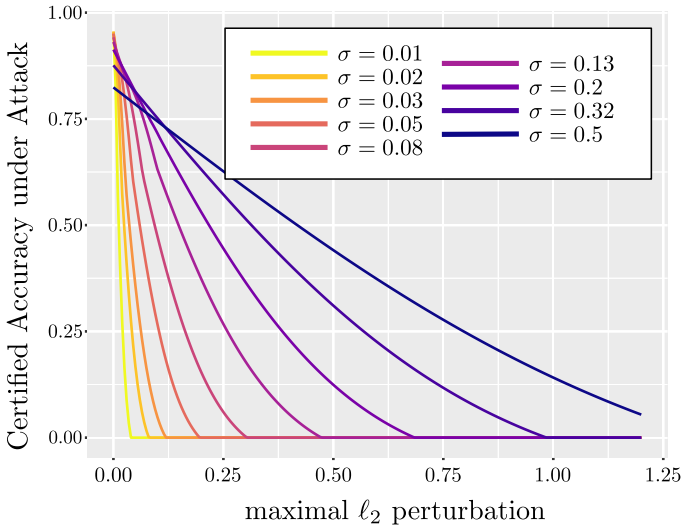
CIFAR-100



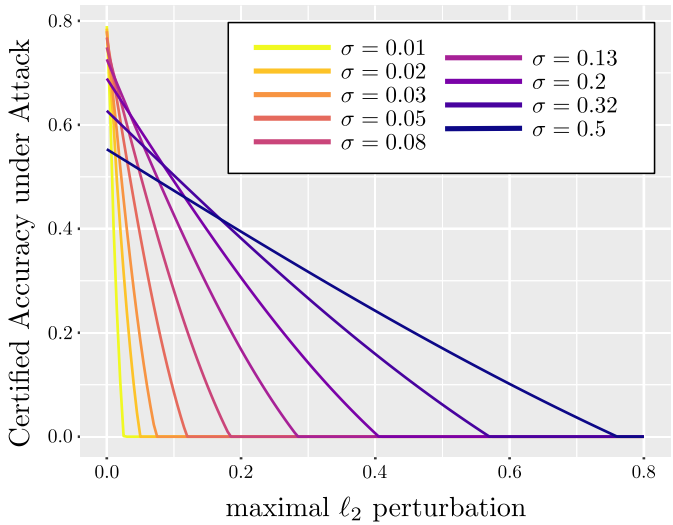
Application to benchmark datasets

Some results for WideResNets 28 (state-of-the-art model + Gaussian noise)

CIFAR-10



CIFAR-100

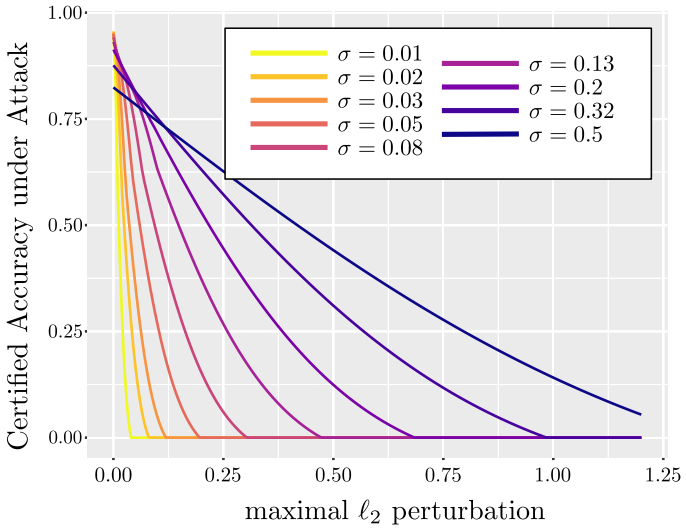


- Good guarantees for **small perturbation** ($\alpha^* \leq 0.5$)

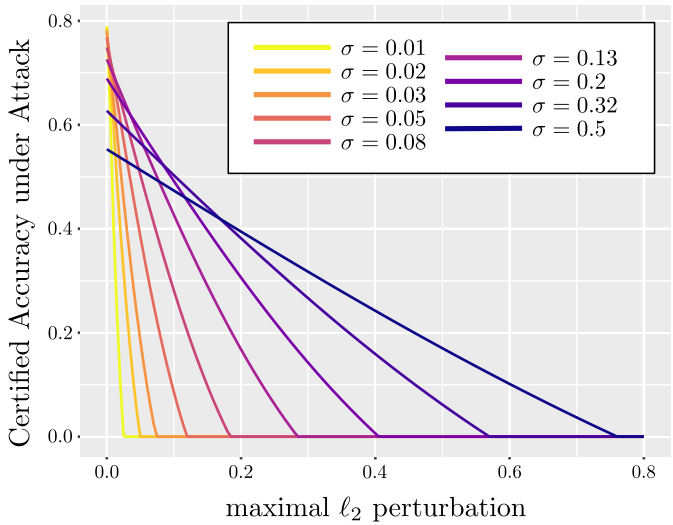
Application to benchmark datasets

Some results for WideResNets 28 (state-of-the-art model + Gaussian noise)

CIFAR-10



CIFAR-100

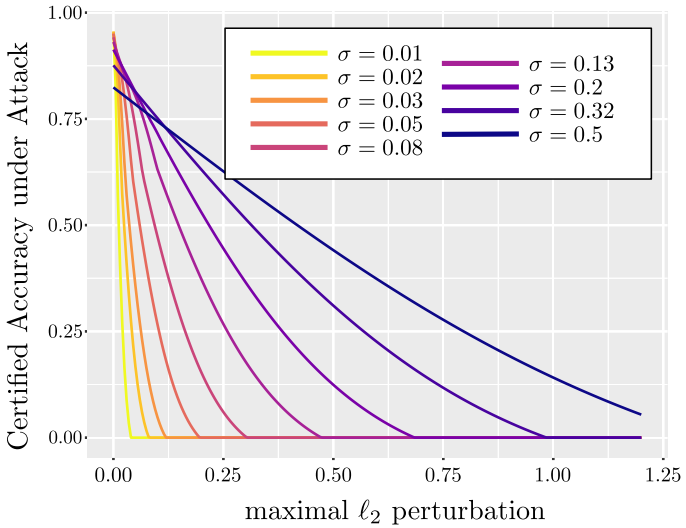


- Good guarantees for **small perturbation** ($\alpha^* \leq 0.5$)
- The attack remains **imperceptible for much bigger values** of α^*

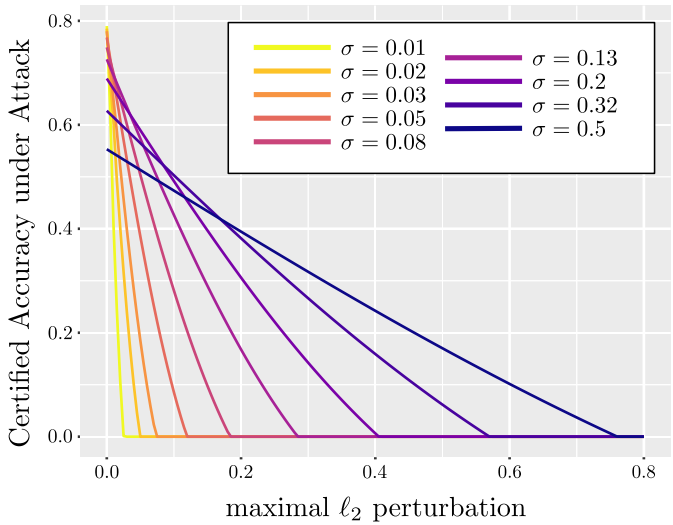
Application to benchmark datasets

Some results for WideResNets 28 (state-of-the-art model + Gaussian noise)

CIFAR-10



CIFAR-100



- Good guarantees for **small perturbation** ($\alpha^* \leq 0.5$)
- The attack remains **imperceptible for much bigger values** of α^*
- For **further improvements:** specify classifier and study the impact of the noise

Conclusion

Conclusion and perspectives

- Adversarial examples are a burning issue with genuine security impacts

Conclusion and perspectives

- Adversarial examples are a burning issue with genuine security impacts
- Adversarial examples are hard to study and to mitigate

Conclusion and perspectives

- Adversarial examples are a burning issue with genuine security impacts
- Adversarial examples are hard to study and to mitigate
- We studied randomized classifiers and presented some interesting results

Conclusion and perspectives

- Adversarial examples are a burning issue with genuine security impacts
- Adversarial examples are hard to study and to mitigate
- We studied randomized classifiers and presented some interesting results

Game theoretical point of view

Conclusion and perspectives

- Adversarial examples are a burning issue with genuine security impacts
- Adversarial examples are hard to study and to mitigate
- We studied randomized classifiers and presented some interesting results

Game theoretical point of view

> No Pure Nash Equilibrium
> Study the duality gap more closely

Conclusion and perspectives

- Adversarial examples are a burning issue with genuine security impacts
- Adversarial examples are hard to study and to mitigate
- We studied randomized classifiers and presented some interesting results

Game theoretical point of view

> No Pure Nash Equilibrium
> Study the duality gap more closely

> Randomization matters for robustness
> Mixed Nash Equilibrium and stability

Conclusion and perspectives

- Adversarial examples are a burning issue with genuine security impacts
- Adversarial examples are hard to study and to mitigate
- We studied randomized classifiers and presented some interesting results

Game theoretical point of view

> No Pure Nash Equilibrium
> Study the duality gap more closely

> Randomization matters for robustness
> Mixed Nash Equilibrium and stability

Accuracy vs robustness trade-off

Conclusion and perspectives

- Adversarial examples are a burning issue with genuine security impacts
- Adversarial examples are hard to study and to mitigate
- We studied randomized classifiers and presented some interesting results

Game theoretical point of view

> No Pure Nash Equilibrium
> Study the duality gap more closely

> Randomization matters for robustness
> Mixed Nash Equilibrium and stability

Accuracy vs robustness trade-off

> Bound on the trade-off (h- agnostic)
> Study the trade-off with assumptions on h

Conclusion and perspectives

- Adversarial examples are a burning issue with genuine security impacts
- Adversarial examples are hard to study and to mitigate
- We studied randomized classifiers and presented some interesting results

Game theoretical point of view

> No Pure Nash Equilibrium
> Study the duality gap more closely

> Randomization matters for robustness
> Mixed Nash Equilibrium and stability

Accuracy vs robustness trade-off

> Bound on the trade-off (h- agnostic)
> Study the trade-off with assumptions on h

> Analysis based on D_β and D_{TV}
> Extend the class of divergences

Some broader perspectives

- Revisiting the adversarial game with **different objectives**

$$\begin{cases} \sup\limits_{\phi \in \mathcal{F}} & \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\ell_{\text{adv}} \left(h \left(\phi(x) \right), y \right) \right], \quad \text{for a given hypothesis } h \\ \inf\limits_{h \in \mathcal{H}} & \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\ell_{\text{def}} \left(h \left(\phi(x) \right), y \right) \right], \quad \text{for a given adversary } \phi \end{cases}$$

Some broader perspectives

- Revisiting the adversarial game with **different objectives**

$$\begin{cases} \sup\limits_{\phi \in \mathcal{F}} & \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\ell_{\text{adv}} \left(h \left(\phi(x) \right), y \right) \right], \quad \text{for a given hypothesis } h \\ \inf\limits_{h \in \mathcal{H}} & \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\ell_{\text{def}} \left(h \left(\phi(x) \right), y \right) \right], \quad \text{for a given adversary } \phi \end{cases}$$

- Revisiting adversarial learning theory with regard to double descent and interpolation regimes (Zhang et al., 2017; Belkin et al., 2019)

Some broader perspectives

- Revisiting the adversarial game with **different objectives**

$$\begin{cases} \sup\limits_{\phi \in \mathcal{F}} & \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\ell_{\text{adv}} \left(h \left(\phi(x) \right), y \right) \right], \quad \text{for a given hypothesis } h \\ \inf\limits_{h \in \mathcal{H}} & \mathbb{E}_{(x,y)\sim\mathcal{D}} \left[\ell_{\text{def}} \left(h \left(\phi(x) \right), y \right) \right], \quad \text{for a given adversary } \phi \end{cases}$$

- Revisiting adversarial learning theory with regard to double descent and interpolation regimes (Zhang et al., 2017; Belkin et al., 2019)

- Transfer existing results from image to sound processing. Adversarial examples also work well on speech-to-text (Carlini and Wagner, 2018)

Published works

On robustness to adversarial examples:

Randomization matters. how to defend against strong adversarial attacks
(Pinot et al., 2020) ICML 2020

Advocating for Multiple Defense Strategies against Adversarial Examples
(Araujo et al., 2020) Workshop at ECML 2020

Theoretical evidence for robustness through randomization (Journal TBA)
(Pinot et al., 2019a) NeurIPS 2019

On privacy preserving data analysis:

SPEED: Secure, PrivatE, and Efficient Deep learning (under review)
(Sébert et al., 2020) Journal track ECML 2021

Unified view on differential privacy and robustness to adversarial examples
(Pinot et al., 2019b) Workshop at ECML 2019

Graph-based Clustering under Differential Privacy
(Pinot et al., 2018) UAI 2018

We also participated to: Teaching, scientific outreach, coding, etc.

Questions

f-divergences (f convex): $D_f(P,Q) := \mathbb{E}_{z\sim Q}\left[f\left(p(z)/q(z)\right)\right]$

- Total variation $f(t) = |t-1|$, Kullback Leibler $f(t) = |t\log t|$
- Pro: Data-processing inequality / Con: accuracy vs robustness trade-off

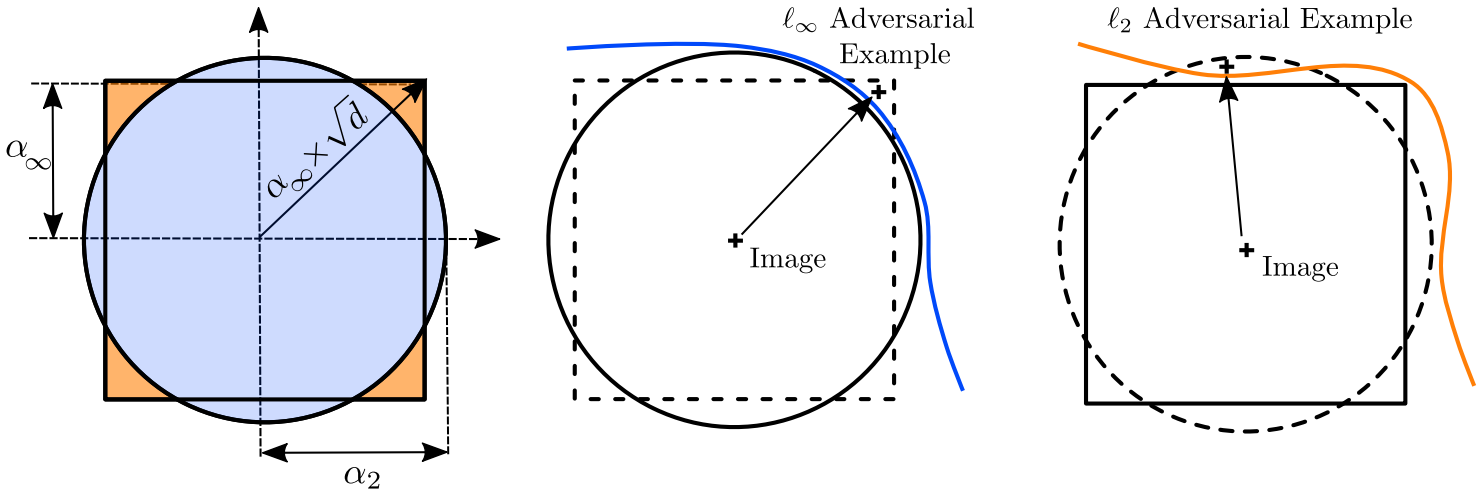
Integral probability metrics: $\gamma_{\mathcal{G}}(P,Q) := \sup_{g\in\mathcal{G}} |\mathbb{E}_{z\sim Q}\left[g(z)\right] - \mathbb{E}_{z\sim P}\left[g(z)\right]|$

- Total variation $\mathcal{G} = \{g \mid ||g||_\infty \leq 1\}$, Wasserstein $\mathcal{G} = \{g \mid ||g||_L \leq 1\}$
- Links with f-divergences (Sriperumbudur et al., 2009)

Bregman divergences (U convex): $D_U(P,Q) := U(p) - U(q) - \langle \nabla U(q), p-q\rangle$

- Kullback Leibler $U(p) = \sum p_i \log(p_i)$, Itakura–Saito $U(p) = -\sum \log(p_i)$
- Asymptotic equivalence with f-divergences (Pardo and Vajda, 2003)

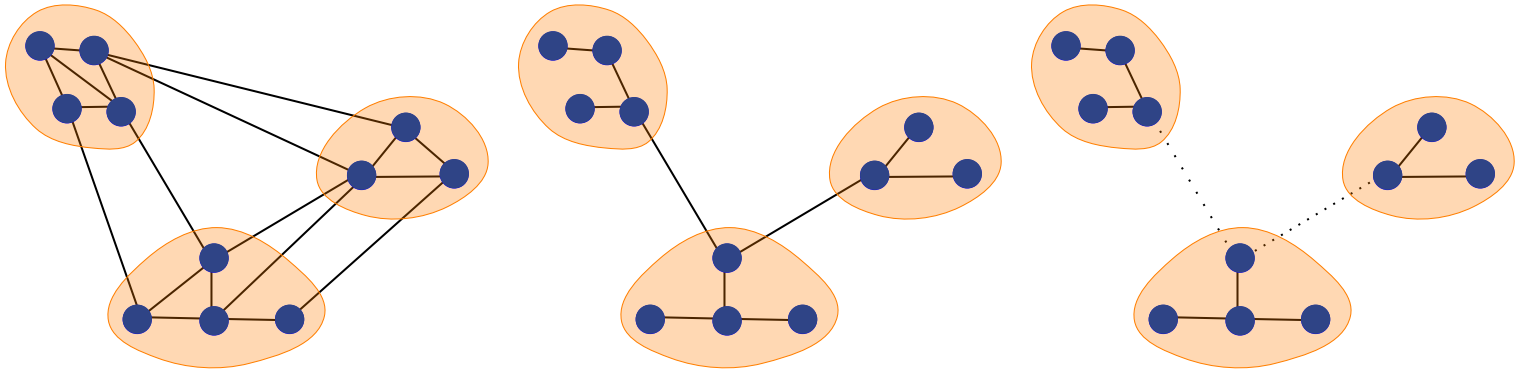
Selection of the norm in high dimension



- The volume of the intersection is negligible when d is sufficiently large
- Results obtained for an ℓ_p norm **do not generalize** to other norms

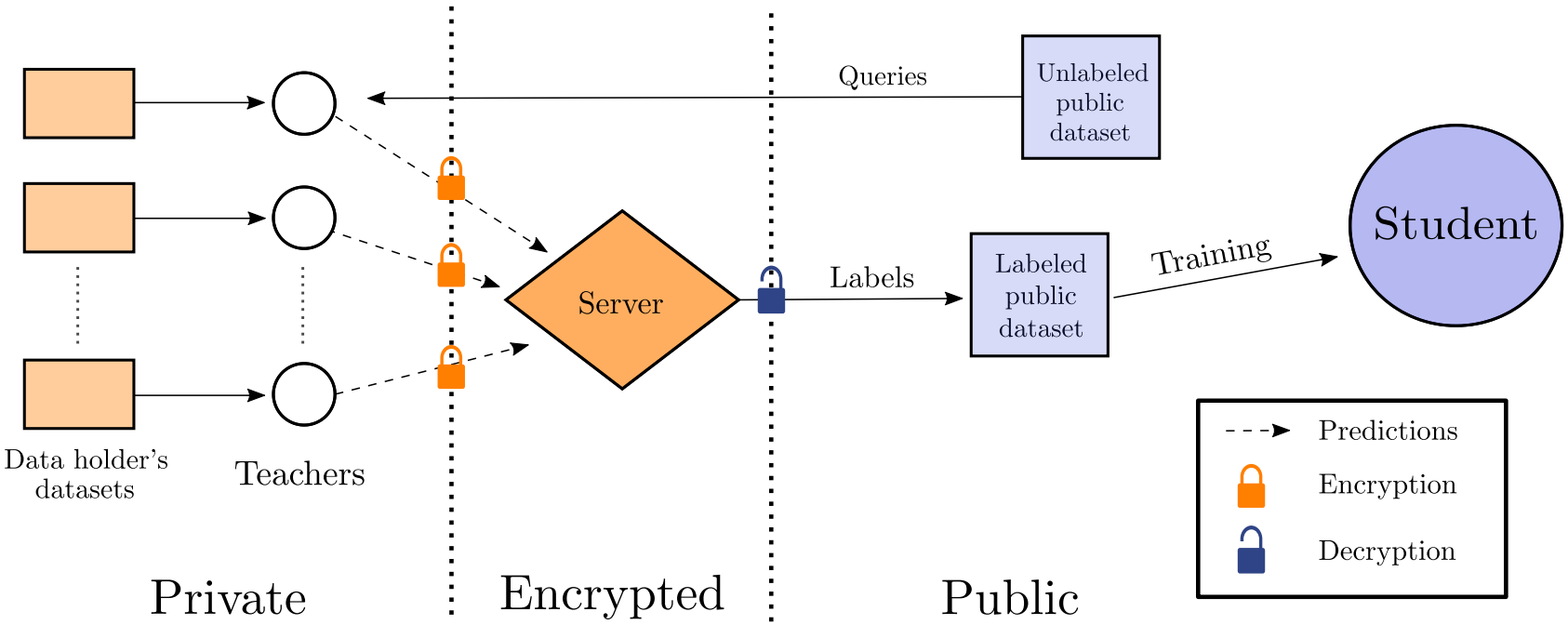
$\rightarrow$ Mixing defense strategies (Tramèr and Boneh, 2019; Maini et al., 2020)

Graph clustering with differential privacy constraints



- Start from a Tree-based clustering algorithm (Schaeffer, 2007)
- Demonstrate that the clustering method provably works (Pinot et al., 2018)
- Change Step 1 with a **privacy preserving** Tree algorithm (Pinot, 2017)
- Use the **data-processing inequality** to get results for Step 2

Secure and private deep learning with encrypted aggregation operator



References

Alexandre Araujo, Laurent Meunier, Rafael Pinot, and Benjamin Negrevergne. Advocating for multiple defense strategies against adversarial examples. *Workshop on Machine Learning for CyberSecurity (MLCS@ECML-PKDD)*, 2020.

Pranjal Awasthi, Natalie Frank, and Mehryar Mohri. Adversarial learning guarantees for linear hypotheses and neural networks. *International Conference on Machine Learning*, 2020.

Normand J. Beaudry and Renato Renner. An intuitive proof of the data processing inequality. *Quantum Info. Comput.*, 12(5-6):432–441, May 2012. ISSN 1533-7146.

Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

References ii

Arjun Nitin Bhagoji, Daniel Cullina, and Prateek Mittal. Lower bounds on adversarial robustness from optimal transport. In *Advances in Neural Information Processing Systems 32*, pages 7496–7508. Curran Associates, Inc., 2019.

Battista Biggio, Igino Corona, Davide Maiorca, Blaine Nelson, Nedim Šrndić, Pavel Laskov, Giorgio Giacinto, and Fabio Roli. Evasion attacks against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pages 387–402. Springer, 2013.

Sébastien Bubeck, Yin Tat Lee, Eric Price, and Ilya Razenshteyn. Adversarial examples from computational constraints. In *International Conference on Machine Learning*, pages 831–840, 2019.

N. Carlini and D. Wagner. Audio adversarial examples: Targeted attacks on speech-to-text. In *2018 IEEE Security and Privacy Workshops (SPW)*, pages 1–7, 2018. doi: 10.1109/SPW.2018.00009.

Jeremy M. Cohen, Elan Rosenfeld, and J. Zico Kolter. Certified adversarial robustness via randomized smoothing. *CoRR*, abs/1902.02918, 2019.

References iii

C. Dwork and A. Roth. The algorithmic foundations of differential privacy. *Foundations and Trends® in Theoretical Computer Science*, 9(3-4):211–407, 2013.

Ian Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

Aleksander Madry, Aleksandar Makelov, Ludwig Schmidt, Dimitris Tsipras, and Adrian Vladu. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Pratyush Maini, Eric Wong, and J Zico Kolter. Adversarial robustness against the union of multiple perturbation models. *International Conference on Machine Learning*, 2020.

References iv

M. C. Pardo and I. Vajda. On asymptotic properties of information-theoretic divergences. *IEEE Transactions on Information Theory*, 49(7):1860–1867, 2003. doi: 10.1109/TIT.2003.813509.

Rafael Pinot. Minimum spanning tree release under differential privacy constraints. Master's thesis, Sorbonne University, 2017.

Rafael Pinot, Anne Morvan, Florian Yger, Cédric Gouy-Pailler, and Jamal Atif. Graph-based clustering under differential privacy. *Uncertainty in Artificial Intelligence*, 2018.

Rafael Pinot, Laurent Meunier, Alexandre Araujo, Hisashi Kashima, Florian Yger, Cédric Gouy-Pailler, and Jamal Atif. Theoretical evidence for adversarial robustness through randomization. In *Advances in Neural Information Processing Systems*, pages 11838–11848, 2019a.

Rafael Pinot, Florian Yger, Cédric Gouy-Pailler, and Jamal Atif. A unified view on differential privacy and robustness to adversarial examples. *MLCS Workshop at ECML*, 2019b.

References v

Rafael Pinot, Raphael Ettedgui, Geovani Rizk, Yann Chevaleyre, and Jamal Atif. Randomization matters. how to defend against strong adversarial attacks. *International Conference on Machine Learning*, 2020.

Muni Sreenivas Pydi and Varun Jog. Adversarial risk via optimal transport and optimal couplings. In *International Conference on Machine Learning*. 2020.

Hadi Salman, Jerry Li, Ilya Razenshteyn, Pengchuan Zhang, Huan Zhang, Sebastien Bubeck, and Greg Yang. Provably robust deep learning via adversarially trained smoothed classifiers. In *Advances in Neural Information Processing Systems*, pages 11289–11300, 2019.

S. E. Schaeffer. Graph clustering. *Computer Science Review*, 1(1):27 – 64, 2007.

Bharath K. Sriperumbudur, Arthur Gretton, Kenji Fukumizu, Gert R. G. Lanckriet, and Bernhard Schölkopf. A note on integral probability metrics and ϕ-divergences. *CoRR*, abs/0901.2698, 2009. URL *http://arxiv.org/abs/0901.2698*.

References vi

Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Arnaud Grivet Sébert, Rafael Pinot, Martin Zuber, Cédric Gouy-Pailler, and Renaud Sirdey. Speed: Secure, private, and efficient deep learning. *arive preprint, 2006.09475*, 2020.

Florian Tramèr and Dan Boneh. Adversarial training and robustness for multiple perturbations. In *Advances in Neural Information Processing Systems*, pages 5866–5876, 2019.

Dong Yin, Ramchandran Kannan, and Peter Bartlett. Rademacher complexity for adversarially robust generalization. In *International Conference on Machine Learning*, pages 7085–7094, 2019.

Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *International Conference on Learning Representation*, 2017.

References vii

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. *International conference on Machine Learning*, 2019.

Picture License: Pictures and photos in slides 1, 3, 6 and 21 have been designed by pch.vector / Freepik