Classifier-independent Lower-Bounds for Adversarial Robustness Error

Elvis Dohmatob
Criteo AI Lab

Abstract

We theoretically analyse the limits of robustness to test-time adversarial examples. Our work focuses on deriving bounds which uniformly apply to all classifiers (i.e all measurable functions from features to labels) for a given problem. Our contributions are two-fold. (1) We use optimal transport theory to derive variational formulae for the Bayes-optimal error a classifier can make on a given classification problem, subject to adversarial attacks. (2) We derive explicit lower-bounds on the Bayes-optimal error in the case of the popular distance-based attacks. These bounds are universal in the sense that they depend on the geometry of the class-conditional distributions of the data, but not on a particular classifier. Our results are in sharp contrast with the existing literature, wherein adversarial vulnerability of classifiers is derived as a consequence of nonzero ordinary test error.

1 Introduction

1.1 Context

Despite their popularization, machine-learning powered systems (assisted-driving, natural language processing, facial recognition, etc.) are not likely to be deployed for critical tasks which require a stringent error margin, in a closed-loop regime any time soon. One of the main blockers which has been identified by practitioners and ML researchers alike is the phenomenon of *adversarial examples* (Szegedy et al., 2013). There is now an arms race (Athalye et al., 2018) between adversarial attack developers and defenders, and there is some speculation that adversarial examples in machine-learning might simply be inevitable.

In a nutshell an adversarial (evasion) attack operates as follows. A classifier is trained and deployed (e.g the road traffic sign recognition sub-system on an AI-assisted car). At test / inference time, an attacker (aka adversary) may submit queries to the classifier by sampling a data point x with true label y, and modifying it $x \to x^{\text{adv}}$ according to a prescribed threat model. For example, modifying a few pixels on a road traffic sign (Su et al., 2017), modifying intensity of pixels by a limited amount determined by a prescribed tolerance level variance per pixel, etc. The goal of the attacker is to fool the classifier into classifying x^{adv} with a label different from y. A robust classifier tries to limit this failure mode, for a prescribed attack model.

In this manuscript, we establish universal lower-bounds on the test error any classifier can attain under adversarial attacks.

1.2 Overview of related works

Questions around adversarial examples and fundamental limits of defense mechanisms, are an active area of research in machine-learning, with a large body of scientific literature.

Classifier-dependent lower-bounds. There is now a rich array of works which study adversarial examples as a natural consequence of nonzero test error. In particular, let us mention (Tsipras et al., 2018), (Schmidt et al., 2018), (Shafahi et al., 2018), (Gilmer et al., 2018), (Mahloujifar et al., 2018), (Dohmatob, 2019). These all use a form of the *Gaussian isoperimetric inequality* (Boucheron et al., 2013): in these theories, adversarial examples exist as a consequence of ordinary test-error in high-dimensional problems with concentrated class-conditional distributions. On such problems, for a classifier which does not attain 100% on clean test examples (which is likely to be the case in practice), every test example will be close to a misclassified example, i.e can be misclassified by adding a small perturbation. Still using Gaussian isoperimetry, (Gilmer et al., 2019) has studied the relationship between robustness to adversaries and robustness to random noise. The authors argued that adversarial examples are a natural consequence of errors made by a classifier on noisy images.

One should also mention some works which exploit curvature of the decision boundary of neural networks to exhibit the existence of vectors in low-dimensional subspaces, which when added to every example in a target class, can fool a classifier on a fraction of the samples (Moosavi-

. Correspondence to: Elvis Dohmatob <e.dohmatob@criteo.com>.

Preprint (December 2020)

Dezfooli et al., 2017; Moosavi-Dezfooli et al., 2017).

Universal / classifier-independent bounds. To our knowledge, (Fawzi et al., 2018; Mahloujifar et al., 2018; Bhagoji et al., 2019) are the only works to derive universal / classifier-independent lower-bounds for adversarial robustness. Particularly, (Bhagoji et al., 2019) and (Pydi & Jog, 2019), are the most related to ours. In (Bhagoji et al., 2019), the authors considered general adversarial attacks (i.e beyond distance-based models of attack), and show that Bayes-optimal error for the resulting classification problem under such adversaries is linked to a certain transport distance between the class-conditional distributions (see our Theorem 3.1 for a generalization of the result). This result is singularly different from the previous literature as it applies even to classifiers which have zero test-error in the normal / non-adversarial sense. Thus, there adversarial examples that exist solely as a consequence of the geometry of the problem. The results in section 3 of our paper are strict extension of the bounds in (Bhagoji et al., 2019). The main idea in (Bhagoji et al., 2019; Pydi & Jog, 2019) is to construct an optimal-transport metric (w.r.t a certain binary cost-function induced by the attack), and then use Kantorovich-Rubenstein duality to relate this metric to the infimal adversarial error a classifier can attain under the adversarial attack. 3.

Finally, one should mention (Cranko et al., 2019) which studies vulnerability of hypothesis classes in connection to loss functions used.

1.3 Summary of our main contributions

Our main contributions can be summarized as follows.

• In section 3 (after developing some background material in section 2), we use optimal transport theory to derive variational formulae for the Bayes-optimal error (aka smallest possible test error) of a classifier under adversarial attack, as a function of the "budget" of the attacker, These formulae suggest that instead of doing adversarial training, practitioners should rather do normal training on adversarially augmented data. Incidentally, this is a well-known trick to boost up the adversarial robustness of classifiers to known attacks, and is usually used in practice under the umbrella name of "adversarial data-augmentation". See (Yang et al., 2019), for example. In our manuscript, this principle appears as a natural consequence of our variational formulae.

• For the special case of distance-based attacks, we proceed in 4 to *(1)* Establish universal lower-bounds on the adversarial Bayes-optimal error. These bounds are a consequence of concentration properties of light-tailed class-conditional distributions of the features (e.g sub-Gaussianity, etc.). *(2)* Establish universal bounds under more general moment constraints conditions on the class-conditional distributions (e.g existence of covariance matrices for the class-conditional distributions of the features).

2 Preliminaries

2.1 Classification framework

All through this manuscript, the feature space will be denoted $\mathcal{X}$. Except otherwise stated, $\mathcal{X}$ will be an abstract set with no topological or geometric structure at all. The *label (aka classification target)* is a random variable Y with values in $\mathcal{Y} = \{1, 2\}$, and random variable X called the *features*, with values in $\mathcal{X}$. We only consider binary classification problems in this work. The goal is to predict Y given X. This corresponds to prescribing a measurable function $h : \mathcal{X} \to \{1, 2\}$, called a *classifier*. The joint distribution $P_{X,Y}$ of (X, Y) is unkown. The goal of learning is to find a classifier h (e.g a deep neural net) such that $h(X) = Y$ as often as possible, possibly under additional constraints.

We will only consider binary-classification problems. Multi-class problems can be considered in one-versus-all fashion. For each label $k \in \{1, 2\}$, we define the (unnormalized) probability measure P^k on the feature space $\mathcal{X}$ by

$$\begin{aligned} P^k(A) &:= \mathbb{P}(X \in A, Y = k) \\ &= \mathbb{P}(Y = k)\mathbb{P}(X \in A | Y = k) = \pi_k P_k(A), \end{aligned} \tag{1}$$

for every measurable $A \subseteq X$. Thus, P^k is an unnormalized probability distribution on the feature space $\mathcal{X}$ which integrates to $\pi_k := \mathbb{P}(Y = k)$, and $P_k := P_{X|Y=k}$ is the probability distribution of X conditioned on the label being k. The classification problem is therefore entirely captured by the pair $P = (P^1, P^2)$, also called a *binary experiment* (Reid & Williamson, 2011).

2.2 An abstraction for adversarial attacks

In full generality, an *adversarial attack model* on the feature space $\mathcal{X}$ is any subset $\Omega \subseteq \mathcal{X}^2$. Given points $x', x \in \mathcal{X}$, we call x' an *adversarial example* of x if $(x, x') \in \Omega$. The subset $\mathsf{diag}(\mathcal{X}^2) := \{(x, x) \mid x \in \mathcal{X}\}$ corresponds to classical / standard classification theory where there is no adversary. A nontrivial example is the case of so-called distance-based attacks, where d is a metric on $\mathcal{X}$ and the attack model is $\Omega = D_\varepsilon$, where

$$D_\varepsilon = \{(x, x') \in \mathcal{X}^2 \mid d(x, x') \leq \varepsilon\}, \tag{2}$$

with $\varepsilon \geq 0$ being the budget of the attacker. These include the well-known ℓ_p-norm attacks in finite-dimensional euclidean spaces usually studied in the literature (e.g (Szegedy et al., 2013; Tsipras et al., 2018; Schmidt et al., 2018; Shafahi et al., 2018; Gilmer et al., 2018)).

Another instance of our general formulation is when $\Omega = \mathcal{A}^\times$, where $\mathcal{A}^\times := \{(x, x') \in \mathcal{X}^2 \mid \mathcal{A}_x \cap \mathcal{A}_{x'} \neq \emptyset\}$, for

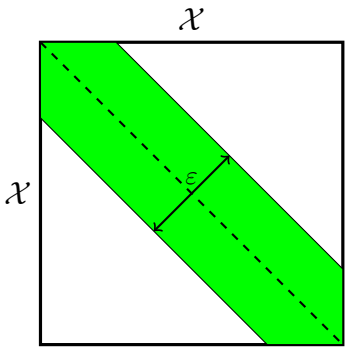

Figure 1. Showing a generic distance-based attack model. The green region corresponds to the set $D_\varepsilon \subseteq \mathcal{X}^2$ defined in (2). An attacker is allowed to swap any point $x \in \mathcal{X}$ with another point $x' \in \mathcal{X}$, provided the pair (x, x') lies in the green region.

a system $(\mathcal{A}_x)_{x\in\mathcal{X}}$ of subsets of $\mathcal{X}$. This framework is already much more general than the distance-based framework (which is the default setting in the literature), and corresponds to the setting considered in (Bhagoji et al., 2019). Working at this level of generallity allows the possibility to study general attacks like pixel-erasure attacks (Su et al., 2017), for example, which cannot be metrically expressed.

A *type$-\Omega$ adversarial attacker* on the feature space $\mathcal{X}$ is then a measurable mapping $a : \mathcal{X} \to \mathcal{X}$, such that $(x, a(x)) \in \Omega$ for all $x \in \mathcal{X}$. For example, in the distance-based attacks, such an attack corresponds to a measurable selection for every $x \in \mathcal{X}$, of some $x' = a(x) \in \mathcal{X}$ with $d(x, x') \leq \varepsilon$.

Adversarial error of a classifier. For any input example $x \in \mathcal{X}$, denote $\Omega(x) := \{x' \in \mathcal{X} \mid (x, x') \in \Omega\}$. Given a classifier $h : \mathcal{X} \to \{1, 2\}$, and a label $k \in \{1, 2\}$, define

$$\Omega^{h,k} := \{x \in \mathcal{X} \mid \exists x' \in \Omega(x) \text{ with } h(x') \neq k\}, \tag{3}$$

the set of examples with a "neighbor" whose predicted label is different from k. Conditioned on the event $Y = k$, the "size" of the set $\Omega^{h,k}$ is the adversarial error / risk of the classifier h, on the class k. This will be made precise in the passage. The *adversarial error / risk* of the classifier h under type-Ω adversarial attacks, is defined by

$$\begin{aligned}\mathsf{R}_\Omega(h; P^1, P^2) &:= \mathbb{P}_{X,Y}(\exists x' \in \Omega(X) \mid h(x') \neq Y) \\ &= \sum_{k=1}^{2} P^k(\Omega^{h,k}).\end{aligned} \tag{4}$$

Thus, $\mathsf{R}_\Omega(h; P^1, P^2)$ is the least possible classification error suffered by h under type-Ω attacks. For the special case of distance-based attacks with budget ε, where the attack model is $\Omega = D_\varepsilon$ (defined in Eq. (2)), we will simply write $\mathsf{R}^*_\varepsilon(P^1, P^2)$ in lieu of $\mathsf{R}_{D_\varepsilon}(P^1, P^2)$, that is

$$\begin{aligned}\mathsf{R}_\varepsilon(h; P^1, P^2) &:= \mathsf{R}^*_{D_\varepsilon}(P^1, P^2) := \sum_{k=1}^{2} P^k(D^{h,k}_\varepsilon) \\ &= \sum_{k=1}^{2} P^k(\{x \in \mathcal{X} \mid \exists x' \in \mathsf{Ball}(x; \varepsilon),\ h(x') \neq k\}).\end{aligned} \tag{5}$$

Adversarial Bayes-optimal error. The *adversarial Bayes-optimal error* for type-Ω attacks, denoted $\mathsf{R}^*_\Omega(P^1, P^2)$, is defined by

$$\mathsf{R}^*_\Omega(P^1, P^2) := \inf_h \mathsf{R}_\Omega(h; P^1, P^2), \tag{6}$$

where the infimum is taken over all measurable functions $h : \mathcal{X} \to \{1, 2\}$, i.e over all classifiers. Econometrically, adversarial Bayes-optimal error $\mathsf{R}^*_\Omega(P^1, P^2)$ corresponds to the maximal payoff of a type-Ω adversarial attacker who tries to uniformly "blunt" all classifiers at the task of solving the classification problem (P^1, P^2). In general, if $\mathcal{H}$ is a set of classifiers $\mathcal{X} \to \{1, 2\}$ (i.e a hypothesis class), then we define the restricted adversarial Bayes-optimal error denoted $\mathsf{R}^*_\Omega(\mathcal{H}; P^1, P^2)$ by

$$\mathsf{R}^*_\Omega(\mathcal{H}; P^1, P^2) := \inf_{h\in\mathcal{H}} \mathsf{R}_\Omega(h; P^1, P^2). \tag{7}$$

Of course, it holds that $\mathsf{R}^*\,\Omega(P^1, P^2) \leq \mathsf{R}^*_\Omega(\mathcal{H}; P^1, P^2) \leq \mathsf{R}_\Omega(P^1, P^2)$ for all $h \in \mathcal{H}$.

2.3 A universal No Free Lunch Theorem for adversarial robustness

For our contributions, we start with the following elementary but powerful theorem motivates which shows a trade-off between the normal classification error due to normal examples, on the one hand; and the classification error due only to adversarial examples, on the other hand. The Bayes-optimal adversarial classification error $\mathsf{R}^*_\Omega(P^1, P^2)$ then appears as a critical point below which a classifier faces diminishes returns: decrease in normal error can only be at the expense of increased "pure" adversarial error. This theorem motivates our study of the universal quantity $\mathsf{R}^*_\Omega(P^1, P^2)$ in the rest of our manuscript. Viz

Theorem 2.1 (Universal No Free Lunch Theorem for adversarial robustness)**.** *Consider a binary-classification problem (P^1, P^2) on $\mathcal{X}$. Let $\mathcal{H}$ be a set of classifiers $\mathcal{X} \to \{1, 2\}$ (i.e a hypothesis class on $\mathcal{X}$) and let $\Omega \subseteq \mathcal{X}^2$ be an attack model on $\mathcal{X}$. For any classifier $h \in \mathcal{H}$, we have the lower-bound*

$$\underbrace{\mathsf{R}_{\Omega\setminus\mathsf{diag}(\mathcal{X}^2)}(h; P^1, P^2)}_{\text{"pure" adversarial error of } h} \geq \underbrace{\mathsf{R}^*_\Omega(\mathcal{H}; P^1, P^2)}_{\text{global offset}} - \underbrace{\mathsf{R}(h; P^1, P^2)}_{\text{normal error of } h},$$

where $\mathsf{diag}(\mathcal{X}^2) := \{(x, x) \mid x \in \mathcal{X}\}$. *In other words, the error of h due to adversarial examples alone increases*

as soon as its normal error decreases beyond the universal level $\mathsf{R}^*_\Omega(\mathcal{H}; P^1, P^2)$. *This is a* No free lunch theorem, *which trades between the normal error and the "pure" adversarial error of any classifier.*

Proof. We may partition the attack model Ω as $\Omega = (\Omega \cap \mathsf{diag}(\mathcal{X}^2)) \cup (\Omega \backslash \mathsf{diag}(\mathcal{X}^2))$. Noting that $\mathsf{R}^*_\Omega(\mathcal{H}; P^1, P^2) := \inf_{h' \in \mathcal{H}} \mathsf{R}_\Omega(h'; P^1, P^2) \leq \mathsf{R}_\Omega(h; P^1, P^2)$, direct computation gives

$$\begin{aligned} &\mathsf{R}^*_\Omega(\mathcal{H}; P^1, P^2) \leq \mathsf{R}_\Omega(h; P^1, P^2) \\ &= \mathsf{R}_{\Omega \cap \mathsf{diag}(\mathcal{X}^2)}(h; P^1, P^2) + \mathsf{R}_{\Omega \backslash \mathsf{diag}(\mathcal{X}^2)}(h; P^1, P^2) \quad (8) \\ &\leq \mathsf{R}_{\mathsf{diag}(\mathcal{X}^2)}(h; P^1, P^2) + \mathsf{R}_{\Omega \backslash \mathsf{diag}(\mathcal{X}^2)}(h; P^1, P^2), \end{aligned}$$

where the inequality in the third line is because $\Omega \cap \mathsf{diag}(\mathcal{X}^2) \subseteq \mathsf{diag}(\mathcal{X}^2)$. Finally, noting that $\mathsf{R}_{\mathsf{diag}(\mathcal{X}^2)}(h; P^1, P^2) = \mathsf{R}(h; P^1, P^2)$, the normal error of h, immediately completes the proof. □

In view of the above theorem, accuracy might be at odds with adversarial robustness as first hypothesized in (Tsipras et al., 2018). We will return to this matter later in section ... when we give a one-line proof to the main theorem of (Tsipras et al., 2018), namely their Theorem 2.1.

3 Optimal transport characterization of adversarial vulnerability

3.1 Adversarial attacks as transport plans

Let the feature space $\mathcal{X}$ have the structure of a topological space, and let $\Omega \subseteq \mathcal{X}^2$ be an attack model. As before, we also demand that Ω be closed in the product topology on $\mathcal{X}^2$. Consider the binary cost-function $c_\Omega : \mathcal{X}^2 \to \{0, 1\}$ defined by

$$c_\Omega(x, x') := \begin{cases} 0, & \text{if } (x, x') \in \Omega, \\ 1, & \text{else.} \end{cases} \quad (9)$$

This cost-function is special in that, for every $(x, x') \in \Omega$, one can transport x to x' without incurring any cost at all. If x and x' happen to belong to different classes, then an adversarial attack which replaces x with x' will be perfectly undetectable. As in (Bhagoji et al., 2019), we start with a variational formula for measuring the cost of a type-Ω for the task of "blunting" the Bayes-optimal classifier for the classification problem (P^1, P^2).

Definition 3.1 (Adversarial total-variation)**.** *Let* $\mathsf{OT}_\Omega(P^1, P^2)$ *be the optimal transport distance between* P^1 *and* P^2 *w.r.t to the ground cost* c_Ω *defined in Eq.* (9)*, i.e*

$$\begin{aligned} \mathsf{OT}_\Omega(P^1, P^2) &:= \inf_{\gamma \in \Pi(P^1, P^2)} \int_{\mathcal{X}^2} c_\Omega(x_1, x_2) \, \mathrm{d}\gamma(x_1, x_2) \\ &= \inf_{\gamma \in \Pi(P^1, P^2)} \mathbb{E}_\gamma[c_\Omega(X_1, X_2)], \end{aligned} \quad (10)$$

where $\Pi(P^1, P^2)$ *is the set of all couplings of* P^1 *and* P^2*, i.e the set of all measures on* $\mathcal{X}^2$ *with marginals* P^1 *and* P^2*, and* (X_1, X_2) *is a pair of r.vs on* $\mathcal{X}$ *with joint distribution* γ*.*

If γ is a coupling of P^1 and P^2 and $(X_1, X_2) \sim \gamma$, with abuse of language we shall also refer to (X_1, X_2) as a coupling of P^1 and P^2.

Lemma 3.1. *The ground-cost function* c_Ω *is lower-semicontinuous (l.s.c).*

Proof. In fact, we proof that c_Ω is l.s.c iff Ω is closed in Ω. Recall that the definition of lower-semicontinuity c_Ω is that the set $S_t := \{(x, x') \in \mathcal{X}^2 \mid c_\Omega(x, x') \leq t\}$ is closed in $\mathcal{X}^2$ for every $t \in \mathbb{R}$. A simple calculation reveals that

$$S_t = \begin{cases} \emptyset, & \text{if } t < 0, \\ \Omega, & \text{if } 0 \leq t < 1, \\ \mathcal{X}^2, & \text{if } t \geq 1. \end{cases}$$

Thus, S_t is closed in $\mathcal{X}^2$ $\forall t \in \mathbb{R}$ iff Ω is closed in $\mathcal{X}^2$. □

Thus, $\mathsf{OT}_\Omega(\cdot, \cdot)$ defines a distance over measures on the feature space $\mathcal{X}$. In the particular case of distance-based attacks, we have $\Omega = D_\varepsilon$ as defined in Eq. (2), and formula (10) can be equivalently written as

$$\mathsf{OT}_{D_\varepsilon} = \mathsf{TV}_\varepsilon(P^1, P^2) := \inf_{(X_1, X_2)} \mathbb{P}(d(X_1, X_2) > \varepsilon), \quad (11)$$

where the infimum is taken over all couplings (X_1, X_2) of P^1 and P^2. The joint distribution γ_ε of (X_1, X_2) is then an optimal adversarial attack plan for the classification problem (P^1, P^2). Note that the case $\varepsilon = 0$ conveniently corresponds to the usual definition of total-variation, namely

$$\begin{aligned} \mathsf{TV}(P^1, P^2) : &= \sup_{A \subseteq \mathcal{X} \text{ measurable}} P^1(A) - P^2(A) \\ &= \inf_{(X_1, X_2)} \mathbb{P}(X_1 \neq X_2), \end{aligned} \quad (12)$$

The RHS of the above formula is usually referred to as *Strassen's* formula for total-variation.

Coincidentally, the metric TV_ε in (11) has been studied in context of statistical testing, under the name "perturbed variation" (Harel & Mannor, 2015) as robust replace for usual total-variation. Moreover, the authors proposed an efficient algorithm for computing both the optimal plan γ_ε as a maximal graph matching in a bipartite graph. In has also been studied in (Yang et al., 2019) in the context of adversarial attacks.

Link to classical theory of classification. It is well-known (Reid & Williamson, 2011) in standard classification theory that the Bayes-optimal error is exactly equal to

$$\mathsf{R}^*(P^1, P^2) := \frac{1}{2} - \mathsf{TV}(P^1, P^2). \tag{13}$$

Thus, one might expect that the adversarial total-variation metric $\mathsf{TV}_\Omega(\cdot,\cdot)$ defined in Eq. (10) would play a role in control of the adversarial Bayes-optimal error $\mathsf{R}^*_\Omega(\cdot,\cdot)$ (defined in Eq. (6)) which is similar to the role played by ordinary total-variation $\mathsf{TV}(\cdot,\cdot)$ (defined in Eq. (12)) plays in formula (13) for the classical / standard Bayes-optimal error. This is indeed the case.

Remark 1. The adversarial Bayes-optimal error $\mathsf{R}^*_\Omega(P^1, P^2)$ under type-Ω adversarial attacks should not be confused with the adversarial error of the standard / classical Bayes-optimal classifier $h^* := \operatorname{argmin}_h \mathsf{R}(P^1, P^2)$ for the unattacked classification problem. In fact $\mathsf{R}^*_\Omega(P^1, P^2) \leq \mathsf{R}^*_\Omega(h^*; P^1, P^2)$, and we can construct explicit scenarios in which the inequality is strict. For example, consider a one-dimensional binary-classification problem whose class-conditional distributions are gaussians with different means and same variance, under the distance-based attack model $\Omega = D_\varepsilon := \{(x, x') \in \mathbb{R}^2 \mid |x - x'| \leq \varepsilon\}$.

Theorem 3.1 (Extension of Theorem 1 of (Bhagoji et al., 2019))**.** *For any attack model Ω on the feature space $\mathcal{X}$, the adversarial Bayes-optimal error under type-Ω attacks satisfies $\mathsf{R}^*_\Omega(P^1, P^2) \geq \frac{1}{2} - \mathsf{OT}_\Omega(P^1, P^2)$.*

Note that the reverse inequality does not hold in general. A remarkable exception is the case of distance-based attacks with a distance d that turns the feature space $\mathcal{X}$ into a *complete separable* metric space with the *midpoint property*[1]. Examples of such spaces include complete riemannian manifolds and any closed convex subset of a separable Banach space. We shall return to such spaces in section 3.3.

3.2 Characterizing the adversarial error via optimal transport

Henceforth, assume the feature space $\mathcal{X}$ is *Polish* (i.e $\mathcal{X}$ is metrizable, complete, and separable). Finally, given a subset $U \subseteq \mathcal{X}$, define its Ω-closure $\overline{U}_\Omega$ by

$$\overline{U}_\Omega := \{x \in \mathcal{X} \mid (x, x') \in \Omega \text{ for some } x' \in U\}. \tag{14}$$

In the case of metric attacks where $\Omega = D_\varepsilon := \{(x, x') \in \mathcal{X}^2 \mid d(x, x') \leq \varepsilon\}$, we have $\overline{U}_\Omega = U^\varepsilon$, where U^ε is the ε-neighborhood of U defined by

$$U^\varepsilon := \{x \in \mathcal{X} \mid d(x, x') \leq \varepsilon \text{ for some } x' \in U\}. \tag{15}$$

[1]That is, for every $x, x' \in \mathcal{X}$, there exists $z \in \mathcal{X}$ such that $d(x, z) = d(x', z) = d(x, x')/2$.

The following theorem is a direct application of *Strassen's Marriage Theorem* (see Theorem 1.27 of (Villani, 2003)), and is as a first simplification of the complicated distance OT_Ω that appears in Theorem 3.1. For distance-based attacks has been, a special case of our result has also been obtained in (Pydi & Jog, 2019).

Theorem 3.2. *Let Ω be an attack model on $\mathcal{X}$. Then we have the identity*

$$\mathsf{OT}_\Omega(P^1, P^2) = \sup_{U \subseteq \mathcal{X} \text{ closed}} P^1(U) - P^2(\overline{U}_\Omega). \tag{16}$$

In particular, for distance-based attacks we have

$$\mathsf{OT}_\varepsilon(P^1, P^2) = \sup_{U \subseteq \mathcal{X} \text{ closed}} P^2(U) - P^1(U^\varepsilon). \tag{17}$$

We now present a lemma which allows us to rewrite the optimal transport distance OT_Ω as a linear program over partial transport plans will be one of the main ingredients in the proof of Thm 3.3 below. The lemma is important in its own right.

Lemma 3.2. *Let Ω be an attack model on the feature space $\mathcal{X}$. Then, we have the formula*

$$\mathsf{TV}_\Omega(P^1, P^2) = \inf_{\gamma \in \Pi_\leq(P^1, P^2),\ \mathsf{supp}(\gamma) \subseteq \Omega} 1 - \gamma(\mathcal{X}^2), \tag{18}$$

where $\Pi_\leq(P^1, P^2)$ is the set of partial couplings of P^1 and P^2, i.e Borel measures on $\mathcal{X}^2$ whose marginals are dominated by the P^k's.

Proof. Define the quantity

$$E(P^1, P^2) := \inf_{\gamma \in \Pi_\leq(P^1, P^2),\ \mathsf{supp}(\gamma) \subseteq \Omega} 1 - \gamma(\mathcal{X}^2),$$

where $\Pi_\leq$ denotes the set of partial transport plans, i.e. probabilities on $\mathcal{X}^2$ with marginals smaller than P^1 and P^2 respectively. First, let us show that $\mathrm{OT}_\Omega = E$. Let $\gamma \in \Pi(P^1, P^2)$ and let $\tilde{\gamma}$ be its restriction to ε'. Then $\tilde{\gamma}$ is feasible for E and it holds $\gamma(\Omega) = 1 - \tilde{\gamma}(\mathcal{X}^2)$ so $E \leq \mathrm{OT}_\Omega$. Conversely, let γ be feasible for E and consider any $\tilde{\gamma} \in \Pi(P^1 - \mathsf{proj}^1_\# \gamma, P^2 - \mathsf{proj}^2_\# \gamma)$. Then $\gamma + \tilde{\gamma}$ is feasible for OT_Ω and $(\gamma + \tilde{\gamma})(\Omega) = \tilde{\gamma}(\Omega) \leq \tilde{\gamma}(\mathcal{X}^2) = 1 - \gamma(\mathcal{X}^2)$. So, $\mathrm{OT}_\Omega \leq E$ and thus $\mathrm{OT}_\Omega = E$. □

3.3 Adversarial couplings

We now turn to distance-based attacks and refine representation presented in the previous lemma.

Definition 3.2 (Metric and pseudo-metric spaces)**.** *A mapping $d : \mathcal{X}^2 \to \mathcal{X}$ is called a pseudo-metric on $\mathcal{X}$ iff for all $x, x', z \in \mathcal{X}$, the following hold:*

- ***Reflexivity:*** $d(x, x) = 0$.

- ***Symmetry:*** $d(x, x') = d(x', x)$.
- ***Triangle inequality:*** $d(x, x') \leq d(x, z) + d(z, x')$.

The pair $(\mathcal{X}, d)$ is then called a pseudo-metric space. If in addition, $d(x, x') = 0 \implies x = x'$, then we say d is a metric (or distance) on $\mathcal{X}$, and the pair $(\mathcal{X}, d)$ is called a metric space.

Recall that a (pseudo)metric space is said to have the midpoint property if for every pair of points z and z', there is a point $\eta(z, z')$ in the space which seats exactly halfway between them. Examples of such spaces include normed vector-spaces and riemannian manifolds. For our next result, it will be important to be able to select the midpoint $\eta(z, z')$ in a measurable manner almost-everywhere.

Condition 3.1 (Measurable Midpoint (MM) property)**.** *A metric space $\mathcal{Z} = (\mathcal{Z}, d)$ is said to satify the measurable midpoint (MM) property if for every Borel measure Q on $\mathcal{Z}^2$ there exists a Q-measurable map $\eta : \mathcal{Z}^2 \to \mathcal{Z}$ such that $d(z, \eta(z, z')) = d(z', \eta(z, z')) = d(z, z')/2$ for all $z, z' \in \mathcal{Z}$.*

The feature space for most problems in machine learning together with the distances usually used in adversarial attacks, satisfies the measurable midpoint property 3.1. Indeed, generic examples of metric spaces which satisfy this condition include:cHilbert spaces; closed convex subsets of Banach spaces; complete riemannian manifolds (equipped with the geodesic distance); complete separable metric spaces with the midpoint property.

In fact, in the first two examples, the midpoint mapping η can be chosen to be continuous everywhere. The last example, which can be proved via the classical *Kuratowski-Ryll-Nardzewski measurable selection theorem*, is the most general and most remarkable, and deserves an explicit restatement.

Lemma 3.3. *Every complete separable metric space which has the midpoint property also has the measurable midpoint property.*

The following theorem, which is proved in the appendix (as are all the other theorems in this manuscript), is one of our main results.

Theorem 3.3 (Adversarially augmented data, a proxy for adversarial robustness)**.** *Consider a classification problem (P^1, P^2). Suppose d is a distance on the feature space $\mathcal{X}$ with the MM property 3.1, and consider the distance-based attack model $D_\varepsilon := \{(x, x') \in \mathcal{X}^2 \mid d(x, x') \leq \varepsilon\}$. Recall the definition of $\mathsf{TV}_\varepsilon(P^1, P^2)$ from Eq.* (11)*. Define*

$$\begin{aligned} \widetilde{\mathsf{TV}}_\varepsilon(P^1, P^2) &:= \inf_{\gamma_1, \gamma_2} \mathsf{TV}(\mathsf{proj}^2{}_\#\gamma_1, \mathsf{proj}^1{}_\#\gamma_2), \\ \widetilde{\widetilde{\mathsf{TV}}}_\varepsilon(P^1, P^2) &:= \inf_{a_1, a_2 \ \textit{type-}D_{\varepsilon/2}} \mathsf{TV}(a_{1\#}P^2, a_{2\#}P^1), \end{aligned} \tag{19}$$

where "#" denotes pushfoward of measures and the 1st inf. is taken over all pairs of distributions (γ_1, γ_2) on $\mathcal{X}^2$ concentrated on $D_{\varepsilon/2}$ such that $\mathsf{proj}^1{}_\#\gamma_1 = P^2$ and $\mathsf{proj}^2{}_\#\gamma_2 = P^1$. It holds that

$$\mathsf{TV}_\varepsilon(P^1, P^2) = \widetilde{\mathsf{TV}}_\varepsilon(P^1, P^2) \leq \widetilde{\widetilde{\mathsf{TV}}}_\varepsilon(P^1, P^2), \tag{20}$$

and there is equality if P^1 and P^2 have densities w.r.t the Borel measure on $\mathcal{X}$.

Consequently, we have the following lower-bound for the adversarial Bayes-optimal error:

$$\begin{aligned} \mathsf{R}^*_\varepsilon(P^1, P^2) &\geq \frac{1}{2} - \mathsf{TV}_\varepsilon(P^1, P^2) \geq \frac{1}{2} - \widetilde{\mathsf{TV}}_\varepsilon(P^1, P^2) \\ &\geq \frac{1}{2} - \widetilde{\widetilde{\mathsf{TV}}}_\varepsilon(P^1, P^2). \end{aligned} \tag{21}$$

3.4 Case study: (separable) Banach spaces

Theorem 3.3 has several important consequences, which will be heavily explored in the sequel. A particularly simple consequence is the Consider the special case where $\mathcal{X} = (\mathcal{X}, \|\cdot\|)$, a separable Banach space Given a point $z \in \mathcal{X}$, let $P^1 + z$ be the translation of P^1 by z. For $z, z' \in \mathsf{Ball}_\mathcal{X}(\varepsilon/2)$, consider the type-$D_{\varepsilon/2}$ distance-based attacks $a_1^{z,z'}, a_2^{z,z'} : \mathcal{X} \to \mathcal{X}$ defined by $a_1^{z,z'}(x) = x - z$ and $a_2^{z,z'}(x) = x + z'$. One computes

$$\begin{aligned} \mathsf{TV}_\varepsilon(P^1, P^2) &:= \inf_{a_1, a_2 \ \textit{type-}D_{\varepsilon/2}} \mathsf{TV}(a_{1\#}P^1, a_{2\#}P^2) \\ &\leq \inf_{\|z\| \leq \varepsilon/2, \ \|z'\| \leq \varepsilon/2} \mathsf{TV}(a_1^{z,z'}{}_\#P^2, a_2^{z,z'}{}_\#P^1) \\ &= \inf_{\|z\| \leq \varepsilon/2, \ \|z'\| \leq \varepsilon/2} \mathsf{TV}(P^1 - z, P^2 + z') \\ &\leq \inf_{\|z\| \leq \varepsilon} \mathsf{TV}(P^1, P^2 + z), \end{aligned}$$

where $P^2 + z$ is the translation of distribution P^2 by the vector z. Note that in the above upper bound, the LHS can be made very concrete in case the distributions are prototypical (e.g multivariate Gaussians with same covariance matrix; etc.). Thus we have the following result

Corollary 3.1. *Let the feature space $\mathcal{X}$ be a normed vector space and consider a distance-based attack model $D_\varepsilon = \{(x, x') \in \mathcal{X}^2 \mid \|x' - x\| \leq \varepsilon\}$. Then, it holds that*

$$\mathsf{R}^*_\varepsilon(P^1, P^2) \geq \frac{1}{2} - \sup_{\|z\| \leq \varepsilon} \mathsf{TV}(P^1, P^2 + z). \tag{22}$$

A solution in z^* to optimization problem in the RHS of (22) would be a (*doubly*) universal adversarial perturbation: a single fixed small vector which fools all classifiers on proportion of test samples. Such a phenomenon has been reported in (Moosavi-Dezfooli et al., 2017).

Application: multivariate gaussians. As a concrete application of Corollary 3.1, consider the following problem inspired by (Tsipras et al., 2018) were the classification target is uniformly distributed on $\{1, 2\}$ and the class-conditional distribution of the features are multivariate Gaussians with the same covariance matrix for both classes. More formerly, the joint distribution of X and Y is given by $Y \sim \mathsf{Bern}(1/2)$, $P_{X|Y=k} = \mathcal{N}(\mu_k, \Sigma)$, where $\mu_1, \mu_2 \in \mathbb{R}^m$ and Σ is an m-by-m psd matrix with eigenvalues $\sigma_1^2 \geq \ldots \geq \sigma_m^2$. We have the following corollary.

Corollary 3.2. *For any L_∞-norm adversarial budget $\varepsilon \geq 0$, it holds that $\mathsf{R}^*_\varepsilon(P^1, P^2) \geq 1 - 2\Phi\left(\|s(\varepsilon)\|_2\right)$, where the vector $s(\varepsilon) \in \mathbb{R}^m$ is defined by setting $s(\varepsilon)_j := (\Delta_j/2 - \varepsilon)_+/\sigma_j$, with $\Delta_j := |\mu_{1,j} - \mu_{2,j}|$ for all $j \in [\![m]\!]$.*

Equipped with the above Corollary, we can give a one-line proof to **Theorem 2.1** of (Tsipras et al., 2018), restated here for completeness.

Proposition 1 (Theorem 2.1 of (Tsipras et al., 2018))**.** *Consider the binary-classification problem studied in Corollary 3.2. Let $\varepsilon \geq 0$ and $\delta \in [0, 1)$. Any classifier which attains a normal accuracy at least $1 - \delta$ has adversarial accuracy at most $1 - 2\Phi(\|s(\varepsilon)\|_2) - \delta$ against L_∞-attacks with budget ε. In particular, if $\varepsilon \geq \|\mu_1 - \mu_2\|_\infty = \max_{j \in [\![m]\!]} \Delta_j$, then the adversarial accuracy of the classifier is at most* $\min(1/2 + \delta, 1 - \delta)$.

Proof. Follows directly from Corollary 3.2 and Theorem 2.1 applied to the distance-based attack model $\Omega = D_\varepsilon := \{(x, x') \mid \mathbb{R}^m \mid \|x - x'\|_\infty \leq 2\varepsilon\}$. □

4 Universal bounds for general distance-based attacks

We now turn to the special case distance-based attacks on a metricized feature space $\mathcal{X} = (\mathcal{X}, d)$. We will exploit geometric properties of the class-conditional distributions P^k to obtain upper-bounds on $\mathsf{TV}_\varepsilon(P^1, P^2)$, which will in turn imply lower lower bounds on optimal error (thanks to Theorems 3.1 and 3.3).

4.1 Bounds for light-tailed class-conditional distributions

We now establish a series of upper-bounds on $\mathsf{TV}_\varepsilon(P^1, P^2)$, which in turn provide hard lower bounds for the adversarial robustness error on any classifier for the binary classification experiment (P^1, P^2), namely $\mathsf{R}^*_\varepsilon(P^1, P^2)$. These bounds are a consequence of ligh-tailed class conditional distributions.

Name of the game. We always have the upper-bound $\mathsf{R}^*_\varepsilon(P^1, P^2) \leq 1/2$ (attained by random guessing). Thus, the real challenge is to show that $\mathsf{R}^*_\varepsilon(P^1, P^2) = 1/2 + o_\varepsilon(1)$, where $o_\varepsilon(1)$ goes to zero as the attack "budget" ε is increased.

Definition 4.1 (Bounded tails)**.** *Let $\alpha : [0, \infty) \to [0, 1]$ be a function. We say the a distribution Q on $(\mathcal{X}, d)$ has α-light tail about the point $x_0 \in \mathcal{X}$ if $\mathbb{P}_{x \sim Q}(d(x, x_0) > t) \leq \alpha(t)\ \forall t \geq 0$.*

Theorem 4.1 (The curse of light-tailed class-conditional distributions)**.** *Suppose P^1 and P^2 have α-light tails about a points $\mu_1 \in \mathcal{X}$ and $\mu_2 \in \mathcal{X}$ resp. Then*

$$\mathsf{R}^*_\varepsilon(P^1, P^2) \geq 1/2 - \alpha\left((\varepsilon - d(\mu_1, \mu_2))/2\right), \tag{23}$$

holds for every $\varepsilon \geq d(\mu_1, \mu_2)$.

Proof of Theorem 4.1. Define $\widetilde{\varepsilon} := (\varepsilon - d(\mu_1, \mu_2))/2$ and let (X_1, X_2) be a any coupling of P^1 and P^2. By definition of $\mathsf{TV}_\varepsilon(P^1, P^2)$, we have

$$\begin{aligned}
\mathsf{TV}_\varepsilon(P^1, P^2) &\leq \mathbb{P}(d(X_1, X_2) > \varepsilon) \\
&\leq \mathbb{P}(d(X_1, \mu_1) + d(X_2, \mu_2) > \varepsilon - d(\mu_1, \mu_2)) \\
&\leq \mathbb{P}\left(d(X_1, \mu_1) > \widetilde{\varepsilon}\right) + \mathbb{P}\left(d(X_2, \mu_2) > \widetilde{\varepsilon}\right) \\
&\leq \alpha(\widetilde{\varepsilon}) + \alpha(\widetilde{\varepsilon}) = 2\alpha(\widetilde{\varepsilon}),
\end{aligned}$$

where the 1st inequality is the triangle inequality and the 2nd is a union bound. The result then follows by minimizing over the coupling (X_1, X_2). □

4.2 Bounds under general moment and tail constraints

The following condition will be central for the rest of the manuscript.

Condition 4.1 (Moment constraints)**.** *There exists $\alpha > 0$ and $M : \mathbb{R}_+ \to \mathbb{R}_+$ be an increasing convex function such that $M(0) = 0$. We will occasionally assume that there exist $\mu_1, \mu_2 \in \mathcal{X}$ such that the following moment condition is satisfied*

$$\mathop{\mathbb{E}}_{x_1 \sim P^1}[M(d(x_1, \mu_1))] + \mathop{\mathbb{E}}_{x_2 \sim P^2}[M(d(x_2, \mu_2))] \leq 2\alpha. \tag{24}$$

The function M in Condition 4.1 is called a *Young function*. For example, if each P^k is σ-subGaussian about $\mu_k \in \mathbb{R}^m$, then we may take $M(r) := e^{r^2/\sigma^2} - 1$ to satisfy the condition. More generally, recall that the Orlicz M-norm of a random variable $X_k \sim P^k$ (relative to the reference point μ_k) is defined by

$$\|X_k\|_M := \inf\{C > 0 \mid \mathbb{E}[M(d(X_k, \mu_k)/C)] \leq 1\}. \tag{25}$$

Thus, Condition 4.1 is more general than demanding that both P^1 and P^2 have Orlicz M-norm at most α.

Theorem 4.2 (The curse of bounded moments)**.** *Suppose* (P^1, P^2) *satisfies Condition 4.1. Then*

$$\mathsf{R}^*_\varepsilon(P^1, P^2) \geq \frac{1}{2} - \frac{\alpha}{M(\widetilde{\varepsilon})},\ \forall \varepsilon \geq 0. \tag{26}$$

Proof. Define $\widetilde{\varepsilon} := (\varepsilon - d(\mu_1, \mu_2))/2$. Let (X_1, X_2) be a coupling of P^1 and P^2. Then, by the definition of $\mathsf{TV}_\varepsilon(P^1, P^2)$, we have

$$\begin{aligned}
\mathsf{TV}_\varepsilon(P^1, P^2) &\leq \mathbb{P}(d(X_1, X_2) > \varepsilon) \\
&\leq \mathbb{P}\left(d(X_1, \mu_1) + d(X_2, \mu_2) > \varepsilon - d(\mu_1, \mu_2)\right) \\
&\leq \mathbb{P}\left(M\left(\frac{d(X_1, \mu_1)}{2} + \frac{d(X_2, \mu_2))}{2}\right) > M(\widetilde{\varepsilon})\right) \\
&\leq \mathbb{P}\left(\frac{M(d(X_1, \mu_1)) + M(d(X_2, \mu_2))}{2} > M(\widetilde{\varepsilon})\right) \\
&\leq \frac{\alpha}{M(\widetilde{\varepsilon})},
\end{aligned}$$

where the 2nd inequality is the triangle inequality; the 3rd inequality is because M is increasing; the 4th is because M is convex; the 5th is Markov's inequality and the moment the assumption. □

A variety of corollaries to Theorem 4.2 can be obtained by considering different choices for the moment function M and the parameter α. More are presented in the supplementary materials. For example if P^1 and P^2 have $d(\cdot, x_0) \in L^p(P^1) \cap L^p(P^2)$ for some (and therefore all) $x_0 \in \mathcal{X}$, we may take $M(r) := r^p$ and $\alpha = W_{d,p}(P^1, P^2)^p$, where $W_{d,p}(P^1, P^2)$ is the order-p Wasserstein distance between P^1 and P^2, and obtain the following corollary, which was also obtained independently in the recent paper (Pydi & Jog, 2019). We have

Corollary 4.1 (Lower-bound from Wasserstein distance)**.** *Under the conditions in the previous paragraph, we have*

$$\mathsf{R}^*_\varepsilon(P^1, P^2) \geq \frac{1}{2} - \left(\frac{W_{d,p}(P^1, P^2)}{\varepsilon}\right)^p. \tag{27}$$

Proof. Follows from Theorem 4.2 with $\mu_1 = \mu_2 = x_0 \in \mathcal{X}$ (any point!), $M(r) \equiv r^p$, $\alpha = W_{d,p}(P^1, P^2)^p$. □

Our extension from L^p spaces to general *Orlicz spaces* allows for a more effective exploitation of geometric structure of the data, via other *Young functions* which may grow much faster than the polynomials $r \mapsto r^p$ associated to L^p spaces. As an example, consider (again) the Gaussian binary-classification problem $\mathbb{R}^m$ given by $P^k := \frac{1}{2}P_{X|Y=k} = \frac{1}{2}\mathcal{N}(\mu^{(k)}, I_m)$. Let $\Delta := \|\mu^{(1)} - \mu^{(2)}\|_2$ be the separation distance of the means (which measures the difficulty of this problem). Each P^k satisfies Condition 4.1 with the choices *(A)* $M_p(r) \equiv r^p$ (for any $p \geq 1$) with $\alpha = W_{\|\cdot\|_2,p}(P^1, P^2)^p = \Delta^p$, and *(B)* $M_{\text{subG}}(r) \equiv e^{r^2} - 1$ with $\alpha = 1$. Thus, thanks to our Theorem 4.2 and Corollary 4.1, M_p leads the lower-bound $\mathsf{R}^*_\varepsilon(P^1, P^2) \geq 1 - (\frac{\Delta}{\varepsilon})^p$ for any L_2-attack budget $\varepsilon \geq \Delta$, while M_{subG} leads to the much improved (i.e larger) lower-bound $\mathsf{R}^*_\varepsilon(P^1, P^2) \geq 1 - e^{-\frac{(\varepsilon-\Delta)^2}{4}}$.

5 Concluding remarks

Our results extend the current theory on the limitations of adversarial robustness in machine learning. Using techniques from optimal transport theory, we have obtained explicitly variational formulae and lower-bounds on the Bayes-optimal error classifiers can attain under adversarial attack. These formulae suggest that instead of doing adversarial training on normal data, practitioners should strive to do normal training on adversarially augmented data. Going further, in the case of metric attacks, we have obtained explicit bounds which exploit the high-dimensional geometry of the class-conditional distribution of the data. These bounds are universal in that the are classifier-independent; they only depend on the geometric properties of the class-conditional distribution of the data (e.g light-tailed distributions, distribution with finite-moments, Orlicz spaces, etc.).

References

Athalye, A., Carlini, N., and Wagner, D. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In Dy, J. and Krause, A. (eds.), *ICML*, volume 80, pp. 274–283. PMLR, 2018.

Bhagoji, A. N., Cullina, D., and Mittal, P. Lower bounds on adversarial robustness from optimal transport, 2019.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. OUP Oxford, 2013. ISBN 9780199535255.

Cranko, Z., Menon, A., Nock, R., Ong, C. S., Shi, Z., and Walder, C. Monge blunts bayes: Hardness results for adversarial training. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1406–1415, Long Beach, California, USA, 09–15 Jun 2019. PMLR.

Dohmatob, E. Generalized no free lunch theorem for adversarial robustness. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings*

of Machine Learning Research, pp. 1646–1654. PMLR, 2019.

Fawzi, A., Fawzi, H., and Fawzi, O. Adversarial vulnerability for any classifier. *CoRR*, abs/1802.08686, 2018.

Gilmer, J., Metz, L., Faghri, F., Schoenholz, S. S., Raghu, M., Wattenberg, M., and Goodfellow, I. J. Adversarial spheres. *CoRR*, abs/1801.02774, 2018.

Gilmer, J., Ford, N., Carlini, N., and Cubuk, E. D. Adversarial examples are a natural consequence of test error in noise. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning, ICML 2019, 9-15 June 2019, Long Beach, California, USA*, volume 97 of *Proceedings of Machine Learning Research*, pp. 2280–2289. PMLR, 2019.

Harel, M. and Mannor, S. The perturbed variation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 37(10):2119–2130, 2015.

Mahloujifar, S., Diochnos, D. I., and Mahmoody, M. The curse of concentration in robust learning: Evasion and poisoning attacks from concentration of measure. *CoRR*, abs/1809.03063, 2018.

Moosavi-Dezfooli, S., Fawzi, A., Fawzi, O., and Frossard, P. Universal adversarial perturbations. In *2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 86–94, 2017.

Moosavi-Dezfooli, S., Fawzi, A., Fawzi, O., Frossard, P., and Soatto, S. Analysis of universal adversarial perturbations. abs/1705.09554, 2017. URL `http://arxiv.org/abs/1705.09554`.

Pydi, M. S. and Jog, V. Adversarial risk via optimal transport and optimal couplings. In *ArXiv preprint (to appear in ICML 2020)*. PMLR, 2019. URL `http://arxiv.org/abs/1705.09554`.

Reid, M. D. and Williamson, R. C. Information, divergence and risk for binary experiments. *J. Mach. Learn. Res.*, 12: 731–817, 2011.

Schmidt, L., Santurkar, S., Tsipras, D., Talwar, K., and Madry, A. Adversarially robust generalization requires more data. *CoRR*, abs/1804.11285, 2018.

Shafahi, A., Huang, W. R., Studer, C., Feizi, S., and Goldstein, T. Are adversarial examples inevitable? *CoRR*, abs/1809.02104, 2018.

Su, J., Vargas, D. V., and Sakurai, K. One pixel attack for fooling deep neural networks. *CoRR*, abs/1710.08864, 2017.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. *CoRR*, abs/1805.12152, 2018.

Villani, C. *Topics in Optimal Transportation*. American Mathematical Society, 2003.

Yang, Y., Rashtchian, C., Wang, Y., and Chaudhuri, K. Adversarial examples for non-parametric methods: Attacks, defenses and large sample limits. *CoRR*, abs/1906.03310, 2019.

A Recap of main results

For the convenience of the reader, let us begin by informally summarizing the main contributions of our paper. Rigorus restatements (and proofs) of the results will follow.

Our main contributions can be summarized as follows.

• In section 3 (after developing some background material in section 2), we use optimal transport theory to derive variational formulae for the Bayes-optimal error (aka smallest possible test error) of a classifier under adversarial attack, as a function of the "budget" of the attacker, These formulae suggest that instead of doing adversarial training, practitioners should rather do normal training on adversarially augmented data. Incidentally, this is a well-known trick to boost up the adversarial robustness of classifiers to known attacks, and is usually used in practice under the umbrella name of "adversarial data-augmentation". See (Yang et al., 2019), for example. In our manuscript, this principle appears as a natural consequence of our variational formulae.

• For the special case of distance-based attacks, we proceed in 4 to *(1)* Establish universal lower-bounds on the adversarial Bayes-optimal error. These bounds are a consequence of concentration properties of light-tailed class-conditional distributions of the features (e.g sub-Gaussianity, etc.). *(2)* Establish universal bounds under more general moment constraints conditions on the class-conditional distributions (e.g existence of covariance matrices for the class-conditional distributions of the features).

B Proofs of lemmas, propositions, theorems, and corollaries

In this appendix we provide complete proofs for the theorems, corollaries, etc. which were stated without proof in the manuscript. For clarity, each result from the manuscript (theorems, corollaries, etc.) is restated in this supplemental before proved.

B.1 Proofs for results in section 3

Theorem 3.1 (Extension of Theorem 1 of (Bhagoji et al., 2019))**.** *For any attack model Ω on the feature space $\mathcal{X}$, the adversarial Bayes-optimal error under type-Ω attacks satisfies $\mathsf{R}^*_\Omega(P^1, P^2) \geq \frac{1}{2} - \mathsf{OT}_\Omega(P^1, P^2)$.*

First note that the Ω defined is not automatically a closed subset of $\mathcal{X}^2$. A sufficient condition is that the metric space $(\mathcal{X}, d)$ has the *mid-point property*.

Proof of Theorem 3.1. For $x \in \mathcal{X}$, define $\Omega(x) := \{x' \in \mathcal{X} \mid (x', x) \in \Omega\}$. For a classifier h, consider the derived randomized classifier $\widetilde{h} : \mathcal{X} \to \{1, 2\}$ defined by

$$\widetilde{h}(x) := \begin{cases} y, & \text{if } h(x') = y \ \forall x' \in \Omega(x), \\ Y', & \text{else.} \end{cases} \tag{28}$$

where Y' is a random variable with the same distribution as the label Y, which encodes an *"I don't know, I'm just going to random-guess the label!"* decision. Let $\mathbb{P}(Y = 1) = \pi$ be the class probability for class 1. Let X_k be a random variable that has the same distribution as X conditioned on the event $Y = k$. By construction, it is easy to see that the adversarial accuracy of h is exactly equal to the normal accuracy of $\widetilde{h}$, plus an additional bonus of $|\pi - (1 - \pi)|/2 = |2\pi - 1|/2$ due to the *"I don't know!"* decisions encoded by Y'. That is

$$1 - \mathsf{R}_\Omega(h; P^1, P^2) = \pi\mathbb{E}[\mathbb{1}[\widetilde{h}(X_1) = 1]] + (1 - \pi)\mathbb{E}[\mathbb{1}[\widetilde{h}(X_2) = 2]] + \frac{1}{2}|2\pi - 1|. \tag{29}$$

WLOG, suppose $0 \leq \pi \leq 1/2$, and define $g_0(x') := \mathbb{1}[\widetilde{h}(x') = 1]$ and $f_0(x) = \mathbb{1}[\widetilde{h}(x) \neq 2] = 1 - \mathbb{1}[\widetilde{h}(x) = 2])$. It is clear that g_0 (resp. f_0) is bounded and P^1-measurable (resp. P^2-measurable). Moreover, if $x', x \in \mathcal{X}$ with $c_\Omega(x', x) = 0$, then $x' \notin \Omega(x)$. Since $h^{-1}(\{1\})$ and $h^{-1}(\{2\})$ partition $\mathcal{X}$, at most one of $\Omega(x') \subseteq h^{-1}(\{1\})$ and $\Omega(x) \subseteq h^{-1}(\{2\})$ holds. Thus, $\mathbb{1}[\widetilde{h}(x') = 1] + \mathbb{1}[\widetilde{h}(x) = 2] \leq 1$, and so $g_0(x') - f_0(x) = \mathbb{1}[\widetilde{h}(x') = 1] + \mathbb{1}[\widetilde{h}(x') = 2] - 1 \leq c_\Omega(x', x) \ \forall x, x' \in \mathcal{X}$. Thus, (f_0, g_0) is a pair of *Kantorovich potentials* for the cost-function c_Ω. Consequently, invoking the *Kantorovich-Rubinstein*

duality formula for optimal transport, and recalling the definition of X_k and P^k, we compute

$$\begin{aligned}\mathsf{OT}_\Omega(P^1,P^2) &= \sup_{\text{potentials } f,g} \int_{\mathcal{X}} g(x)dP^1(x) - \int_{\mathcal{X}} f(x)dP^2(x) = \sup_{\text{potentials } f,g} \pi\mathbb{E}[g(X_1)] - (1-\pi)\mathbb{E}[f(X_2)] \\ &\geq \pi\mathbb{E}[g_0(X_1)] - (1-\pi)\mathbb{E}[f_0(X_2)] = \pi\mathbb{E}[\mathbb{1}[\widetilde{h}(X_1)=1]] + (1-\pi)\mathbb{E}[\mathbb{1}[\widetilde{h}(X_2)=2]] - (1-\pi) \\ &\overset{(29)}{=} 1 - \frac{|2\pi-1|}{2} - (1-\pi) - \mathsf{R}_\Omega(h;P^1,P^2) = \frac{1}{2} - \mathsf{R}_\Omega(h;P^1,P^2),\end{aligned}$$

where the last equality is because $0 \leq \pi \leq 1/2$. Maximizing the RHS over h, we obtain $\mathsf{R}^*_\Omega(P^1,P^2) := \inf_h \mathsf{R}_\Omega(h;P^1,P^2) \geq 1/2 - \mathsf{OT}_\Omega(P^1,P^2)$ as claimed. □

Theorem 3.2. *Let Ω be an attack model on $\mathcal{X}$. Then we have the identity*

$$\mathsf{OT}_\Omega(P^1,P^2) = \sup_{U\subseteq\mathcal{X} \text{ closed}} P^1(U) - P^2(\overline{U}_\Omega). \tag{16}$$

In particular, for distance-based attacks we have

$$\mathsf{OT}_\varepsilon(P^1,P^2) = \sup_{U\subseteq\mathcal{X} \text{ closed}} P^2(U) - P^1(U^\varepsilon). \tag{17}$$

Proof. One directly computes

$$\begin{aligned}\mathsf{OT}_\Omega(P^1,P^2) &:= \inf_{\gamma\in\Pi(P^1,P^2)} \int_{\mathcal{X}^2} c_\Omega(x,x')\,\mathrm{d}\gamma(x,x' \quad = \inf_{\gamma\in\Pi(P^1,P^2)} \int_{\mathcal{X}^2} \mathbb{1}[(x,x')\in\Omega]\,\mathrm{d}\gamma(x,x') \\ &= \inf_{\gamma\in\Pi(P^1,P^2)} \int_\Omega \mathrm{d}\gamma(x,x') = \inf_{\gamma\in\Pi(P^1,P^2)} \gamma(\Omega),\end{aligned} \tag{30}$$

On the other hand, by *Strassen's Marriage Theorem* (see (**?**)Theorem 1.27 of]villaniTopics) and the definition of $\overline{U}_\Omega$ in Eq. (14), one has

$$\inf_{\gamma\in\Pi(P^1,P^2)} \gamma(\Omega) = \sup_{U\subseteq\mathcal{X} \text{ closed}} P^2(U) - P^1(\overline{U}_\Omega),$$

and the result follows. The particular case of distance-based attacks corresponds to letting $\Omega := D_\varepsilon := \{(x,x') \in \mathcal{X}^2 \mid d(x,x') \leq \varepsilon\}$, so that $\overline{U}_\Omega = U^\varepsilon$, the ε-neighborhood of U. □

Theorem 3.3 (Adversarially augmented data, a proxy for adversarial robustness)**.** *Consider a classification problem (P^1,P^2). Suppose d is a distance on the feature space $\mathcal{X}$ with the MM property 3.1, and consider the distance-based attack model $D_\varepsilon := \{(x,x') \in \mathcal{X}^2 \mid d(x,x') \leq \varepsilon\}$. Recall the definition of $\mathsf{TV}_\varepsilon(P^1,P^2)$ from Eq.* (11)*. Define*

$$\begin{aligned}\widetilde{\mathsf{TV}}_\varepsilon(P^1,P^2) &:= \inf_{\gamma_1,\gamma_2} \mathsf{TV}(\mathsf{proj}^2{}_\#\gamma_1, \mathsf{proj}^1{}_\#\gamma_2), \\ \widetilde{\widetilde{\mathsf{TV}}}_\varepsilon(P^1,P^2) &:= \inf_{a_1,a_2 \text{ type-}D_{\varepsilon/2}} \mathsf{TV}(a_{1\#}P^2, a_{2\#}P^1),\end{aligned} \tag{19}$$

where "#" denotes pushfoward of measures and the 1st inf. is taken over all pairs of distributions (γ_1,γ_2) on $\mathcal{X}^2$ concentrated on $D_{\varepsilon/2}$ such that $\mathsf{proj}^1{}_\#\gamma_1 = P^2$ and $\mathsf{proj}^2{}_\#\gamma_2 = P^1$. It holds that

$$\mathsf{TV}_\varepsilon(P^1,P^2) = \widetilde{\mathsf{TV}}_\varepsilon(P^1,P^2) \leq \widetilde{\widetilde{\mathsf{TV}}}_\varepsilon(P^1,P^2), \tag{20}$$

and there is equality if P^1 and P^2 have densities w.r.t the Borel measure on $\mathcal{X}$.

Consequently, we have the following lower-bound for the adversarial Bayes-optimal error:

$$\begin{aligned}\mathsf{R}^*_\varepsilon(P^1,P^2) &\geq \frac{1}{2} - \mathsf{TV}_\varepsilon(P^1,P^2) \geq \frac{1}{2} - \widetilde{\mathsf{TV}}_\varepsilon(P^1,P^2) \\ &\geq \frac{1}{2} - \widetilde{\widetilde{\mathsf{TV}}}_\varepsilon(P^1,P^2).\end{aligned} \tag{21}$$

Proof. Note that each P^k is a Borel measure on $\mathcal{X}$ which integrates to $1/2$. For the convenience of the proof, we rescale each P^k by 2, so that it integrates to 1.

Let $D'_\varepsilon := \mathcal{X}^2 \setminus D_\varepsilon$. To prove the theorem, we consider the following intermediate quantity

$$E(P^1, P^2) := \inf_{\gamma \in \Pi_\leq(P^1,P^2),\, \mathsf{supp}(\gamma) \subseteq D_\varepsilon} 1 - \gamma(\mathcal{X}^2).$$

Applying Lemma 3.2 with $\Omega = D_\varepsilon$, we know that $\mathsf{OT}_\varepsilon = E$. The rest of the proof is divided into separate steps.

Step 1: proving the equality $E = \widetilde{\mathsf{TV}}_\varepsilon$. Let γ be feasible for E. Because $(\mathcal{X}, d)$ satisfies the MM property (Condition 3.1), there exists a γ-measurable map $\eta : \mathcal{X}^2 \to \mathcal{X}$ such that $\eta(x, x')$ is a midpoint of x and x' for all $x, x' \in \mathcal{X}$. Now, consider the γ-measurable maps $T_1, T_2 : \mathcal{X}^2 \to \mathcal{X}^2$, $D : \mathcal{X} \to \mathcal{X}^2$ defined by

$$T_1(x, x') := (x, \eta(x, x')),\ T_2(x, x') := (\eta(x, x'), x'),\ T_3(x) = (x, x). \tag{31}$$

Construct couplings $\gamma_1 = (T_1)_\# \gamma + T_{3\#}(P^1 - \mathsf{proj}^1_\# \gamma)$ and $\gamma_2 = (T_2)_\# \gamma + T_{3\#}(P^2 - \mathsf{proj}^2_\# \gamma)$. Then (γ_1, γ_2) is feasible for $\widetilde{\mathsf{TV}}_\epsilon$ and

$$\mathsf{TV}(\mathsf{proj}^2_\# \gamma_1, \mathsf{proj}^1_\# \gamma_2) \leq (P^1 - \mathsf{proj}^1_\# \gamma)(\mathcal{X}) + (P^2 - \mathsf{proj}^2_\# \gamma)(\mathcal{X}) = 2(1 - \gamma(\mathcal{X}^2))$$

because the second marginal of $(T_1)_\#\gamma$ and the first marginal of $(T_2)_\#\gamma$ agree by construction. Thus $\mathsf{TV}_\varepsilon \leq E$. Conversely, let γ_1, γ_2 be feasible for $\widetilde{\mathsf{TV}}_\varepsilon$, and let $\tilde{\gamma}_1 \leq \gamma_1$ and $\tilde{\gamma}_2 \leq \gamma_2$ be such that $\mathsf{proj}^2_\# \tilde{\gamma}_1 = \mathsf{proj}^1_\# \gamma_2 = (\mathsf{proj}^2_\# \gamma_1) \wedge (\mathsf{proj}^1_\# \gamma_2)$ where $\wedge$ is the "pointwise" minimum of two measures (they can be built with the disintegration theorem). Now build $\tilde{\gamma}$ feasible for E by gluing together $\tilde{\gamma}_1$ and $\tilde{\gamma}_2$. It holds

$$\mathsf{TV}(\mathsf{proj}^2_\# \gamma_1, \mathsf{proj}^1_\# \gamma_2) = 2(1 - \mathsf{proj}^2_\# \gamma_1 \wedge \mathsf{proj}^1_\# \gamma_2)(\mathcal{X}) = 2(1 - \tilde{\gamma}(\mathcal{X}^2)).$$

Thus $E \leq \widetilde{\mathsf{TV}}_\varepsilon$ hence $E = \widetilde{\mathsf{TV}}_\varepsilon$.

Step 2: proving the inequality $\widetilde{\mathsf{TV}}_\varepsilon \leq \widetilde{\widetilde{\mathsf{TV}}}_\varepsilon$. Now, the fact that $\widetilde{\mathsf{TV}}_\varepsilon \leq \widetilde{\widetilde{\mathsf{TV}}}_\varepsilon$ in general is due to the fact to any transport map a satisfying $d(a(x), x) \leq \varepsilon$ corresponds a deterministic transport plan $(\mathrm{id}, a)_\# P^1$ supported on D_ε. In general, equality in the theorem will fail to hold. For example, on the real line, $P^1 = \frac{1}{3}\delta_{-\varepsilon} + \frac{1}{3}\delta_0 + \frac{1}{3}\delta_\varepsilon$ and $P^2 = \frac{1}{2}\delta_{-\varepsilon} + \frac{1}{2}\delta_\varepsilon$ has $\widetilde{\widetilde{\mathsf{TV}}}_\varepsilon(P^1, P^2) = 2/6$ and $\widetilde{\mathsf{TV}}_\varepsilon(P^1, P^2) = 0$.

Step 3: proving the inequality $\widetilde{\mathsf{TV}}_\varepsilon \geq \widetilde{\widetilde{\mathsf{TV}}}_\varepsilon$ for absolutely continuous P^k's. Finally, the fact that $\widetilde{\mathsf{TV}}_\varepsilon \geq \widetilde{\widetilde{\mathsf{TV}}}_\varepsilon$ when P^1 and P^2 are absolutely continuous is a consequence of the existence of an optimal transport map for the W_∞ distance. Indeed, if (γ_1, γ_2) is feasible for $\widetilde{\mathsf{TV}}_\varepsilon$, then $W_\infty(P^1, \mathsf{proj}^2_\# \gamma_1) \leq \varepsilon$ and there exists a measurable map $a_1 : \mathcal{X} \to \mathcal{X}$ such that $d(a_1, x) \leq \varepsilon$ P^1-a.e. and $(a_1)_\# P^1 = \mathsf{proj}^2_\# \gamma_1$ (one can build a_2 similarly). □

Corollary 3.2. *For any L_∞-norm adversarial budget $\varepsilon \geq 0$, it holds that $\mathsf{R}^*_\varepsilon(P^1, P^2) \geq 1 - 2\Phi\left(\|s(\varepsilon)\|_2\right)$, where the vector $s(\varepsilon) \in \mathbb{R}^m$ is defined by setting $s(\varepsilon)_j := (\Delta_j/2 - \varepsilon)_+/\sigma_j$, with $\Delta_j := |\mu_{1,j} - \mu_{2,j}|$ for all $j \in [\![m]\!]$.*

Proof. The first part of the claim follows from a direct application of (**?**)Theorem 1]barsov87:

$$\mathsf{TV}(\mathcal{N}(\mu_1, \Sigma), \mathcal{N}(\mu_2, \Sigma)) = 2\Phi(\|\mu\|_{\Sigma^2}/2) - 1,$$

where $\mu := \mu_1 - \mu_1 \in \mathbb{R}^d$. Thus $\mathsf{R}^*_\Omega \geq 1 - 2\Phi(\Delta(\varepsilon)/2)$, where $\Delta(\varepsilon) := \min_{\|z\|_\mathcal{X} \leq \varepsilon} \|z - \mu\|_{\Sigma^2}$. It now remains to bound $\Phi(\Delta(\varepsilon))$, and we are led to consider the computation of quantities of the following form.

Bounding the quantity $\Delta(\varepsilon)$. We are led to consider problems of the form

$$\alpha \quad := \max_{\|w\|_\Sigma \leq 1} w^T a - \varepsilon \|w\|_1, \tag{32}$$

where $a \in \mathbb{R}^d$ and Σ be a positive definite matrix of size d. Of course, the solution value might not be analytically expressible in general, but there is some hope, when the matrix Σ is diagonal. That notwithstanding, using the dual representation of the ℓ_1-norm, one has

$$\begin{aligned}\alpha &= \max_{\|w\|_\Sigma \le 1} \min_{\|z\|_\infty \le \varepsilon} w^T a - w^T z = \max_{\|z\|_\infty \le \varepsilon} \min_{\|w\|_\Sigma \le 1} w^T (z-a) \\ &= \min_{\|z\|_\infty \le \varepsilon} \left(\max_{\|w\|_\Sigma \le 1} w^T (z-a) \right) = \min_{\|z\|_\infty \le \varepsilon} \left(\max_{\|\tilde{w}\|_2 \le 1} \tilde{w}^T \Sigma^{-1} (z-a) \right) \\ &= \min_{\|z\|_\infty \le \varepsilon} \|z-a\|_{\Sigma^{-1}} = \min_{\|z\|_\infty \le \varepsilon} \|z-a\|_{\Sigma^{-1}},\end{aligned} \tag{33}$$

where we have used *Sion's minimax theorem* to interchange min and max in the first line, and we have introduced the auxiliary variable $\tilde{w} := \Sigma^{-1/2} w$ in the fourth line. We note that given a value for the dual variable z, the optimal value of the primal variable w is

$$w \propto \frac{\Sigma^{-1}(a-z)}{\|\Sigma^{-1}(a-z)\|_2}. \tag{34}$$

The above expression (33) for the optimal objective value α is unlikely to be computable analytically in general, due to the non-separability of the objective (even though the constraint is perfectly separable as a product of 1D constraints). In any case, it follows from the above display that $\alpha \le 0$, with equality iff $\|a\|_\infty \le \varepsilon$.

Exact formula for diagonal Σ. In the special case where $\Sigma = \mathrm{diag}(\sigma_1, \ldots, \sigma_2)$, the square of the optimal objective value α^2 can be separated as

$$\begin{aligned}\alpha \ge 0,\ \alpha^2 = \sum_{j=1}^d \min_{|z_j| \le \varepsilon} \sigma_j^{-2} (z_j - a_j)^2 &= \sum_{j=1}^d \sigma_j^{-2} \begin{cases} (a_j+\varepsilon)^2, & \text{if } a_j \le -\varepsilon, \\ 0, & \text{if } -\varepsilon < a_j \le \varepsilon, \\ (a_j-\varepsilon)^2, & \text{if } a_j > \varepsilon, \end{cases} \\ &= \sum_{j=1}^d \sigma_j^{-2} ((|a_j| - \varepsilon)_+)^2.\end{aligned}$$

Thus $\alpha = \sqrt{\sum_{j=1}^d \sigma_j^{-2} ((|a_j| - \varepsilon)_+)^2}$. By the way, the optimium is attained at

$$\begin{aligned} z_j &= \begin{cases} -\varepsilon, & \text{if } a_j \le -\varepsilon, \\ a_j, & \text{if } -\varepsilon < a_j \le \varepsilon, \\ \varepsilon, & \text{if } a_j > \varepsilon, \end{cases} \\ &= a_j - \mathsf{sign}(a_j)(|a_j| - \varepsilon)_+ \end{aligned} \tag{35}$$

Plugging this into (34) yields the optimal weights

$$w_j \propto \sigma_j^{-2}\, \mathsf{sign}(a_j)(|a_j| - \varepsilon)_+. \tag{36}$$

Upper / lower bounds for general Σ. Let $\sigma_1, \sigma_2, \ldots, \sigma_d > 0$ be the eigenvalues of Σ. Then

$$\|z-a\|_{\Sigma^{-1}}^2 := (z-a)^T \Sigma^{-1} (w-a) \le \sum_{j=1}^d (z_j - a_j)^2 / \sigma_j^2 =: \|z-a\|_{\mathsf{diag}(1/\sigma_1, \ldots, 1/\sigma_d)}^2.$$

Therefore in view of the previous computations for diagonal covariance matrices, one has the bound $\alpha \le \sqrt{\sum_{j=1}^d \sigma_j^{-2} ((|a_j| - \varepsilon_j)_+)^2}$. □