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

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 $\psi^{-1} : \mathbb{R} \to (0,1)$. 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

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

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

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

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

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

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,

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

then if we let

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

then we get

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

It follows from (11) and (14),

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

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

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

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}} \boldsymbol{\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}} \boldsymbol{\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 $\varepsilon \in [0, 1]$,

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

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 $\boldsymbol{\gamma}^g_{\mathcal{H},a} = \boldsymbol{\gamma}_{\mathcal{H},a}$ 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

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

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 $a \in \mathcal{A}$,

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

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

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

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

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

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

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

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.

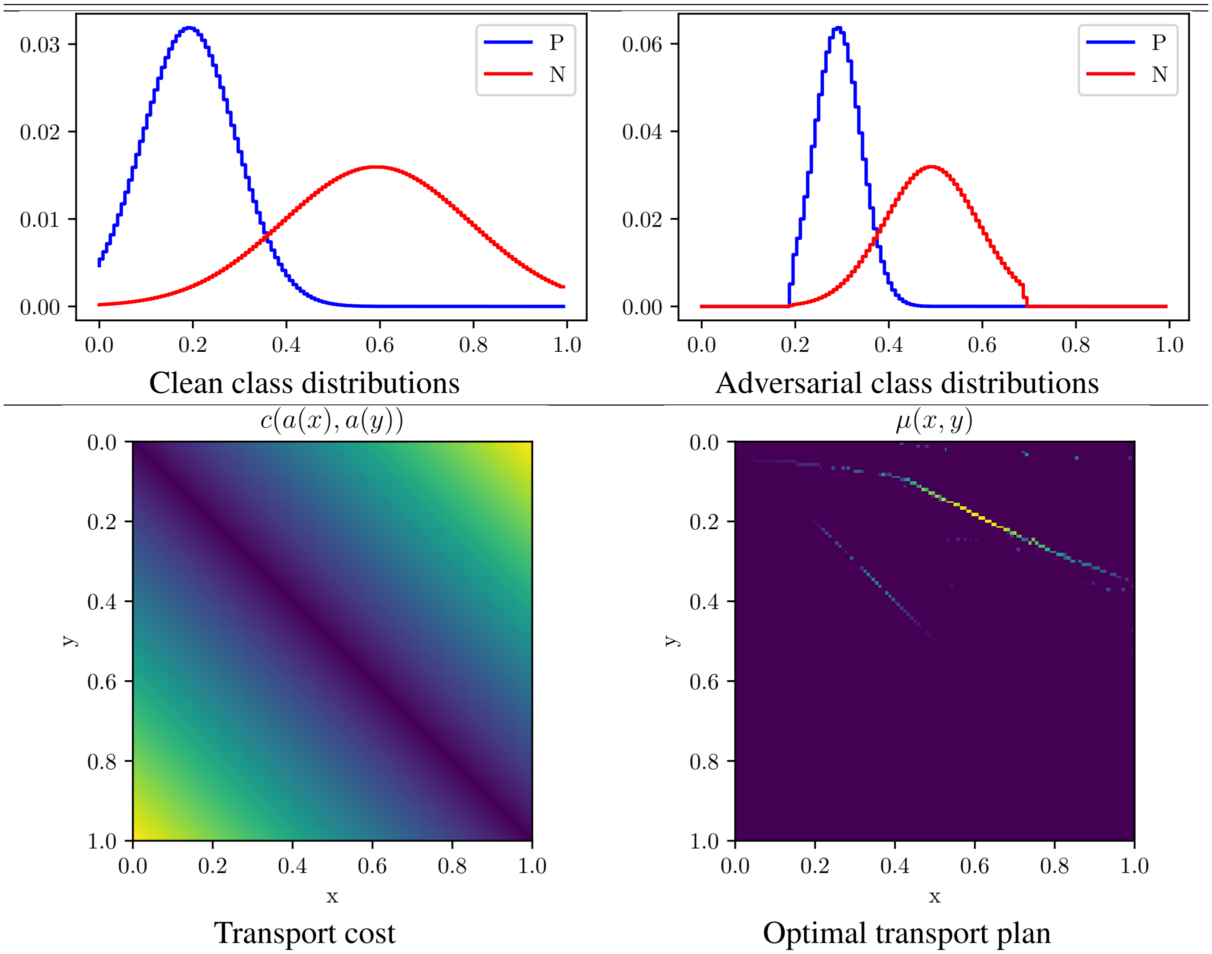

Figure 1: Visualising the toy example for the case $\alpha = 0.5$. Clockwise from top left: (a) the clean class conditional distributions, (b) the class distributions mapped by the adversary a, (c) the transport cost c under the adversarial mapping a, (d) the corresponding optimal transport μ.

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.