Monge blunts Bayes:
Hardness Results for Adversarial Training

Zac Cranko

Australian National University

CSIRO DATA 61

Aditya Krishna Menon

Google AI

Richard Nock

CSIRO DATA 61

Australian National University

THE UNIVERSITY OF SYDNEY

Cheng Soon Ong

CSIRO DATA 61

Australian National University

Zhan Shi

UIC

Christian Walder

CSIRO DATA 61

Australian National University

Overview

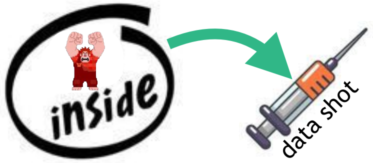
- Hardness results on adversarial training. ***Key result*** applicable to a learner:
  - optimising ***any*** loss satisfying a mild statistical requirement, and
  - learning a classifier from ***any*** class satisfying a mild continuity assumption

- Implementation disentangles adversarial training:
  1. generate adversarial data (***Key result*** solves the compression of an OT plan)

     —//—

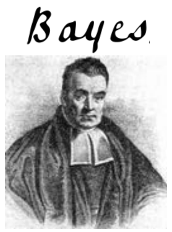
  2. training as usual

- Toy experiments against “weakly activated” adversarial data reveal generalisation improves *on clean data as well*



Key players: Bayes

Bayes



1- Classifiers

$$\mathcal{H} \subseteq \mathbb{R}^{\mathcal{X}}$$

← domain

← sign = class

2- Adversaries

$$\mathcal{A} \subseteq \mathcal{X}^{\mathcal{X}}$$

← domain

3- (differentiable) Loss

$$\ell : \{-1, 1\} \times [0, 1] \to \overline{\mathbb{R}}$$

(conditional) Bayes risk

$$\underline{L}(\pi) \doteq \inf_c \mathsf{E}_{\mathsf{Y}\sim\pi} \ell(\mathsf{Y}, c) \Rightarrow$$ ***proper***

(π in inf)

composite loss

link $\psi : [0, 1] \to \mathbb{R}$

$$\ell_\psi(y, v) \doteq \ell(y, \psi^{-1}(v))$$

canonical loss

$$\psi \doteq -\underline{L}'$$

(ψ “hidden”)

blunt predictor

$$h^\circ \doteq \psi\left(\tfrac{1}{2}\right) \quad = 0$$ (often)

corresponding loss: $\ell^\circ_\psi = \ell^\circ$

4- general adversarial loss

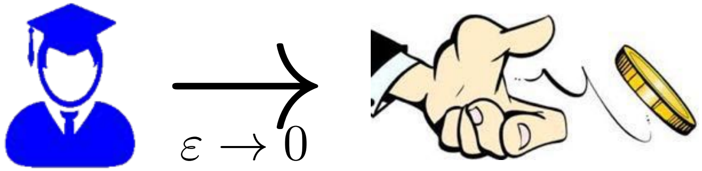
$$\ell(\mathcal{H}, \mathcal{A}, D) \doteq \min_{h \in \mathcal{H}} \mathsf{E}_{(\mathsf{X},\mathsf{Y})\sim D}\left[\max_{a \in \mathcal{A}} \ell(\mathsf{Y}, h \circ a(\mathsf{X}))\right]$$

particular case, Madry *et al.*’18:

$$\boldsymbol{x} \overset{a}{\to} \boldsymbol{x} + \boldsymbol{\delta} \text{ s.t. } \|\boldsymbol{\delta}\| \leq \delta^*$$

$\mathcal{H}$ **ε-defeated by $\mathcal{A}$ on ℓ** iff

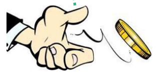
$$\ell(\mathcal{H}, \mathcal{A}, D) \geq (1 - \varepsilon) \cdot \ell^\circ$$



Main negative result

- For any proper composite loss ℓ, classifiers $\mathcal{H}$, adversaries $\mathcal{A}$ (+integrability assumptions),

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



$$\beta_a \doteq \max_{h \in \mathcal{H}} \{ \varphi(\underset{\text{“+” ex.}}{P}, f, \pi, 2\underline{L}(1)) - \varphi(\underset{\text{“-” ex.}}{N}, f, 1-\pi, -2\underline{L}(0)) \} \text{ with } \varphi(Q, f, u, v) \doteq \int_{\mathcal{X}} u \cdot (f(\boldsymbol{x}) + v) \mathrm{d}Q(\boldsymbol{x})$$

and $f \doteq (-\underline{L}') \circ \psi^{-1} \circ h \circ a$

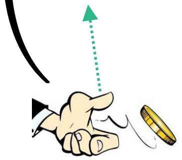
Example: if $\underline{L}(0) = \underline{L}(1)$ and $\pi = 1/2$, then β_a is $\propto$ Integral Probability Metric for class $\{(-\underline{L}') \circ \psi^{-1} \circ h \circ a : h \in \mathcal{H}\}$

(more in paper / poster)

Main negative result — consequence #1

- For any proper composite loss ℓ, classifiers $\mathcal{H}$, adversaries $\mathcal{A}$ (+integrability assumptions),

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



$\beta_a \doteq \max_{h\in\mathcal{H}}\{$ … $u \cdot (f(\boldsymbol{x}) + v)\mathrm{d}Q(\boldsymbol{x})$

… $h \circ a$

Example: if $\underline{L}(0)$ … for class $\{(-\underline{L}') \circ \psi^{-1} \circ h \circ a : h \in \mathcal{H}\}$

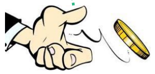
Hence, if $\exists a \in \mathcal{A}$ such that $\beta_a \leq 2\varepsilon\ell^\circ$ then $\mathcal{H}$ is ε-defeated by $\mathcal{A}$ on ℓ

Main negative result — consequence #2

- For any proper composite loss ℓ, classifiers $\mathcal{H}$, adversaries $\mathcal{A}$ (+integrability assumptions),

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

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



$\beta_a \doteq \max_{h \in \mathcal{H}} \{$ … $u \cdot (f(\boldsymbol{x}) + v)\mathrm{d}Q(\boldsymbol{x})$

$h \circ a$

Example: if $\underline{L}(0)$ … ass $\{(-\underline{L}') \circ \psi^{-1} \circ h \circ a : h \in \mathcal{H}\}$

RHS: roles permuted — adversary in *outer* optimisation suggests 2-stages optimisation to train classifier:

(i) (build adversary) craft *adversarial training data*

(ii) **train from adversarial data**

Adversaries 1/3: MMD

- Direct link with Maximum Mean Discrepancy (MMD)
  - Let $\mathcal{H}$ be the unit ball of a RKHS w/ reproducing kernel κ.

Adversarial mean embedding of a on Q

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

(*Adversarial*) MMD between P and N

$$\mathrm{MMD}[P, N|a] \doteq \|\mu_{a,P} - \mu_{a,N}\|_{\mathcal{H}}$$

(if $\underline{L}(0) = \underline{L}(1)$ and $\pi = 1/2$)

$$\beta_a = \frac{1}{4} \cdot \mathrm{MMD}[P, N|a]$$

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

$$\exists a \in \mathcal{A} \text{ s.t. } \mathrm{MMD}[P, N|a] \leq 8\varepsilon\ell^{\circ}$$

(more in paper / poster)

Adversaries 2/3: Monge

vs

- Allows to build efficient adversaries when classifiers are Lipschitz
  - *Solve the compression of an optimal transport plan*

(*Adversarial*) OT plan between P and N

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

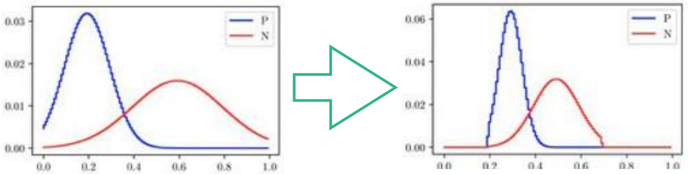
Monge efficiency (for cost $c : \mathcal{X} \times \mathcal{X} \rightarrow \mathbb{R}$)

$\mathcal{A}$ **δ-Monge efficient for c on P, N iff**

$$\exists a \in \mathcal{A} : C(a, P, N) \leq \delta$$

Suppose $\mathcal{H}$ is K-Lipschitz with respect to c, $\mathcal{A}$ is δ-Monge efficient for c on P, N. Suppose $\underline{L}(0) = \underline{L}(1)$, $\pi = 1/2$.

If $\delta \leq 8\varepsilon\ell^\circ / K$, then $\mathcal{H}$ is $\mathcal{E}$-defeated by $\mathcal{A}$ on ℓ



(more general statement in paper / poster)

Adversaries 3/3: Boosting

A. It is possible to $\mathcal{E}$-defeat $\mathcal{H}$ *simultaneously* on a whole set $\mathcal{L}$ of **symmetric** losses

Simple way to defeat strategies learning/tuning the loss
— important case because common losses fit in (log, square, Matsushita, etc.)

B. It is possible to craft **very strong** adversaries from **very weak** ones

RKHS example — suppose there exists a weakly contractive adversary a in a feature map Φ of the RKHS: $\|\Phi \circ a(\boldsymbol{x}) - \Phi \circ a(\boldsymbol{x}')\|_{\mathcal{H}} \leq (1-\eta) \cdot \|\Phi(\boldsymbol{x}) - \Phi(\boldsymbol{x}')\|_{\mathcal{H}}, \forall \boldsymbol{x}, \boldsymbol{x}' \in \mathcal{X}$
Then $\forall \delta > 0$**, composing just** $(1/\eta) \cdot \log(W_1^{\Phi}/\delta)$ **adversaries yields** δ**-Monge efficiency**

W_1^{Φ} = 1-Wasserstein distance between P and N in Φ

(complete statements in paper / poster)

Take home theoretical messages

vs

A. Replace adversarial training by **training from adversarial data**

B. If loss in specific classes, incl. popular losses, adversary can be **loss agnostic**

C. If learner's $\mathcal{H}$ is Lipschitz, use Lipschitz cost in an **OT compression problem**

D. **Adversarial boosting:** craft strong adversaries from weak adversaries

Some Monge efficient adversaries

A. *Mixup* adversaries (named after Zhang, Cissé, Dauphin & Lopez-Paz ’18)

general transformation: $a(\boldsymbol{x}) \doteq (1-\lambda)\cdot\boldsymbol{x} + \lambda\cdot\boldsymbol{x}'$

neighbour in other class

cluster / class centroid

sample centroid, etc.

B. Monge adversary – for a tight control on Monge efficiency, focus on

$$\min_{a} \text{Wasserstein-OT s.t. } d(a(\boldsymbol{x}),\boldsymbol{x}) \leq \underset{\text{“budget”}}{\alpha}, \forall\boldsymbol{x}$$

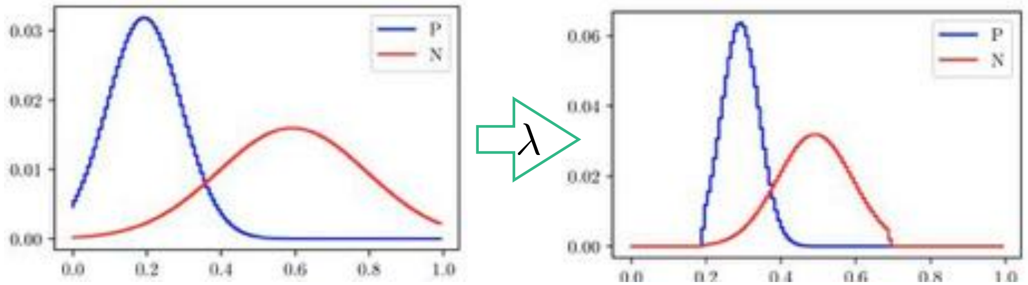
(more in paper / poster)

Toy experiments 1/2 — data & transformations

A. *Mixup*-to-sample-centroid

$$x' = \mathrm{E}_D[\mathrm{X}]$$

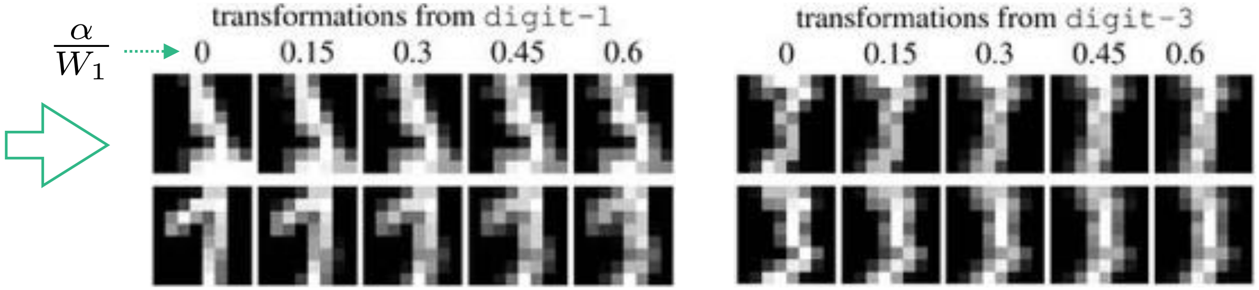
1D normal classes



B. Monge adversary for Wasserstein = W_2^2 and $d = \|.\|_1$ (cvx)

USPS data

classes 1 & 3


transformations from `digit-1`

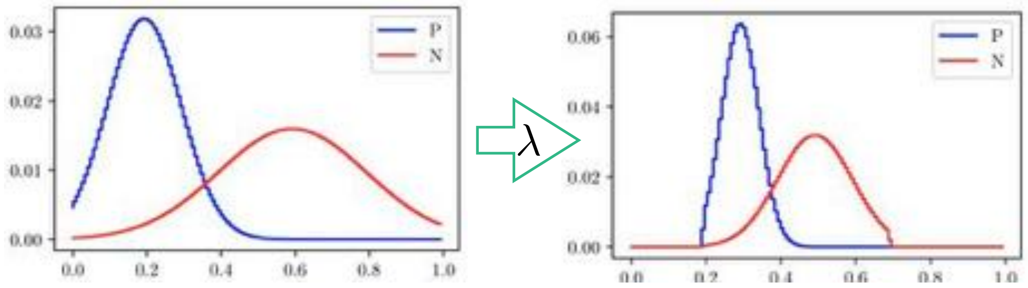
transformations from `digit-3`

Toy experiments 2/2 — findings

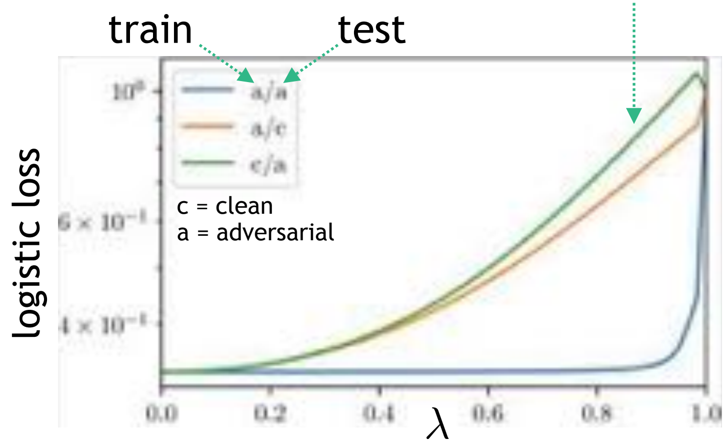
A. *Mixup*-to-sample-centroid

$$x' = \mathrm{E}_D[\mathrm{X}]$$

1D normal classes



worst trains on *c* !

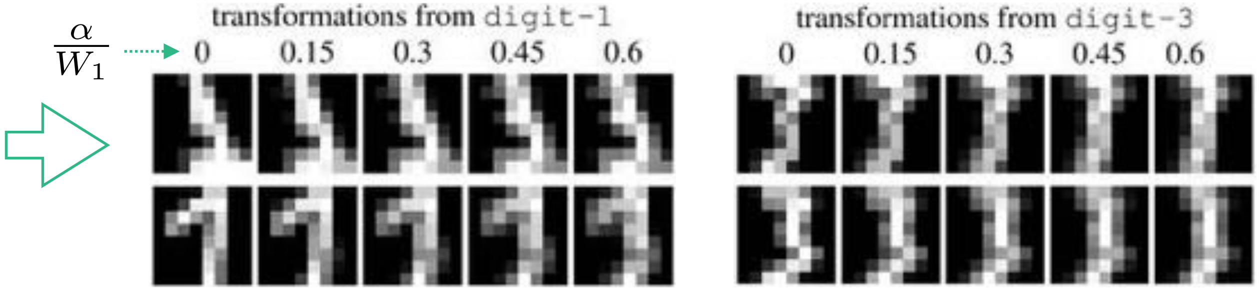


c = clean
a = adversarial

B. Monge adversary for Wasserstein = W_2^2 and $d = \|.\|_1$ (cvx)

USPS data

classes 1 & 3



best trains on (weak) *a*
***(even tested on c)* !**

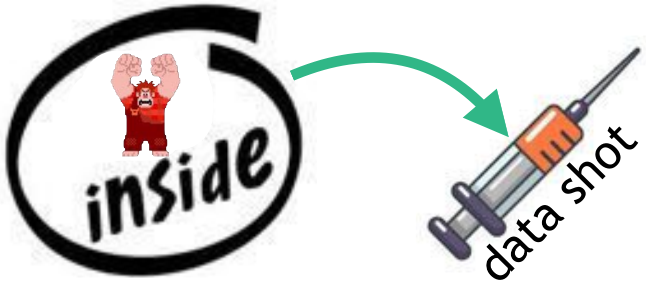
α/d	c/c	c/a	a/c	a/a
0.15	0.03	0.11	**0.00**	**0.02**
0.30	0.03	0.25	**0.00**	0.12
0.45	0.03	0.48	**0.01**	0.55
0.60	0.03	0.74	0.20	0.96

Conclusion

- Replacement of adversarial training by training from adversarial data
- Adversaries that can be effective against wide ranges of $(\mathcal{H}, \ell)$

- Adversarial strategy against Lipschitz classifiers: **compression of OT plans** (between class marginals)

- **Toy experiments** reveal that *sufficiently weak adversarial data* can **improve generalisation on clean data**



- Next step: explain such a “vaccination phenomenon”

Thank you

(get your data shot at poster # 191)