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 7 and Corollary 8 ............................................................... 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), $\ell$ being twice differentiable and proper, its conditional Bayes risk $L$ and partial losses $\ell_1$ and $\ell_{-1}$ are related by:

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

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

$$\ell_1(c) = -\int_c^1 (1-u)L''(u)du,$$ (2)

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

$$\ell_1(c) + L(1) = -[(1-u)L'(u)]_c^1 - \int_c^1 L'(u)du + L(1)$$

$$= (1-c)L(c) + L(c) - L(1) + L(1)$$

$$= -(-L)(c) + c \cdot (-L')(c) - (-L)(c)$$

$$= -(L)(c) + (-L)^*((-L)'(c)).$$ (3)

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

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

Since $(-L)^{-1}(v_{\ell, \psi}) = \psi^{-1}(v)$, the proper composite loss $\ell$ with link $\psi$ on prediction $v$ is the same as the proper composite loss $\ell$ with link $(-L)'$ 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 $c \doteq (-L)^{-1}(v_{\ell, \psi})$, we get that the partial loss satisfies

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

Notice the constant appearing on the right hand side. Notice also that if we see (3) as a Bregman divergence, $\ell_1(c) = (-L)(1) - (-L)(c) - ((1-c)(-L')(c) = D_{-L}(1\|c)$, 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 $\ell_{-1}$, which yields (Reid & Williamson 2010, eq. (5)):

$$\ell_{-1}(c) + L(0) = -\int_0^c u L''(u)du + L(0)$$

$$= -[uL'(u)]_0^c + \int_0^c L'(u)du$$

$$= -cL'(c) + L(c) - L(0) + L(0)$$

$$= c \cdot (-L')(c) - (-L)(c)$$

$$= (-L)^*((-L)'(c)),$$ (8)
and using the canonical link, we get this time
\[ \ell^{-1}(c) = (-L)^*(v_{\ell,\psi}) - L(0). \]  

(9)

We get from (6) and (9) the canonical proper composite loss
\[ \ell(y, v) = (-L)^*(v_{\ell,\psi}) - \frac{y + 1}{2} \cdot v_{\ell,\psi} - \frac{1}{2} \cdot ((1 - y) \cdot L(0) + (1 + y) \cdot L(1)). \]

(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 \((-L)^*(v_{\ell,\psi}) - ((y + 1)/2) \cdot v_{\ell,\psi} \geq L((y + 1)/2)\) and so
\[ \ell(y, v) \geq L \left( \frac{1 + y}{2} \right) - \frac{1}{2} \cdot ((1 - y) \cdot L(0) + (1 + y) \cdot L(1)) \]
\[ = L \left( \frac{1}{2} \cdot (1 - y) \cdot 0 + \frac{1}{2} \cdot (1 + y) \cdot 1 \right) - \frac{1}{2} \cdot ((1 - y) \cdot L(0) + (1 + y) \cdot L(1)) \]
\[ \geq 0, \forall y \in \{-1, 1\}, \forall v \in \mathbb{R}, \]

(11)

from Jensen’s inequality (the conditional Bayes risk \( L \) is always concave (Reid & Williamson, 2010)). Now, if we consider the alternative use of Fenchel-Young inequality,
\[ (-L)^*(v_{\ell,\psi}) - \frac{1}{2} \cdot v_{\ell,\psi} \geq L \left( \frac{1}{2} \right), \]

(12)

then if we let
\[ \Delta(y) \equiv L \left( \frac{1}{2} \right) - \frac{1}{2} \cdot ((1 - y) \cdot L(0) + (1 + y) \cdot L(1)), \]

(13)

then we get
\[ \ell(y, v) \geq \Delta(y) - \frac{y}{2} \cdot v_{\ell,\psi}, \forall y \in \{-1, 1\}, \forall v \in \mathbb{R}. \]

(14)

It follows from (11) and (14),
\[ \ell(y, v) \geq \max \left\{ 0, \Delta(y) - \frac{y}{2} \cdot v_{\ell,\psi} \right\}, \forall y \in \{-1, 1\}, \forall v \in \mathbb{R}, \]

(15)

and we get, \( \forall h \in \mathbb{R}^X, a \in \mathcal{X}^X, \)
\[ E_{(X,Y) \sim D}[\ell(y, h \circ a(X))]
\]
\[ \geq E_{(X,Y) \sim D} \left[ \max \left\{ 0, \Delta(Y) - \frac{Y}{2} \cdot (h \circ a)_{\ell,\psi}(X) \right\} \right]
\]
\[ \geq \max \left\{ 0, E_{(X,Y) \sim D} \left[ \Delta(Y) - \frac{Y}{2} \cdot (h \circ a(X))_{\ell,\psi} \right] \right\}
\]
\[ = \max \left\{ 0, L \left( \frac{1}{2} \right) - \frac{1}{2} \cdot E_{(X,Y) \sim D} [Y \cdot (h \circ a(X))_{\ell,\psi} + (1 - Y) \cdot L(0) + (1 + Y) \cdot L(1)] \right\}
\]
\[ = \max \left\{ 0, L \left( \frac{1}{2} \right) - \frac{1}{2} \cdot \left( \begin{array}{c} \mathbb{E}_{X \sim P} \left[ \pi \cdot ((h \circ a(X))_{\ell,\psi} + 2L(1)) \right] \\ - \mathbb{E}_{X \sim N} \left[ (1 - \pi) \cdot ((h \circ a(X))_{\ell,\psi} - 2L(0)) \right] \end{array} \right) \right\}
\]
\[ = \max \left\{ 0, L \left( \frac{1}{2} \right) - \frac{1}{2} \cdot (\varphi(P, (h \circ a)_{\ell,\psi}, \pi, 2L(1)) - \varphi(N, (h \circ a)_{\ell,\psi}, 1 - \pi, -2L(0))) \right\} \]

(16)
with

$$\varphi(Q, f, b, c) = \int X b \cdot (f(x) + c) dQ(x),$$

and we recall

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

(18)

Hence,

\[
\begin{align*}
\min_{h \in \mathcal{H}} E_{(X,Y) \sim D}[\max_{a \in A} \ell(Y, h \circ a(X))] \\
& \geq \min_{h \in \mathcal{H}} \max_{a \in A} E_{(X,Y) \sim D}[\ell(Y, h \circ a(X))] \\
& \geq \min_{h \in \mathcal{H}} \max_{a \in A} \left\{ 0, L \left( \frac{1}{2} \right) - \frac{1}{2} \cdot (\varphi(P, (h \circ a)_{\ell, \psi}, \pi, 2L(1)) - \varphi(N, (h \circ a)_{\ell, \psi}, 1 - \pi, -2L(0))) \right\} \\
& \geq \max_{a \in A} \min_{h \in \mathcal{H}} \left\{ 0, L \left( \frac{1}{2} \right) - \frac{1}{2} \cdot (\varphi(P, (h \circ a)_{\ell, \psi}, \pi, 2L(1)) - \varphi(N, (h \circ a)_{\ell, \psi}, 1 - \pi, -2L(0))) \right\} \\
& = \max_{a \in A} \left( L \left( \frac{1}{2} \right) - \frac{1}{2} \cdot \min_{h \in \mathcal{H}} (\varphi(P, (h \circ a)_{\ell, \psi}, \pi, 2L(1)) - \varphi(N, (h \circ a)_{\ell, \psi}, 1 - \pi, -2L(0))) \right)_+ \\
& = \left( L \left( \frac{1}{2} \right) - \frac{1}{2} \cdot \min_{a \in A} \gamma^g_{\mathcal{H}, a}(P, N, \pi, 2L(1), 2L(0)) \right)_+ \\
& = \left( \ell^0 - \frac{1}{2} \cdot \min_{a \in A} \beta_{a} \right)_+, \\
\end{align*}
\]

(20)

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

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

(21)

then

$$\begin{align*}
\min_{h \in \mathcal{H}} E_{(X,Y) \sim D}[\max_{a \in A} \ell(Y, h \circ a(X))] & \geq (\ell^0 - \varepsilon \cdot \ell^0)_+ \\
& = (1 - \varepsilon) \cdot \ell^0,
\end{align*}$$

(22)

which ends the proof of Corollary 3 if $\ell$ is proper composite with link $\psi$. If it is proper canonical, then $(-L') \circ \psi^{-1} = \text{Id}$ 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 strengthen 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 $\beta_a = \gamma_{\mathcal{H},a}(P,N,\frac{1}{2},2L(1),2L(0))$. We prove the following, more general result which does not assume $\pi = 1/2$ nor $\gamma_{\text{hard}} = 0$.

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

$$\mu_{a,Q} \equiv \int_X \kappa(a(x),.)dQ(x)$$

(23)

the adversarial mean embedding of $a$ on $Q$. Then

$$2 \cdot \gamma_{\mathcal{H},a}(P,N,\pi,2L(1),2L(0)) = \gamma_{\text{hard}}^\ell + \|\pi \cdot \mu_{a,P} - (1-\pi) \cdot \mu_{a,N}\|_{\mathcal{H}}.$$  

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

$$2 \cdot \gamma_{\mathcal{H},a}(P,N,\pi,2L(1),2L(0))$$

$$= \gamma_{\text{hard}}^\ell + \max_{h \in \mathcal{H}} \left\{ \pi \cdot \int_X h \circ a(x)dP(x) - (1-\pi) \cdot \int_X h \circ a(x)dN(x) \right\}$$

$$= \gamma_{\text{hard}}^\ell + \max_{h \in \mathcal{H}} \left\{ \pi \cdot \left\langle h, \int_X \kappa(a(x),.)dP(x) \right\rangle_{\mathcal{H}} - (1-\pi) \cdot \left\langle h, \int_X \kappa(a(x),.)dN(x) \right\rangle_{\mathcal{H}} \right\}$$

$$= \gamma_{\text{hard}}^\ell + \max_{h \in \mathcal{H}} \left\{ \langle h, \pi \cdot \mu_{a,P} - (1-\pi) \cdot \mu_{a,N} \rangle_{\mathcal{H}} \right\}$$

$$= \gamma_{\text{hard}}^\ell + \|\pi \cdot \mu_{a,P} - (1-\pi) \cdot \mu_{a,N}\|_{\mathcal{H}},$$

(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 of 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 $u \leq v$.  

6
We now prove Theorem thOTA. In fact, we shall prove the following more general Theorem.

\[ u \circ h(x) - v \circ h(y) + u \circ h(y) - v \circ h(x) \leq K \cdot (c(x, y) + c(y, x)), \forall h \in \mathcal{H}, \forall x, y \in \mathcal{X}. \]

\[ (u \circ h - v \circ h)(x) - (u \circ h - v \circ h)(y) \leq 2Kc(x, y), \forall h \in \mathcal{H}, \forall x, y \in \mathcal{X}. \]

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:

\[ u \circ h(x) - v \circ h(y) + u \circ h(y) - v \circ h(z) \leq K \cdot (c(x, y) + c(y, z)), \forall h \in \mathcal{H}, \forall x, y, z \in \mathcal{X}. \]

\[ u \circ h(x) - v \circ h(z) + \Delta(y) \leq Kc(x, z), \forall h \in \mathcal{H}, \forall x, y, z \in \mathcal{X}, \]

where \( \Delta(y) \equiv u \circ h(y) - v \circ h(y) \). If \( c \) is finite for at least one couple \( (x, z) \), then we cannot have \( u - v \) unbounded in \( \bigcup_h \text{Im}(h) \).

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

**Theorem 4** Fix any \( \varepsilon > 0 \) and proper loss \( \ell \) with link \( \psi \). 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. \( A \) is \( \delta \)-Monge efficient for cost \( c \) on marginals \( P, N \) for

\[ \delta \leq 2 \cdot \frac{2\varepsilon c^0 - \gamma^\ell_{\text{hard}}}{K}. \] (25)

Then \( \mathcal{H} \) is \( \varepsilon \)-defeated by \( A \) on \( \ell \).

**Proof** We have for all \( a \in \mathcal{A} \),

\[
\max_{h \in \mathcal{H}} \left( \phi(P, h \circ a, \pi, 2L(1)) - \phi(N, h \circ a, 1 - \pi, -2L(0)) \right) = \gamma^\ell_{\text{hard}} + \frac{1}{2} \max_{h \in \mathcal{H}} \left( \int_{\mathcal{X}} \pi \cdot g \circ h \circ a(x) dP(x) - \int_{\mathcal{X}} (1 - \pi) \cdot g \circ h \circ a(x') dN(x') \right),
\]

(26)

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

\[ \Delta \equiv \max_{h \in \mathcal{H}} \left( \int_{\mathcal{X}} \pi \cdot g \circ h \circ a(x) dP(x) - \int_{\mathcal{X}} (1 - \pi) \cdot g \circ h \circ a(x') dN(x') \right). \] (27)

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

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

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

\[ \Delta \leq \max_{\Psi(x) - \chi(x) \leq Kc(a(x), a(x'))} \left( \int_{\mathcal{X}} \Psi(x) dP(x) - \int_{\mathcal{X}} \chi(x) dN(x) \right) \]

\[ \leq K \cdot \inf_{\mu \in \Pi(P, N)} \int c(a(x), a(x')) d\mu(x, x'). \] (28)
See for example [Villani 2009, Section 4] for the last inequality. Now, if some adversary \( a \in A \) is \( \delta \)-Monge efficient for cost \( c \), then

\[
K \cdot \inf_{\mu \in \Pi(P,N)} \int c(a(x), a(x')) \, d\mu(x, x') \leq K \delta. 
\]  

(29)

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

\[
\gamma_{\text{hard}}^\ell + \frac{1}{2} \cdot K \delta \leq 2\varepsilon \ell^o,
\]

resulting in

\[
\delta \leq \frac{2 \varepsilon \ell^o - \gamma_{\text{hard}}^\ell}{K},
\]

(31)

as claimed.

\[ \blacksquare \]

**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 \( \ell \) is canonical proper as \( g = \text{Id} \) in this case. We take it as a potential difficulty for the adversary which, we recall, cannot act on \( \pi \).

**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 \( \ell \) is proper with link \( \psi \) and furthermore its canonical link satisfies, some \( K_{\ell} > 0 \):

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

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

\[
\delta \leq \frac{4\varepsilon \ell^o - 2\gamma_{\text{hard}}^\ell}{K_{\ell}K_h}.
\]

(32)

Then \( \mathcal{H} \) is \( \varepsilon \)-defeated by \( A \) on \( \ell \).

**Proof** The domination condition on links,

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

(33)

implies \( g \) is Lipschitz and letting \( y \doteq \psi^{-1} \circ h \circ a(x), y' \doteq \psi^{-1} \circ h \circ a(x') \), we obtain equivalently

\[
g \circ h \circ a(x) - g \circ h \circ a(x) \leq K_{\ell} \cdot |h \circ a(x) - h \circ a(x')|, \forall x, x' \in X. \]

But \( \mathcal{H} \) is \( K_h \)-Lipschitz with respect to some non-negative \( c \), so we have \(|h \circ a(x) - h \circ a(x')| \leq K_h c(a(x), a(x')) \). But bringing these two inequalities together, we have from the proof of Theorem 9 that \( \Delta \) now satisfies

\[
\Delta \leq \frac{K_{\ell}K_h}{2} \cdot \inf_{\mu \in \Pi(P,N)} \int c(a(x), a(x')) \, d\mu(x, x'),
\]

(34)
so to be \( \varepsilon \)-defeated by \( \mathcal{A} \) on \( \ell \), we now want that \( a \) satisfies

\[
\gamma_{\text{hard}}^\ell + \frac{K_t K_h}{2} \cdot \delta \leq 2\varepsilon \varepsilon^\ell, \tag{35}
\]
resulting in the statement of the Corollary.

\section{Proof of Theorem 12}

Denote \( a^J = a \circ a \circ \ldots \circ a \) (\( J \) times). We have by definition

\[
C_\Phi(a^J, P, N) = \inf_{\mu \in \Pi(P, N)} \int_X \| \Phi \circ a^J(x) - \Phi \circ a^J(x') \|_{\mathcal{H}} d\mu(x, x')
\]
\[
= \inf_{\mu \in \Pi(P, N)} \int_X \| \Phi \circ a \circ a^{J-1}(x) - \Phi \circ a \circ a^{J-1}(x') \|_{\mathcal{H}} d\mu(x, x')
\]
\[
\leq (1 - \eta) \cdot \inf_{\mu \in \Pi(P, N)} \int_X \| \Phi \circ a^{J-1}(x) - \Phi \circ a^{J-1}(x') \|_{\mathcal{H}} d\mu(x, x')
\]
\[
\vdots
\]
\[
\leq (1 - \eta)^J \cdot \inf_{\mu \in \Pi(P, N)} \int_X \| \Phi(x) - \Phi(x') \|_{\mathcal{H}} d\mu(x, x')
\]
\[
= (1 - \eta)^J \cdot W_1^\Phi, \tag{37}
\]
where we have used the assumption that \( a \) is \( \eta \)-contractive and the definition of \( W_1^\Phi \). There remains to bound the last line by \( \delta \) and solve for \( J \) to get the statement of the Theorem. We can also stop at (36) to conclude that \( \mathcal{A} \) is \( \delta \)-Monge efficient for \( \delta = (1 - \eta) \cdot W_1^\Phi \). The number of iterations for \( \mathcal{A}^J \) to be \( \delta \)-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 \( \log(1/(1 - \eta)) \geq \eta \).

\section{Proof of Lemma 10}

The proof follows from the observation that for any \( x, x' \) in \( S \),

\[
\| a(x) - a(x') \| = \lambda \| x - x' \|, \tag{39}
\]
where \( \| \cdot \| \) is the metric of \( \mathcal{X} \). Thus, letting \( a \) denote a mixup to \( x^* \) adversary for some \( \lambda \in [0, 1] \), we have \( C(a, P, N) = \lambda \cdot W_1(dP, dN) \), where \( W_1(dP, dN) \) denotes the Wasserstein distance of order 1 between the class marginals. \( \delta > 0 \) being fixed, all mixups to \( x^* \) adversaries in \( \mathcal{A} \) that are also \( \delta \)-Monge efficient are those for which:

\[
\lambda \leq \frac{\delta}{W_1(dP, dN)}, \tag{40}
\]
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 $\mu$.

### 7 Experiments

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

### References


