# Generalized No Free Lunch Theorem for Adversarial Robustness **(poster #69)**

**Elvis Dohmatob**

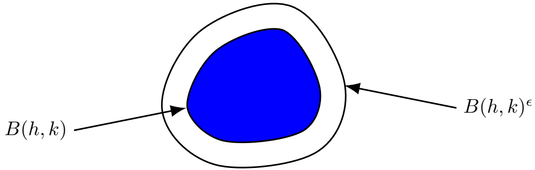
Criteo AI Lab

June 11, 2019

# Bad news: adversarial examples are here to stay :)

# Adversarial attacks are a 'butterfly effect' on data manifold

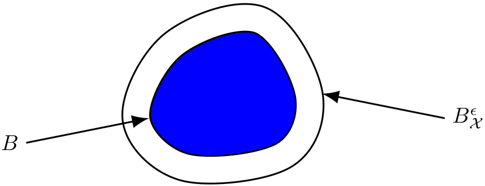
- **Classifier**: Measurable function $h : \mathcal{X} \to \mathcal{Y}$
- **Error set**: $B(h,k) = \{x \in \mathcal{X} \mid h(x) \neq k\}$, $h =$ classifier
- **Almost error set**: $B(h,k)^{\varepsilon} := \{x \in \mathcal{X} \mid \text{dist}(x, B(h,k)) \leq \varepsilon\}$




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

# Isoperimetry in general metric spaces

- $\mathcal{X} = (\mathcal{X}, d)$, $\mu$ is probability measure on $\mathcal{X}$, $B$ is Borel subset




- **Gaussian isoperimetric inequality (GIPI)** means that:

$$\mu(B^{\epsilon}) \geq 1 - e^{-\frac{1}{2c}(\varepsilon - \varepsilon_B)^2}, \ \forall \varepsilon \geq \varepsilon_B$$

- Current works [Tsipras '18; Fawzi et al. 18; Shafahi 18] use elementary GIPI, where $\mathcal{X} = (\mathbb{R}^p, L_2)$, and $\mu = \gamma_p$.

- Arguments not powerful enough for more general geometry!

# Talagrand transportation-cost inequality

**The $T_2(c)$ property**

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

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

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

- $T_2(c) \implies \mathrm{GIPI}(c) \implies$ concentration
- Satisfied by a variety of distributions $\mu$
- Links relative entropy to optimal transport

# Generalized No Free Lunch Theorem for geodesic distance

**Theorem (Generalized "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 $\mathrm{err}(h|k) > 0$, define $\varepsilon(h|k) := \sigma_k \sqrt{2 \log(1/\mathrm{err}(h|k))}$. Then we have the following bounds:*

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

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

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

$$d(h|k) \leq \sigma_k \left( \sqrt{\log(1/\mathrm{err}(h|k))} + \sqrt{\pi/2} \right) \tag{3}$$

# Generalized No Free Lunch Theorem for $\ell_p$ distances

**Corollary (Generalized NFLT for $\ell_\infty$ attacks [Dohmatob '18])**

*In particular, for the $\ell_\infty$ threat model, we have the following bounds:*

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

$$\mathrm{acc}_\varepsilon(h|k) \leq e^{-\frac{p}{2\sigma_k^2}(\varepsilon - \varepsilon(h|k)/\sqrt{p})^2}. \tag{4}$$

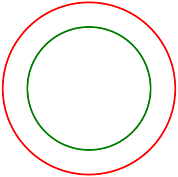
*(B2)* **Average distance to error set**:

$$d(h|k) \leq \frac{\sigma_k}{\sqrt{p}} \left(\sqrt{\log(1/\operatorname{err}(h|k))} + \sqrt{\pi/2}\right) \tag{5}$$

- $\therefore$ **inherent vulnerability** to $\ell_\infty$-perturbations of size $\mathcal{O}(1/\sqrt{p})$.
- Result is largely independent of classifier!

# Special cases of our results

- Log-concave distribs $dP_{X|k} \propto e^{-v_k(x)}dx$ satisfying Emeŕy-Bakry curvature condition: $\text{Hess}_x(v_k) + \text{Ric}_x(\mathcal{X}) \succeq (1/\sigma_k^2)I_p$.
  - e.g Gaussians (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 "adversarial spheres" (considered in [Gilmer '18]), tori, or any compact Lie group.
- Pushforward via a Lipschitz function $f$, of a distribution in $\text{T}_2(\sigma_k^2)$. Indeed, take $\tilde{\sigma}_k = \|f\|_{\text{Lip}}\sigma_k$. E.g [Fawzi 18; Shafahi 18]
- Tensor product $\mu_1 \otimes \mu_2 \otimes \ldots \otimes \mu_p$ of distributions having $\text{T}_2(c)$ also has $\text{T}_2(c)$.
- Etc., etc.

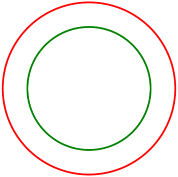


- $Y \sim \text{Bern}(1/2, \{\pm\})$,
- $X|k \sim \text{uniform}(\mathbb{S}^p_{R_k})$, where $R_+ > R_- > 0$.
- $\mathbb{S}^p_{R_k}$ is a compact Riemannian manifold with constant Ricci curvature $(p-1)R_k^{-2}$.
- Thus $P_{X|k}$ satisfies $\text{T}_2(R_k^2/(p-1))$.

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

- Same bounds obtained in [Gilmer '18] "manually"

# Special case: "Adversarial spheres" [Gilmer '18]

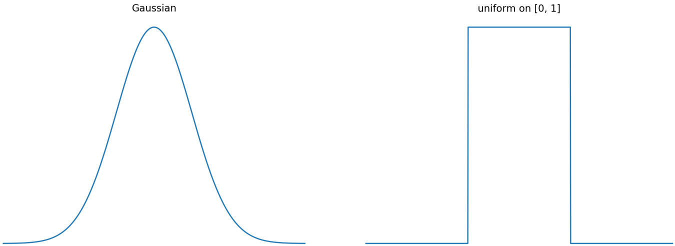


- $Y \sim \text{Bern}(1/2, \{\pm\})$,
- $X|k \sim \text{uniform}(\mathbb{S}^p_{R_k})$, where $R_+ > R_- > 0$.
- $\mathbb{S}^p_{R_k}$ is a compact Riemannian manifold with constant Ricci curvature $(p-1)R_k^{-2}$.
- Thus $P_{X|k}$ satisfies $\text{T}_2(R_k^2/(p-1))$.

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

- Same bounds obtained in [Gilmer '18] "manually"

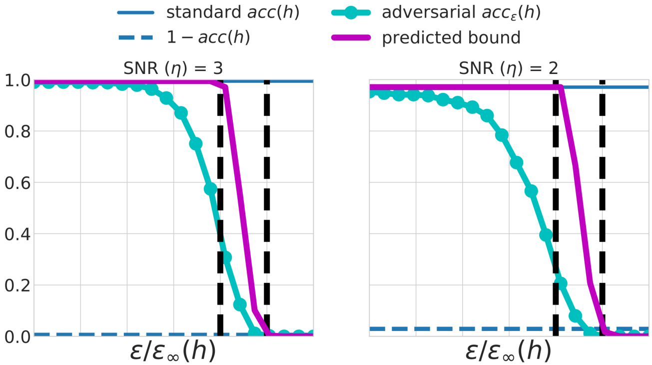
# Special case: Hypercube $[0,1]^p$ [Shafahi 18]




- $U([0,1]) = \text{CDF}(\textit{Gaussian}) := \int_{-\infty}^{z} \gamma(x)dx$, a $\frac{1}{\sqrt{2\pi}}$-Lipschitz map
- Therefore $\mathcal{U}([0,1])$ has concentration property with $c = 2\pi$.

# Simulated data [Tsipras '18] "noisy features" dataset

- $Y \sim \mathrm{Bern}(\{\pm 1\})$, $X|Y \sim \mathcal{N}(Y\eta, 1)^{\times p}$, with $p = 1000$ where $\eta$ is an SNR parameter which controls the difficulty of the problem.




- Phase-transition occurs as predicted by our theorems

# Summary of contributions

- We have shown that on a very broad class of data distributions, any classifier with even a bit of accuracy is vulnerable to adversarial attacks
- We use powerful tools from geometric probability theory to generalize recent impossibility results on adversarial robustness [Fawzi '18, Gilmer '18, Tsipras '18; Shafahi '18; etc.].
- Our predictions are not incompatible with current research endeavors being investing in designing defenses against adversarial attacks.
  - It simply says there is a sharp and definitive limit to the amount of robustness that can be guaranteed
- Full manuscript:

# Questions ? (come to **poster #69**)