Limits on Robustness to Adversarial Examples

Elvis Dohmatob

Criteo AI Lab

October 2, 2019

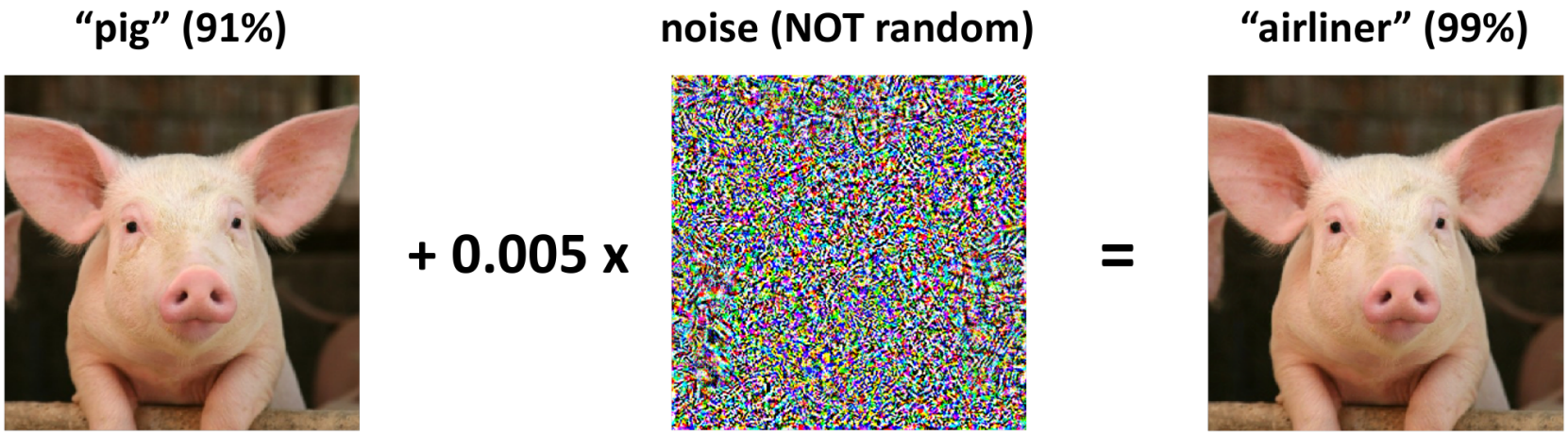
Table of contents

Preliminaries on adversarial robustness

Definition of adversarial attacks

- A **classifier** is trained and deployed (e.g the computer vision system on a self-driving car)
- At test / inference time, an attacker may submit queries to the classifier by
  - sampling a real sample point x with true label k (e.g "pig"),
  - modifying it $x \mapsto x^{\text{adv}}$ given to a prescribed **threat model**.
- Goal of attacker is to make classifier label x^{adv} as $\neq k$ (e.g airliner)

The flying pig!



(Picture is courtesy of https://gradientscience.org/intro_adversarial/)

- $x \mapsto x^{\text{adv}} := x +$ noise , $\|\text{noise}\| \leq \varepsilon = 0.005$ (in example above)
- **Fast Gradient Sign Method**: noise $= \text{sign}(\nabla_x loss(h(x), y))$

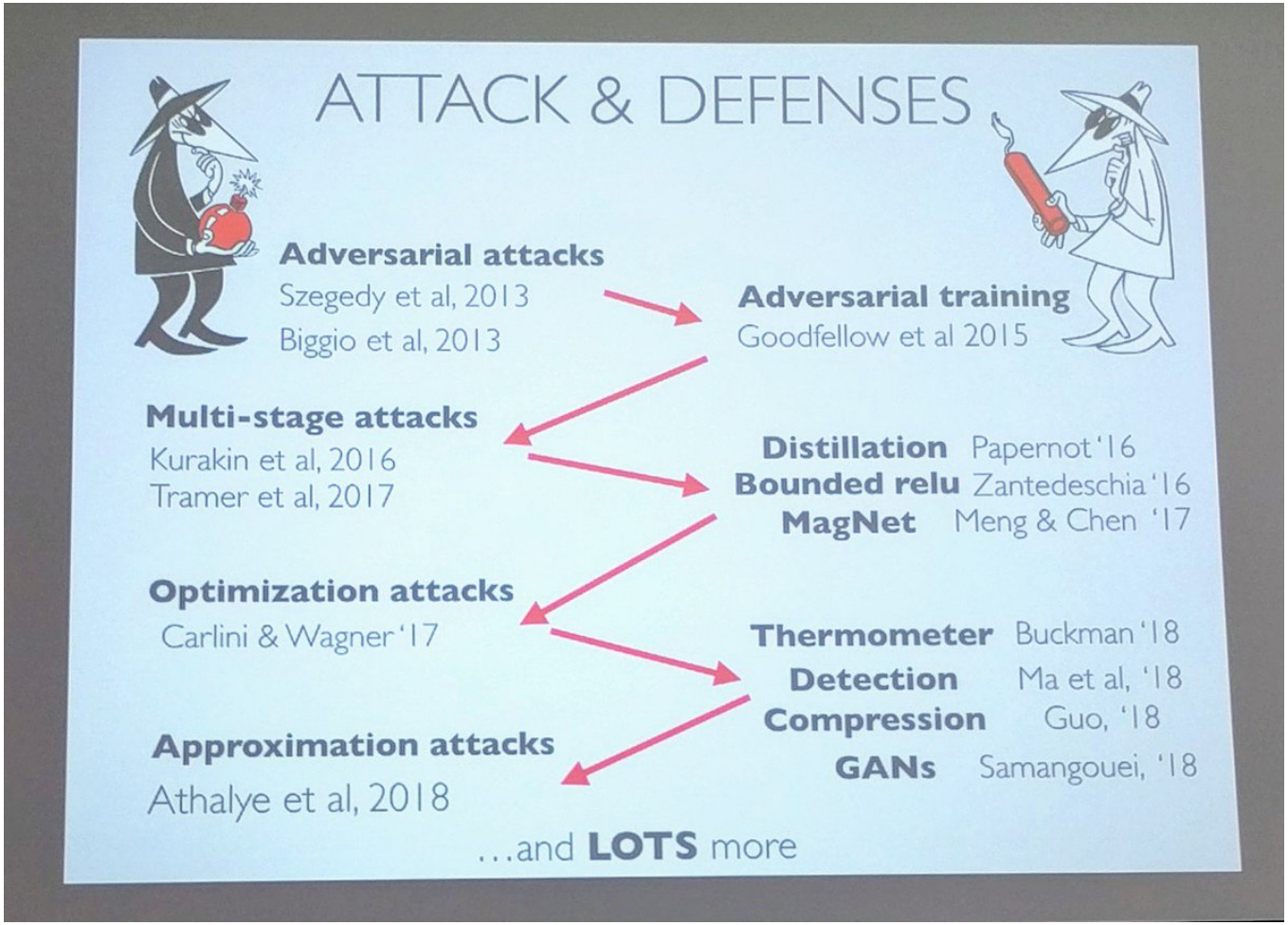
FGSM for generating adversarial examples [Goodfellow '14]

```python
class WhiteBoxAttack(object):
    """
    Gets adversarial perturbation of features (in sup norm threat model)
    via a first-order Taylor approx. of the loss function, w.r.t the features
    """
    def __init__(self, epsilon):
        self.epsilon = epsilon

    def __call__(self, model, true_features, true_labels, loss_func=F.nll_loss):
        true_features = Variable(true_features, requires_grad=True)
        pred_labels = model.forward(true_features)
        loss = loss_func(pred_labels, true_labels)
        grad = autograd.grad(loss, true_features, retain_graph=True)[0]
        return true_features + self.epsilon * torch.sign(grad)  # move uphill
```

- $x \mapsto x^{\text{adv}} := \text{clip}(x + \varepsilon \text{sign}(\nabla_x loss(h(x), y)))$

Adversarial attacks and defenses, an arms race!



- Image courtesy of [Goldstein' 19; Shafahi '19]

Classifier-dependent lower bounds

Problem setup

A **classifier** is simply a Borel-measurable mapping $h : \mathcal{X} \to \mathcal{Y}$ from feature space $\mathcal{X}$ (with metric d) to label space $\mathcal{Y} := \{1, \ldots, K\}$.

- A classifier is trained and deployed (e.g the computer vision system on a self-driving car)
- At test / inference time, an attacker may submit queries to the classifier by sampling a real sample point $x \in \mathcal{X}$ with true label $k \in \mathcal{Y}$, and modifying it $x \mapsto x^{\text{adv}}$ according to a prescribed threat model.
  - For example, modifying a few pixels on a road traffic sign [Su et al. '17]
  - Modifying intensity of pixels by a limited amount determined by a prescribed tolerance level [Tsipras '18], etc., on it.

Problem setup

A **classifier** is simply a Borel-measurable mapping $h : \mathcal{X} \to \mathcal{Y}$ from feature space $\mathcal{X}$ (with metric d) to label space $\mathcal{Y} := \{1, \ldots, K\}$.

- A classifier is trained and deployed (e.g the computer vision system on a self-driving car)
- At test / inference time, an attacker may submit queries to the classifier by sampling a real sample point $x \in \mathcal{X}$ with true label $k \in \mathcal{Y}$, and modifying it $x \mapsto x^{\text{adv}}$ according to a prescribed threat model.
  - For example, modifying a few pixels on a road traffic sign [Su et al. '17]
  - Modifying intensity of pixels by a limited amount determined by a prescribed tolerance level [Tsipras '18], etc., on it.

Problem setup: notations

- **Standard accuracy:** $\mathrm{acc}(h|k) := 1 - \mathrm{err}(h|k)$, where $\mathrm{err}(h|k) := P_{X|k}(h(X) \neq k)$ is the error of h on class k.
  - Small $\mathrm{acc}(h|k) \implies h$ is **inaccurate** on class k.
- **Adversarial robustness accuracy:** $\mathrm{acc}_\varepsilon(h|k) := 1 - \mathrm{err}_\varepsilon(h|k)$, where $\mathrm{err}_\varepsilon(h|k) := P_{X|k}(\exists x' \in \mathrm{Ball}(X;\varepsilon) \mid h(x') \neq k)$ is the adversarial robustness error of h on class k.
  - Small $\mathrm{acc}_\varepsilon(h|k) \implies h$ is **vulnerable** to attacks on class k.

Problem setup: notations

- **Standard accuracy:** $\mathrm{acc}(h|k) := 1 - \mathrm{err}(h|k)$, where $\mathrm{err}(h|k) := P_{X|k}(h(X) \neq k)$ is the error of h on class k.
  - Small $\mathrm{acc}(h|k) \implies h$ is **inaccurate** on class k.
- **Adversarial robustness accuracy:** $\mathrm{acc}_\varepsilon(h|k) := 1 - \mathrm{err}_\varepsilon(h|k)$, where $\mathrm{err}_\varepsilon(h|k) := P_{X|k}(\exists x' \in \mathrm{Ball}(X;\varepsilon) \mid h(x') \neq k)$ is the adversarial robustness error of h on class k.
  - Small $\mathrm{acc}_\varepsilon(h|k) \implies h$ is **vulnerable** to attacks on class k.
- **Distance to error set:** $d(h|k) := \mathbb{E}_{P_{X|k}}[d(X, B(h,k))]$ denotes the average distance of a sample point of true label k, from the error set $B(h,k) := \{x \in \mathcal{X} \mid h(x) \neq k\}$ of samples assigned to another label.
  - Small $d(h|k) \implies h$ is **vulnerable** to attacks on class k.

Problem setup: notations

- **Standard accuracy:** $\mathrm{acc}(h|k) := 1 - \mathrm{err}(h|k)$, where $\mathrm{err}(h|k) := P_{X|k}(h(X) \neq k)$ is the error of h on class k.
  - Small $\mathrm{acc}(h|k) \implies h$ is **inaccurate** on class k.
- **Adversarial robustness accuracy:** $\mathrm{acc}_\varepsilon(h|k) := 1 - \mathrm{err}_\varepsilon(h|k)$, where $\mathrm{err}_\varepsilon(h|k) := P_{X|k}(\exists x' \in \mathrm{Ball}(X;\varepsilon) \mid h(x') \neq k)$ is the adversarial robustness error of h on class k.
  - Small $\mathrm{acc}_\varepsilon(h|k) \implies h$ is **vulnerable** to attacks on class k.
- **Distance to error set:** $d(h|k) := \mathbb{E}_{P_{X|k}}[d(X, B(h,k))]$ denotes the average distance of a sample point of true label k, from the error set $B(h,k) := \{x \in \mathcal{X} \mid h(x) \neq k\}$ of samples assigned to another label.
  - Small $d(h|k) \implies h$ is **vulnerable** to attacks on class k.

A motivating example (from [Tsipras '18])

Consider the following classification problem:

- **Prediction target**: $Y \sim \text{Bern}(1/2, \{\pm 1\})$ based on $p \geq 2$ explanatory variables $X := (X^1, X^2, \ldots, X^p)$ given by
- **Robust feature**: $X^1 \mid Y = +Y$ w.p 70% and $-Y$ w.p. 30%.

A motivating example (from [Tsipras '18])

Consider the following classification problem:

- **Prediction target**: $Y \sim \text{Bern}(1/2, \{\pm 1\})$ based on $p \geq 2$ explanatory variables $X := (X^1, X^2, \ldots, X^p)$ given by
- **Robust feature**: $X^1 \mid Y = +Y$ w.p 70% and $-Y$ w.p. 30%.
- **Non-robust features**: $X^j \mid Y \sim \mathcal{N}(\eta Y, 1)$, for $j = 2, \ldots, p$, where $\eta \sim p^{-1/2}$ is a fixed scalar which controls the difficulty.

A motivating example (from [Tsipras '18])

Consider the following classification problem:

- **Prediction target**: $Y \sim \text{Bern}(1/2, \{\pm 1\})$ based on $p \geq 2$ explanatory variables $X := (X^1, X^2, \ldots, X^p)$ given by
- **Robust feature**: $X^1 \mid Y = +Y$ w.p 70% and $-Y$ w.p. 30%.
- **Non-robust features**: $X^j \mid Y \sim \mathcal{N}(\eta Y, 1)$, for $j = 2, \ldots, p$, where $\eta \sim p^{-1/2}$ is a fixed scalar which controls the difficulty.

The linear classifier $h_{\text{lin}}(x) \equiv \text{sign}(w^T x)$ with $w = (0, 1/p, \ldots, 1/p)$, where we allow ℓ_∞-perturbations of maximum size $\varepsilon \geq 2\eta$, **solves the problem perfectly** (100% accuracy) but its **adversarial robustness is zero**!

A motivating example (from [Tsipras '18])

Consider the following classification problem:

- **Prediction target**: $Y \sim \text{Bern}(1/2, \{\pm 1\})$ based on $p \geq 2$ explanatory variables $X := (X^1, X^2, \ldots, X^p)$ given by
- **Robust feature**: $X^1 \mid Y = +Y$ w.p 70% and $-Y$ w.p. 30%.
- **Non-robust features**: $X^j \mid Y \sim \mathcal{N}(\eta Y, 1)$, for $j = 2, \ldots, p$, where $\eta \sim p^{-1/2}$ is a fixed scalar which controls the difficulty.

The linear classifier $h_{\text{lin}}(x) \equiv \text{sign}(w^T x)$ with $w = (0, 1/p, \ldots, 1/p)$, where we allow ℓ_∞-perturbations of maximum size $\varepsilon \geq 2\eta$, **solves the problem perfectly** (100% accuracy) but its **adversarial robustness is zero**!

Proof.

The **standard accuracy** of the classifier writes

$$\begin{aligned}\mathrm{acc}(h_{\mathrm{lin}}) &:= \mathbb{P}_{(X,Y)}\left(h_{\mathrm{lin}}(X) = Y\right) = \mathbb{P}\left(Yw^T X \geq 0\right) \\ &= \mathbb{P}_Y\left((Y/(p-1))\sum_{j\geq 2}\mathcal{N}(\eta Y, 1) \geq 0\right) \\ &= \mathbb{P}\left(\mathcal{N}(\eta, 1/(p-1)) \geq 0\right) = \mathbb{P}\left(\mathcal{N}(0, 1/(p-1)) \geq -\eta\right) \\ &= \mathbb{P}\left(\mathcal{N}(0, 1/(p-1)) \leq \eta\right) \geq 1 - e^{-(p-1)\eta^2/2},\end{aligned}$$

which is $\geq 1-\delta$ if $\eta \geq \sqrt{2\log(1/\delta)/(p-1)}$.

- $\Longrightarrow$ h_{lin} is quasi-perfect!

Proof.

The **standard accuracy** of the classifier writes

$$\begin{aligned}\mathrm{acc}(h_{\mathrm{lin}}) &:= \mathbb{P}_{(X,Y)}\left(h_{\mathrm{lin}}(X) = Y\right) = \mathbb{P}\left(Yw^T X \geq 0\right) \\ &= \mathbb{P}_Y\left((Y/(p-1))\sum_{j\geq 2}\mathcal{N}(\eta Y, 1) \geq 0\right) \\ &= \mathbb{P}\left(\mathcal{N}(\eta, 1/(p-1)) \geq 0\right) = \mathbb{P}\left(\mathcal{N}(0, 1/(p-1)) \geq -\eta\right) \\ &= \mathbb{P}\left(\mathcal{N}(0, 1/(p-1)) \leq \eta\right) \geq 1 - e^{-(p-1)\eta^2/2},\end{aligned}$$

which is $\geq 1-\delta$ if $\eta \geq \sqrt{2\log(1/\delta)/(p-1)}$.

- $\Longrightarrow$ h_{lin} is quasi-perfect!

Proof.

the **adversarial robustness accuracy** of h_{lin} writes

$$\begin{aligned}
&\text{acc}_\varepsilon(h_{\text{lin}}) := \mathbb{P}_{(X,Y)}\left(Yh_{\text{lin}}(X+\Delta x) \geq 0 \ \forall \|\Delta x\|_\infty \leq \varepsilon\right) \\
&= \mathbb{P}_{(X,Y)}\left(\inf_{\|\Delta x\|_\infty \leq \varepsilon} Yw^T(X+\Delta x) \geq 0\right) \\
&= \mathbb{P}_{(X,Y)}\left(Yw^TX - \sup_{\|\Delta x\|_\infty \leq \varepsilon} Yw^T\Delta x \geq 0\right) \\
&= \mathbb{P}_{(X,Y)}\left(Yw^TX - \varepsilon\|Yw\|_1 \geq 0\right) = \mathbb{P}_{(X,Y)}\left(Yw^TX - \varepsilon \geq 0\right) \\
&= \mathbb{P}(\mathcal{N}(0, 1/(p-1)) \geq \varepsilon - \eta) \leq e^{-(p-1)(\varepsilon-\eta)^2/2}.
\end{aligned}$$

Thus $\text{acc}_\varepsilon(h_{\text{lin}}) \leq \delta$ for $\varepsilon \geq \eta + \sqrt{2\log(1/\delta)/(p-1)}$.

Proof.

the **adversarial robustness accuracy** of h_{lin} writes

$$\begin{aligned}
&\text{acc}_\varepsilon(h_{\text{lin}}) := \mathbb{P}_{(X,Y)}\left(Yh_{\text{lin}}(X+\Delta x) \geq 0 \ \forall \|\Delta x\|_\infty \leq \varepsilon\right) \\
&= \mathbb{P}_{(X,Y)}\left(\inf_{\|\Delta x\|_\infty \leq \varepsilon} Yw^T(X+\Delta x) \geq 0\right) \\
&= \mathbb{P}_{(X,Y)}\left(Yw^TX - \sup_{\|\Delta x\|_\infty \leq \varepsilon} Yw^T\Delta x \geq 0\right) \\
&= \mathbb{P}_{(X,Y)}\left(Yw^TX - \varepsilon\|Yw\|_1 \geq 0\right) = \mathbb{P}_{(X,Y)}\left(Yw^TX - \varepsilon \geq 0\right) \\
&= \mathbb{P}(\mathcal{N}(0, 1/(p-1)) \geq \varepsilon - \eta) \leq e^{-(p-1)(\varepsilon-\eta)^2/2}.
\end{aligned}$$

Thus $\text{acc}_\varepsilon(h_{\text{lin}}) \leq \delta$ for $\varepsilon \geq \eta + \sqrt{2\log(1/\delta)/(p-1)}$.

Proof.

the **adversarial robustness accuracy** of h_{lin} writes

$$\begin{aligned}
&\text{acc}_\varepsilon(h_{\text{lin}}) := \mathbb{P}_{(X,Y)}\left(Yh_{\text{lin}}(X+\Delta x) \geq 0 \ \forall \|\Delta x\|_\infty \leq \varepsilon\right) \\
&= \mathbb{P}_{(X,Y)}\left(\inf_{\|\Delta x\|_\infty \leq \varepsilon} Yw^T(X+\Delta x) \geq 0\right) \\
&= \mathbb{P}_{(X,Y)}\left(Yw^TX - \sup_{\|\Delta x\|_\infty \leq \varepsilon} Yw^T\Delta x \geq 0\right) \\
&= \mathbb{P}_{(X,Y)}\left(Yw^TX - \varepsilon\|Yw\|_1 \geq 0\right) = \mathbb{P}_{(X,Y)}\left(Yw^TX - \varepsilon \geq 0\right) \\
&= \mathbb{P}(\mathcal{N}(0, 1/(p-1)) \geq \varepsilon - \eta) \leq e^{-(p-1)(\varepsilon-\eta)^2/2}.
\end{aligned}$$

Thus $\text{acc}_\varepsilon(h_{\text{lin}}) \leq \delta$ for $\varepsilon \geq \eta + \sqrt{2\log(1/\delta)/(p-1)}$.

Proof.

the **adversarial robustness accuracy** of h_{lin} writes

$$\begin{aligned}
&\text{acc}_\varepsilon(h_{\text{lin}}) := \mathbb{P}_{(X,Y)}\left(Yh_{\text{lin}}(X+\Delta x) \geq 0 \ \forall \|\Delta x\|_\infty \leq \varepsilon\right) \\
&= \mathbb{P}_{(X,Y)}\left(\inf_{\|\Delta x\|_\infty \leq \varepsilon} Yw^T(X+\Delta x) \geq 0\right) \\
&= \mathbb{P}_{(X,Y)}\left(Yw^TX - \sup_{\|\Delta x\|_\infty \leq \varepsilon} Yw^T\Delta x \geq 0\right) \\
&= \mathbb{P}_{(X,Y)}\left(Yw^TX - \varepsilon\|Yw\|_1 \geq 0\right) = \mathbb{P}_{(X,Y)}\left(Yw^TX - \varepsilon \geq 0\right) \\
&= \mathbb{P}(\mathcal{N}(0, 1/(p-1)) \geq \varepsilon - \eta) \leq e^{-(p-1)(\varepsilon-\eta)^2/2}.
\end{aligned}$$

Thus $\text{acc}_\varepsilon(h_{\text{lin}}) \leq \delta$ for $\varepsilon \geq \eta + \sqrt{2\log(1/\delta)/(p-1)}$.

Proof.

the **adversarial robustness accuracy** of h_{lin} writes

$$\begin{aligned}
&\text{acc}_\varepsilon(h_{\text{lin}}) := \mathbb{P}_{(X,Y)}\left(Yh_{\text{lin}}(X+\Delta x) \geq 0 \ \forall \|\Delta x\|_\infty \leq \varepsilon\right) \\
&= \mathbb{P}_{(X,Y)}\left(\inf_{\|\Delta x\|_\infty \leq \varepsilon} Yw^T(X+\Delta x) \geq 0\right) \\
&= \mathbb{P}_{(X,Y)}\left(Yw^TX - \sup_{\|\Delta x\|_\infty \leq \varepsilon} Yw^T\Delta x \geq 0\right) \\
&= \mathbb{P}_{(X,Y)}\left(Yw^TX - \varepsilon\|Yw\|_1 \geq 0\right) = \mathbb{P}_{(X,Y)}\left(Yw^TX - \varepsilon \geq 0\right) \\
&= \mathbb{P}(\mathcal{N}(0, 1/(p-1)) \geq \varepsilon - \eta) \leq e^{-(p-1)(\varepsilon-\eta)^2/2}.
\end{aligned}$$

Thus $\text{acc}_\varepsilon(h_{\text{lin}}) \leq \delta$ for $\varepsilon \geq \eta + \sqrt{2\log(1/\delta)/(p-1)}$.

- That is, the **adversarial accuracy** of h_{lin} **is** close to **zero**!

Proof.

the **adversarial robustness accuracy** of h_{lin} writes

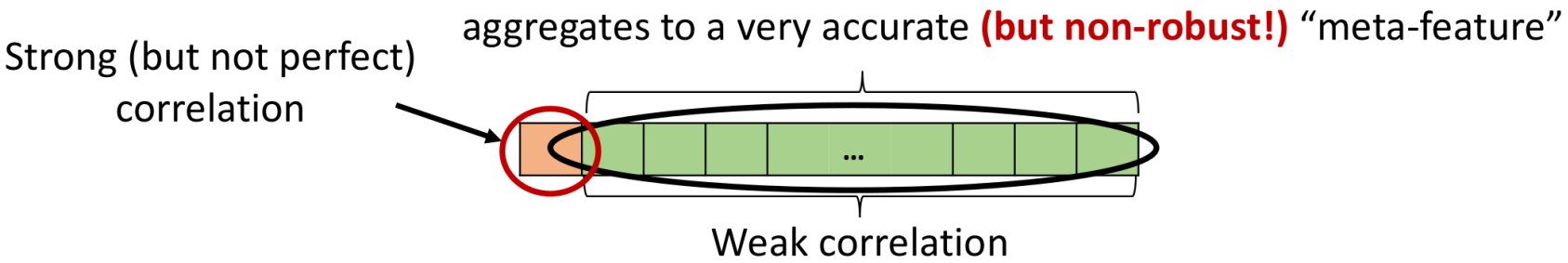
$$\begin{aligned}
&\text{acc}_\varepsilon(h_{\text{lin}}) := \mathbb{P}_{(X,Y)}\left(Yh_{\text{lin}}(X+\Delta x) \geq 0 \ \forall \|\Delta x\|_\infty \leq \varepsilon\right) \\
&= \mathbb{P}_{(X,Y)}\left(\inf_{\|\Delta x\|_\infty \leq \varepsilon} Yw^T(X+\Delta x) \geq 0\right) \\
&= \mathbb{P}_{(X,Y)}\left(Yw^TX - \sup_{\|\Delta x\|_\infty \leq \varepsilon} Yw^T\Delta x \geq 0\right) \\
&= \mathbb{P}_{(X,Y)}\left(Yw^TX - \varepsilon\|Yw\|_1 \geq 0\right) = \mathbb{P}_{(X,Y)}\left(Yw^TX - \varepsilon \geq 0\right) \\
&= \mathbb{P}(\mathcal{N}(0, 1/(p-1)) \geq \varepsilon - \eta) \leq e^{-(p-1)(\varepsilon-\eta)^2/2}.
\end{aligned}$$

Thus $\text{acc}_\varepsilon(h_{\text{lin}}) \leq \delta$ for $\varepsilon \geq \eta + \sqrt{2\log(1/\delta)/(p-1)}$.

- That is, the **adversarial accuracy** of h_{lin} **is** close to **zero**!

What could be going on ? [Intuition from Tsipras and co.]

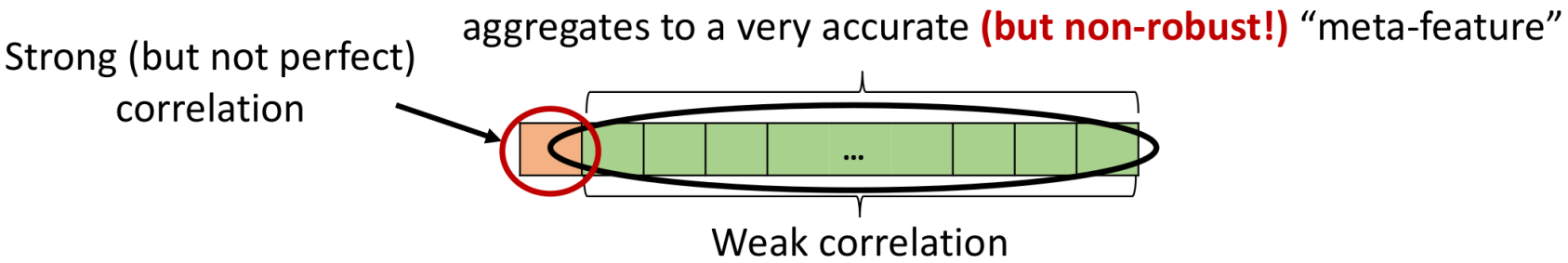
- **Prediction target**: $Y \sim \text{Bern}(1/2, \{\pm 1\})$
- **Robust feature**: $X^1 \mid Y = +Y$ w.p 70% and $-Y$ w.p. 30%.
- **Non-robust features**: $X^j \mid Y \sim \mathcal{N}(\eta Y, 1)$, for $j = 2, \ldots, p$



- BTW, we note that an optimal adversarial attack can be done by taking $\Delta x^1 = 0$ and $\Delta x^j = -\varepsilon y$ for all $j = 2, \ldots, p$.

What could be going on ? [Intuition from Tsipras and co.]

- **Prediction target**: $Y \sim \text{Bern}(1/2, \{\pm 1\})$
- **Robust feature**: $X^1 \mid Y = +Y$ w.p 70% and $-Y$ w.p. 30%.
- **Non-robust features**: $X^j \mid Y \sim \mathcal{N}(\eta Y, 1)$, for $j = 2, \ldots, p$



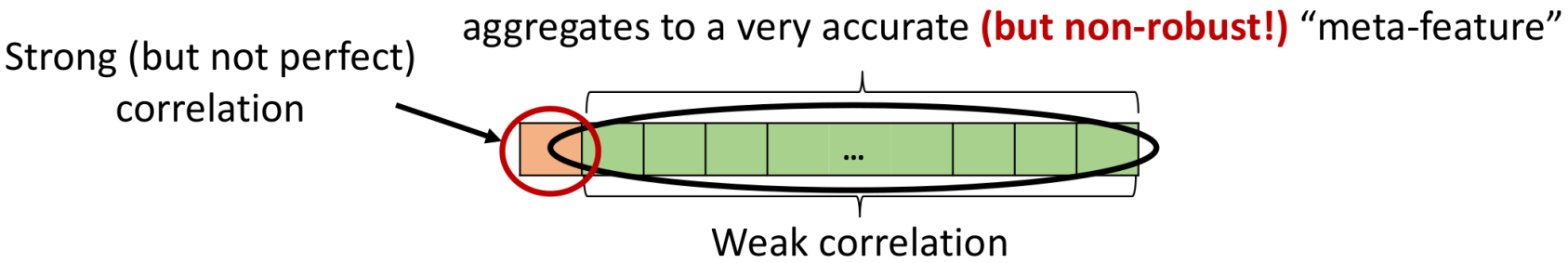
- BTW, we note that an optimal adversarial attack can be done by taking $\Delta x^1 = 0$ and $\Delta x^j = -\varepsilon y$ for all $j = 2, \ldots, p$.

Basic intuition:

- In standard training, all correlation is good correlation

What could be going on ? [Intuition from Tsipras and co.]

- **Prediction target**: $Y \sim \text{Bern}(1/2, \{\pm 1\})$
- **Robust feature**: $X^1 \mid Y = +Y$ w.p 70% and $-Y$ w.p. 30%.
- **Non-robust features**: $X^j \mid Y \sim \mathcal{N}(\eta Y, 1)$, for $j = 2, \ldots, p$



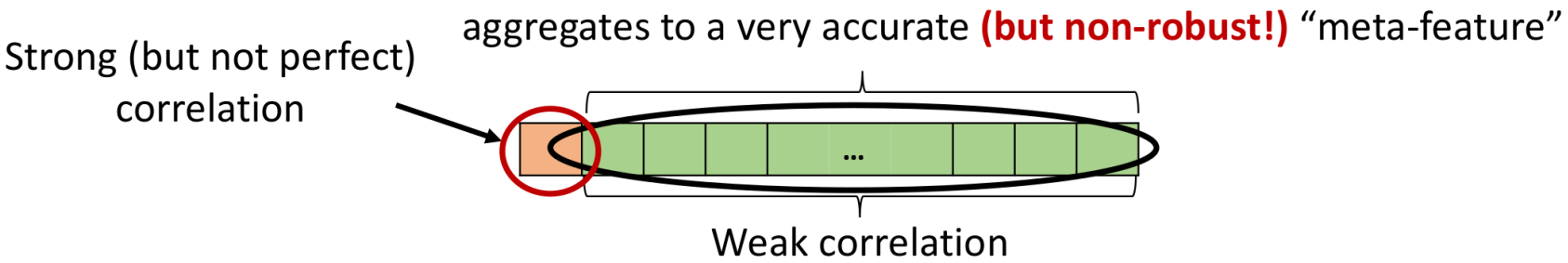
- BTW, we note that an optimal adversarial attack can be done by taking $\Delta x^1 = 0$ and $\Delta x^j = -\varepsilon y$ for all $j = 2, \ldots, p$.

Basic intuition:

- In standard training, all correlation is good correlation
- If we want robustness, must **avoid weakly correlated features**
  - $\Longrightarrow$ learn **causal features** ?

What could be going on ? [Intuition from Tsipras and co.]

- **Prediction target**: $Y \sim \text{Bern}(1/2, \{\pm 1\})$
- **Robust feature**: $X^1 \mid Y = +Y$ w.p 70% and $-Y$ w.p. 30%.
- **Non-robust features**: $X^j \mid Y \sim \mathcal{N}(\eta Y, 1)$, for $j = 2, \ldots, p$

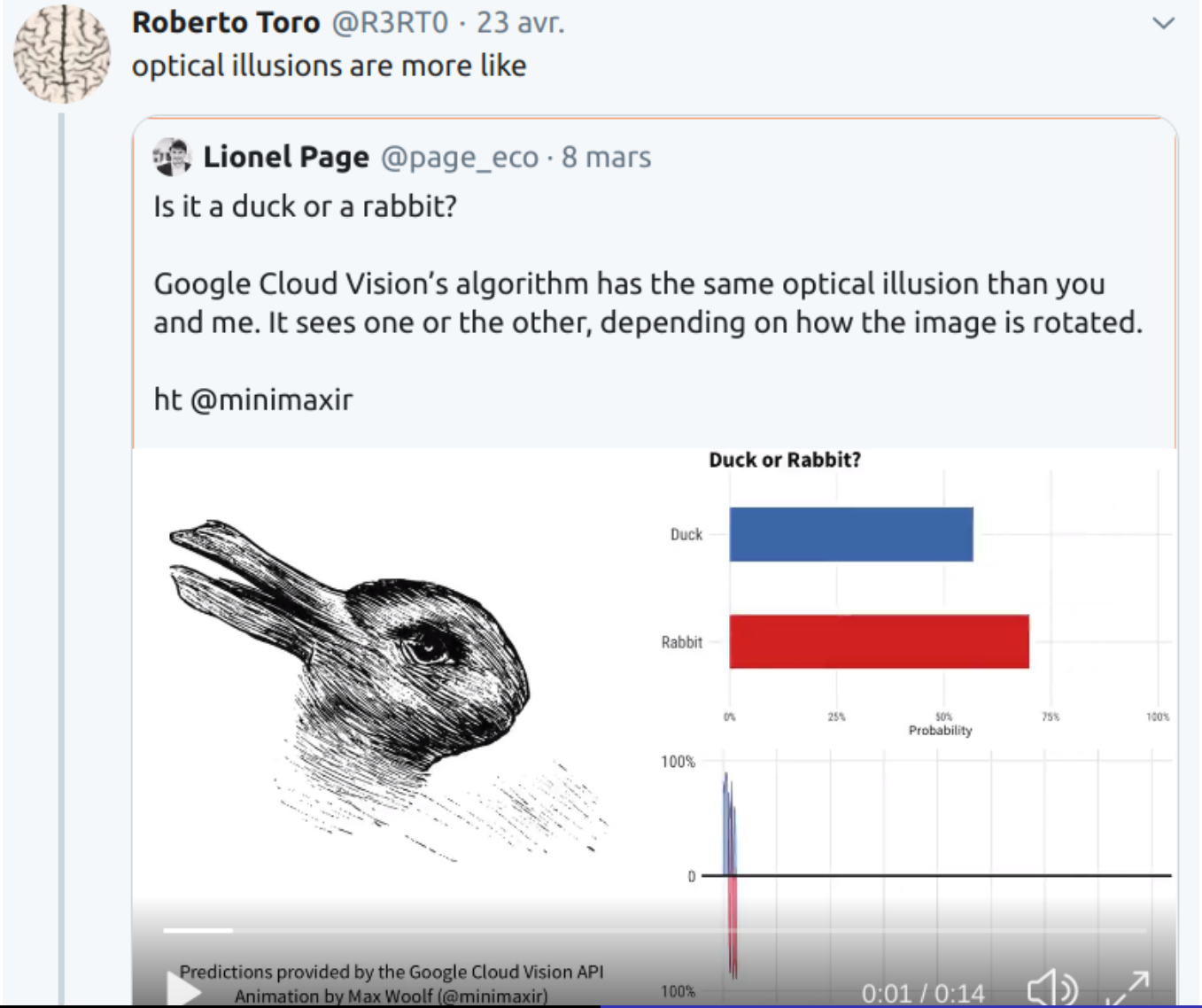


- BTW, we note that an optimal adversarial attack can be done by taking $\Delta x^1 = 0$ and $\Delta x^j = -\varepsilon y$ for all $j = 2, \ldots, p$.

Basic intuition:

- In standard training, all correlation is good correlation
- If we want robustness, must **avoid weakly correlated features**
  - $\Longrightarrow$ learn **causal features** ?

BTW, humans are not "perfect"



Talagrand transportation-cost inequality

The $T_2(c)$ property

Given $c \geq 0$, a distribution μ on $\mathcal{X}$ is said to satisfy $T_2(c)$ if for every distribution ν on $\mathcal{X}$ with $\nu \ll \mu$, one has

$$W_2(\nu, \mu) \leq \sqrt{2c \, \mathrm{kl}(\nu \| \mu)}, \tag{1}$$

where $\mathrm{kl}(\nu \| \mu) := \int_{\mathcal{X}} \log(d\nu/d\mu) d\mu$, entropy of ν relative to μ.

- Generalizes the well-known **Pinsker's inequality** for the total variation distance between probability measures (take $2c = 1/2$).

Talagrand transportation-cost inequality

The $T_2(c)$ property

Given $c \geq 0$, a distribution μ on $\mathcal{X}$ is said to satisfy $T_2(c)$ if for every distribution ν on $\mathcal{X}$ with $\nu \ll \mu$, one has

$$W_2(\nu, \mu) \leq \sqrt{2c \, \mathrm{kl}(\nu \| \mu)}, \tag{1}$$

where $\mathrm{kl}(\nu \| \mu) := \int_{\mathcal{X}} \log(d\nu/d\mu) d\mu$, entropy of ν relative to μ.

- Generalizes the well-known **Pinsker's inequality** for the total variation distance between probability measures (take $2c = 1/2$).
- Unlike Pinsker's inequality which holds unconditionally, the inequality $T_2(c)$ is a privilege only enjoyed by special classes of reference distributions μ.

Talagrand transportation-cost inequality

The $\mathsf{T}_2(c)$ property

Given $c \geq 0$, a distribution μ on $\mathcal{X}$ is said to satisfy $\mathsf{T}_2(c)$ if for every distribution ν on $\mathcal{X}$ with $\nu \ll \mu$, one has

$$W_2(\nu, \mu) \leq \sqrt{2c\, \mathsf{kl}(\nu \| \mu)}, \tag{1}$$

where $\mathsf{kl}(\nu\|\mu) := \int_{\mathcal{X}} \log(d\nu/d\mu) d\mu$, entropy of ν relative to μ.

- Generalizes the well-known **Pinsker's inequality** for the total variation distance between probability measures (take $2c = 1/2$).
- Unlike Pinsker's inequality which holds unconditionally, the inequality $\mathsf{T}_2(c)$ is a privilege only enjoyed by special classes of reference distributions μ.

BLOWUP / aka concentration of measure

The BLOWUP(c) property

μ is said to satisfy BLOWUP(c) if for every Borel $B \subseteq \mathcal{X}$ with $\mu(B) > 0$ and for every $\varepsilon \geq \sqrt{2c \log(1/\mu(B))}$, it holds that

$$\mu(B^{\varepsilon}) \geq 1 - e^{-\frac{1}{2c}(\varepsilon - \sqrt{2c \log(1/\mu(B))})^2}. \tag{2}$$

- It is a classical result that the Gaussian distribution on $\mathbb{R}^p$ has BLOWUP(1) and $T_2(1)$, a phenomenon known as **Gaussian isoperimetry**.

BLOWUP / aka concentration of measure

The BLOWUP(c) property

μ is said to satisfy BLOWUP(c) if for every Borel $B \subseteq \mathcal{X}$ with $\mu(B) > 0$ and for every $\varepsilon \geq \sqrt{2c \log(1/\mu(B))}$, it holds that

$$\mu(B^{\varepsilon}) \geq 1 - e^{-\frac{1}{2c}(\varepsilon - \sqrt{2c \log(1/\mu(B))})^2}. \tag{2}$$

- It is a classical result that the Gaussian distribution on $\mathbb{R}^p$ has BLOWUP(1) and $\mathrm{T}_2(1)$, a phenomenon known as **Gaussian isoperimetry**.
  - These result dates back to works of Borel, Lévy, Talagrand and of Marton (see [Boucheron '13] textbook)

BLOWUP / aka concentration of measure

The BLOWUP(c) property

μ is said to satisfy BLOWUP(c) if for every Borel $B \subseteq \mathcal{X}$ with $\mu(B) > 0$ and for every $\varepsilon \geq \sqrt{2c \log(1/\mu(B))}$, it holds that

$$\mu(B^{\varepsilon}) \geq 1 - e^{-\frac{1}{2c}(\varepsilon - \sqrt{2c \log(1/\mu(B))})^2}. \tag{2}$$

- It is a classical result that the Gaussian distribution on $\mathbb{R}^p$ has BLOWUP(1) and $\mathrm{T}_2(1)$, a phenomenon known as **Gaussian isoperimetry**.
  - These result dates back to works of Borel, Lévy, Talagrand and of Marton (see [Boucheron '13] textbook)

Marton's Blowup lemma

Lemma (Marton's BLOWUP Lemma)

On a metric space, it holds that $T_2(c) \subseteq BLOWUP(c)$.

Proof. **Fact**: $\mathrm{kl}(\mu|_B \| \mu) = \log(1/\mu(B))$, where $\mu|_B(A) := \frac{\mu(A \cap B)}{\mu(B)}$.

Marton's Blowup lemma

Lemma (Marton's BLOWUP Lemma)

On a metric space, it holds that $T_2(c) \subseteq BLOWUP(c)$.

Proof. **Fact**: $\mathrm{kl}(\mu|_B \| \mu) = \log(1/\mu(B))$, where $\mu|_B(A) := \frac{\mu(A \cap B)}{\mu(B)}$.

Thus $\varepsilon \leq W_2(\mu|_B, \mu_{\mathcal{X} \setminus B^\varepsilon}) \leq W_2(\mu|_B, \mu) + W_2(\mu|_{\mathcal{X} \setminus B^\varepsilon}, \mu)$

Marton's Blowup lemma

Lemma (Marton's BLOWUP Lemma)

On a metric space, it holds that $T_2(c) \subseteq BLOWUP(c)$.

Proof. **Fact**: $\mathrm{kl}(\mu|_B \| \mu) = \log(1/\mu(B))$, where $\mu|_B(A) := \frac{\mu(A \cap B)}{\mu(B)}$.

Thus $\varepsilon \leq W_2(\mu|_B, \mu_{\mathcal{X} \setminus B^\varepsilon}) \leq W_2(\mu|_B, \mu) + W_2(\mu|_{\mathcal{X} \setminus B^\varepsilon}, \mu)$

$$\leq \sqrt{2c\,\mathrm{kl}(\mu|_B \| \mu)} + \sqrt{2c\,\mathrm{kl}(\mu|_{\mathcal{X} \setminus B^\varepsilon} \| \mu)}$$

Marton's Blowup lemma

Lemma (Marton's BLOWUP Lemma)

On a metric space, it holds that $T_2(c) \subseteq BLOWUP(c)$.

Proof. **Fact**: $\mathrm{kl}(\mu|_B \| \mu) = \log(1/\mu(B))$, where $\mu|_B(A) := \frac{\mu(A \cap B)}{\mu(B)}$.

$$\begin{aligned}\text{Thus } \varepsilon &\leq W_2(\mu|_B, \mu_{\mathcal{X} \setminus B^\varepsilon}) \leq W_2(\mu|_B, \mu) + W_2(\mu|_{\mathcal{X} \setminus B^\varepsilon}, \mu) \\ &\leq \sqrt{2c\,\mathrm{kl}(\mu|_B \| \mu)} + \sqrt{2c\,\mathrm{kl}(\mu|_{\mathcal{X} \setminus B^\varepsilon} \| \mu)} \\ &\leq \sqrt{2c \log(1/\mu(B))} + \sqrt{2c \log(1/\mu(\mathcal{X} \setminus B^\varepsilon))}\end{aligned}$$

Marton's Blowup lemma

Lemma (Marton's BLOWUP Lemma)

On a metric space, it holds that $T_2(c) \subseteq BLOWUP(c)$.

Proof. **Fact**: $\mathrm{kl}(\mu|_B \| \mu) = \log(1/\mu(B))$, where $\mu|_B(A) := \frac{\mu(A \cap B)}{\mu(B)}$.

$$\begin{aligned}
\text{Thus } \varepsilon &\leq W_2(\mu|_B, \mu_{\mathcal{X} \setminus B^\varepsilon}) \leq W_2(\mu|_B, \mu) + W_2(\mu|_{\mathcal{X} \setminus B^\varepsilon}, \mu) \\
&\leq \sqrt{2c\,\mathrm{kl}(\mu|_B \| \mu)} + \sqrt{2c\,\mathrm{kl}(\mu|_{\mathcal{X} \setminus B^\varepsilon} \| \mu)} \\
&\leq \sqrt{2c \log(1/\mu(B))} + \sqrt{2c \log(1/\mu(\mathcal{X} \setminus B^\varepsilon))} \\
&= \sqrt{2c \log(1/\mu(B))} + \sqrt{2c \log(1/(1 - \mu(B^\varepsilon))}.
\end{aligned}$$

Marton's Blowup lemma

Lemma (Marton's BLOWUP Lemma)

On a metric space, it holds that $T_2(c) \subseteq BLOWUP(c)$.

Proof. **Fact**: $\mathsf{kl}(\mu|_B \| \mu) = \log(1/\mu(B))$, where $\mu|_B(A) := \frac{\mu(A \cap B)}{\mu(B)}$.

$$\begin{aligned}
\text{Thus } \varepsilon &\leq W_2(\mu|_B, \mu_{\mathcal{X} \setminus B^\varepsilon}) \leq W_2(\mu|_B, \mu) + W_2(\mu|_{\mathcal{X} \setminus B^\varepsilon}, \mu) \\
&\leq \sqrt{2c\, \mathsf{kl}(\mu|_B \| \mu)} + \sqrt{2c\, \mathsf{kl}(\mu|_{\mathcal{X} \setminus B^\varepsilon} \| \mu)} \\
&\leq \sqrt{2c \log(1/\mu(B))} + \sqrt{2c \log(1/\mu(\mathcal{X} \setminus B^\varepsilon))} \\
&= \sqrt{2c \log(1/\mu(B))} + \sqrt{2c \log(1/(1-\mu(B^\varepsilon))}.
\end{aligned}$$

Rearranging the above inequality gives

$$\sqrt{2c \log(1/(1-\mu(B^\varepsilon)))} \geq (\varepsilon - \sqrt{2c \log(1/\mu(B))})_+,$$

and the result follows after squaring & exponentiating. □

Marton's Blowup lemma

Lemma (Marton's BLOWUP Lemma)

On a metric space, it holds that $T_2(c) \subseteq BLOWUP(c)$.

Proof. **Fact**: $\mathrm{kl}(\mu|_B \| \mu) = \log(1/\mu(B))$, where $\mu|_B(A) := \frac{\mu(A \cap B)}{\mu(B)}$.

$$\begin{aligned}
\text{Thus } \varepsilon &\leq W_2(\mu|_B, \mu_{\mathcal{X}\setminus B^\varepsilon}) \leq W_2(\mu|_B, \mu) + W_2(\mu|_{\mathcal{X}\setminus B^\varepsilon}, \mu) \\
&\leq \sqrt{2c\,\mathrm{kl}(\mu|_B \| \mu)} + \sqrt{2c\,\mathrm{kl}(\mu|_{\mathcal{X}\setminus B^\varepsilon} \| \mu)} \\
&\leq \sqrt{2c\log(1/\mu(B))} + \sqrt{2c\log(1/\mu(\mathcal{X}\setminus B^\varepsilon))} \\
&= \sqrt{2c\log(1/\mu(B))} + \sqrt{2c\log(1/(1-\mu(B^\varepsilon))}.
\end{aligned}$$

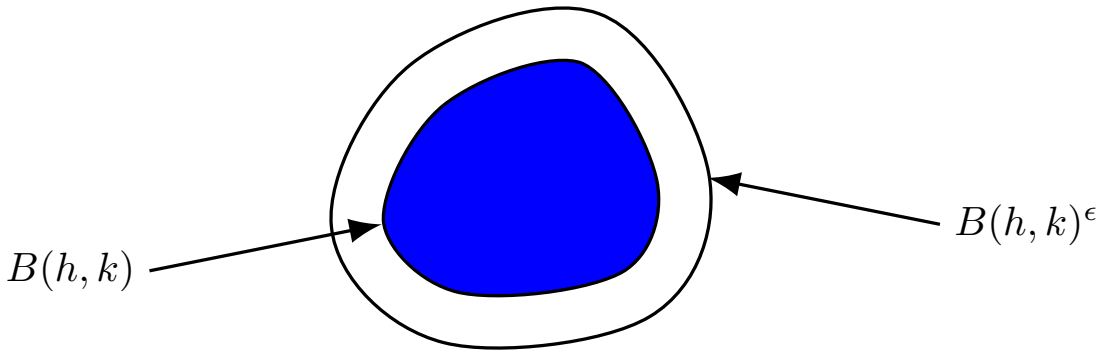
Rearranging the above inequality gives

$$\sqrt{2c\log(1/(1-\mu(B^\varepsilon)))} \geq (\varepsilon - \sqrt{2c\log(1/\mu(B))})_+,$$

and the result follows after squaring & exponentiating. □

Adversarial attacks are a 'butterfly effect' on data manifold

- **Error set**: $B(h,k) = \{x \in \mathcal{X} \mid h(x) \neq k\}$, $h =$ classifier
- **Neighbors of error set**: $B(h,k)^{\varepsilon} := \{x \in \mathcal{X} \mid d(x, B(h,k)) \leq \varepsilon\}$



- $\mathrm{err}(h|k) := P_{X|k}(B(h,k)) > 0$ if h is **not perfect** on class k.
- Consequence is that $\mathrm{acc}_{\varepsilon}(h|k) \searrow 0$ expo. fast as function of ε.
- **Thus adversarial robustness is impossible in general!**
- Manuscript: `https://arxiv.org/pdf/1810.04065.pdf`

Strong No Free Lunch Theorem

Theorem (Strong "No Free Lunch" [Dohmatob '18])

Suppose that conditional distribution $P_{X|k}$ has the $T_2(\sigma_k^2)$ property. Given a classifier $h : \mathcal{X} \mapsto \mathcal{Y}$ such that $\text{err}(h|k) > 0$, define $\varepsilon(h|k) := \sigma_k \sqrt{2 \log(1/\text{err}(h|k))}$. Then we have the following bounds:

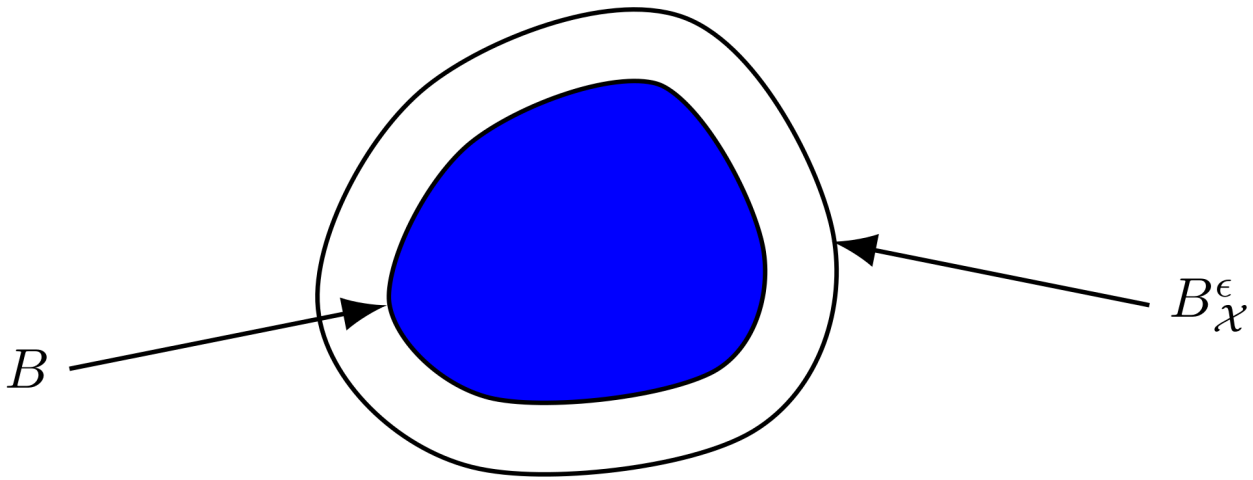
(A) **Adversarial robustness accuracy**: *if $\varepsilon \geq \varepsilon(h|k)$, then*

$$\text{acc}_\varepsilon(h|k) \leq e^{-\frac{1}{2\sigma_k^2}(\varepsilon - \varepsilon(h|k))^2}. \tag{3}$$

(B) **Average distance to error set**:

$$d(h|k) \leq \sigma_k \left(\sqrt{\log(1/\text{err}(h|k))} + \sqrt{\pi/2} \right) \tag{4}$$

Proof



- Use Marton's Lemma: $\text{BLOWUP}(\sigma_k^2) \subseteq \mathsf{T}_2(\sigma_k^2)$ with $B := B(h,k) := \{x \in \mathcal{X} \mid h(x) \neq k\}$ and $\mu = P_{X|k}$. □

Corollary (Strong "No Free Lunch" Theorem on flat space)

Let $1 \leq q \leq \infty$. Define $\epsilon_q(h|k) := \varepsilon(h|k)p^{1/q-1/2}$. If in addition to the assumptions of Strong No Free Lunch Theorem, and suppose the feature space is flat, i.e $\mathrm{Ric}_{\mathcal{X}} = 0$, then for the ℓ_q threat model, we have the following bounds:

(A1) **Adversarial robustness accuracy**: *if $\varepsilon \geq \epsilon_q(h|k)$, then*

$$\mathrm{acc}_{\varepsilon}(h|k) \leq e^{-\frac{p^{1-2/q}}{2\sigma_k^2}(\varepsilon - \epsilon_q(h|k))^2}. \tag{5}$$

(A2) **Average distance to error set**:

$$d(h|k) \leq \sigma_k p^{1/q-1/2}\left(\sqrt{\log(1/\,\mathrm{err}(h|k))} + \sqrt{\pi/2}\right). \tag{6}$$

- Note that the case $q = 1$ is a proxy for "few-pixel" attack models [Su et a. '18].

Strong No Free Lunch Theorem

Corollary (Strong NFLT for ℓ_∞ attacks [Dohmatob '18])

In particular, for the ℓ_∞ threat model, we have the following bounds:

(B1) **Adversarial robustness accuracy**: *If $\varepsilon \geq \varepsilon(h|k)/\sqrt{p}$, then*

$$\mathrm{acc}_\varepsilon(h|k) \leq e^{-\frac{p}{2\sigma_k^2}(\varepsilon - \varepsilon(h|k)/\sqrt{p})^2}. \tag{7}$$

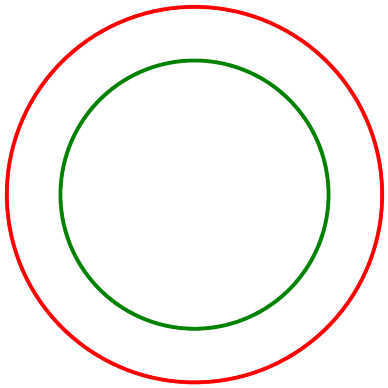
(B2) **Average distance to error set**:

$$d(h|k) \leq \frac{\sigma_k}{\sqrt{p}}\left(\sqrt{\log(1/\,\mathrm{err}(h|k))} + \sqrt{\pi/2}\right) \tag{8}$$

Special cases of our results

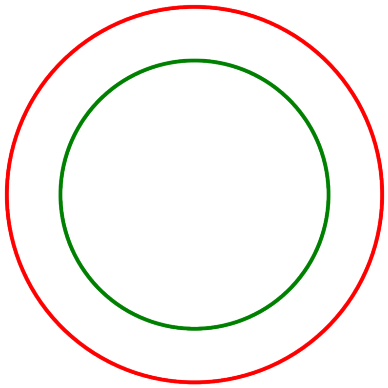
- Log-concave distribs $dP_{X|k} \propto e^{-v_k(x)}dx$ satisfying Emeŕy-Bakry curvature condition: $\mathrm{Hess}_x(v_k) + \mathrm{Ric}_x(\mathcal{X}) \succeq (1/\sigma_k^2)I_p$.
  - e.g multi-variate Gaussian (considered in [Tsipras '18, Fawzi et al. 18])
- Perturbed log-concave distribs (via Holley-Shroock Theorem)
- The uniform measure on compact Riemannian manifolds of positive Ricci curvature, e.g spheres (considered in [Gilmer '18]), tori, or any compact Lie group.
- Pushforward via a Lipschitz function f, of a distribution in $\mathrm{T}_2(\sigma_k^2)$. Indeed, take $\tilde{\sigma}_k = \|f\|_{\mathrm{Lip}}\sigma_k$.
- etc.

Worked example: Adversarial spheres [Gilmer '18]



- $Y \sim \mathrm{Bern}(1/2, \{\pm\})$,
- $X|k \sim \mathrm{uniform}(\mathbb{S}^p_{R_k})$, where $R_+ > R_- > 0$.
- $\mathbb{S}^p_{R_k}$ is a compact Riemannian manifold with constant Ricci curvature $(p-1)R_k^{-2}$.
- Thus $P_{X|k}$ satisfies $\mathrm{T}_2(R_k^2/(p-1))$.

Worked example: Adversarial spheres [Gilmer '18]

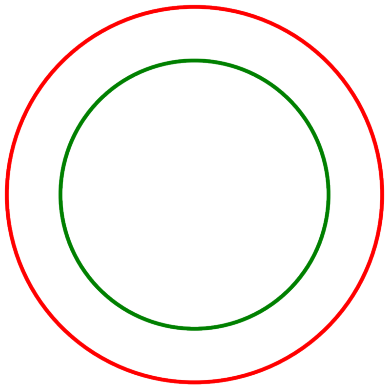


- $Y \sim \text{Bern}(1/2, \{\pm\})$,
- $X|k \sim \text{uniform}(\mathbb{S}^p_{R_k})$, where $R_+ > R_- > 0$.
- $\mathbb{S}^p_{R_k}$ is a compact Riemannian manifold with constant Ricci curvature $(p-1)R_k^{-2}$.
- Thus $P_{X|k}$ satisfies $\mathsf{T}_2(R_k^2/(p-1))$.

$$\therefore \mathbb{E}_{X|k}[d_{\text{geo}}(X, B(h,k))] \leq \frac{R_k}{\sqrt{p-1}}\left(\sqrt{2\log(1/err(h|k))} + \sqrt{\pi/2}\right)$$

$$\sim \frac{R_k}{\sqrt{p}}\Phi^{-1}(\text{acc}(h|k))$$

Worked example: Adversarial spheres [Gilmer '18]

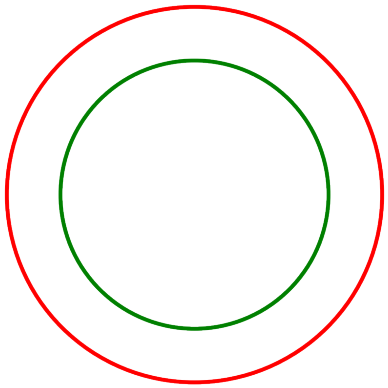


- $Y \sim \text{Bern}(1/2, \{\pm\})$,
- $X|k \sim \text{uniform}(\mathbb{S}^p_{R_k})$, where $R_+ > R_- > 0$.
- $\mathbb{S}^p_{R_k}$ is a compact Riemannian manifold with constant Ricci curvature $(p-1)R_k^{-2}$.
- Thus $P_{X|k}$ satisfies $\mathsf{T}_2(R_k^2/(p-1))$.

$$\therefore \mathbb{E}_{X|k}[d_{\text{geo}}(X, B(h,k))] \leq \frac{R_k}{\sqrt{p-1}}\left(\sqrt{2\log(1/err(h|k))} + \sqrt{\pi/2}\right)$$
$$\sim \frac{R_k}{\sqrt{p}}\Phi^{-1}(\text{acc}(h|k)$$

- This is the same bound obtained in "manually" [Gilmer '18].

Worked example: Adversarial spheres [Gilmer '18]

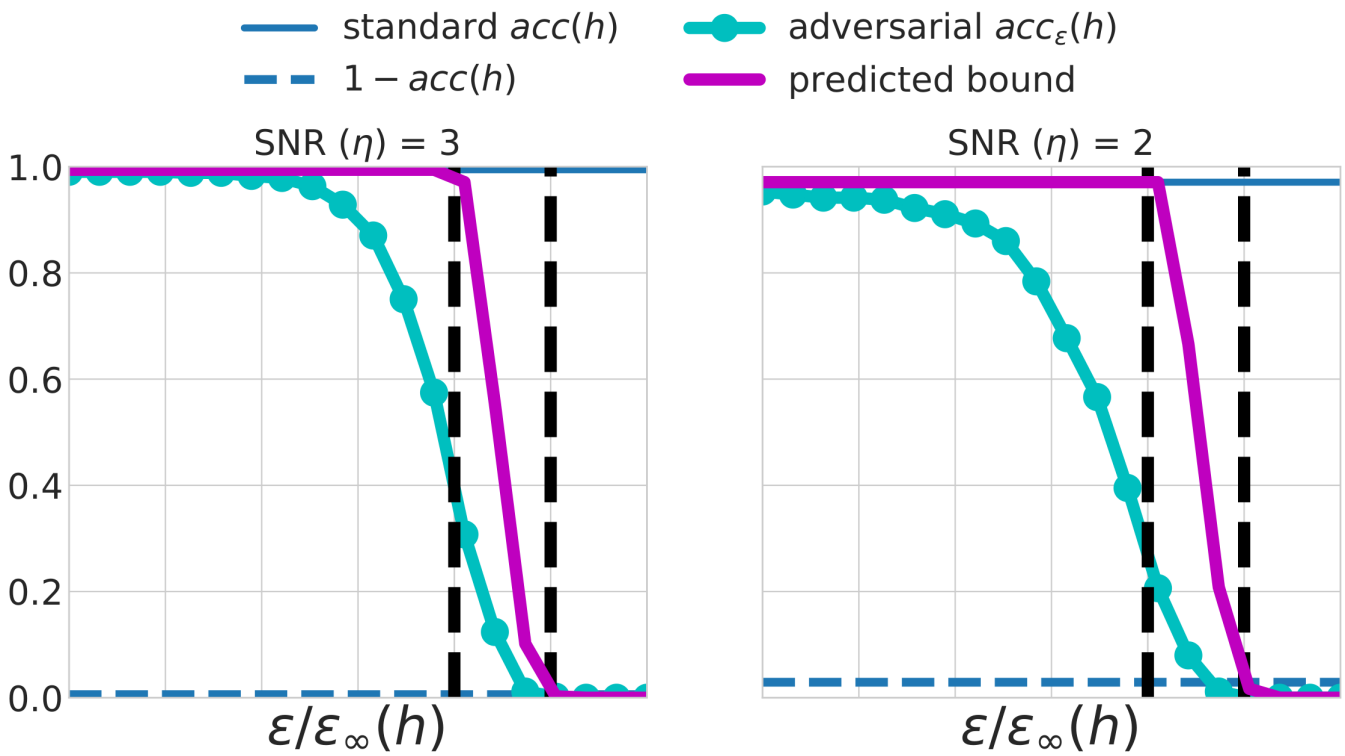


- $Y \sim \text{Bern}(1/2, \{\pm\})$,
- $X|k \sim \text{uniform}(\mathbb{S}^p_{R_k})$, where $R_+ > R_- > 0$.
- $\mathbb{S}^p_{R_k}$ is a compact Riemannian manifold with constant Ricci curvature $(p-1)R_k^{-2}$.
- Thus $P_{X|k}$ satisfies $\mathsf{T}_2(R_k^2/(p-1))$.

$$\therefore \mathbb{E}_{X|k}[d_{\text{geo}}(X, B(h,k))] \leq \frac{R_k}{\sqrt{p-1}}(\sqrt{2\log(1/err(h|k))} + \sqrt{\pi/2})$$
$$\sim \frac{R_k}{\sqrt{p}}\Phi^{-1}(\text{acc}(h|k)$$

- This is the same bound obtained in "manually" [Gilmer '18].

Some empirical confirmation



- Phase-transition occurs as predicted by our theorems

Key papers

- [Tsipras '18] *There is no free lunch in adversarial robustness*
- [Gilmer '18] *Adversarial spheres*
- [Fawzi '18] *Adversarial vulnerability for any classifier*
- [Athalye '18] *Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples*
- [Dohmatob '19] Generalized No Free Lunch Theorem for Adversarial Robustness
- [Shafahi '19] Are adversarial examples inevitable?

Universal lower bounds

Overview

- Previous **classifier-dependent** bounds make very strong assumptions on both the data and the classifier (e.g the theory fails for perfect classifiers)
- It would be nice to have **universal** bounds which only depend on the geomtry of the class-conditional distributions P_+ and P_-
- This is very very recent work, started by [Bhagoji '19] (to appear in NeurIPS!)
- My own work builds on [Bhagoji '19] as is still largely ongoing (AISTATS ???)

References

- [Bhagoji '19] *Lower Bounds on Adversarial Robustness from Optimal Transport*

Abstract view of adversarial attacks

- The **feature space** $\mathcal{X}$ is an abstract measure space, and the target space is $\{\pm 1\}$ (binary classification). E.g $\mathcal{X} = (\mathbb{R}^p, \text{Borell})$.
- Let P be an unknown probability distribution on the product space $\mathcal{X} \times \{\pm 1\}$.
- A **classifier** is any measurable function $h : \mathcal{X} \to \{\pm 1\}$.
- An **attack-model** $\mathcal{A}$ is the prescription of a closed neighborhood $\mathcal{A}_x$ for each point x of $\mathcal{X}$. E.g $\mathcal{A}_x = \text{Ball}_{\ell_\infty}(x; \varepsilon)$. The case $\mathcal{A}_x = \{x\}\ \forall x \in \mathcal{X}$ corresponds to the attackless model.
- A **type-$\mathcal{A}$ attack** is any measurable function $a : \mathcal{X} \times \{\pm 1\} \to \mathcal{X}$ such that $a(x, y) \in \mathcal{A}_x\ \forall (x, y) \in \mathcal{X} \times \{\pm 1\}$. With abuse of notation, we'll also write $a \in \mathcal{A}$. E.g $a(x, y) := x - yz$ for some fixed $z \in \text{Ball}_{\ell_\infty}(0; \varepsilon)$.

Abstract view of adversarial attacks

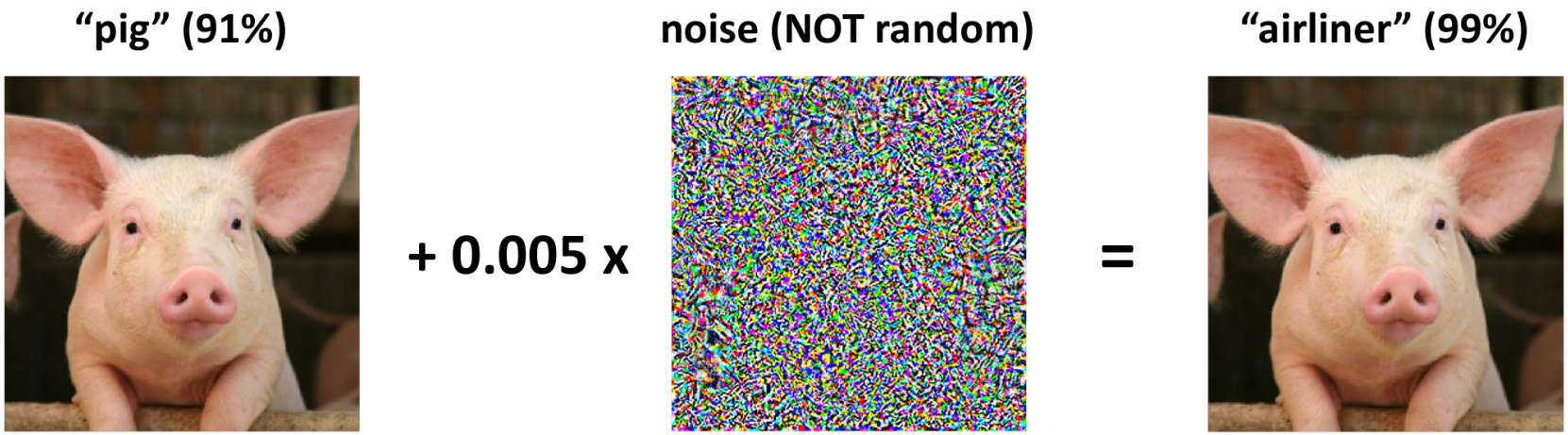
- The **robustness error** of h to type-$\mathcal{A}$ attacks is

$$\mathrm{err}_{\mathcal{A}}(h) := \mathbb{E}_{(x,y)\sim P}[\exists x' \in \mathcal{A}_x \text{ s.t } h(x') \neq y] \tag{9}$$

- The **Bayes-optimal robustness error** for type-$\mathcal{A}$ attacks is

$$\mathrm{err}^*_{\mathcal{A}} := \inf_h \mathrm{err}_{\mathcal{A}}(h) \tag{10}$$

The flying pig example again!



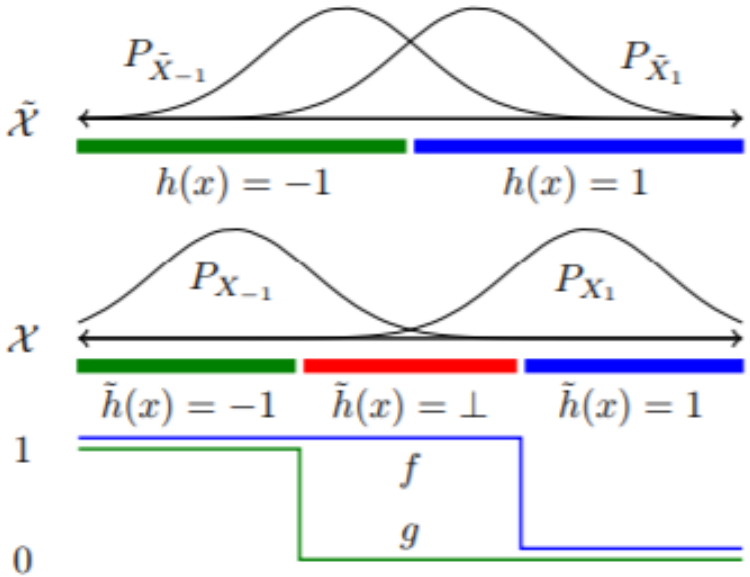
(Picture is courtesy of https://gradientscience.org/intro_adversarial/)

- $x \mapsto x^{\text{adv}} := x +$ noise , $\|\text{noise}\| \leq \varepsilon = 0.005$ (in example above)
- $\mathcal{X} = \mathbb{R}^{\#\textit{pixels}}$, $\mathcal{A}_x = \text{Ball}_{\ell_\infty}(x; 0.005)$

Adversarial attacks as optimal transport [Bhagoji '19]

Given a classifier h, consider the derived classifier $\tilde{h} : \mathcal{X} \to \{\pm 1\}$

$$\tilde{h}(x) := \begin{cases} y, & \text{if } \exists y \in \{\pm 1\} \text{ s.t } h(x') = y \ \forall x' \in \mathcal{A}_x, \\ \perp, & \text{else.} \end{cases} \tag{11}$$



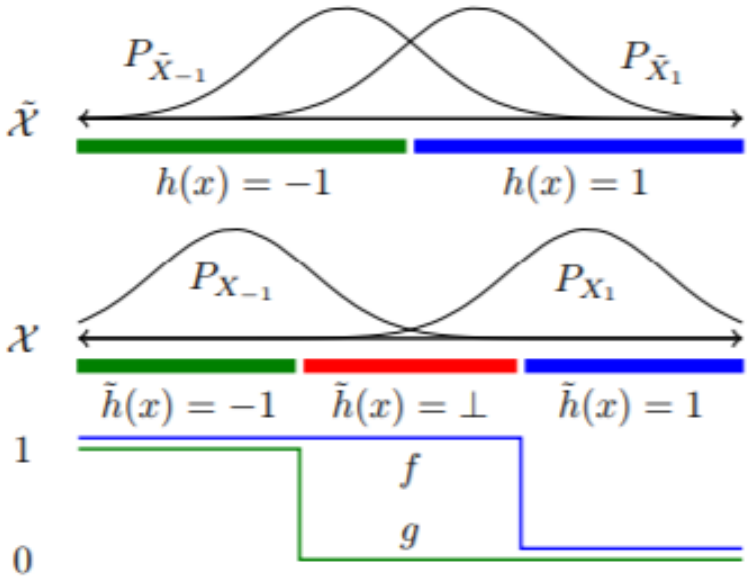
- Define the transport ground-cost

$$c_{\mathcal{A}}(x, x') = \begin{cases} 1, & \text{if } \mathcal{A}_x \cap \mathcal{A}_{x'} = \emptyset, \\ 0, & \text{else,} \end{cases}$$

Adversarial attacks as optimal transport [Bhagoji '19]

Given a classifier h, consider the derived classifier $\tilde{h} : \mathcal{X} \to \{\pm 1\}$

$$\tilde{h}(x) := \begin{cases} y, & \text{if } \exists y \in \{\pm 1\} \text{ s.t } h(x') = y \ \forall x' \in \mathcal{A}_x, \\ \perp, & \text{else.} \end{cases} \tag{11}$$



- Define the transport ground-cost

$$c_{\mathcal{A}}(x, x') = \begin{cases} 1, & \text{if } \mathcal{A}_x \cap \mathcal{A}_{x'} = \emptyset, \\ 0, & \text{else,} \end{cases}$$

and note that $\forall x, x' \in \mathcal{X}$, one has

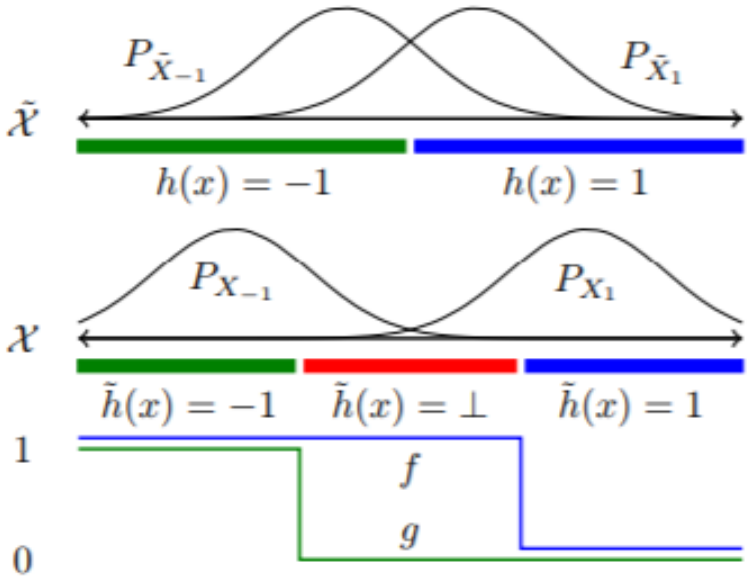
$$\mathbb{1}_{\{\tilde{h}(x)=1\}} + \mathbb{1}_{\{\tilde{h}(x)=-1\}} \leq c_{\mathcal{A}}(x, x') + 1,$$

- i.e $f(x) - g(x') \leq c_{\mathcal{A}}(x, x')$.

Adversarial attacks as optimal transport [Bhagoji '19]

Given a classifier h, consider the derived classifier $\tilde{h} : \mathcal{X} \to \{\pm 1\}$

$$\tilde{h}(x) := \begin{cases} y, & \text{if } \exists y \in \{\pm 1\} \text{ s.t } h(x') = y \ \forall x' \in \mathcal{A}_x, \\ \perp, & \text{else.} \end{cases} \tag{11}$$



- Define the transport ground-cost

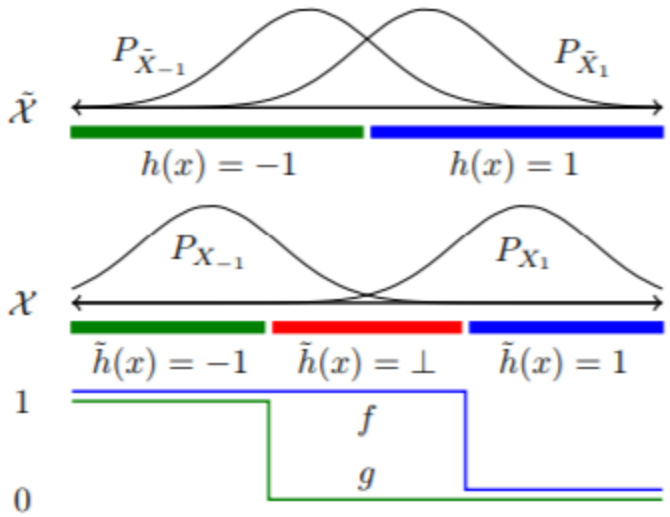
$$c_{\mathcal{A}}(x, x') = \begin{cases} 1, & \text{if } \mathcal{A}_x \cap \mathcal{A}_{x'} = \emptyset, \\ 0, & \text{else,} \end{cases}$$

and note that $\forall x, x' \in \mathcal{X}$, one has

$$\mathbb{1}_{\{\tilde{h}(x)=1\}} + \mathbb{1}_{\{\tilde{h}(x)=-1\}} \leq c_{\mathcal{A}}(x, x') + 1,$$

- i.e $f(x) - g(x') \leq c_{\mathcal{A}}(x, x')$.

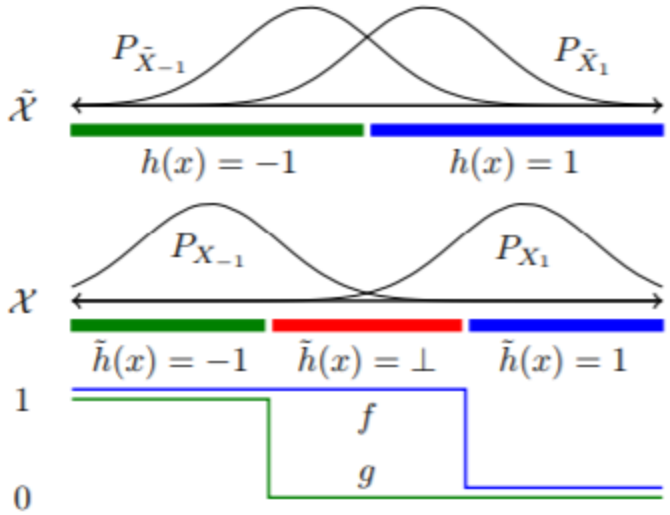
Adversarial attacks as optimal transport [Bhagoji '19]



- $c_{\mathcal{A}}(x, x') = \begin{cases} 1, & \text{if } \mathcal{A}_x \cap \mathcal{A}_{x'} = \emptyset, \\ 0, & \text{else,} \end{cases}$
- $f_h(x) - g_h(x') \leq c_{\mathcal{A}}(x, x') \; \forall x, x' \in \mathcal{X}$, and so (f_h, g_h) is a pair of Kantorovich potentials for OT with ground-cost $c_{\mathcal{A}}$.

$$\therefore OT_{c_{\mathcal{A}}}(P_-, P_+) := \sup_{K-\text{potentials } \phi,\psi} \mathbb{E}_{P_-}[\phi(x)] - \mathbb{E}_{P_+}[\psi(x)]$$

Adversarial attacks as optimal transport [Bhagoji '19]

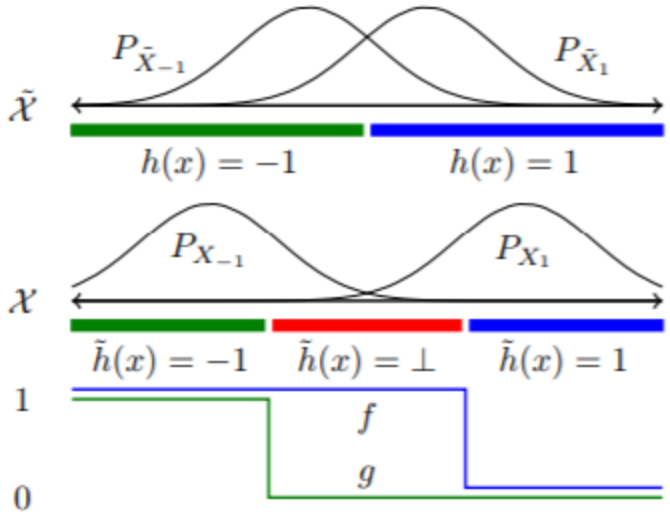


- $c_{\mathcal{A}}(x, x') = \begin{cases} 1, & \text{if } \mathcal{A}_x \cap \mathcal{A}_{x'} = \emptyset, \\ 0, & \text{else,} \end{cases}$
- $f_h(x) - g_h(x') \leq c_{\mathcal{A}}(x, x') \; \forall x, x' \in \mathcal{X}$, and so (f_h, g_h) is a pair of Kantorovich potentials for OT with ground-cost $c_{\mathcal{A}}$.

$$\therefore OT_{c_{\mathcal{A}}}(P_-, P_+) := \sup_{K-\text{potentials } \phi,\psi} \mathbb{E}_{P_-}[\phi(x)] - \mathbb{E}_{P_+}[\psi(x)]$$

$$\geq \sup_h \mathbb{E}_{P_-}[g_h(x)] - \mathbb{E}_{P_+}[f_h(x)]$$

Adversarial attacks as optimal transport [Bhagoji '19]



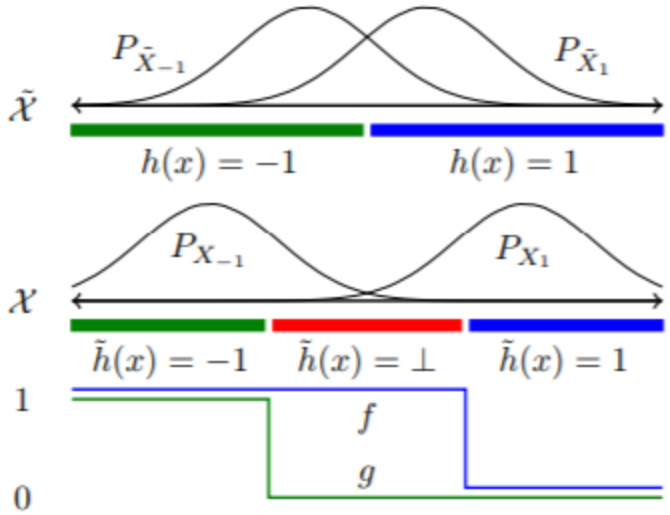
- $c_{\mathcal{A}}(x, x') = \begin{cases} 1, & \text{if } \mathcal{A}_x \cap \mathcal{A}_{x'} = \emptyset, \\ 0, & \text{else,} \end{cases}$
- $f_h(x) - g_h(x') \leq c_{\mathcal{A}}(x, x') \; \forall x, x' \in \mathcal{X}$, and so (f_h, g_h) is a pair of Kantorovich potentials for OT with ground-cost $c_{\mathcal{A}}$.

$$\therefore OT_{c_{\mathcal{A}}}(P_-, P_+) := \sup_{K-\text{potentials } \phi,\psi} \mathbb{E}_{P_-}[\phi(x)] - \mathbb{E}_{P_+}[\psi(x)]$$

$$\geq \sup_h \mathbb{E}_{P_-}[g_h(x)] - \mathbb{E}_{P_+}[f_h(x)]$$

$$= \sup_h 2(1 - \text{err}_{\mathcal{A}}(h)) - 1 = 1 - \text{err}^*_{\mathcal{A}}$$

Adversarial attacks as optimal transport [Bhagoji '19]



- $c_{\mathcal{A}}(x, x') = \begin{cases} 1, & \text{if } \mathcal{A}_x \cap \mathcal{A}_{x'} = \emptyset, \\ 0, & \text{else,} \end{cases}$
- $f_h(x) - g_h(x') \leq c_{\mathcal{A}}(x, x') \; \forall x, x' \in \mathcal{X}$, and so (f_h, g_h) is a pair of Kantorovich potentials for OT with ground-cost $c_{\mathcal{A}}$.

$$\therefore OT_{c_{\mathcal{A}}}(P_-, P_+) := \sup_{K-\text{potentials } \phi,\psi} \mathbb{E}_{P_-}[\phi(x)] - \mathbb{E}_{P_+}[\psi(x)]$$

$$\geq \sup_h \mathbb{E}_{P_-}[g_h(x)] - \mathbb{E}_{P_+}[f_h(x)]$$

$$= \sup_h 2(1 - \text{err}_{\mathcal{A}}(h)) - 1 = 1 - \text{err}^*_{\mathcal{A}}$$

Universal lower bound on adversarial robustness error

Theorem ([Bhagoji '19])

Given an attack model $\mathcal{A}$, let $OT_{\mathcal{A}}(P_+, P_-)$ be the **optimal transport** *distance between the +ve and -ve class-conditonal distributions of the samples, with the ground cost given by $c_{\mathcal{A}}(x, x') = \mathbb{1}_{\{\mathcal{A}_x \cap \mathcal{A}_{x'} = \emptyset\}}$. Then we have he following lower bound on the classification error against $\mathcal{A}$-attacks*

$$\text{err}^*_{\mathcal{A}} \geq \frac{1}{2}(1 - OT_{\mathcal{A}}(P_+, P_-)) \tag{12}$$

- In particular, for the **attackless case** where $\mathcal{A}_x = \{x\}\ \forall x \in \mathcal{X}$, one has $c_{\mathcal{A}}(x, x') = \mathbb{1}_{x \neq x'}$ and so $OT_{\mathcal{A}}(P_+, P_-) = TV(P_+, P_-)$. The theorem then reduces to the following well-known result

$$\text{err}^* \geq \frac{1}{2}(1 - TV(P_+, P_-)). \tag{13}$$

Total-Variational reformulation of the bounds

Theorem ([Dohmatob '20 ?] / ongoing work)

- *Let $\mathcal{A}$ be an attack, and for $a \in \mathcal{A}$ define $a_+(x) :\equiv a(x, +1)$. Define $\Omega := \{(x, x') \in \mathcal{X}^2 \mid \mathcal{A}_x \cap \mathcal{A}_{x'} \neq \emptyset\}$, and*

$$\begin{aligned} TV_{\mathcal{A}}(P_-, P_+) &:= \inf_{a \in \mathcal{A}} TV(a_{-\#}P_-, a_{+\#}P_+), \\ \widetilde{TV}_{\mathcal{A}}(P_-, P_+) &:= \inf_{\gamma_1, \gamma_2} TV(\text{proj}_{2\#}\gamma_1, \text{proj}_{1\#}\gamma_2), \end{aligned} \tag{14}$$

where the inf is taken over all distributions on $\mathcal{X}^2$ which are concentrated on Ω s.t $\text{proj}_{1\#}\gamma_1 = P_-$ and $\text{proj}_{2\#}\gamma_2 = P_+$. Then,

$$OT_{\mathcal{A}}(P_-, P_+) = \widetilde{TV}_{\mathcal{A}}(P_-, P_+) \leq TV_{\mathcal{A}}(P_-, P_+), \tag{15}$$

and there is equality if P_- and P_+ have densities w.r.t Lebesgue.

- Above bound suggest that rather than doing adversarial training, we'd rather do normal training on adversarially augmented data!

Worked example: hierarchical Gaussian classification

(Example from [Schmidt '18])

- $\mu \sim \mathcal{N}(0, I_p)$, $y \sim \text{Bern}(\{\pm 1\})$, $X|(Y = y) \sim \mathcal{N}(y\mu, \sigma^2 I_p)$,
- Consider the ℓ_∞-norm attack model $\mathcal{A}$ give by $\mathcal{A}_x = \text{Ball}_{\ell_\infty}(x; \varepsilon)$.
- Given n samples $S_n = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ from this model, how small can the robust error of a classifier be ?
- More precisely, lets bound

$$\mathbb{E}_{\mu \sim \mathcal{N}(0,1)} \inf_{\hat{h}} \mathbb{E}_{S_n \sim P^n} \mathbb{E}_{\hat{h}_n \sim \hat{h}(S_n)} [\text{err}_{\mathcal{A}}(\hat{h}_n; \mu)], \tag{16}$$

where $\text{err}_{\mathcal{A}}(\hat{h}_n; \mu)$ is the adversarial robust error of $\hat{h}_n$ defined by

$$\text{err}_{\mathcal{A}}(\hat{h}_n; \mu) := \mathbb{E}_{y \sim \text{Bern}(\{\pm 1\})} \mathbb{E}_{x \sim \mathcal{N}(y\mu, \sigma^2 I_p)} [\exists x' \in \mathcal{A}_x \text{ s.t } \hat{h}_n(x') \neq y].$$

Worked example: hierarchical Gaussian classification

$$\mathrm{err}_{\mathcal{A}}(\hat{h}_n; \mu) := \mathbb{E}_{y \sim \mathrm{Bern}(\{\pm 1\})} \mathbb{E}_{x \sim \mathcal{N}(y\mu, \sigma^2 I_p)} [\exists x' \in \mathcal{A}_x \text{ s.t } \hat{h}_n(x') \neq y].$$

- The posterior distribution of the model parameter is $\mathcal{N}(\hat{\mu}_n, \hat{\sigma}_n^2)$, with $\hat{\sigma}_n^2 = \frac{\sigma^2}{\sigma^2 + n}$, and $\hat{\mu}_n = \frac{n}{\sigma^2+n}\bar{x}$ with $\bar{x} = \frac{1}{n}\sum_{i=1}^n x_i$

$$\therefore \mathrm{err}_{\mathcal{A}}(\hat{h}_n) = \inf_h \mathrm{err}_{\mathcal{A}}(h; \hat{\mu}_n, \hat{\sigma}_n^2) \geq \dots$$

$$\geq \mathbb{E}_\mu \mathbb{E}_{S_n} \frac{1}{2} \left(1 - \inf_{\|z\|_\infty \leq \varepsilon} \mathrm{TV}(\mathcal{N}(-\hat{\mu}_n + z, \hat{\sigma}_n^2), \mathcal{N}(\hat{\mu}_n - z, \hat{\sigma}_n^2)) \right)$$

$$\geq \mathbb{E}_\mu \mathbb{E}_{S_n} \Phi \left(\frac{\sqrt{p}}{\hat{\sigma}_n} (\|\hat{\mu}_n\| - \varepsilon)_+ \right) \geq \mathbb{E}_\mu \mathbb{E}_{S_n} \mathbb{P}(\|\hat{\mu}_n\|_\infty \leq \varepsilon) \Phi(0)$$

$$\approx \frac{1}{2} \mathbb{P}_{u \sim \mathcal{N}(0, I_p)} \left(\frac{n}{n + \sigma^2} \|u\|_\infty \leq \varepsilon \right)$$

Worked example: hierarchical Gaussian classification

- Thus, if $n \leq \frac{\varepsilon^2\sigma^2}{8\log(d)}$, then $\frac{n}{\sigma^2+n} \leq \frac{\varepsilon}{2\sqrt{2\log(d)}}$, and so

$$\begin{aligned}
\therefore \mathrm{err}^*_{\mathcal{A},\,n} &\geq \frac{1}{2}\mathbb{P}_{u\sim\mathcal{N}(0,I_p)}\left(\frac{n}{n+\sigma^2}\|u\|_\infty \leq \varepsilon\right) \\
&\geq \frac{1}{2}\mathbb{P}_{u\sim\mathcal{N}(0,I_p)}\left(\|u\|_\infty \leq 2\sqrt{2\log(d)}\right) \\
&\vdots \\
&\geq \frac{1}{2}(1-1/d) \approx \frac{1}{2} \text{ in high dimensions!}
\end{aligned}$$

Main references

- [Dohmatob '19] *Generalized No Free Lunch Theorem for Adversarial Robustness*
- [Bhagoji '19] *Lower Bounds on Adversarial Robustness from Optimal Transport*
- [Tsipras '18] *There is no free lunch in adversarial robustness*
- [Gilmer '18] *Adversarial spheres*
- [Goodfellow '14] *Explaining and harnessing adversarial examples*
- [Su '17] *One pixel attack for fooling deep neural networks*
- [Fawzi '18] *Adversarial vulnerability for any classifier*
- [Athalye '18] *Obfuscated Gradients Give a False Sense of Security: Circumventing Defenses to Adversarial Examples*

Questions ?