Monge blunts Bayes: Hardness Results for Adversarial Training

Zac Cranko [1 2] **Aditya Krishna Menon** [3] **Richard Nock** [2 1 4] **Cheng Soon Ong** [2 1] **Zhan Shi** [5] **Christian Walder** [2 1]

Abstract

The last few years have seen a staggering number of empirical studies of the robustness of neural networks in a model of adversarial perturbations of their inputs. Most rely on an adversary which carries out local modifications within prescribed balls. None however has so far questioned the broader picture: how to frame a *resource-bounded* adversary so that it can be *severely detrimental* to learning, a non-trivial problem which entails at a minimum the choice of loss and classifiers.

We suggest a formal answer for losses that satisfy the minimal statistical requirement of being *proper*. We pin down a simple sufficient property for any given class of adversaries to be detrimental to learning, involving a central measure of "harmfulness" which generalizes the well-known class of integral probability metrics. A key feature of our result is that it holds for *all* proper losses, and for a popular subset of these, the optimisation of this central measure appears to be *independent of the loss*. When classifiers are Lipschitz – a now popular approach in adversarial training –, this optimisation resorts to *optimal transport* to make a low-budget compression of class marginals. Toy experiments reveal a finding recently separately observed: training against a sufficiently budgeted adversary of this kind *improves* generalization.

1. Introduction

Starting from the observation that deep nets are sensitive to imperceptible perturbations of their inputs (Szegedy et al., 2013), a surge of recent work has focussed on new *adversarial training* approaches to supervised learning (Athalye et al., 2018a;b; Bastani et al., 2016; Buckman et al., 2018; Bubeck et al., 2018; Cai et al., 2018; Dhillon et al., 2018; Fawzi et al., 2018; Gilmer et al., 2018; Goswami et al., 2018; Guo et al., 2018; Ilias et al., 2018; Kurakin et al., 2017; Ma et al., 2018; Madry et al., 2018; Samangouei et al., 2018; Song et al., 2018; Tramèr et al., 2018; Tsipras et al., 2019; Uesato et al., 2018; Wang et al., 2018; Wong & Zico Kolter, 2018) (and references within). In the now popular model of Madry et al. (2018), we want to learn a classifier from a set $\mathcal{H}$, given a distribution of clean examples D and loss ℓ. Adversarial training then seeks to find

$$\arg\min_{h\in\mathcal{H}} \mathsf{E}_{(\mathsf{X},\mathsf{Y})\sim D}\left[\max_{\boldsymbol{\delta}:\|\boldsymbol{\delta}\|\leq\delta^*} \ell(\mathsf{Y}, h(\mathsf{X}+\boldsymbol{\delta}))\right], \quad (1)$$

where $\|.\|$ is a norm and δ^* is the budget of the adversary. It has recently been observed that adversarial training damages standard accuracy as data size and adversary's budget (δ^*) increases (Tsipras et al., 2019). A Bayesian explanation is given for a particular $\{D, \|.\|, \ell\}$ in Tsipras et al. (2019), and the authors conclude their findings questioning the interplay between adversarial robustness and standard accuracy.

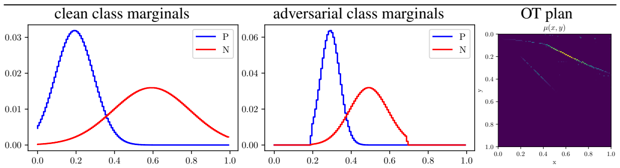
In this paper, we dig into this relationship (i) by casting the standard accuracy and loss in (1) in the broad context of Bayesian decision theory (Grünwald & Dawid, 2004) and (ii) by considering a general form of adversaries, not restricted to the ones used in (1). In particular, we assume that the loss is *proper*, which is just a general form of statistical unbiasedness that many popular choices meet (Hendrickson & Buehler, 1971; Reid & Williamson, 2010). The minimization of a proper loss gives guarantees on the accuracy (for example, Kearns & Mansour (1996)), so it directly connects to the setting of Tsipras et al. (2019). Regarding the adversaries, instead of relying on the local adversarial modification $\mathsf{X} \to \mathsf{X} + \boldsymbol{\delta}$ for some $\|\boldsymbol{\delta}\| \leq \delta^*$, we consider a set of possible local modifications $\mathsf{X} \to a(\mathsf{X})$ for some $a \in \mathcal{A} \subseteq \mathcal{X}^{\mathcal{X}}$ ($\mathcal{A}$ fixed). We then analyze the conditions on $\mathcal{A}$ under which, for some $\varepsilon > 0$,

$$\boxed{\min_{h\in\mathcal{H}} \mathsf{E}_{(\mathsf{X},\mathsf{Y})\sim D}\left[\max_{a\in\mathcal{A}} \ell(\mathsf{Y}, h\circ a(\mathsf{X}))\right] \geq (1-\varepsilon)\ell_0}\ , \quad (2)$$

where ℓ_0 is the loss of the "blunt" predictor which predicts nothing. If h has range $\mathbb{R}$, this blunt predictor is in general 0 (for the log loss, square loss, etc), which translates into

[1]The Australian National University (Australia) [2]Data61 (Australia) [3]Google Research (USA) [4]The University of Sydney (Australia) [5]University of Illinois at Chicago (USA). Authors in alphabetical order. Correspondence to: <first.last@{data61.csiro.au,anu.edu.au}, zshi22@uic.edu, adityakmenon@google.com>.

Proceedings of the 36th International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.



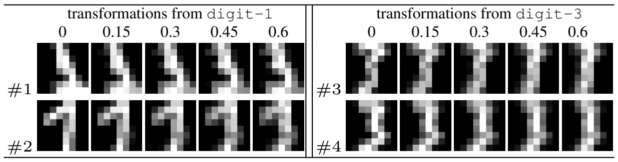


Table 1: *Top table*: compression of the optimal transport (OT) plan for a Mixup adversary on a toy 1D data. *Bottom table*: transformations performed by a Monge adversary for the `digit-1` vs `digit-3` classification problem on four USPS digits (noted #1 to #4), for various adversarial budgets (0 = clean data, see Section 7 for details).

a class probability estimate of $1/2$ for all observations and global accuracy of 50% for two classes, *i.e.* that of an unbiased coin. We see the connection of (2) to the accuracy: as $\varepsilon \searrow$, the learner will be tricked into converging to an extremely poorly accurate predictor. How one can design such provably efficient adversaries, furthermore under tight budget constraints, is the starting point of our paper.

Our first contribution (Section 4) analyzes budgeted adversaries that can enforce (2). Our main finding shows that (2) is implied by a very simple condition involving a central quantity γ generalizing the celebrated integral probability metrics (Sriperumbudur et al., 2009). Furthermore, under some additional condition on the loss, satisfied by the log, square and Matsushita losses, the adversarial optimization of γ *does not depend on the loss*. In other words,

> the adversary can attack the learner disregarding its loss.

Our second contribution (Section 5) considers the adversarial optimisation of γ when the classifiers in $\mathcal{H}$ satisfy a generalized form of Lipschitz continuity. Controlling Lipschitz continuity has recently emerged as a solution to limit the impact of adversarial examples (Cissé et al., 2017). In this context, efficient budgeted adversaries take a particular form: we show that, for an adversary to minimize γ,

> it is sufficient to compress the optimal transport plan between class marginals using the Lipschitz function as transportation cost, disregarding the learner's $\mathcal{H}$.

This result brings the machinery of optimal transport (OT) to the table of adversarial design (Villani, 2009), with a new purpose (the compression of OT plans). These two findings turn out to be very useful from an experimental standpoint: we have implemented two kinds of adversaries inspired by our theory (called Mixup and Monge for their respective links with Zhang et al. (2018); Villani (2009)); Table 1 displays their behaviour on two simple problems. We have observed that training a learner against a "weak" (severely budgeted) adversary improves *generalization* on clean data, a phenomenon also observed elsewhere (Tsipras et al., 2019; Zhang et al., 2018). The `digit` experiment displays how our adversaries progressively transform observations of one class into credible observations of the other (See Section 7, and Supplementary Material, SM).

Our third contribution (Section 6) is an adversarial boosting result: it answers the question as to whether one can efficiently craft an arbitrarily *strong* adversary from the sole access to a black box *weak* adversary. In the framework of reproducing kernel Hilbert spaces (RKHS), we show that

> this "weak adversary" ⇒ "strong adversary" design does exist, and our proof is constructive: we build one.

Our proof revolves around a standard concept of fixed point theory: contractive mappings. We insist on the computational efficiency of this design, linear in the coding size of the Wasserstein distance between class marginals. It shows that, on some adversarial training problems, the existence of the weakest forms of adversaries implies that much stronger ones may be available at cheap (computational) cost.

2. Related work

Formal approaches to the problem of adversarial training are sparse compared to the growing literature on the arms race of experimental results. The formal trend has started on adversarial changes to a loss to be optimized (Sinha et al., 2018) or more directly on a classifier's output (Hein & Andriushchenko, 2017; Raghunathan et al., 2018). For example, (Sinha et al., 2018) add a Wasserstein penalty to a loss, computing the distance between the true and adversarial distributions. They provide smoothness guarantees for the loss and robustness in generalization for its minimization. (Raghunathan et al., 2018) directly penalize the classifier's output (not a loss per se), in the context of shallow networks, and compute adversarial perturbations in a bounded L_∞ ball. A similar approach (but in L_p-norm) is taken in (Hein & Andriushchenko, 2017) for kernel methods and shallow networks. Recent ones also focus on introducing general robustness constraints (Bastani et al., 2016).

More recently, a handful of work have started to investigate the *limits* of learning in an adversarial training setting, but they are limited in that they address particular simulated domains with a particular loss to be optimized, and consider particular adversaries (Bubeck et al., 2018; Fawzi et al., 2018; Gilmer et al., 2018). The distribution can involve Gaussians of mixtures (Bubeck et al., 2018; Fawzi et al., 2018) or the data lies on concentric spheres (Gilmer et al., 2018). The loss involves a distance based on a norm for all, and the adversary makes local shifts to data of bounded radius. In the case of (Bubeck et al., 2018), the access to the data is restricted to statistical queries. The essential results are either that robustness requires too much information compared to not requiring robustness (Bubeck et al., 2018), or the "safety" radius of inoffensive modifications is in fact small relative to some of the problem's parameters, meaning even "cheap" adversaries can sometimes generate damaging adversarial examples (Fawzi et al., 2018; Gilmer et al., 2018). This depicts a pretty negative picture of adversarial training — negative but *local*: all these results share the same common design pattern of relying on particular choices for all key components of the problem: domain, loss and adversaries (and eventually classifiers). There is no approach to date that would relax any of these choices, even less so one that would simultaneously relax all.

3. Definitions and notations

We present some important definitions and notations.

▷ *Proper losses*. Many of our notations follow (Reid & Williamson, 2010). Suppose we have a prediction problem with binary labels. We let $\ell : \{-1, 1\} \times [0, 1] \rightarrow \overline{\mathbb{R}}$ denote a general loss function to be minimized, where the left argument is a class $\mathsf{Y} \in \{-1, 1\}$ and the right argument is a class probability estimate ($\overline{\mathbb{R}}$ is the closure of $\mathbb{R}$). Its *conditional Bayes risk* function is the best achievable loss when labels are drawn with a particular positive base-rate,

$$\underline{L}(\pi) \quad \doteq \quad \inf_c \mathsf{E}_{\mathsf{Y}\sim\pi} \ell(\mathsf{Y}, c), \tag{3}$$

where $\pi \in [0, 1]$, so that $\Pr[\mathsf{Y} = 1] = \pi$ and $\Pr[\mathsf{Y} = -1] \doteq 1 - \pi$. We call the loss ***proper*** iff Bayes prediction locally achieves the minimum everywhere[1]: $\underline{L}(\pi) = \mathsf{E}_{\mathsf{Y}} \ell(\mathsf{Y}, \pi), \forall \pi \in [0, 1]$. One value of $\underline{L}$ is interesting in our context, the one which corresponds to Bayes rule returning maximal "uncertainty", *i.e.* for $\pi = 1/2$,

$$\ell^\circ \quad \doteq \quad \underline{L}\left(\frac{1}{2}\right). \tag{4}$$

Without further ado, we give the key definition which makes more precise the framework sketched in (2).

Definition 1 *For any proper loss ℓ and $(\mathcal{H}, \mathcal{A})$ integrable with respect to some distribution D, the* ***adversarial loss*** *$\ell(\mathcal{H}, \mathcal{A}, D)$ is defined as*

$$\ell(\mathcal{H}, \mathcal{A}, D) \quad \doteq \quad \min_{h\in\mathcal{H}} \mathsf{E}_{(\mathsf{X},\mathsf{Y})\sim D}\left[\max_{a\in\mathcal{A}} \ell(\mathsf{Y}, h \circ a(\mathsf{X}))\right] \tag{5}$$

*For any $\varepsilon \in [0, 1]$, we say that $\mathcal{H}$ is ε-**defeated** by $\mathcal{A}$ on ℓ iff*

$$\ell(\mathcal{H}, \mathcal{A}, D) \quad \geq \quad (1 - \varepsilon) \cdot \ell^\circ. \tag{6}$$

Intuitively, if the adversary can modify instances such that the learner does not do much better than a trivial blunt constant predictor, the adversary can declare victory. The additional quantities (such as the integrability condition) are given later in this section. To finish up with general proper losses, as an example, the log-loss given by $\ell(+1, c) = -\log c$ and $\ell(-1, c) = -\log(1 - c)$ is proper, with conditional Bayes risk given by the Shannon entropy $\underline{L}(\pi) = -\pi \cdot \log \pi - (1 - \pi) \cdot \log(1 - \pi)$.

▷ *Composite, canonical proper losses*. We let $\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$ denote a set of classifiers. To convert real valued predictions into class probability estimates (McCullagh & Nelder, 1989), one traditionally uses an invertible *link* function $\psi : [0, 1] \rightarrow \mathbb{R}$, forming a *composite* loss $\ell_\psi(y, v) \doteq \ell(y, \psi^{-1}(v))$ (Reid & Williamson, 2010). We shall leave hereafter the adjective composite for simplicity, and the link implicit from context whenever appropriate. The unique (up to multiplication or addition by a scalar (Buja et al., 2005)) *canonical link* for a proper loss $\ell : \{-1, 1\} \times [0, 1] \rightarrow \mathbb{R}$ is defined from the conditional Bayes risk as $\psi \doteq -\underline{L}'$ (Reid & Williamson, 2010, Section 6.1), (Buja et al., 2005). As an example, for log-loss we find the canonical link $\psi(u) = \log \frac{u}{1-u}$, with inverse the well-known sigmoid $\psi^{-1}(v) = (1 + e^{-v})^{-1}$. A proper loss will also be assumed to be twice differentiable. Twice differentiability is a technical convenience to simplify derivations. It can be removed (Reid & Williamson, 2010, Footnote 6). A canonical proper loss is a proper loss using the canonical link.

▷ *Adversaries*. Let $\mathcal{A} \subseteq \mathcal{X}^{\mathcal{X}}$ denote a set of adversaries, so that any $a \in \mathcal{A}$ is allowed to transform instances in some way (e.g., change pixel values on an image). Suppose D (fixed) denotes a distribution over $\mathcal{X} \times \{-1, 1\}$ and P (resp. N) is the corresponding distribution conditioned on $\mathsf{Y} = 1$ (resp. $\mathsf{Y} = -1$). The only assumption we make about adversaries is a measurability one. We assume that $\forall h \in \mathcal{H}, \forall a \in \mathcal{A}$, $h \circ a$ is integrable with respect to P and N: $h \circ a \in L^1(\mathcal{X}, \mathrm{d}P) \cap L^1(\mathcal{X}, \mathrm{d}N)$. For the sake of simplicity, we shall denote the tuple $(\mathcal{H}, \mathcal{A})$ *integrable* with respect to D. Assuming loss ℓ is proper composite with link ψ, there is one interesting constant $h^\circ \in \mathbb{R}$:

$$h^\circ \quad \doteq \quad \psi\left(\frac{1}{2}\right), \tag{7}$$

because this value delivers the real valued prediction corresponding to maximal uncertainty in (4). For example, when the loss is proper canonical and furthermore required to be *symmetric*, *i.e.* there is no class-dependent misclassification cost, we have (Nock & Nielsen, 2008)

$$h^\circ \quad = \quad 0, \tag{8}$$

[1] Losses for which *properness* makes particular sense are called class probability estimation losses (Reid & Williamson, 2010).

which corresponds to a classifier always abstaining and indeed delivering maximal uncertainty on prediction. It is not hard to check that $\ell^\circ = \mathsf{E}_{(\mathsf{X},\mathsf{Y})\sim D}[\ell(\mathsf{Y}, h^\circ)]$ is the loss of constant h°. So we can now see that in Definition 1, as $\varepsilon \searrow 0$, training against the adversarial loss essentially produces a classifier no better than predicting nothing. We do not assume that $h^\circ \in \mathcal{H}$, but keep in mind that such prediction with maximal uncertainty is the baseline against which a learner has to compete to "learn" something.

▷ *The adversarial distortion parameter* γ. We now unveil the key parameter used earlier in the Introduction. For any $f \in L^1(\mathcal{X}, \mathrm{d}Q)$, $u, v \in \mathbb{R}$, we let:

$$\varphi(Q, f, u, v) \doteq \int_{\mathcal{X}} u \cdot (f(\boldsymbol{x}) + v)\mathrm{d}Q(\boldsymbol{x}). \quad (9)$$

For any $g : \mathbb{R} \to \mathbb{R}$, the adversarial distortion γ is:

$$\gamma^g_{\mathcal{H},a}(P, N, \pi, b, c) \doteq \max_{h\in\mathcal{H}} \{\varphi(P, g \circ h \circ a, \pi, b) - \varphi(N, g \circ h \circ a, 1 - \pi, -c)\}. \quad (10)$$

Finally, $\gamma_{\mathcal{H},a} \doteq \gamma^{\mathrm{Id}}_{\mathcal{H},a}$. While abstract, we shall shortly see that quantities $\gamma_{\mathcal{H},a}, \gamma^g_{\mathcal{H},a}$ relate to a well-known object in the study of distances between probability distributions. Let

$$\gamma^\ell_{\text{hard}} \doteq \pi\underline{L}(1) + (1 - \pi)\underline{L}(0). \quad (11)$$

As an example, we have for the the log-loss $\gamma^\ell_{\text{hard}} = 0, \forall \pi$, with the convention $0 \log 0 = 0$. We remark that $\gamma^\ell_{\text{hard}}$ in (11) is related to $\gamma^g_{\mathcal{H},a}$ in (10):

$$\gamma^\ell_{\text{hard}} = \gamma^{g^*}_{\mathcal{H}^*,a}(P, N, \pi, 0, 0), \quad (12)$$

for $g^* \doteq \mathsf{Y} \cdot \underline{L}$ and $\mathcal{H}^*$ the singleton classifier which makes the hard prediction 0 over N and 1 over P (Hereafter, we note $\gamma^{g^*}_{\mathcal{H}^*,a}$ instead of $\gamma^{g^*}_{\mathcal{H}^*,a}(P, N, \pi, 0, 0)$ for short). Remark that such a classifier is not affected by a particular adversary, but it is not implementable in the general case as it would require to know the class of an observation.

4. Main result: the hardness theorem

We now show a lower bound on the adversarial loss of (5).

Theorem 2 *For any proper loss ℓ, link ψ and any $(\mathcal{H}, \mathcal{A})$ integrable with respect to D, the following holds true:*

$$\ell(\mathcal{H}, \mathcal{A}, D) \geq \left(\ell^\circ - \frac{1}{2} \cdot \min_{a\in\mathcal{A}} \beta_a\right)_+, \quad (13)$$

where:

$$\begin{aligned} x_+ &\doteq \max\{0, x\}, \\ \beta_a &\doteq \gamma^g_{\mathcal{H},a}(P, N, \pi, 2\underline{L}(1), 2\underline{L}(0)), \\ g &\doteq (-\underline{L}') \circ \psi^{-1}. \end{aligned} \quad (14)$$

(all other parameters implicit in the definition of β_a, Proof in SM, Section 2) This pins down a simple condition for the adversary to defeat $\mathcal{H}$.

Corollary 3 *Under the conditions and with notations of Theorem 2, if there exists $\varepsilon \in [0, 1]$ and $a \in \mathcal{A}$ such that*

$$\beta_a \leq 2\varepsilon\ell^\circ, \quad (15)$$

then $\mathcal{H}$ is ε-defeated by $\mathcal{A}$ on ℓ.

(Proof in SM, Section 2) We remark that whenever ℓ is canonical, $g = \mathrm{Id}$ and so

$$\beta_a = \gamma_{\mathcal{H},a}(P, N, \pi, 2\underline{L}(1), 2\underline{L}(0)). \quad (16)$$

We also note that constants $\underline{L}(0), \underline{L}(1)$ get out of the maximization problem in (10) so when ℓ is canonical, the *optimisation* of $\gamma_{\mathcal{H},a}$ does *not* depend on the loss at hand — hence, its optimisation by an adversary could be done without knowing the loss that the learner is going to minimise. We also remark that the condition for $\mathcal{H}$ to be ε-defeated by $\mathcal{A}$ does not involve an algorithmic component: it means that *any* learning algorithm minimising loss ℓ will end up with a poor predictor if (15) is satisfied, regardless of its computational resources.

▷ *Relationships with integral probability metrics*. In a special case, the somewhat abstract quantity $\gamma^g_{\mathcal{H},a}$ can be related to the more familiar class of integral probability metrics (IPMs) (Sriperumbudur et al., 2009). The latter are a class of metrics on probability distributions, capturing *e.g.* the total variation divergence, Wasserstein distance, and maximum mean discrepancy. The proof of the following Corollary is immediate.

Corollary 4 *Suppose $\gamma^\ell_{hard} = 0$ and $\mathcal{H}$ is closed by negation. Then*

$$2 \cdot \beta_a = \max_{h\in\mathcal{H}} \left| \int_{\mathcal{X}} g \circ h \circ a(\boldsymbol{x})\mathrm{d}P(\boldsymbol{x}) - \int_{\mathcal{X}} g \circ h \circ a(\boldsymbol{x})\mathrm{d}N(\boldsymbol{x}) \right|,$$

which is the integral probability metric for the class $\{g \circ h \circ a : h \in \mathcal{H}\}$ on P and N. Here, g is defined in (14).

We may now interpret Theorem 2 as saying: for an adversary to defeat a learner minimising a proper loss, it suffices to make a suitable IPM between the class-conditionals P, N small. The particular choice of IPM arises from the learner's choice of hypothesis class, $\mathcal{H}$. Of particular interest is when this comprises kernelized scorers, as we now detail.

▷ *Relationships with the maximum mean discrepancy*. The maximum mean discrepancy (MMD) (Gretton et al., 2006) corresponds to an IPM where $\mathcal{H}$ is the unit-ball in an RKHS. We have the following re-expression of $\gamma_{\mathcal{H},a}$ for this hypothesis class, which turns out to involve the MMD.

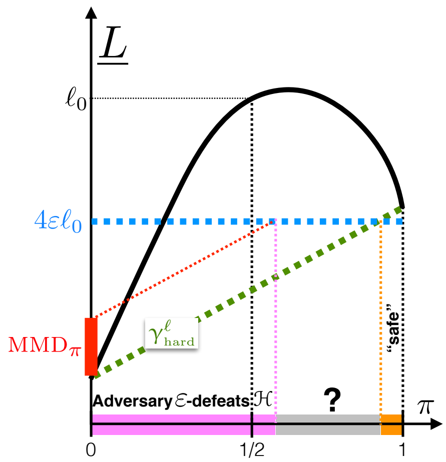


Figure 1: Suppose an adversary a can guarantee an upper-bound on MMD_π as displayed in thick red. For some fixed $\ell(\underline{L})$ and ε, we display the range of π values (in pink) for which a ε-defeats $\mathcal{H}$. Notice that outside this interval, it may not be possible for a to ε-defeat $\mathcal{H}$ (in grey, tagged "?"), and if π is large enough (orange, tagged "safe"), then it is not possible for condition (15) to be satisfied anymore.

Corollary 5 *Suppose ℓ is proper canonical and let $\mathcal{H}$ denote the unit ball of a reproducing kernel Hilbert space (RKHS) of functions with reproducing kernel κ. Denote*

$$\mu_{a,Q} \doteq \int_{\mathcal{X}} \kappa(a(\boldsymbol{x}), .)\mathrm{d}Q(\boldsymbol{x}) \quad (17)$$

the adversarial mean embedding of a on Q. If $\pi = 1/2$ and $\gamma^{\ell}_{hard} = 0$, then

$$2 \cdot \beta_a = \frac{1}{2} \cdot \|\mu_{a,P} - \mu_{a,N}\|_{\mathcal{H}}. \quad (18)$$

The constraints on $\pi, \gamma^{\ell}_{\mathrm{hard}}$ are for readability: the proof (in SM, Section 3) shows a more general result, with unrestricted $\pi, \gamma^{\ell}_{\mathrm{hard}}$. The right-hand side of (18) is proportional to the MMD between P and N. In the more general case, the right-hand side of (18) is replaced by $\mathrm{MMD}_\pi \doteq \|\pi \cdot \mu_{a,P} - (1-\pi) \cdot \mu_{a,N}\|_{\mathcal{H}}$. Figure 1 displays an example picture (for unrestricted $\pi, \gamma^{\ell}_{\mathrm{hard}}$) for some canonical proper but asymmetric loss ($\underline{L}(0) \neq \underline{L}(1)$) when an adversary with a given upperbound guarantee on MMD_π can indeed ε-defeat some $\mathcal{H}$. We remark that while this may be possible for a whole range of π, this may not be possible for all. The picture would be different if the loss were symmetric (Corollary 6 below), since in this case a guarantee to ε-defeat $\mathcal{H}$ for *some* π would imply a guarantee for *all*. Loss asymmetry thus brings a difficulty for the adversary which, we recall, cannot act on π.

▷ *Simultaneously defeating $\mathcal{H}$ over* ***sets*** *of losses.* Satisfying (15) involves at least the knowledge of one value of the loss, if not of the loss itself. It turns out that if the loss is canonical and the adversary has just a partial knowledge of it, it may in fact still be possible for him to guess whether (15) can be satisfied over this set, as we now show.

Corollary 6 *Let $\mathcal{L}$ be a set of canonical proper losses satisfying the following property: $\forall \ell \in \mathcal{L}, \exists \underline{L}^\dagger \in \mathbb{R}$ such that $\underline{L}(1) = \underline{L}(0) \doteq \underline{L}^\dagger$. Assuming $(\mathcal{H}, \mathcal{A})$ integrable with respect to D, if*

$$\exists a \in \mathcal{A} : \gamma_{\mathcal{H},a}(P, N, \pi, 0, 0) \leq \varepsilon \cdot \inf_{\ell \in \mathcal{L}} \ell^\circ - \underline{L}^\dagger (19)$$

then $\mathcal{H}$ is jointly ε-defeated by $\mathcal{A}$ on ***all*** *losses of $\mathcal{L}$.*

Notice that all the adversary needs to know is $\mathcal{L}$. The result easily follows from remarking that we have in this case:

$$\beta_a = 2\underline{L}^\dagger + \gamma_{\mathcal{H},a}(P, N, \pi, 0, 0),$$

which we then plug in (15) to get the statement of the Corollary. Corollary 6 is interesting for two reasons. First, it applies to all proper symmetric losses (Nock & Nielsen, 2008; Reid & Williamson, 2010), which includes popular losses like the square, logistic and Matsushita losses. Finally, it does not just offer the adversarial strategy to defeat classifiers that would be learned on any of such losses, it also applies to more sophisticated learning strategies that would *tune* the loss at learning time (Nock & Nielsen, 2008; Reid & Williamson, 2010) or *tailor* the loss to specific constraints (Buja et al., 2005).

5. Monge efficient adversaries

We now highlight a sufficient condition on adversaries for (15) to be satisfied, which considers classifiers in the increasingly popular framework of "Lipschitz classification" for adversarial training (Cissé et al., 2017), and turns out to frame adversaries in optimal transport (OT) theory (Villani, 2009). We proceed in three steps, first framing OT adversaries, then Lipschitz classifiers and finally showing how the former defeats the latter.

Definition 7 *Given any $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ and some $\delta \in \mathrm{Im}(c)$, we say that $\mathcal{A}$ is δ-**Monge efficient** for cost c on marginals P, N iff $\exists a \in \mathcal{A} : C(a, P, N) \leq \delta$, with*

$$C(a, P, N) \doteq \inf_{\mu \in \Pi(P,N)} \int c(a(\boldsymbol{x}), a(\boldsymbol{x}'))\mathrm{d}\mu(\boldsymbol{x}, \boldsymbol{x}'),$$

and Π is the set of all joint probability measures whose marginals are P and N.

Hence, Monge efficiency relates to an efficient compression of the transport plan between class marginals. In fact, we should require c to satisfy some mild additional assumptions for the existence of optimal couplings (Villani, 2009, Theorem 4.1), such as lower semicontinuity. We skip them for the sake of simplicity, but note that infinite costs are possible without endangering the existence of optimal couplings of (P, N) (Villani, 2009), which is convenient for the following generalized notion of Lipschitz continuity.

Definition 8 *Let $c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$. For some $K > 0$ and $u, v : \mathbb{R} \to \mathbb{R}$, set $\mathcal{H}$ is said to be (u, v, K)-Lipschitz with respect to c iff*

$$u \circ h(\boldsymbol{x}) - v \circ h(\boldsymbol{y}) \leq K \cdot c(\boldsymbol{x}, \boldsymbol{y}), \forall h \in \mathcal{H}, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}. \quad (20)$$

We shall also write that $\mathcal{H}$ is K-Lipschitz if Definition 8 holds for $u = v = \mathrm{Id}$ (c implicit). Actual Lipschitz continuity would restrict c to involve a distance, and the state of the art of adversarial training would restrict further the distance to be based on a norm (Cissé et al., 2017). Equipped with this, we obtain the main result of this Section.

Theorem 9 *Fix any $\varepsilon > 0$ and proper canonical loss ℓ. Suppose $\exists c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that:*

(1) $\mathcal{H}$ is $2K$-Lipschitz with respect to c;

(2) $\mathcal{A}$ is δ-Monge efficient for cost c on marginals P, N for

$$\delta \quad \leq \quad \frac{4\varepsilon\ell^\circ - 2\gamma^\ell_{hard}}{K}. \quad (21)$$

Then $\mathcal{H}$ is ε-defeated by $\mathcal{A}$ on ℓ.

The proof (in SM, Section 4) is given for the more general case where π is not necessarily $1/2$ and any proper loss, not necessarily canonical. We also show in the proof that unless $\pi = 1/2$, c cannot be a distance in the general case. We take it as a potential difficulty for the adversary which, we recall, cannot act on π.

Theorem 9 is particularly interesting with respect to the current developing strategies around adversarial training that "Lipschitzify" classifiers (Cissé et al., 2017). Such strategies assume that the loss ℓ is Lipschitz (remark that we do not make such an assumption). In short, if we rename ℓ_{adv} the inner part (within [.]) in (5), those strategies exploit the fact that (omitting key parameters for readability)

$$\ell_{\text{adv}}(h) \leq \ell_{\text{clean}}(h) + K_\ell K_h, \forall h \in \mathcal{H}, \quad (22)$$

where ℓ_{clean} is the adversary-free loss and $K_.$ is the Lipschitz constant of the loss (ℓ) or classifier learned (h). One might think that minimizing (22) is not a good strategy in the light of Theorem 9 because the regularization enforces a minimization of K_h (K in Theorem 9), so we seemingly alleviate constraints on the adversary to be δ-Monge efficient in (21) and can end up being more easily defeated. This is however a too simplistic conclusion that does not take into account the other parameters at play, as we now explain in the context of Cissé et al. (2017). Consider the logistic loss (Cissé et al., 2017), for which:

$$\ell^\circ = K_\ell = 1, \gamma^\ell_{\text{hard}} = 0. \quad (23)$$

Suppose we can reduce *both* $\ell_{\text{clean}}(h)$ and K_h (which is in fact not hard to ensure for deep architectures (Miyato et al., 2018, Section 2.1), (Cranko et al., 2018)) so that $K_h \leq (1-\ell_{\text{clean}}(h))/2 = (\ell^\circ - \ell_{\text{clean}}(h))/(K_\ell + \ell^\circ)$. Reorganizing, we get $\ell_{\text{clean}}(h) + K_\ell K_h \leq (1 - K_h)\ell^\circ$, so for $\mathcal{H}$ to be ε-defeated, we in fact get a constraint on ε: $\varepsilon \geq K_h$, which reframes the constraint on δ in (21) as (see also SM, (32)),

$$\delta \quad \leq \quad 4\ell^\circ - \frac{\gamma^\ell_{\text{hard}}}{K_h} = 4, \quad (24)$$

which does not depend anymore on K_h.

The proof of Theorem 9 is followed in SM by a proof of an interesting generalization in the light of those recent results (Cissé et al., 2017; Cranko et al., 2018; Miyato et al., 2018): the Monge efficieny requirement can be weakened under a form of dominance (similar to a Lipschitz condition) of the canonical link with respect to the chosen link of the loss. We now provide a simple family of Monge efficient adversaries.

▷ *Mixup adversaries.* Very recently, it was experimentally demonstrated how a simple modification of a training sample yields models more likely to be robust to adversarial examples and generalize better (Zhang et al., 2018). The process can be summarized in a simple way: perform random interpolation between two randomly chosen training examples to create a new example (repeat as necessary). Since we do not allow the adversary to tamper with the class, we define as λ-*mixup* (for $\lambda \in [0, 1]$) the process which creates for two observations $\boldsymbol{x}$ and $\boldsymbol{x}'$ having a different class the following adversarial observation (same class as $\boldsymbol{x}$):

$$a(\boldsymbol{x}) \quad \doteq \quad \lambda \cdot \boldsymbol{x} + (1 - \lambda) \cdot \boldsymbol{x}'. \quad (25)$$

We make the assumption that $\mathcal{X}$ is metric with an associated distance that stems from this metric. We analyze a very simple case of λ-mixup, which we call λ-*mixup to* $\boldsymbol{x}^*$, which replaces $\boldsymbol{x}'$ by some $\boldsymbol{x}^*$ in $\mathcal{X}$ in (25). Notice that as $\lambda \to 0$, we converge to the maximally harmful adversary mentioned in the introduction. The intuition thus suggests that the set $\mathcal{A}$ of all λ-mixups to some $\boldsymbol{x}^*$ (where we vary λ) designs in fact an arbitrarily Monge efficient adversary, where the optimal transport problem involves the associated distance of $\mathcal{X}$. This is indeed true and in fact simple to show.

Lemma 10 *For any $\delta > 0$ the set of all λ-mixups to $\boldsymbol{x}^*$ is δ-Monge efficient for $\lambda \leq \delta / W_1$, where W_1 is the 1-Wasserstein distance between the class marginals.*

(Proof in SM, Section 6) The mixup methodology as defined in (Zhang et al., 2018) can be specialized in numerous ways: for example, instead of mixing up with a single observation, we could perform all possible mixups within $\mathcal{S}$ in a spirit closer to (Zhang et al., 2018), or mixups with several distinguished observations (*e.g.* after clustering), etc. . Many choices like these would be eligible to be at least Monge efficient, but while they can be computationally simple to compute, they are just surrogates for Monge efficiency: tackling directly the compression of the optimal transport plan is a more direct option to Monge efficiency.

6. From weak to strong Monge efficiency

In Theorem 9, we showed how Monge efficiency for adversaries can "take over" Lipschitz classifiers and defeat them for some $\varepsilon > 0$. Suppose now that the $\mathcal{A}$ we have is *weak* in that all its elements are Monge efficient but for large values of δ. In other words, we cannot satisfy condition (2) in Theorem 9. Is there another set of adversaries, $\mathcal{A}^\star$, whose elements would combine the elements of $\mathcal{A}$ is a computationally savvy way, and which would achieve any desired level of Monge efficiency? Such a question parallels that of the boosting framework in supervised learning, in which one combines classifiers just different from random to achieve a combination arbitrarily accurate (Schapire & Freund, 2012).

We now answer our question by the affirmative, in the context of kernel machines. Let $\mathcal{H}$ denote a RKHS and Φ a feature map of the RKHS. $\forall f : \mathcal{X} \rightarrow \mathcal{X}$, define cost

$$
\begin{aligned}
&C_\Phi(f, P, N)\\
&\quad \doteq \inf_{\upmu \in \Pi(P,N)} \int_{\mathcal{X}} \|\Phi \circ f(\boldsymbol{x}) - \Phi \circ f(\boldsymbol{x}')\|_{\mathcal{H}} \mathrm{d}\upmu(\boldsymbol{x}, \boldsymbol{x}').
\end{aligned}
$$

Definition 11 *Function $a : \mathcal{X} \rightarrow \mathcal{X}$ is said η-contractive for Φ, for some $\eta > 0$ iff $\|\Phi \circ a(\boldsymbol{x}) - \Phi \circ a(\boldsymbol{x}')\|_{\mathcal{H}} \leq (1-\eta) \cdot \|\Phi(\boldsymbol{x}) - \Phi(\boldsymbol{x}')\|_{\mathcal{H}}, \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$.*

Set $\mathcal{A}$ is said η-contractive for Φ iff it contains at least one adversary η-contractive for Φ (and we make no assumption on the others). Define now $\mathcal{A}^J \doteq \{a \circ a \circ ... \circ a \ (J \text{ times}) : a \in \mathcal{A}\}$ for any $J \in \mathbb{N}_*$, and $W_1^\Phi \doteq \inf_{\upmu \in \Pi(P,N)} \int_{\mathcal{X}} \|\Phi(\boldsymbol{x}) - \Phi(\boldsymbol{x}')\|_{\mathcal{H}} \mathrm{d}\upmu(\boldsymbol{x}, \boldsymbol{x}')$, the 1-Wasserstein distance between class marginals in the feature map.

Theorem 12 *Let $\mathcal{H}$ denote a RKHS with feature map Φ and $\mathcal{A}$ be η-contractive for Φ. Then $\mathcal{A}$ is δ-Monge efficient for $\delta = (1-\eta) \cdot W_1^\Phi$. Furthermore, $\forall \delta > 0$, $\mathcal{A}^J$ is δ-Monge efficient when $J \geq (1/\eta) \cdot \log(W_1^\Phi / \delta)$.*

(Proof in SM, Section 5) To amplify the difference between $\mathcal{A}$ and $\mathcal{A}^J$, remark that the worst case of Monge efficiency is $\delta = W_1^\Phi$, since it is just the Monge efficiency for contracting nothing. So, as $\eta \rightarrow 0$, there is barely any guarantee we can get from the η-contractive $\mathcal{A}$ while $\mathcal{A}^J$ can still be arbitrarily Monge efficient for a J *linear* in the coding size of the Wasserstein distance between class marginals.

α/d	c/c	c/a	a/c	a/a
0.15	0.03	0.11	**0.00**	**0.02**
0.30	0.03	0.25	**0.00**	0.12
0.45	0.03	0.48	**0.01**	0.55
0.60	0.03	0.74	0.20	0.96

Table 2: log loss USPS results. α/d is the strength of the adversary. The convention {a,c}/{a,c} follows Figure 2. **Bold faces** denote results better than the c/c baseline.

7. Experiments

We have performed toy experiments to demonstrate our new setting. Our objective is not to investigate the competition with respect to the wealth of results that have been recently published in the field, but rather to touch upon the interest that such a novel setting might have for further experimental investigations. Compared to the state-of-the-art, ours is a clear two-stage setting where we first compute the adversaries assuming relevant knowledge of the learner (in our case, we rely on Theorem 12 and therefore assume that the adversary knows at least the cost c, see below), and then we learn based on an adversarially transformed set of examples. This process has the advantage over the direct minimization of (2) that it extracts the computation of the adversarial examples from the training loop: we can generate *once* the adversarial examples, then store them and / or share / reuse them to robustly train various models (recall that under a general Lipschitz assumptions on classifiers, such examples can fit the adversarial training of different kinds of models, see Theorem 9). This process is also reminiscent of the training process for invariant support vector machines (DeCoste & Schölkopf, 2002) and can also be viewed as a particular form of vicinal risk minimization (Chapelle et al., 2000). We have performed two experiments: a 1D experiment involving a particular Mixup adversary and a USPS experiment involving a closer proxy of the optimal transport compression that we call Monge adversary.

▷ *1D experiment, mixup adversary*. Our example involves the unit interval $\mathcal{X} = [0, 1]$ with $P(x) \propto \exp(-(x - 0.2)^2/0.1^2)$ and $N(x) \propto \exp(-(x-0.6)^2/0.2^2)$. We let $\mathcal{A}$ contain a single deterministic mapping parametrised by α as $\mathcal{A} \doteq \{a(x) \doteq (1-\alpha)x + \alpha \mathsf{E}_{(\mathsf{X},\mathsf{Y})\sim D}\mathsf{X}\}$. Notice that this adversary is just the $(1-\alpha)$-mixup to the unconditional mean, following Section 6. We further let $\mathcal{H}$ be the space of linear functions $h(x) = \boldsymbol{w} \cdot (x, 1)^\top$, $\boldsymbol{w} \in \mathbb{R}^2$, which is the RKHS with linear kernel $\kappa(\boldsymbol{x}, \boldsymbol{y}) = \boldsymbol{x} \cdot \boldsymbol{y}$ (assuming that

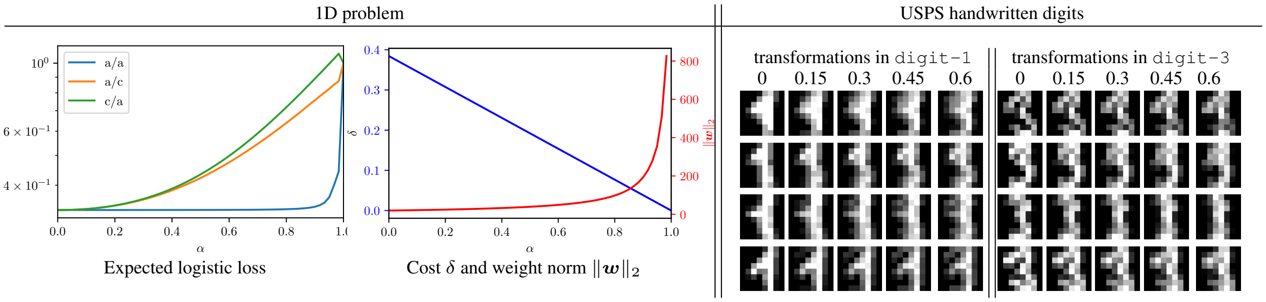


Figure 2: **Left**: results for the 1D toy problem as a function of α. *Left plot:* the expected log loss for the training/testing distribution pairs a/a, a/c, and c/a, where a (respectively c) denotes the adversarial (clean) data distribution. Hence for a/c we optimised the logistic regression classifier on the adversarial distribution, and computed the log loss on the clean distribution. *Right plot:* the optimal transport cost δ (left scale) and the norm of the logistic regression weights $\|\boldsymbol{w}\|_2$ (right scale). **Right**: sample results of `digits` as they are transformed by the OT adversary (convention follows Figure 1).

$\boldsymbol{x}$ and $\boldsymbol{y}$ include the constant 1), and $\|h\|_{\mathcal{H}} = \|\boldsymbol{w}\|_2$. The transport cost function of interest is $c(\boldsymbol{x}, \boldsymbol{y}) = \|\boldsymbol{x} - \boldsymbol{y}\|_2$. We discretize $\mathcal{X}$ to simplify the computation of the OT cost. Results are summarized in Figure 2 (and SM). We theoretically achieve loss ℓ_0 as $\alpha \to 1$. There are several interesting observations from Figure 2: first, the mixup adversary indeed works like a Monge efficient adversary: by tuning α, we can achieve any desired level of Monge efficiency. The left plot completes in this simple case observations of Tsipras et al. (2019); Zhang et al. (2018): the *worst* result is consistently obtained for training on clean data and testing on adversarial data, which indicates that our adversaries may be useful to get robustness using adversarial training.

▷ *USPS digits, Monge adversary*. We have picked 100 examples of each of the "`1`" and "`3`" classes of the 8×8 pixel greyscale USPS handwritten digit dataset. The set of *Monge* adversaries is $\mathcal{A} \doteq \{a : \mathbb{R}^{64} \mapsto \mathbb{R}^{64} \,|\, \|a(x) - x\|_1 \leq \alpha\}$, in which, under the L_1 budget constraint, we optimize the Wasserstein distance W_2^2 between the empirical class marginals. This budgeted optimisation problem is convex and we solve it by combining a generic gradient-free optimiser with a linear program solver. We learn using logistic regression. We demonstrate three strengths of adversary — namely $\alpha/d = 0.15, 0.30, 0.45, 0.6$ where d is L_1 distance between the (clean) class conditional means. Sample transformations as obtained by the Monge adversary are displayed in Figure 2 (more in SM), and Table 2 provides log loss values for different training / test schemes, following the scheme of the 1D data. It clearly emerges two facts: (i) as the budget increases, the Monge adversary smoothly transforms digits in credible adversarial examples, and (ii), as previously observed, training over a tight budget adversary tends to increase generalization abilities (Tsipras et al., 2019; Zhang et al., 2018).

8. Conclusion

It has been observed over the past years that classifiers can be extremely sensitive to small changes in inputs. How such *resource*-constrained changes can affect and be so damaging to machine learning and how to find a cure has been growing as a very intensive area of research. There is so far little understanding on the formal side and some experimental approaches would rely on adversarial data that, in some way, shrinks the gap between classes in a controlled way.

In this paper, we make this intuition formal. Our answer involves a simple, sufficient (and sometimes loss-independent) property for any given class of adversaries to be detrimental to learning. This property involves a measure of "harmfulness", which relates to (and generalizes) integral probability metrics and the maximum mean discrepancy. When classifiers are Lipschitz, this further translates to an optimisation problem (Monge efficiency) which amounts to compressing optimal transport plans, which we believe also defines a new avenue for optimal transport. We also delivered a negative boosting result which shows how weakly contractive adversaries for a RKHS can be combined to build a maximally detrimental adversary.

On the experimental side, we provided a simple toy assessment of the ways one can compute and then use such adversaries in a two-stage process. Our adversaries are simple Monge efficient adversaries that we built to exemplify the use and impact of our theory on toy domains, so we expect that significant improvement can be obtained from the standpoint of designing such efficient adversaries. This is important because training against a weakly activated form of such adversary can improve generalization performances, and it goes with the significant advantage that the examples we generate are reusable and provably adversarial for any class of models satisfying mild conditions.

Acknowledgments and code availability

The authors warmly thank Kamalika Chaudhuri, Giorgio Patrini, Bob Williamson, Xinhua Zhang for numerous remarks and stimulating discussions around this material. Thanks are also due to the reviewers for many suggestions. Code at: https://gitlab.com/machlearn/monge_image_example.

References

Athalye, A., Carlini, N., and Wagner, D.-A. Obfuscated gradients give a false sense of security: Circumventing defenses to adversarial examples. In *35^{th} ICML*, 2018a.

Athalye, A., Engstrom, L., Ilyas, A., and Kwok, K. Synthesizing robust adversarial examples. In *35^{th} ICML*, 2018b.

Bastani, O., Ioannou, Y., Lampropoulos, L., Vytiniotis, D., Nori, A.-V., and Criminisi, A. Measuring neural net robustness with constraints. In *NIPS*29*, 2016.

Bubeck, S., Price, E., and Razenshteyn, I. Adversarial examples from computational constraints. *CoRR*, abs/1805.10204, 2018.

Buckman, J., Roy, A., Raffel, C., and Goodfellow, I. Thermometer encoding: one hot way to resist adversarial examples. In *6^{th} ICLR*, 2018.

Buja, A., Stuetzle, W., and Shen, Y. Loss functions for binary class probability estimation ans classification: structure and applications, 2005. Technical Report, University of Pennsylvania.

Cai, Q.-Z., Du, M., Liu, C., and Song, D. Curriculum adversarial training. In *IJCAI-ECAI'18*, 2018.

Chapelle, O., Weston, J., Bottou, L., and Vapnik, V. Vicinal risk minimization. In *Advances in Neural Information Processing Systems*13*, 2000.

Cissé, M., Bojanowski, P., Grave, E., Dauphin, Y., and Usunier, N. Parseval networks: improving robustness to adversarial examples. In *34^{th} ICML*, 2017.

Cranko, Z., Kornblith, S., Shi, Z., and Nock, R. Lipschitz networks and distributional robustness. *CoRR*, abs/1809.01129, 2018.

DeCoste, D. and Schölkopf, B. Training invariant support vector machines. *Machine Learning*, 46:161–190, 2002.

Dhillon, G., Azizzadenesheli, K., Lipton, Z.-C., Bernstein, J., Kossaifi, J., Khanna, A., and Anandkumar, A. Stochastic activation pruning for robust adversarial defense. In *6^{th} ICLR*, 2018.

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. Adversarial spheres. In *6^{th} ICLR Workshops*, 2018.

Goswami, G., Ratha, N., Agarwal, A., Singh, R., and Vatsa, M. Unravelling robustness of deep learning based face recognition against adversarial attacks. In *AAAI'18*, 2018.

Gretton, A., Borgwardt, K.-M., Rasch, M.-J., Schölkopf, B., and Smola, A.-J. A kernel method for the two-sample-problem. In *NIPS*19*, pp. 513–520, 2006.

Grünwald, P. and Dawid, P. Game theory, maximum entropy, minimum discrepancy and robust Bayesian decision theory. *Ann. of Stat.*, 32:1367–1433, 2004.

Guo, C., Rana, M., Cisse, M., and van der Maaten, L. Countering adversarial images using input transformations. In *6^{th} ICLR*, 2018.

Hein, M. and Andriushchenko, M. Formal guarantees on the robustness of a classifier against adversarial manipulation. In *NIPS*30*, 2017.

Hendrickson, A.-D. and Buehler, R.-J. Proper scores for probability forecasters. *Annals of Mathematical Statistics*, 42:1916–11921, 1971.

Ilias, A., Engstrom, L., Athalye, A., and Lin, J. Adversarial attacks under restricted threat models. In *35^{th} ICML*, 2018.

Kearns, M. and Mansour, Y. On the boosting ability of top-down decision tree learning algorithms. In *Proc. of the 28 th ACM STOC*, pp. 459–468, 1996.

Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial machine learning at scale. In *5^{th} ICLR*, 2017.

Ma, X., Li, B., Wang, Y., Erfani, S.-M., Wijewickrema, S., Schoenebeck, G., Song, D., Houle, M.-E., and Bayley, J. Characterizing adversarial subspaces using local intrinsic dimensionality. In *6^{th} ICLR*, 2018.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *6^{th} ICLR*, 2018.

McCullagh, P. and Nelder, J. *Generalized Linear Models*. Chapman & Hall/CRC, 1989.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. In *ICLR'18*, 2018.

Nock, R. and Nielsen, F. On the efficient minimization of classification-calibrated surrogates. In *NIPS*21*, pp. 1201–1208, 2008.

Raghunathan, A., Steinhardt, J., and Liang, P. Certified defenses against adversarial examples. In *6th ICLR*, 2018.

Reid, M.-D. and Williamson, R.-C. Composite binary losses. *JMLR*, 11:2387–2422, 2010.

Samangouei, P., Kabkab, M., and Chellappa, R. Defense-GAN: protecting classifiers against adversarial attacks using generative models. In *6th ICLR*, 2018.

Schapire, R.-E. and Freund, Y. *Boosting, Foundations and Algorithms*. MIT Press, 2012.

Sinha, A., Namkoong, H., and Duchi, J. Certifying some distributional robustness with principled adversarial training. In *6th ICLR*, 2018.

Song, Y., Kim, T., Nowozin, S., Ermon, S., and Kushman, N. PixelDefend: leveraging generative models to understand and defend against adversarial examples. In *6th ICLR*, 2018.

Sriperumbudur, B.-K., Fukumizu, K., Gretton, A., Schölkopf, B., and Lanckriet, G.-R.-G. On integral probability metrics, φ-divergences and binary classification. *CoRR*, abs/0901.2698, 2009.

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

Tramèr, F., Kurakin, A., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. Ensemble adversarial training: attacks and defenses. In *6th ICLR*, 2018.

Tsipras, D., Santurkar, S., Engstrom, L., Turner, A., and Madry, A. Robustness may be at odds with accuracy. In *7th ICLR*, 2019.

Uesato, J., O'Donoghue, B., Kohli, P., and van den Oord, A. Adversarial risk and the dangers of evaluating against weak attacks. In *35th ICML*, 2018.

Villani, C. *Optimal transport, old and new*. Springer, 2009.

Wang, Y., Jha, S., and Chaudhuri, K. Analyzing the robustness of nearest neighbors to adversarial examples. In *35th ICML*, 2018.

Wong, E. and Zico Kolter, J. Provable defense against adversarial examples via the outer adversarial polytope. In *35th ICML*, 2018.

Zhang, H., Cisse, M., Dauphin, Y.-D., and Lopez-Paz, D. *mixup*: beyond empirical risk minimization. In *6th ICLR*, 2018.

Monge blunts Bayes: Hardness Results for Adversarial Training
— Supplementary Material —

Zac Cranko[♠,†] Aditya Krishna Menon[♡] Richard Nock[†,♠,♣]
Cheng Soon Ong[†,♠] Zhan Shi[◇] Christian Walder[†,♠]

[†]Data61, [♠]the Australian National University, [♡]Google Research
[♣]the University of Sydney, [◇]University of Illinois at Chicago
firstname.lastname@{data61.csiro.au,anu.edu.au}; zshi22@uic.edu

Abstract

This is the Supplementary Material to Paper " Monge blunts Bayes: Hardness Results for Adversarial Training", appearing in the proceedings of ICML 2019.

1 Table of contents

Supplementary material on proofs Pg 3
Proof of Theorem 2 and Corollary 3 Pg 3
Proof sketch of Corollary 5 Pg 6
Proof of Theorem 9 Pg 6
Proof of Theorem 12 Pg 9
Proof of Lemma 10 Pg 9

Supplementary material on experiments Pg 10

2 Proof of Theorem 2 and Corollary 3

Our proof assumes basic knowledge about proper losses (see for example Reid & Williamson (2010)). From (Reid & Williamson, 2010, Theorem 1, Corollary 3) and Shuford et al. (1966), ℓ being twice differentiable and proper, its conditional Bayes risk $\underline{L}$ and partial losses ℓ_1 and ℓ_{-1} are related by:

$$-\underline{L}''(c) = \frac{\ell'_{-1}(c)}{c} = -\frac{\ell'_1(c)}{1-c} \ , \forall c \in (0,1). \tag{1}$$

The weight function (Reid & Williamson, 2010, Theorem 1) being also $w = -\underline{L}''$, we get from the integral representation of partial losses (Reid & Williamson, 2010, eq. (5)),

$$\ell_1(c) \ = \ -\int_c^1 (1-u)\underline{L}''(u)\mathrm{d}u, \tag{2}$$

from which we derive by integrating by parts and then using the Legendre conjugate of $-\underline{L}$,

$$\begin{aligned}
\ell_1(c) + \underline{L}(1) &= -\left[(1-u)\underline{L}'(u)\right]_c^1 - \int_c^1 \underline{L}'(u)\mathrm{d}u + \underline{L}(1) \\
&= (1-c)\underline{L}'(c) + \underline{L}(c) - \underline{L}(1) + \underline{L}(1) \qquad (3)\\
&= -(-\underline{L}')(c) + c\cdot(-\underline{L}')(c) - (-\underline{L})(c) \\
&= -(-\underline{L}')(c) + (-\underline{L})^\star((-\underline{L})'(c)). \qquad (4)
\end{aligned}$$

Now, suppose that the way a real-valued prediction v is fit in the loss is through a general inverse link $\psi^{-1} : \mathbb{R} \rightarrow (0,1)$. Let

$$v_{\ell,\psi} \ \doteq \ (-\underline{L}') \circ \psi^{-1}(v). \tag{5}$$

Since $(-\underline{L})'^{-1}(v_{\ell,\psi}) = \psi^{-1}(v)$, the proper composite loss ℓ with link ψ on prediction v is the same as the proper composite loss ℓ with link $(-\underline{L})'$ on prediction $v_{\ell,\psi}$. This last loss is in fact using its canonical link and so is proper canonical (Reid & Williamson, 2010, Section 6.1), (Buja et al., 2005). Letting in this case $c \doteq (-\underline{L})'^{-1}(v_{\ell,\psi})$, we get that the partial loss satisfies

$$\ell_1(c) \ = \ -v_{\ell,\psi} + (-\underline{L})^\star(v_{\ell,\psi}) - \underline{L}(1). \tag{6}$$

Notice the constant appearing on the right hand side. Notice also that if we see (3) as a Bregman divergence, $\ell_1(c) = (-\underline{L})(1) - (-\underline{L})(c) - ((1-c)(-\underline{L}')(c) = D_{-\underline{L}}(1\|c)$, then the canonical link is the function that defines uniquely the dual affine coordinate system of the divergence (Amari & Nagaoka, 2000) (see also (Reid & Williamson, 2010, Appendix B)).

We can repeat the derivations for the partial loss ℓ_{-1}, which yields (Reid & Williamson, 2010, eq. (5)):

$$\begin{aligned}
\ell_{-1}(c) + \underline{L}(0) &= -\int_0^c u\underline{L}''(u)\mathrm{d}u + \underline{L}(0) \\
&= -\left[u\underline{L}'(u)\right]_0^c + \int_0^c \underline{L}'(u)\mathrm{d}u \\
&= -c\underline{L}'(c) + \underline{L}(c) - \underline{L}(0) + \underline{L}(0) \qquad (7)\\
&= c\cdot(-\underline{L}')(c) - (-\underline{L})(c) \\
&= (-\underline{L})^\star((-\underline{L})'(c)), \qquad (8)
\end{aligned}$$

and using the canonical link, we get this time

$$\ell_{-1}(c) \quad = \quad (-\underline{L})^\star(v_{\ell,\psi}) - \underline{L}(0). \tag{9}$$

We get from (6) and (9) the canonical proper composite loss

$$\ell(y,v) \quad = \quad (-\underline{L})^\star(v_{\ell,\psi}) - \frac{y+1}{2}\cdot v_{\ell,\psi} - \frac{1}{2}\cdot((1-y)\cdot\underline{L}(0) + (1+y)\cdot\underline{L}(1)). \tag{10}$$

Note that for the optimisation of $\ell(y,v)$ for v, we could discount the right-hand side parenthesis, which acts just like a constant with respect to v. Using Fenchel-Young inequality yields the non-negativity of $\ell(y,v)$ as it brings $(-\underline{L})^\star(v_{\ell,\psi}) - ((y+1)/2)\cdot v_{\ell,\psi} \geq \underline{L}((y+1)/2)$ and so

$$\begin{aligned}
\ell(y,v) \quad &\geq \quad \underline{L}\left(\frac{1+y}{2}\right) - \frac{1}{2}\cdot((1-y)\cdot\underline{L}(0) + (1+y)\cdot\underline{L}(1)) \\
&= \underline{L}\left(\frac{1}{2}\cdot(1-y)\cdot 0 + \frac{1}{2}\cdot(1+y)\cdot 1\right) - \frac{1}{2}\cdot((1-y)\cdot\underline{L}(0) + (1+y)\cdot\underline{L}(1)) \\
&\geq \quad 0, \forall y \in \{-1,1\}, \forall v \in \mathbb{R},
\end{aligned} \tag{11}$$

from Jensen's inequality (the conditional Bayes risk $\underline{L}$ is always concave (Reid & Williamson, 2010)). Now, if we consider the alternative use of Fenchel-Young inequality,

$$(-\underline{L})^\star(v_{\ell,\psi}) - \frac{1}{2}\cdot v_{\ell,\psi} \quad \geq \quad \underline{L}\left(\frac{1}{2}\right), \tag{12}$$

then if we let

$$\Delta(y) \quad \doteq \quad \underline{L}\left(\frac{1}{2}\right) - \frac{1}{2}\cdot((1-y)\cdot\underline{L}(0) + (1+y)\cdot\underline{L}(1)), \tag{13}$$

then we get

$$\ell(y,v) \quad \geq \quad \Delta(y) - \frac{y}{2}\cdot v_{\ell,\psi}, \forall y \in \{-1,1\}, \forall v \in \mathbb{R}. \tag{14}$$

It follows from (11) and (14),

$$\ell(y,v) \quad \geq \quad \max\left\{0, \Delta(y) - \frac{y}{2}\cdot v_{\ell,\psi}\right\}, \forall y \in \{-1,1\}, \forall v \in \mathbb{R}, \tag{15}$$

and we get, $\forall h \in \mathbb{R}^{\mathcal{X}}, a \in \mathcal{X}^{\mathcal{X}}$,

$$\begin{aligned}
&\mathsf{E}_{(\mathsf{X},\mathsf{Y})\sim D}[\ell(y, h\circ a(\mathsf{X}))] \\
&\quad \geq \quad \mathsf{E}_{(\mathsf{X},\mathsf{Y})\sim D}\left[\max\left\{0, \Delta(\mathsf{Y}) - \frac{\mathsf{Y}}{2}\cdot(h\circ a)_{\ell,\psi}(\mathsf{X})\right\}\right] \\
&\quad \geq \quad \max\left\{0, \mathsf{E}_{(\mathsf{X},\mathsf{Y})\sim D}\left[\Delta(\mathsf{Y}) - \frac{\mathsf{Y}}{2}\cdot(h\circ a(\mathsf{X}))_{\ell,\psi}\right]\right\} \\
&\quad = \max\left\{0, \underline{L}\left(\frac{1}{2}\right) - \frac{1}{2}\cdot\mathsf{E}_{(\mathsf{X},\mathsf{Y})\sim D}\left[\mathsf{Y}\cdot(h\circ a(\mathsf{X}))_{\ell,\psi} + (1-\mathsf{Y})\cdot\underline{L}(0) + (1+\mathsf{Y})\cdot\underline{L}(1)\right]\right\} \\
&\quad = \quad \max\left\{0, \underline{L}\left(\frac{1}{2}\right) - \frac{1}{2}\cdot\left(\begin{array}{c}\mathsf{E}_{\mathsf{X}\sim P}\left[\pi\cdot((h\circ a(\mathsf{X}))_{\ell,\psi} + 2\underline{L}(1))\right] \\ -\mathsf{E}_{\mathsf{X}\sim N}\left[(1-\pi)\cdot((h\circ a(\mathsf{X}))_{\ell,\psi} - 2\underline{L}(0))\right]\end{array}\right)\right\} \\
&\quad = \quad \max\left\{0, \underline{L}\left(\frac{1}{2}\right) - \frac{1}{2}\cdot(\varphi(P, (h\circ a)_{\ell,\psi}, \pi, 2\underline{L}(1)) - \varphi(N, (h\circ a)_{\ell,\psi}, 1-\pi, -2\underline{L}(0)))\right\},
\end{aligned} \tag{16}$$

with

$$\varphi(Q, f, b, c) \doteq \int_{\mathcal{X}} b \cdot (f(\boldsymbol{x}) + c) \mathrm{d}Q(\boldsymbol{x}), \tag{17}$$

and we recall

$$(h \circ a)_{\ell,\psi} = (-\underline{L}') \circ \psi^{-1} \circ h \circ a. \tag{18}$$

Hence,

$$\begin{aligned}
&\min_{h\in\mathcal{H}} \mathsf{E}_{(\mathsf{X},\mathsf{Y})\sim D}[\max_{a\in\mathcal{A}} \ell(\mathsf{Y}, h\circ a(\mathsf{X}))] \\
&\geq \min_{h\in\mathcal{H}} \max_{a\in\mathcal{A}} \mathsf{E}_{(\mathsf{X},\mathsf{Y})\sim D}[\ell(\mathsf{Y}, h\circ a(\mathsf{X}))] \qquad (19)\\
&\geq \min_{h\in\mathcal{H}} \max_{a\in\mathcal{A}} \max\left\{0, \underline{L}\left(\frac{1}{2}\right) - \frac{1}{2}\cdot(\varphi(P, (h\circ a)_{\ell,\psi}, \pi, 2\underline{L}(1)) - \varphi(N, (h\circ a)_{\ell,\psi}, 1-\pi, -2\underline{L}(0)))\right\}\\
&\geq \max_{a\in\mathcal{A}} \min_{h\in\mathcal{H}} \max\left\{0, \underline{L}\left(\frac{1}{2}\right) - \frac{1}{2}\cdot(\varphi(P, (h\circ a)_{\ell,\psi}, \pi, 2\underline{L}(1)) - \varphi(N, (h\circ a)_{\ell,\psi}, 1-\pi, -2\underline{L}(0)))\right\}\\
&= \max_{a\in\mathcal{A}} \max\left\{0, \min_{h\in\mathcal{H}}\left(\underline{L}\left(\frac{1}{2}\right) - \frac{1}{2}\cdot(\varphi(P, (h\circ a)_{\ell,\psi}, \pi, 2\underline{L}(1)) - \varphi(N, (h\circ a)_{\ell,\psi}, 1-\pi, -2\underline{L}(0)))\right)\right\}\\
&= \max_{a\in\mathcal{A}} \max\left\{0, \underline{L}\left(\frac{1}{2}\right) - \frac{1}{2}\cdot\max_{h\in\mathcal{H}}(\varphi(P, (h\circ a)_{\ell,\psi}, \pi, 2\underline{L}(1)) - \varphi(N, (h\circ a)_{\ell,\psi}, 1-\pi, -2\underline{L}(0)))\right\}\\
&= \max_{a\in\mathcal{A}} \left(\underline{L}\left(\frac{1}{2}\right) - \frac{1}{2}\cdot\max_{h\in\mathcal{H}}(\varphi(P, (h\circ a)_{\ell,\psi}, \pi, 2\underline{L}(1)) - \varphi(N, (h\circ a)_{\ell,\psi}, 1-\pi, -2\underline{L}(0)))\right)_+\\
&= \left(\underline{L}\left(\frac{1}{2}\right) - \frac{1}{2}\cdot\min_{a\in\mathcal{A}}\max_{h\in\mathcal{H}}(\varphi(P, (h\circ a)_{\ell,\psi}, \pi, 2\underline{L}(1)) - \varphi(N, (h\circ a)_{\ell,\psi}, 1-\pi, -2\underline{L}(0)))\right)_+\\
&= \left(\underline{L}\left(\frac{1}{2}\right) - \frac{1}{2}\cdot\min_{a\in\mathcal{A}} \gamma^g_{\mathcal{H},a}(P, N, \pi, 2\underline{L}(1), 2\underline{L}(0))\right)_+\\
&= \left(\ell^\circ - \frac{1}{2}\cdot\min_{a\in\mathcal{A}} \gamma^g_{\mathcal{H},a}(P, N, \pi, 2\underline{L}(1), 2\underline{L}(0))\right)_+\\
&= \left(\ell^\circ - \frac{1}{2}\cdot\min_{a\in\mathcal{A}} \beta_a\right)_+, \qquad (20)
\end{aligned}$$

as claimed for the statement of Theorem 2 (we have let $g \doteq (-\underline{L}') \circ \psi^{-1}$). Hence, if, for some $\varepsilon \in [0, 1]$,

$$\exists a \in \mathcal{A} : \gamma^g_{\mathcal{H},a}(P, N, \pi, 2\underline{L}(1), 2\underline{L}(0)) \leq 2\varepsilon \cdot \ell^\circ, \tag{21}$$

then

$$\begin{aligned}
\min_{h\in\mathcal{H}} \mathsf{E}_{(\mathsf{X},\mathsf{Y})\sim D}[\max_{a\in\mathcal{A}} \ell(\mathsf{Y}, h\circ a(\mathsf{X}))] &\geq (\ell^\circ - \varepsilon\cdot\ell^\circ)_+\\
&= (1-\varepsilon)\cdot\ell^\circ, \qquad (22)
\end{aligned}$$

which ends the proof of Corollary 3 if ℓ is proper composite with link ψ. If it is proper canonical, then $(-\underline{L}') \circ \psi^{-1} = \mathrm{Id}$ and so $\gamma^g_{\mathcal{H},a} = \gamma_{\mathcal{H},a}$ in (21).

Remark 1 *Theorem 2 and Corollary 3 are very general, which naturally questions the optimality of the condition in Corollary 3 to defeat $\mathcal{H}$ – and therefore the optimality of the Monge adversaries to appear later. Inspecting their proof shows that suboptimality comes essentially from the use of Fenchel-Young inequality in (12). There are ways to strenghten this result for subclasses of losses, which might result in fine in the characterisation of different but arguably more specific adversaries.*

3 Proof sketch of Corollary 5

Recall that $\beta_a = \gamma_{\mathcal{H},a}\big(P, N, \frac{1}{2}, 2\underline{L}(1), 2\underline{L}(0)\big)$. We prove the following, more general result which does not assume $\pi = 1/2$ nor $\gamma^{\ell}_{\text{hard}} = 0$.

Corollary 2 *Suppose ℓ is canonical proper and let $\mathcal{H}$ denote the unit ball of a reproducing kernel Hilbert space (RKHS) of functions with reproducing kernel κ. Denote*

$$\mu_{a,Q} \doteq \int_{\mathcal{X}} \kappa(a(\boldsymbol{x}), .)\mathrm{d}Q(\boldsymbol{x}) \tag{23}$$

the adversarial mean embedding of a on Q. Then

$$\begin{aligned}&2 \cdot \gamma_{\mathcal{H},a}(P, N, \pi, 2\underline{L}(1), 2\underline{L}(0))\\ &\quad= \gamma^{\ell}_{\mathit{hard}} + \|\pi \cdot \mu_{a,P} - (1-\pi)\cdot \mu_{a,N}\|_{\mathcal{H}}.\end{aligned}$$

Proof It comes from the reproducing property of $\mathcal{H}$,

$$\begin{aligned}
&2 \cdot \gamma_{\mathcal{H},a}(P, N, \pi, 2\underline{L}(1), 2\underline{L}(0))\\
&= \gamma^{\ell}_{\text{hard}} + \max_{h\in\mathcal{H}} \left\{\pi \cdot \int_{\mathcal{X}} h\circ a(\boldsymbol{x})\mathrm{d}P(\boldsymbol{x}) - (1-\pi)\cdot \int_{\mathcal{X}} h\circ a(\boldsymbol{x})\mathrm{d}N(\boldsymbol{x})\right\}\\
&= \gamma^{\ell}_{\text{hard}} + \max_{h\in\mathcal{H}} \left\{\pi \cdot \left\langle h, \int_{\mathcal{X}} \kappa(a(\boldsymbol{x}), .)\mathrm{d}P(\boldsymbol{x})\right\rangle_{\mathcal{H}} - (1-\pi)\cdot \left\langle h, \int_{\mathcal{X}} \kappa(a(\boldsymbol{x}), .)\mathrm{d}N(\boldsymbol{x})\right\rangle_{\mathcal{H}}\right\}\\
&= \gamma^{\ell}_{\text{hard}} + \max_{h\in\mathcal{H}} \left\{\langle h, \pi\cdot\mu_{a,P} - (1-\pi)\cdot\mu_{a,N}\rangle_{\mathcal{H}}\right\}\\
&= \gamma^{\ell}_{\text{hard}} + \|\pi \cdot \mu_{a,P} - (1-\pi)\cdot \mu_{a,N}\|_{\mathcal{H}},
\end{aligned} \tag{24}$$

as claimed, where the last equality holds for the unit ball. ■

4 Proof of Theorem 9

We first show a Lemma giving some additional properties on our definition os Lipschitzness.

Lemma 3 *Suppose $\mathcal{H}$ is (u, v, K)-Lipschitz. If c is symmetric, then $\{u\circ h - v \circ h\}_{h\in\mathcal{H}}$ is $2K$-Lipschitz. If c satisfies the triangle inequality, then $u - v$ is bounded. If c satisfies the identity of indiscernibles, then $u \leq v$.*

Proof If c is symmetric, then we just add two instances of (20) with $\boldsymbol{x}$ and $\boldsymbol{y}$ permuted, reorganize and get:

$$\begin{aligned} u \circ h(\boldsymbol{x}) - v \circ h(\boldsymbol{y}) + u \circ h(\boldsymbol{y}) - v \circ h(\boldsymbol{x}) &\leq K \cdot (c(\boldsymbol{x}, \boldsymbol{y}) + c(\boldsymbol{y}, \boldsymbol{x})), \forall h \in \mathcal{H}, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}. \\ \Leftrightarrow (u \circ h - v \circ h)(\boldsymbol{x}) - (u \circ h - v \circ h)(\boldsymbol{y}) &\leq 2Kc(\boldsymbol{x}, \boldsymbol{y}), \forall h \in \mathcal{H}, \forall \boldsymbol{x}, \boldsymbol{y} \in \mathcal{X}. \end{aligned}$$

and we get the statement of the Lemma. If c satisfies the triangle inequality, then we add again two instances of (20) but this time as follows:

$$\begin{aligned} u \circ h(\boldsymbol{x}) - v \circ h(\boldsymbol{y}) + u \circ h(\boldsymbol{y}) - v \circ h(\boldsymbol{z}) &\leq K \cdot (c(\boldsymbol{x}, \boldsymbol{y}) + c(\boldsymbol{y}, \boldsymbol{z})), \forall h \in \mathcal{H}, \forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathcal{X}. \\ \Leftrightarrow u \circ h(\boldsymbol{x}) - v \circ h(\boldsymbol{z}) + \Delta(\boldsymbol{y}) &\leq Kc(\boldsymbol{x}, \boldsymbol{z}), \forall h \in \mathcal{H}, \forall \boldsymbol{x}, \boldsymbol{y}, \boldsymbol{z} \in \mathcal{X}, \end{aligned}$$

where $\Delta(\boldsymbol{y}) \doteq u \circ h(\boldsymbol{y}) - v \circ h(\boldsymbol{y})$. If c is finite for at least one couple $(\boldsymbol{x}, \boldsymbol{z})$, then we cannot have $u - v$ unbounded in $\cup_h \mathrm{Im}(h)$. Finally, if c satisfies the identity of indiscernibles, then picking $\boldsymbol{x} = \boldsymbol{y}$ in (20) yields $u \circ h(\boldsymbol{x}) - v \circ h(\boldsymbol{x}) \leq 0, \forall h \in \mathcal{H}, \forall \boldsymbol{x} \in \mathcal{X}$ and so $(u - v)(\cup_h \mathrm{Im}(h)) \cap \mathbb{R}_+ \subseteq \{0\}$, which, disregarding the images in $\mathcal{H}$ for simplicity, yields $u \leq v$. ■

We now prove TheoremthOTA. In fact, we shall prove the following more general Theorem.

Theorem 4 *Fix any $\varepsilon > 0$ and proper loss ℓ with link ψ. Suppose $\exists c : \mathcal{X} \times \mathcal{X} \to \mathbb{R}$ such that:*

(1) $\mathcal{H}$ is $(\pi \cdot g, (1 - \pi) \cdot g, K)$-Lipschitz with respect to c, where g is defined in (14);

(2) $\mathcal{A}$ is δ-Monge efficient for cost c on marginals P, N for

$$\delta \leq 2 \cdot \frac{2\varepsilon\ell^\circ - \gamma^\ell_{hard}}{K}. \tag{25}$$

Then $\mathcal{H}$ is ε-defeated by $\mathcal{A}$ on ℓ.

Proof We have for all $a \in \mathcal{A}$,

$$\begin{aligned} &\max_{h \in \mathcal{H}} (\varphi(P, h \circ a, \pi, 2\underline{L}(1)) - \varphi(N, h \circ a, 1 - \pi, -2\underline{L}(0))) \\ &= \gamma^\ell_{\mathrm{hard}} + \frac{1}{2} \cdot \max_{h \in \mathcal{H}} \left(\int_{\mathcal{X}} \pi \cdot g \circ h \circ a(\boldsymbol{x}) \mathrm{d}P(\boldsymbol{x}) - \int_{\mathcal{X}} (1 - \pi) \cdot g \circ h \circ a(\boldsymbol{x}') \mathrm{d}N(\boldsymbol{x}') \right), \end{aligned} \tag{26}$$

where we recall $g \doteq (-\underline{L}') \circ \psi^{-1}$. Let us denote for short

$$\Delta \doteq \max_{h \in \mathcal{H}} \left(\int_{\mathcal{X}} \pi \cdot g \circ h \circ a(\boldsymbol{x}) \mathrm{d}P(\boldsymbol{x}) - \int_{\mathcal{X}} (1 - \pi) \cdot g \circ h \circ a(\boldsymbol{x}') \mathrm{d}N(\boldsymbol{x}') \right). \tag{27}$$

$\mathcal{H}$ being $(\pi \cdot g, (1 - \pi) \cdot g, K)$-Lipschitz for cost c, since

$$\mathcal{H} \subseteq \{h \in \mathbb{R}^{\mathcal{X}} : \pi g \circ h \circ a(\boldsymbol{x}) - (1 - \pi) g \circ h \circ a(\boldsymbol{x}') \leq Kc(a(\boldsymbol{x}), a(\boldsymbol{x}')), \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}\},$$

it comes after letting for short $\Psi \doteq \pi g \circ h \circ a, \chi \doteq (1 - \pi) g \circ h \circ a$,

$$\begin{aligned} \Delta &\leq \max_{\Psi(\boldsymbol{x}) - \chi(\boldsymbol{x}') \leq Kc(a(\boldsymbol{x}), a(\boldsymbol{x}'))} \left(\int_{\mathcal{X}} \Psi(\boldsymbol{x}) \mathrm{d}P(\boldsymbol{x}) - \int_{\mathcal{X}} \chi(\boldsymbol{x}) \mathrm{d}N(\boldsymbol{x}) \right) \\ &\leq K \cdot \inf_{\mu \in \Pi(P, N)} \int c(a(\boldsymbol{x}), a(\boldsymbol{x}')) \mathrm{d}\mu(\boldsymbol{x}, \boldsymbol{x}'). \end{aligned} \tag{28}$$

See for example (Villani, 2009, Section 4) for the last inequality. Now, if some adversary $a \in \mathcal{A}$ is δ-Monge efficient for cost c, then

$$K \cdot \inf_{\upmu \in \Pi(P,N)} \int c(a(\boldsymbol{x}), a(\boldsymbol{x}'))\mathrm{d}\upmu(\boldsymbol{x}, \boldsymbol{x}') \quad \leq \quad K\delta. \tag{29}$$

From Theorem 2, if we want $\mathcal{H}$ to be ε-defeated by $\mathcal{A}$, then it is sufficient from (26) that a satisfies

$$\gamma^{\ell}_{\text{hard}} + \frac{1}{2} \cdot K\delta \quad \leq \quad 2\varepsilon\ell^{\circ}, \tag{30}$$

resulting in

$$\delta \quad \leq \quad 2 \cdot \frac{2\varepsilon\ell^{\circ} - \gamma^{\ell}_{\text{hard}}}{K}, \tag{31}$$

as claimed. ■

Remark 1 note that unless $\pi = 1/2$, c cannot be a distance in the general case fot Theorem 9: indeed, the identity of indiscernibles and Lemma 2 enforce $(1 - 2\pi) \cdot g \geq 0$ and so g cannot take both signs, which is impossible whenever ℓ is canonical proper as $g = \mathrm{Id}$ in this case. We take it as a potential difficulty for the adversary which, we recall, cannot act on π.

Remark 2 In the light of recent results (Cissé et al., 2017; Cranko et al., 2018; Miyato et al., 2018), there is an interesting corollary to Theorem 9 when $\pi = 1/2$ using a form of Lipschitz continuity of the *link* of the loss .

Corollary 5 *Suppose loss ℓ is proper with link ψ and furthermore its canonical link satisfies, some $K_\ell > 0$:*

$$(\underline{L})'(y) - (\underline{L})'(y') \leq K_\ell \cdot |\psi(y) - \psi(y')|, \forall y, y' \in [0, 1].$$

Suppose furthermore that (i) $\pi = 1/2$, (ii) $\mathcal{H}$ is K_h-Lipschitz with respect to some non-negative c and (iii) $\mathcal{A}$ is δ-Monge efficient for cost c on marginals P, N for

$$\delta \quad \leq \quad \frac{4\varepsilon\ell^{\circ} - 2\gamma^{\ell}_{\textit{hard}}}{K_\ell K_h}. \tag{32}$$

Then $\mathcal{H}$ is ε-defeated by $\mathcal{A}$ on ℓ.

Proof The domination condition on links,

$$(\underline{L})'(y) - (\underline{L})'(y') \quad \leq \quad K_\ell \cdot |\psi(y) - \psi(y')|, \forall y, y' \in [0, 1], \tag{33}$$

implies g is Lipschitz and letting $y \doteq \psi^{-1} \circ h \circ a(\boldsymbol{x})$, $y' \doteq \psi^{-1} \circ h \circ a(\boldsymbol{x}')$, we obtain equivalently $g \circ h \circ a(\boldsymbol{x}) - g \circ h \circ a(\boldsymbol{x}) \leq K_\ell \cdot |h \circ a(\boldsymbol{x}) - h \circ a(\boldsymbol{x}')|, \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$. But $\mathcal{H}$ is K_h-Lipschitz with respect to some non-negative c, so we have $|h \circ a(\boldsymbol{x}) - h \circ a(\boldsymbol{x}')| \leq K_h c(a(\boldsymbol{x}), a(\boldsymbol{x}'))$, and so bringing these two inequalities together, we have from the proof of Theorem 9 that Δ now satisfies

$$\Delta \quad \leq \quad \frac{K_\ell K_h}{2} \cdot \inf_{\upmu \in \Pi(P,N)} \int c(a(\boldsymbol{x}), a(\boldsymbol{x}'))\mathrm{d}\upmu(\boldsymbol{x}, \boldsymbol{x}'), \tag{34}$$

so to be ε-defeated by $\mathcal{A}$ on ℓ, we now want that a satisfies

$$\gamma^{\ell}_{\text{hard}} + \frac{K_\ell K_h}{2} \cdot \delta \ \ \leq \ \ 2\varepsilon\ell^\circ, \tag{35}$$

resulting in the statement of the Corollary. ■

5 Proof of Theorem 12

Denote $a^J \doteq a \circ a \circ ... \circ a$ (J times). We have by definition

$$\begin{aligned}
C_\Phi(a^J, P, N) &\doteq \inf_{\mu \in \Pi(P,N)} \int_{\mathcal{X}} \|\Phi \circ a^J(\boldsymbol{x}) - \Phi \circ a^J(\boldsymbol{x}')\|_{\mathcal{H}} \mathrm{d}\mu(\boldsymbol{x}, \boldsymbol{x}') \\
&= \inf_{\mu \in \Pi(P,N)} \int_{\mathcal{X}} \|\Phi \circ a \circ a^{J-1}(\boldsymbol{x}) - \Phi \circ a \circ a^{J-1}(\boldsymbol{x}')\|_{\mathcal{H}} \mathrm{d}\mu(\boldsymbol{x}, \boldsymbol{x}') \qquad (36) \\
&\leq (1-\eta) \cdot \inf_{\mu \in \Pi(P,N)} \int_{\mathcal{X}} \|\Phi \circ a^{J-1}(\boldsymbol{x}) - \Phi \circ a^{J-1}(\boldsymbol{x}')\|_{\mathcal{H}} \mathrm{d}\mu(\boldsymbol{x}, \boldsymbol{x}') \\
&\ \vdots \\
&\leq (1-\eta)^J \cdot \inf_{\mu \in \Pi(P,N)} \int_{\mathcal{X}} \|\Phi(\boldsymbol{x}) - \Phi(\boldsymbol{x}')\|_{\mathcal{H}} \mathrm{d}\mu(\boldsymbol{x}, \boldsymbol{x}') \\
&\quad = (1-\eta)^J \cdot W_1^\Phi, \qquad (37)
\end{aligned}$$

where we have used the assumption that a is η-contractive and the definition of W_1^Φ. There remains to bound the last line by δ and solve for J to get the statement of the Theorem. We can also stop at (36) to conclude that $\mathcal{A}$ is δ-Monge efficient for $\delta = (1-\eta) \cdot W_1^\Phi$. The number of iterations for $\mathcal{A}^J$ to be δ-Monge efficient is obtained from (37) as

$$J \ \ \geq \ \ \frac{1}{\log\left(\frac{1}{1-\eta}\right)} \cdot \log \frac{W_1^\Phi}{\delta}, \tag{38}$$

which gives the statement of the Theorem once we remark that $\log(1/(1-\eta)) \geq \eta$.

6 Proof of Lemma 10

The proof follows from the observation that for any $\boldsymbol{x}, \boldsymbol{x}'$ in $\mathcal{S}$,

$$\|a(\boldsymbol{x}) - a(\boldsymbol{x}')\| \ \ = \ \ \lambda \|\boldsymbol{x} - \boldsymbol{x}'\|, \tag{39}$$

where $\|.\|$ is the metric of $\mathcal{X}$. Thus, letting a denote a mixup to $\boldsymbol{x}^*$ adversary for some $\lambda \in [0,1]$, we have $C(a, P, N) = \lambda \cdot W_1(\mathrm{d}P, \mathrm{d}N)$, where $W_1(\mathrm{d}P, \mathrm{d}N)$ denotes the Wasserstein distance of order 1 between the class marginals. $\delta > 0$ being fixed, all mixups to $\boldsymbol{x}^*$ adversaries in $\mathcal{A}$ that are also δ-Monge efficient are those for which:

$$\lambda \ \ \leq \ \ \frac{\delta}{W_1(\mathrm{d}P, \mathrm{d}N)}, \tag{40}$$

and we get the statement of the Lemma.

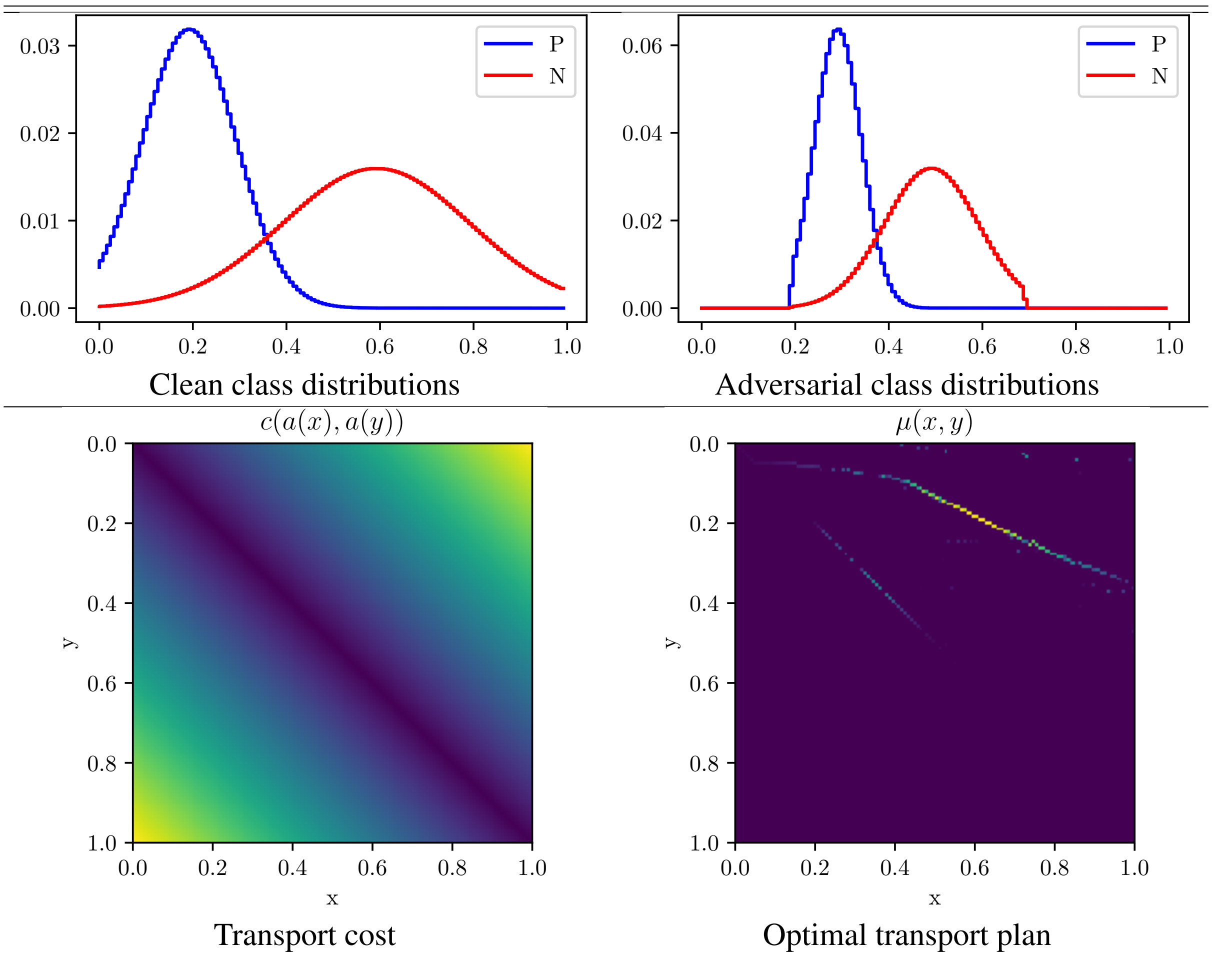


Figure 1: Visualising the toy example for the case $\alpha = 0.5$. Clockwise from top left: (a) the clean class conditional distributions, (b) the class distributions mapped by the adversary a, (c) the transport cost c under the adversarial mapping a, (d) the corresponding optimal transport μ.

7 Experiments

Figure 1 includes detailed plots for the $\alpha = 0.5$ case of the numerical toy example.

References

Amari, S.-I. and Nagaoka, H. *Methods of Information Geometry*. Oxford University Press, 2000.

Buja, A., Stuetzle, W., and Shen, Y. Loss functions for binary class probability estimation ans classification: structure and applications, 2005. Technical Report, University of Pennsylvania.

Cissé, M., Bojanowski, P., Grave, E., Dauphin, Y., and Usunier, N. Parseval networks: improving robustness to adversarial examples. In *34th ICML*, 2017.

Cranko, Z., Kornblith, S., Shi, Z., and Nock, R. Lipschitz networks and distributional robustness. *CoRR*, abs/1809.01129, 2018.

Miyato, T., Kataoka, T., Koyama, M., and Yoshida, Y. Spectral normalization for generative adversarial networks. In *ICLR'18*, 2018.

Reid, M.-D. and Williamson, R.-C. Composite binary losses. *JMLR*, 11:2387–2422, 2010.

Shuford, E., Albert, A., and Massengil, H.-E. Admissible probability measurement procedures. *Psychometrika*, pp. 125–145, 1966.

Villani, C. *Optimal transport, old and new*. Springer, 2009.