Limitations of Adversarial Robustness: Strong No Free Lunch Theorem

Elvis Dohmatob

Abstract

This manuscript presents some new impossibility results on adversarial robustness in machine learning, a very important yet largely open problem. We show that if conditioned on a class label the data distribution satisfies the $W_2$ Talagrand transportation-cost inequality (for example, this condition is satisfied if the conditional distribution has density which is log-concave; is the uniform measure on a compact Riemannian manifold with positive Ricci curvature; etc.) any classifier can be adversarially fooled with high probability once the perturbations are slightly greater than the natural noise level in the problem. We call this result The Strong "No Free Lunch" Theorem as some recent results (Tsipras et al. 2018, etc.) on the subject can be immediately recovered as very particular cases. Our theoretical bounds are demonstrated on both simulated and real data (MNIST). We conclude the manuscript with some speculation on possible future research directions.

1. Introduction

An adversarial attack operates as follows:

- The goal of the attacker is to fool the classifier into classifying $x^{\text{adv}}$ as label different from $k$.
- A robust classifier tries to limit this failure mode, at a prescribed tolerance $\epsilon$.

1.1. A toy example illustrating the fundamental issue

To motivate things, consider the following "toy" problem from (Tsipras et al., 2018), which consists of classifying a target $Y \sim \text{Bern}(1/2, \{\pm 1\})$ based on $p \ge 2$ explanatory variables $X := (X^1, X^2, \ldots, X^p)$ given by

$$X^1|Y = \begin{cases} +Y, & \text{w.p } 70\%, \\ -Y, & \text{w.p } 30\%, \end{cases}$$

and $X^j|Y \sim \mathcal{N}(\eta Y, 1)$, for $j = 2, \ldots, p$, where $\eta \sim p^{-1/2}$ is a fixed scalar which (as we will see) controls the difficulty of the problem. Now, as was shown in (Tsipras et al., 2018), the above problem can be solved perfectly with generalization accuracy $\approx 100\%$, but the "champion" estimator can also be fooled, perfectly! Indeed, the linear estimator given by $h_{\text{avg}}(x) := \text{sign}(w^T x)$ with $w = (0, 1/(p - 1), \ldots, 1/(p - 1)) \in \mathbb{R}^p$, where we allow an attacker to modify each feature by an amount at most $\epsilon \approx 2\eta$, has the afore-mentioned properties. Indeed, using basic tail bounds for the Gaussian distributions, one can show that for any $\delta \in (0, 1]$ the following hold

- The standard accuracy of the linear model $h_{\text{avg}}$ is at least $1 - \delta$ if $\eta \geq \sqrt{2 \log(1/\delta)}/{(p - 1)}$, and
- This same model’s adversarial accuracy is at most $\delta$ for $\epsilon \geq \eta + \sqrt{2 \log(1/\delta)}/{(p - 1)}$

See (Tsipras et al., 2018) for details (or see supplemental).

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

An autopsy of what is going on. Recall that the entropy of a univariate Gaussian is $\text{Ent}(\mathcal{N}(\mu, \sigma^{2})) = \ln(\sqrt{2\pi\sigma^{2}})$ nats. Now, for $j = 2, 3, \ldots, p$, the distribution of feature $X^j$ is a Gaussian mixture $\frac{1}{2} \sum_{Y = \pm 1} \mathcal{N}(\eta Y, 1)$ and so one

1 Criteo, Paris, France. Correspondence to: Elvis Dohmatob <e.dohmatob@criteo.com>.

computes the mutual information between $X^j$ and the class label $Y$ as

$$MI(X^j; Y) := \text{Ent}(X^j) - \text{Ent}(X^j | Y)$$

$$= \text{Ent} \left( \frac{1}{2} \sum_{y=\pm 1} \mathcal{N}(\eta y, 1) \right) - \frac{1}{2} \sum_{y=\pm 1} \text{Ent}(\mathcal{N}(\eta y, 1))$$

$$= \ln(\sqrt{\pi} e) + \eta^2 - r - 2(1/2) \ln(\sqrt{\pi} e) = \eta^2 - r \leq \eta^2,$$

where (see (Michalowicz et al., 2008) for the details)

$$r := \frac{2}{\sqrt{2\pi\eta}} e^{-\eta^2/2} \int_0^\infty e^{z^2/2} \cosh(z) \ln(\cosh(z)) dz \geq 0.$$  

Thus $MI(X^j; Y) \leq \eta^2$. Since $\eta^2 \sim 1/p$, we conclude that these features barely share any information with the target variable $Y$. Indeed, (Tsipras et al., 2018) showed improved robustness on the above problem, with feature-selection based on mutual information.

**Basic “No Free Lunch” Theorem.** Reading the information calculations above, a skeptic could point out that the underlying issue here is that the estimator $h_{avg}$ over-exploits the fragile / non-robust variables $X^2, \ldots, X^p$ to boost ordinary generalization accuracy, at the expense of adversarial robustness. However, it was rigorously shown in (Tsipras et al., 2018) that on this particular problem, every estimator is vulnerable. Precisely, the authors proved the following basic “No Free Lunch” theorem.

**Theorem 1** (Basic No Free Lunch, (Tsipras et al., 2018)).

For the problem above, any estimator which has ordinary accuracy at least $1 - \delta$ must have robust adversarial robustness accuracy at most $7\delta/3$ against $\ell_\infty$-perturbations of maximum size $\epsilon \geq 2\eta$.

**1.2. Highlight of our main contributions**

In this manuscript, we prove that under some “curvature conditions” (to be precised later) on the conditional density of the data, it holds that

For geodesic / faithful attacks:

- Every (non-perfect) classifier can be adversarially fooled with high probability by moving sample points an amount less than a critical value, namely

$$\epsilon(h|k) := \sigma_k \sqrt{2 \log(1/err(h|k))} \approx \sigma_k \Phi^{-1}(\text{acc}(h|k))$$

along the data manifold, where $\sigma_k$ is the “natural noise level” in the data points with class label $k$ and $\text{err}(h|k)$ generalization error of the classifier in the non-adversarial setting

- Moreover, the average distance of a sample point of true label $k$ to the error set is upper-bounded by

$$\epsilon(h|k) + \sigma_k \sqrt{\frac{\pi}{2}} = \sigma_k \left( \Phi^{-1}(\text{acc}(h|k)) + \sqrt{\frac{\pi}{2}} \right)$$

For attacks in flat space $\mathbb{R}^p$:

- In particular, if the data points live in $\mathbb{R}^p$, where $p$ is the number of features), then every classifier can be adversarially fooled with high probability, by changing each feature by an amount less than a critical value, namely

$$\epsilon_\infty(h|k) := \sigma_k \sqrt{2 \log(1/\text{err}(h|k))/p} \approx \sigma_k \Phi^{-1}(\text{acc}(h|k)).$$

- Moreover, we have the bound

$$d(h|k) \leq \epsilon_\infty(h|k) + \frac{\sigma_k}{\sqrt{\pi}} \sqrt{\frac{2}{\pi}} \approx \sigma_k \left( \Phi^{-1}(\text{acc}(h|k)) + \sqrt{\frac{\pi}{2}} \right).$$

In fact, we prove similar results for $\ell_1$ (reminiscent of few-pixel attacks (Su et al., 2017)) and even any $\ell_s$ norm on $\mathbb{R}^p$. We call these results The Strong “No Free Lunch” Theorem as some recent results (e.g. (Tsipras et al., 2018; Fawzi et al., 2018a; Gilmer et al., 2018b)), etc.) on the subject can be immediately recovered as very particular cases. Thus adversarial (non)-robustness should really be thought of as a measure of complexity of a problem. A similar remark has been recently made in (Bubeck et al., 2018).

The sufficient “curvature conditions” alluded to above imply concentration of measure phenomena, which in turn imply our impossibility bounds. These conditions are satisfied in a large number of situations, including cases where the class-conditional distribution is the volume element of a compact Riemannian manifold with positive Ricci curvature; the class-conditional data distribution is supported on a smooth manifold and has log-concave density w.r.t the curvature of the manifold; or the manifold is compact; is the pushforward via a Lipschitz continuous map, of another distribution which verifies these curvature conditions; etc.

**1.3. Notation and terminology**

$\mathcal{X}$ will denote the feature space and $\mathcal{Y} := \{1, 2, \ldots, K\}$ will be the set of class labels, where $K \geq 2$ is the number of classes, with $K = 2$ for binary classification. $P$ will be the (unknown) joint probability distribution over $\mathcal{X} \times \mathcal{Y}$, of two prototypical random variables $X$ and $Y$ referred to the features and the target variable, which take values in $\mathcal{X}$ and $\mathcal{Y}$ respectively. Random variables will be denoted by capital letters $X$, $Y$, $Z$, etc., and realizations thereof will be denoted $x$, $y$, $z$, etc. respectively.

For a given class label $k \in \mathcal{Y}$, $X_k \subseteq \mathcal{X}$ will denote the set of all samples whose label is $k$ with positive probability under $P$. It is the support of the restriction of $P$ onto the plane
Let $\mathcal{X} \times \{k\}$. This restriction is denoted $P_{X|Y=k}$ or just $P_{X|k}$, and defines the conditional distribution of the features $X$ given that the class label has the value $k$. We will assume that all the $X_i$’s are finite-dimensional smooth Riemannian manifolds. This is the so-called manifold assumption, and is not unpopular in machine learning literature. A classifier is just a measurable mapping $h : \mathcal{X} \rightarrow \mathcal{Y}$, from features to class labels.

In our bounds, $\tilde{O}(\ldots)$ will be used to mean the usual big-O notation, except for multiplicative constants which only depend on the performance of the classifier.

**Threat models.** Let $d_{\mathcal{X}}$ be a distance / metric on the input space $\mathcal{X}$ and $\epsilon \geq 0$ be a tolerance level. The $d_{\mathcal{X}}$ threat model at tolerance $\epsilon$ is a scenario where the attacker is allowed to perturb any input point $x \mapsto x^{\text{adv}}$ with the constraint that $d_{\mathcal{X}}(x^{\text{adv}}, x) \leq \epsilon$. When $\mathcal{X}$ is a manifold, the threat model considered will be that induced by the geodesic distance, and will be naturally referred to as the geodesic threat model.

**Flat threat models.** In the special case of euclidean space $\mathcal{X} = \mathbb{R}^n$, we will always consider the distances defined for $q \in [1, \infty]$ by $d(x, z) = \|x - z\|_q$, where

$$\|a\|_q := \begin{cases} \left( \sum_{j=1}^{p} |a_j|^q \right)^{1/q}, & \text{if } 1 \leq q < \infty, \\ \max\{|a|^1, \ldots, |a|^p\}, & \text{if } q = \infty. \end{cases}$$

(1)

The $\ell_\infty$ / sup case where $q = \infty$ (Tsipras et al., 2018) is particularly important: the corresponding threat model allows the adversary to separately increase or decrease each feature by an amount at most $\epsilon$. The sparse case $q = 1$ is a convex proxy for so-called “few-pixel” attacks (Su et al., 2017) wherein the total number of features that can be tampered-with by the adversary is limited.

**Adversarial robustness accuracy and error.** The adversarial robustness accuracy of $h$ at tolerance $\epsilon$ for a class label $k \in \mathcal{Y}$ and w.r.t the $d_{\mathcal{X}}$ threat model, denoted $\text{acc}_{d_{\mathcal{X}}, \epsilon}(h|k)$, is defined by

$$\text{acc}_{d_{\mathcal{X}}, \epsilon}(h|k) := P_{X|k}(h(x') = k \forall x' \in \text{Ball}_{\mathcal{X}}(X; \epsilon)).$$

(2)

This is simply the probability no sample point $x$ with true class label $k$ can be perturbed by an amount $\leq \epsilon$ measured by the distance $d_{\mathcal{X}}$, so that it get misclassified by $h$. This is an adversarial version of the standard class-conditional accuracy $\text{acc}(h) = P_{(X, Y)}(h(X) = Y)$ corresponding to $\epsilon = 0$. The corresponding adversarial robustness error is then $\text{err}_\epsilon(h|k) := 1 - \text{acc}_{\epsilon}(h|k)$. This is the adversarial analogue of the standard notion of the class-conditional generalization / test error, corresponding to $\epsilon = 0$.

Similarly, one defines the unconditional adversarial accuracy

$$\text{acc}_{\epsilon}(h) = P_{(X, Y)}(h(x') = Y \forall x' \in \text{Ball}_{\mathcal{X}}(X; \epsilon)),$$

(3)

which is an adversarial version of the standard accuracy $\text{acc}(h) = P_{(X, Y)}(h(X) = Y)$. Finally, adversarial robustness radius of $h$ on class $k$

$$d(h|k) := \mathbb{E}_{X|k}[d(X, B(h, k))],$$

(4)

where $B(h, k) := \{x \in \mathcal{X} | h(x) \neq k\}$ is the set of inputs classified by $h$ as being of label other than $k$. Measureability of $h$ implies that $B(h, k)$ is a Borel subset of $\mathcal{X}$, $d(x, k)$ is nothing but the average distance of a sample point $x \in \mathcal{X}$ with true label $k$, from the set of samples classified by $h$ as being of another label. The smaller the value of $d(h|k)$, the less robust the classifier $h$ is to adversarial attacks on samples of class $k$.

**Remark 1.** By the properties of expectation and conditioning, it holds that $\min_k \text{acc}_{\epsilon}(h|k) \leq \text{acc}_{\epsilon}(h) = \mathbb{E}_Y[\text{acc}_{\epsilon}(h|Y)] = \sum_{k=1}^{K} \pi_k \text{acc}_{\epsilon}(h|k) \leq \max_k \text{acc}_{\epsilon}(h|k)$, where $\pi_k := P(k)$. Thus, bounds on the $\text{acc}_{\epsilon}(h|k)$’s imply bounds on $\text{acc}_{\epsilon}(h)$.

1.4. Rough organization of the manuscript

In section 1.1, we start off by presenting a simple motivating classification problem from (Tsipras et al., 2018), which as shown by the authors, already exhibits the “No Free Lunch” issue. In section 2.1 we present some relevant notions from geometric probability theory which will be relevant for our work, especially Talagrand’s transportation-cost inequality and also Marton’s blowup Lemma. Then in section 2.3, we present the main result of this manuscript, namely, that on a rich set of distributions no classifier can be robust even to modest perturbations (comparable to the natural noise level in the problem). This generalizes the results of (Tsipras et al., 2018; Gilmer et al., 2018b) and to some extent, (Fawzi et al., 2018a). Our results also extend to the distributional robustness setting (no highlighted here but is presented in the Appendix B).

An in-depth presentation of related works is given in section 3. Section 4 presents experiments on both simulated and real data that confirm our theoretical results. Finally, section 5 concludes the manuscript with possible future research directions.

All proofs are presented in Appendix A.
2. Strong “No Free Lunch” Theorem for adversarial robustness

2.1. Terminology and background

Neighborhood of a set in a metric space. The $\epsilon$-blowup (aka $\epsilon$-neighborhood, aka $\epsilon$-fattening, aka $\epsilon$-enlargement) of a subset $B$ of a metric space $\mathcal{X}$ is denoted $B^\epsilon_{\mathcal{X}}$, defined by $B^\epsilon_{\mathcal{X}} := \{x \in \mathcal{X} | d_\mathcal{X}(x,B) \leq \epsilon\}$, where $d_\mathcal{X}(x,B) := \inf \{d_\mathcal{X}(x,y) | y \in B\}$ is the distance of $x$ from $B$. Note that $B^\epsilon_{\mathcal{X}}$ is an increasing function of both $B$ and $\epsilon$; that is, if $A \subseteq B \subseteq \mathcal{X}$ and $0 \leq \epsilon_1 \leq \epsilon_2$, then $A \subseteq A^\epsilon_1_{\mathcal{X}} \subseteq B^\epsilon_1_{\mathcal{X}} \subseteq B^\epsilon_2_{\mathcal{X}}$. In particular, $B^0_{\mathcal{X}} = B$ and $B^\infty_{\mathcal{X}} = \mathcal{X}$. Also observe that each $B^\epsilon_{\mathcal{X}}$ can be rewritten in the form $B^\epsilon_{\mathcal{X}} = \bigcup_{x \in B} \text{Ball}_\mathcal{X}(x;\epsilon)$, where $\text{Ball}_\mathcal{X}(x;\epsilon) := \{x' \in \mathcal{X} | d_\mathcal{X}(x',x) \leq \epsilon\}$ the closed ball in $\mathcal{X}$ with center $x$ and radius $\epsilon$. Refer to Fig. 1. In a bid to simplify notation, when there is no confusion about the underlying metric space, we will simply write $B^\epsilon$ for $B^\epsilon_{\mathcal{X}}$. When there is no confusion about the the underlying set $\mathcal{X}$ but not the metric thereupon, we will write $B^\epsilon_x$. For example, in the metric space $(\mathbb{R}^p, \ell_q)$, we will write $B^\epsilon_{\mathcal{X}}$ instead of $B^\epsilon_{(\mathbb{R}^p, \ell_q)}$ for the $\epsilon$-blowup of $B \subseteq \mathbb{R}^p$.

An example which will be central to us is when $h : \mathcal{X} \to \mathcal{Y}$ is a classifier, $k \in \mathcal{Y}$ is a class label, and we take $B$ to be the “bad set” $B(h,k) \subseteq \mathcal{X}$ of inputs which are assigned a label different from $k$, i.e.

$$B(h,k) := \{x | h(x) \neq k\} = \bigcup_{k' \neq k} \{x | h(x) = k'\}. \quad (5)$$

$B^\epsilon_{\mathcal{X}} = B(h,k)^\epsilon$ is then nothing but the event that there is data point with a “bad $\epsilon$-neighbor”, i.e the example can be misclassified by applying a small perturbation of size $\leq \epsilon$. This interpretation of blowups will be central in the sequel, and we will be concerned with lower-bounding the probability of the event $B(h,k)^\epsilon$ under the conditional measure $P_{B(h,k)|k}$. This is the proportion of points $x \in \mathcal{X}$ with true class label $k$, such that $h$ assigns a label $\neq k$ to some $\epsilon$-neighbor $x' \in \mathcal{X}$ of $x$. Alternatively, one could study the local robustness radii $r_\mathcal{X}(x,k) := \inf \{d(x',x) | x' \in \mathcal{X}, h(x') \neq k\} := d(x,B(h,k))$, for $(x,k) \sim P$, as was done in (Fawzi et al., 2018a), albeit for a very specific problem setting (generative models with Gaussian noise). More on this in section 3. Indeed $r_\mathcal{X}(x,k) \leq \epsilon \iff x \in B(h,k)^\epsilon$.

2.2. Measure concentration on metric spaces

For our main results, we will need some classical inequalities from optimal transport theory, mainly the Talagrand transportation-cost inequality and Marton’s Blowup inequality (see definitions below). Fix a reference measure $\mu$ in $\mathcal{P}(\mathcal{X})$, the Wasserstein space of all probability measures $\mu$ on the metric space $\mathcal{X}$ with finite order moment, i.e such that there exists a point $a \in \mathcal{X}$ with $\mathbb{E}_{X \sim \mu}[d_\mathcal{X}(a,X)^2] < \infty$.

**Definition 1** (T2(c) property –a.k.a Talagrand W2 transportation-cost inequality). $\mu$ is said to satisfy T2(c) if for every other distribution $\nu$ on $\mathcal{X}$, which is absolutely continuous w.r.t $\mu$ (written $\nu \ll \mu$), one has

$$W_2(\nu,\mu) \leq \sqrt{2c \cdot kl(\nu||\mu)}, \quad (6)$$

where

- $W_2(\nu,\mu)$ is the Wasserstein 2-distance between $\nu$ and $\mu$ defined by
  $$W_2(\nu,\mu) := \left(\inf_{\pi \in \Pi(\nu,\mu)} \mathbb{E}_\pi[d_\mathcal{X}(X',X)^2]\right)^{1/2}, \quad (7)$$

  with $\Pi(\nu,\mu)$ being the set of all couplings of $\nu$ and $\mu$.

- $kl(\nu||\mu)$ is the entropy of $\nu$ relative to $\mu$, defined by $kl(\nu||\mu) = \int_X \log \left(\frac{d\nu}{d\mu}\right) d\mu$ if $\nu \ll \mu$ and $+\infty$ else.

Note that if $0 < c < c'$, then T2(c) $\subseteq$ T2(c'). The inequality (6) in the above definition is a generalization of the well-known Pinsker’s inequality for the total variation distance between probability measures. Unlike Pinsker’s inequality which holds unconditionally, (6) is a privilege only enjoyed by special classes of reference distributions $\mu$. These include: log-concave distributions on manifolds (e.g multi-variate Gaussian), distributions on compact Riemannian manifolds of positive Ricci curvature (e.g spheres, tori, etc.), pushforwards of distributions that satisfy some T2 inequality, etc. In section 2.5, these classes of distributions will be discussed in detail as sufficient conditions for our impossibility theorems.

**Definition 2** (BLOWUP(c) property). $\mu$ is said to satisfy BLO- \text{BLOWUP}(c) if for every Borel $B \subseteq \mathcal{X}$ with $\mu(B) > 0$ and for every $\epsilon \geq \sqrt{2c \cdot \log(1/\mu(B))}$, it holds that

$$\mu(B^\epsilon) \geq 1 - e^{-\frac{1}{c} \left(\sqrt{2c \cdot \log(1/\mu(B))}\right)^2}. \quad (8)$$

It is a classical result that the Gaussian distribution on $\mathbb{R}^p$ has BLO-\text{BLOWUP}(1) and T2(1), a phenomenon known as
**Limitations of Adversarial Robustness: Strong No Free Lunch Theorem**

Gaussian isoperimetry. These results date back to at least works of E. Borel, P. Lévy, M. Talagrand and of K. Marton (Boucheron et al., 2013).

The following lemma is the most important tool we will use to derive our bounds.

**Lemma 1** (Marton’s Blowup lemma). On a fixed metric space, it holds that $T_2(c) \subseteq BLOWUP(c)$.

**Proof.** The proof is classical, and is a variation of original arguments by Marton. We provide it in Appendix A, for the sake of completeness. □

### 2.3. Generalized “No Free Lunch” Theorem

It is now ripe to present the main results of this manuscript.

**Theorem 2** (Strong “No Free Lunch” on curved space). Suppose that for some $\sigma_k > 0$, $P_{X|h}$ has the $T_2(\sigma_k^2)$ property on the conditional manifold $X|h := \text{supp}(P_{X|h}) \subseteq X$. Given a classifier $h : X \rightarrow \{1, 2, \ldots, K\}$ for which $\text{acc}(h|k) < 1$ (i.e the classifier is not perfect on the class k), define

$$
\epsilon(h|k) := \sigma_k \sqrt{2 \log(1/\text{err}(h|k))} = \sigma_k \Phi^{-1}(\text{acc}(h|k)).
$$

Then for the geodesic threat model, we have the bounds

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

$$
\text{acc}_\epsilon(h|k) \leq \min(\text{acc}(h|k), e^{-\frac{1}{2} \sigma_k^2 (\epsilon - \epsilon(h|k))^2}).
$$

(B) **Bound on average distance to error set:**

$$
d(h|k) \leq \sigma_k \left( \sqrt{\log(1/\text{err}(h|k))} + \sqrt{\frac{\pi}{2}} \right)
= \sigma_k \left( \Phi^{-1}(\text{acc}(h|k)) + \sqrt{\frac{\pi}{2}} \right).
$$

**Proof.** The main idea is to invoke Lemma 1, and then apply the bound (8) with $B = B(h, k) := \{x \in X|h(x) \neq k\}$, $\mu = P_{X|h}$, and $e = \sigma_k^2$. See Appendix A for details. □

In the particular case of attacks happening in euclidean space (this is the default setting in the literature), the above theorem has the following corollary.

**Corollary 1** (Strong “No Free Lunch” Theorem on flat space). Let $1 \leq q \leq \infty$, and define

$$
\epsilon_q(h|k) := \epsilon(h|k)p^{\frac{1}{q} - \frac{1}{2}} \approx p^{\frac{1}{q} - \frac{1}{2}} \sigma_k \Phi^{-1}(\text{acc}(h|k)).
$$

If in addition to the assumptions of Theorem 2 the conditional data manifold $X_k$ is flat, i.e Ric$X_k = 0$, then for the $\ell_q$ threat model, we have

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

$$
\text{acc}_\epsilon(h|k) \leq \min(\text{acc}(h|k), e^{-\frac{1}{2} \sigma_k^2 (\epsilon - \epsilon_q(h|k))^2}).
$$

(A2) **Average distance to error set:** if $\epsilon \geq \epsilon_q(h|k)$, then

$$
d(h|k) \leq \frac{\sigma_k}{p^{1/2-1/q}} \left( \sqrt{\log(1/\text{err}(h|k))} + \sqrt{\frac{\pi}{2}} \right)
= \frac{\sigma_k}{p^{1/2-1/q}} \left( \Phi^{-1}(\text{acc}(h|k)) + \sqrt{\frac{\pi}{2}} \right).
$$

In particular, for the $\ell_\infty$ threat model, we have

(B1) **Adversarial robustness accuracy:** if $\epsilon \geq \epsilon_q(h|k)$, then

$$
\text{acc}_\epsilon(h|k) \leq \min(\text{acc}(h|k), e^{-\frac{1}{2} \sigma_k^2 (\epsilon - \epsilon_q(h|k))/\sqrt{p}}^2).
$$

(B2) **Bound on average distance to error set:**

$$
d(h|k) \leq \frac{\sigma_k}{\sqrt{p}} \left( \sqrt{\log(1/\text{err}(h|k))} + \sqrt{\frac{\pi}{2}} \right)
= \frac{\sigma_k}{\sqrt{p}} \left( \Phi^{-1}(\text{acc}(h|k)) + \sqrt{\frac{\pi}{2}} \right).
$$

**Proof.** See Appendix A. □

### 2.4. Making sense of the theorems

Fig. 2 gives an instructive illustration of bounds in the above theorems. For perfect classifiers, the test error $\text{err}(h|k) := 1 - \text{acc}(h|k)$ is zero and so the factor $\sqrt{\log(1/\text{err}(h|k))}$ appearing in definitions for $\epsilon(h|k)$ and $\epsilon_q(h|k)$ is $\infty$; else this classifier-specific factor grows only very slowly (the log function grows very slowly) as $\text{acc}(h|k)$ increases towards the perfect limit where $\text{acc}(h|k) = 1$. As predicted by Corollary 1, we observe in Fig. 2 that beyond the critical value $\epsilon = \epsilon_\infty(h|k) := \sqrt{2 \log(1/\text{err}(h|k))}/\sqrt{p}$, the adversarial accuracy $\text{acc}_\epsilon(h|k)$ decays at a Gaussian rate, and eventually $\text{acc}_\epsilon(h|k) \leq \text{err}(h|k)$ as soon as $\epsilon \geq 2\epsilon_\infty(h|k)$.

Comparing to the Gaussian special case (see section 2.5 below), we see that the curvature parameter $\sigma_k$ appearing in the theorems is an analogue to the natural noise-level in the problem. The flat case $X_k = \mathbb{R}^p$ with an $\ell_q$ threat model is particularly instructive. The critical values of $\epsilon$, namely $\epsilon_\infty(h|k)$ and $2\epsilon_\infty(h|k)$ beyond which the compromising conclusions of the Corollary 1 come into play is proportional to $\sigma_k/\sqrt{p}$.

Finally note that the $\ell_1$ threat model corresponding to $q = 1$ in Corollary 1, is a convex proxy for the “few-pixel” threat model which was investigated in (Su et al., 2017).
2.5. Some applications of the bounds

Consider the conditional data model of the form \( P_{X|k} \propto e^{-v_k(x)/v_k} dx \) on a \( d \)-dimensional Riemannian manifold \( \mathcal{X}_k \subseteq \mathcal{X} \) satisfying the Bakry-Émery curvature condition (Bakry & Émery, 1985)

\[
\text{Hess}_x(v_k) + \text{Ric}_x(\mathcal{X}) \geq (1/\sigma_k^2)I_p,
\]
for some \( \sigma_k > 0 \). Such a distribution is called log-concave. By Corollary 1.1 of (Otto & Villani, 2000) (and Corollary 3.2 of (Bobkov & Goetze, 1999)), \( P_{X|k} \) has the \( T_2(\sigma_k^2) \) property and therefore by Lemma 1, the BLOWUP(\( \sigma_k \)) for some \( \theta \) are \( \leq \) as in (Tsipras et al., 2018), with exactly the same constants in \( m \). By Corollary 1.1 of (Otto & Villani, 2000) (and Corollary 1 for flat space) applies.

The Holley-Stroock perturbation Theorem ensures that if \( P_{X|k} \propto e^{-v_k(x)-u_k(x)\|dx/\| \) where \( u_k \) is bounded, then Theorem 2 (and Corollary 1 for flat space) holds with the noise parameter \( \sigma_k \) degraded to \( \sigma_k := \sigma_k e^{\text{osc}(u_k)} \), where \( \text{osc}(u_k) := \sup_x u_k(x) - \inf_x u_k(x) \geq 0 \).

2.5.1. Log-concave on a Riemannian manifold

Consider a conditional data model of the form \( P_{X|k} \propto e^{-v_k(x)/v_k} dx \) on a \( d \)-dimensional Riemannian manifold \( \mathcal{X}_k \subseteq \mathcal{X} \) satisfying the Bakry-Émery curvature condition (Bakry & Émery, 1985)

\[
\text{Hess}_x(v_k) + \text{Ric}_x(\mathcal{X}) \geq (1/\sigma_k^2)I_p,
\]
for some \( \sigma_k > 0 \). Such a distribution is called log-concave. By Corollary 1.1 of (Otto & Villani, 2000) (and Corollary 3.2 of (Bobkov & Goetze, 1999)), \( P_{X|k} \) has the \( T_2(\sigma_k^2) \) property and therefore by Lemma 1, the BLOWUP(\( \sigma_k \)) for some \( \theta \) are \( \leq \) as in (Tsipras et al., 2018), with exactly the same constants in \( m \). By Corollary 1.1 of (Otto & Villani, 2000) (and Corollary 1 for flat space) applies.

The Holley-Stroock perturbation Theorem ensures that if \( P_{X|k} \propto e^{-v_k(x)-u_k(x)\|dx/\| \) where \( u_k \) is bounded, then Theorem 2 (and Corollary 1 for flat space) holds with the noise parameter \( \sigma_k \) degraded to \( \sigma_k := \sigma_k e^{\text{osc}(u_k)} \), where \( \text{osc}(u_k) := \sup_x u_k(x) - \inf_x u_k(x) \geq 0 \).

2.5.2. Elliptical Gaussian in Euclidean space

Consider the flat manifold \( \mathcal{X}_k = \mathbb{R}^p \) and multi-variate Gaussian distribution \( P_{X|k} \propto e^{-v_k(x)/v_k} dx \) thereupon, where \( v_k(x) = \frac{1}{2}(x - m_k)^T \Sigma_k^{-1}(x - m_k) \), for some vector \( m_k \in \mathbb{R}^p \) (called the mean) and positive-definite matrix \( \Sigma_k \) (called the covariance matrix) all of whose eigenvalues are \( \leq \sigma_k^2 \). A direct computation gives \( \text{Hess}(v_k) + \text{Ric}_x \geq 1/\sigma_k^2 + 0 = 1/\sigma_k^2 \) for all \( x \in \mathbb{R}^p \). So this is an instance of the above log-concave example, and so the same bounds hold. Thus we get an elliptical version (and therefore a strict generalization) of the basic “No Free Lunch” theorem in (Tsipras et al., 2018), with exactly the same constants in the bounds. These results are also confirmed empirically in section 4.1.

2.5.3. Uniform compact Riemannian manifold with positive Ricci curvature

Indeed the distribution \( dP_{X|k} = d\text{vol}(\mathcal{X}_k)/\text{vol}(\mathcal{X}_k) \) is log-concave on \( \mathcal{X}_k \) since it satisfies the Bakry-Émery curvature condition (17) with \( v_k = 0 \) and \( \sigma_k = 1/\sqrt{R_k} \) where \( R_k > 0 \) is the minimum of the Ricci curvature on \( \mathcal{X}_k \). Thus by section 2.5.1, it follows that our theorems hold. A prime example of such a manifold is a \( p \)-sphere of radius \( r > 0 \), with constant Ricci curvature \( (p - 1)/r \). For this example (Gilmer et al., 2018b) is an instance (more on this in section 3).

Application to the “Adversarial Spheres” problem. In the recent research “Adversarial Spheres” paper (Gilmer et al., 2018b), wherein the authors consider a 2-class problem on a so-called “concentric spheres” dataset. This problem can be described in our notation as: \( P_{X|+} = \) uniform distribution on \( p \)-dimensional sphere of radius \( r_+ \) and \( P_{X|-} = \) uniform distribution on \( p \)-dimensional sphere of radius \( r_- \). The classification problem is to decide which of the two concentric spheres a sampled point came from. The authors (see Theorem 5.1 of mentioned paper) show that for this problem, for large \( p \) and for any non-perfect classifier \( h \), the average \( \ell_2 \) distance between a point and the set of misclassified points is bounded as follows

\[
\ell_2(h) = O((\Phi^{-1}(\text{acc}(h|k)))/\sqrt{p}).
\]

We now show how to obtain the above bound via a direct application of our theorem 2. Indeed, since these spaces are clearly compact Riemannian manifolds with constant curvature \( (p - 1)/r_k^2 \), each \( P_{X|k} \) is and instances of 2.5.3, and so satisfies \( T_2 \left( \sqrt{r_k^{-2}} \right) \). Consequently, our Theorem 2 kicks-in and bound the average distance of sample points with true label \( k \in \{ \pm \} \), to the error set (set of misclassified samples):

- \( d_{\text{geo}}(h|k) \leq \frac{r_k}{\sqrt{p}} \left( \sqrt{2 \log(1/\text{err}(h|k))} + \sqrt{\pi/2} \right) \) for the geodesic threat model (induced by the “great circle” distance between points), and
- \( \ell_2(h|k) \leq \frac{r_k}{\sqrt{p}} \left( \sqrt{2 \log(1/\text{err}(h|k))} + \sqrt{\pi/2} \right) \) for the \( \ell_2 \) threat model. This follows from the previous inequality because the geodesic (aka great circle) distance between two points on a sphere is always larger than the euclidean \( \ell_2 \) distance between points.

To link more explicitly with the bound (18) proposed in (Gilmer et al., 2018b), one notes the following elementary (and very crude) approximation of Gaussian quantile function \( \Phi^{-1}(a) \approx \sqrt{2 \log(1/(1 - a))} \) for \( a \in [0, 1) \). Thus, \( \Phi^{-1}(1 - \text{err}(h|k))/\sqrt{p} \) and \( \sqrt{2 \log(1/\text{err}(h|k))/(p - 1)} \) are of the same order, for large \( p \). Consequently, our bounds can be seen as a strict generalization of the bounds in (Gilmer et al., 2018b).

2.5.4. Lipschitz pushforward of a \( T_2 \) distribution

Lemma 2.1 of (Djellout et al., 2004) ensures that if \( P_{X|k} \) is the pushforward via an \( L_k \)-Lipschitz map \( 0 \leq L_k < \infty \) \( \mathcal{X}_k \to \mathcal{X}_k \) between metric spaces (an assumption which is implicitly made when machine learning practitioners
model images using generative neural networks\(^1\), for example, of a distribution \(\mu_x\) which satisfies \(T_2(\hat{\sigma}_k^2)\) on \(Z_k\) for some \(\hat{\sigma}_k > 0\), then \(P_{X|k}\) satisfies \(T_2(L_k^2\hat{\sigma}_k^2)\) on \(X_k\), and so Theorem 2 (and Corollary 1 for flat space) holds with \(\sigma_k = L_k\hat{\sigma}_k\). This is precisely the data model assumed by (Fawzi et al., 2018a), with \(Z_k := \mathbb{R}^p\) and \(\mu_k = \mathcal{N}(0, \sigma I_p)\) for all \(k\).

3. Related works

There is now a rich literature trying to understand adversarial robustness. Just to name a few, let us mention (Tsipras et al., 2018; Schmidt et al., 2018; Bubeck et al., 2018; Gilmer et al., 2018b; Fawzi et al., 2018a; Mahloujifar et al., 2018; Sinha et al., 2017; Blanchet & Murthy, 2016; Mohajerin Esfahani & Kuhn, 2017). Below, we discuss a representative subset of these works, which is most relevant to our own contributions presented in this manuscript. These all use some kind of Gaussian isoperimetric inequality (Boucheron et al., 2013), and turn out to be very special cases of the general bounds presented in Theorem 2 and Corollary 1. See section 2.5 for a detailed discussion on generality our results.

Gaussian and Bernoulli models. We have already mentioned the work (Tsipras et al., 2018), which first showed that motivating problem presented in section 1.1, every classifier can be fooled with high probability. In a follow-up paper (Schmidt et al., 2018), the authors have also suggested that the sample complexity for robust generalization is much higher than for standard generalization. These observations are also strengthened by independent works of (Bubeck et al., 2018).

Adversarial spheres. The work which is most similar in flavor to ours is the recent “Adversarial Spheres” paper (Gilmer et al., 2018b), wherein the authors consider a 2-class problem on classifying two concentric spheres in \(\mathbb{R}^p\) of different radii. The authors showed that the distance of each point to the set of misclassified images is of order \(O(1/\sqrt{p})\). We discussed this work in detail in section 2.5 and showed that it follows directly from our Theorem 2.

Generative models. In (Fawzi et al., 2018a), the authors considered a scenario where data-generating process is via passing a multivariate Gaussian distribution through a Lipschitz continuous mapping \(g : \mathbb{R}^m \to \mathcal{X}\), called the generator. The authors then studied the per-sample robustness radius defined by \(r_X(x,k) := \inf\{||x' - x||_2 \text{ s.t.} x' \in \mathcal{X}, h(x') \neq k\}\). In the notation of our manuscript, this can be rewritten as \(r_X(x,k) := d_X(x,B(h,k))\), from which it is clear that \(r_X(x,k) \leq \epsilon\) iff \(x \in B(h,k)^c\). Using the basic Gaussian isoperimetric inequality (Boucheron et al., 2013), the authors then proceed to obtain bounds on the probability that the classifier changes its output on an \(\epsilon\)-perturbation of some point on manifold the data manifold, namely \(\text{acc}_{\text{switch}}(h) := 1 - \sum_k \pi_k \text{err}_{\text{switch}}(h|k)\), where \(\text{err}_{\text{switch}}(h|k) := P_{X|k}(C_{k\to \epsilon}) = \text{acc}(h|k) \text{err}_{\epsilon}(h|k)\) and \(C_{k\to \epsilon} := B(h,k)^c - B(h,k)\) is the annulus in Fig. 1. Our bounds in Theorem 2 and Corollary 1 can then be seen as generalizing the methods and bounds in (Fawzi et al., 2018a) to more general data distributions satisfying \(W_2\) transportation-cost inequalities \(T_2(c)\), with \(c > 0\).

Distributional robustness and regularization. On a completely different footing, (Blanchet & Murthy, 2016; Mohajerin Esfahani & Kuhn, 2017; Sinha et al., 2017) have linked distributional robustness to robust estimation theory from classical statistics and regularization. An interesting bi-product of these developments is that penalized regression problems like the square-root Lasso and sparse logistic regression have been recovered as distributional robust counterparts of the unregularized problems.

4. Experimental evaluation

We now present some empirical validation for our theoretical results.

4.1. Simulated data

The simulated data are discussed in section 1.1: \(Y \sim \text{Bern}(\{\pm 1\}), X|Y \sim \mathcal{N}(Y\eta,1^{\times p})\), with \(p = 1000\) where \(\eta\) is an SNR parameter which controls the difficulty of the problem. Here, The classifier \(h\) is a multi-layer perceptron with architecture \(1000 \to 200 \to 100 \to 2\) and ReLU activations. The results are are shown in Fig. 2 (Left). As predicted by our theorems, we observe that beyond the critical value \(\epsilon = \epsilon_\infty(h) := \sigma \sqrt{2\log(1/\text{err}(h))/p} = O(\sigma/\sqrt{p})\), where \(\text{err}(h) := 1 - \text{acc}(h)\), the adversarial accuracy \(\text{acc}_{\epsilon}(h)\) decays exponential fast, and passes below the horizontal line \(\text{err}(h)\) as soon as \(\epsilon \geq 2\epsilon_\infty(h)\).

4.2. Real data

Wondering whether the phase transition and bounds predicted by Theorem 2 and Corollary 2 holds for real data, we trained a deep feed-forward CNN (architecture: \(\text{Conv2d} \to \text{Conv2d} \to 320 \to 10\)) for classification on the MNIST dataset (LeCun & Cortes, 2010), a standard benchmark problem in supervised machine-learning. The results are shown in Fig. 2. This model attains a classification accuracy of 98% on held-out data. We consider the performance of the model on adversarially modified images according to the \(\ell_\infty\) threat model, at a given toler-
Limitations of Adversarial Robustness: Strong No Free Lunch Theorem

5. Concluding remarks

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. Our work uses powerful tools from geometric probability theory to generalize all the main impossibility results that have appeared in adversarial robustness literature. Moreover, our results would encourage one to conjecture that the modulus of concentration of probability distribution (e.g. in $T_2$ inequalities) on a manifold completely characterizes the adversarial or distributional robust accuracy in classification problems.

5.1. Redefine the rules of the game?

A limitation for adversarial robustness, as universal our strong No Free Lunch Theorem we have developed in this paper could indicate that the attack models currently being considered in the literature, namely additive perturbations measured in the $\ell_0$, $\ell_1$, $\ell_2$, $\ell_\infty$, etc. norms, and in which the attacker can make as many queries as they which, may be too lax. Just like in coding theory where a rethinking of the constraints on a channel leads to the Shannon limit to be improved, one could hope that a careful rethink of the constraints put on the adversarial attacker might alleviate the pessimistic effect of our impossibility results. As remarked in (Gilmer et al., 2018a), it is not even clear if the current existing attack models are most plausible.

5.2. Future directions

One could consider the following open questions, as natural continuation of our work:

- Extend Theorem 2 and Corollary 1 to more general data distributions.
- Study more complex threat models, e.g. small deformations.
- Fine grained analysis of sample complexity and complexity of hypotheses class, with respect to adversarial and distributional robustness. This question has been partially studied in (Schmidt et al., 2018; Bubeck et al., 2018) in the adversarial case, and (Sinha et al., 2017) in the distributional robust scenario.
- Study more general threat models. (Gilmer et al., 2018a) has argued that most of the proof-of-concept problems studied in theory papers might not be completely aligned with real security concerns faced by machine learning applications. It would be interesting to see how the theoretical bounds presented in our manuscript translate on real-world datasets, beyond the MNIST on which we showed some preliminary experimental results.
- Develop more geometric insights linking adversarial robustness and curvature of decision boundaries. This view was first introduced in (Fawzi et al., 2018b).

Acknowledgments. I would wish to thank Noureddine El Karoui for stimulating discussions; Alberto Bietti and Albert Thomas for their useful comments and remarks.
References


