Understanding adversarial robustness via optimal transport perspective

Jakwang Kim
PIMS Kantotrovich Initiative, University of British Columbia
Joint work with: Nicolás García Trillos(UW-Madison), Matt Jacobs (UC-Santa Barbara), Matt Werenski(Tufts)

Kantorovich Initiative seminar series
September 28, 2023

Outline

1. What is learning
2. Adversarial attack
3. Connection between adversarial training and optimal transport
4. Geometry of adversarial learning: generalized barycenter problem
5. The existence of optimal robust classifiers
6. Efficient numerics
7. Conclusions and future works

What is learning

What is learning?

Let $\{(X_i, Y_i) \in \mathcal{X} \times \mathcal{Y} : i = 1 \dots, n\}$ be i.i.d. samples drawn by unknown distribution μ, $\mathcal{F} = \{f : \mathcal{X} \to \mathcal{Y}\}$ be a certain family of functions and $\ell : \mathcal{Y} \times \mathcal{Y} \to \mathbb{R}$ be a given loss function. A (supervised) learning problem is to obtain a (potentially) nice function f^* by solving the minimization of *the empirical risk*:

$$\min_{f \in \mathcal{F}} \frac{1}{n} \sum_{i=1}^{n} \ell(f(X_i), Y_i).$$

Since we don't know μ, this is the best we can do. Our hope is that there is a solution f^* for the above and it is a true solution in the sense that

$$\mathbb{E}_{(X,Y)\sim\mu}[\ell(f^*(X), Y)] \approx \min_{f \in \mathcal{F}} \mathbb{E}_{(X,Y)\sim\mu}[\ell(f(X), Y)].$$

Better than human?

Since deep learning revolution, there have been enormous progresses in machine learning. In particular, state-of-art neural networks outperform humans in image classification.

Suppose there are some images of dogs and cats. We call each image as a *feature (vector)* and dog/cat as *class*. Image classification problem is given image, to match it to the correct class(dog/cat). In other words, one should answer the question: "Is this picture dog or cat?"

According to Dodge and Karam[1], Human top-5 classification accuracy[2] on the large scale ImageNet dataset has been reported to be 94.9%, while 2023 best performance show 99% accuracy.

[1] Dodge and Karam, "A study and comparison of human and deep learning recognition performance under visual distortions".

[2] Top-5 accuracy measures whether top 5 predictions include the correct label.

ImageNet

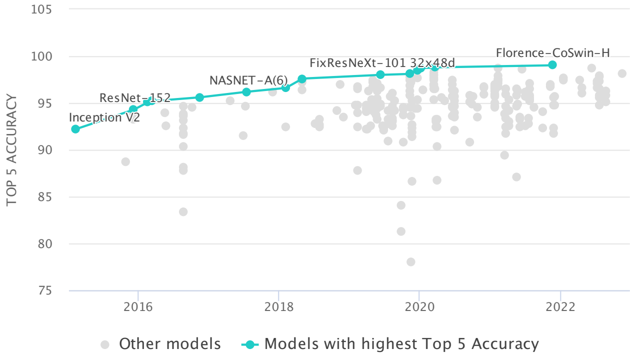


Figure: Image Classification on ImageNet: top 5 accuracy[3],

[3] Yuan et al., “Florence: A new foundation model for computer vision”.

Adversarial attack

Adversarial attack

Researchers observed that neural networks are sometimes very sensitive to a small noise, and their performance completely breaks down by this well-designed noise, called *adversarial attack*.

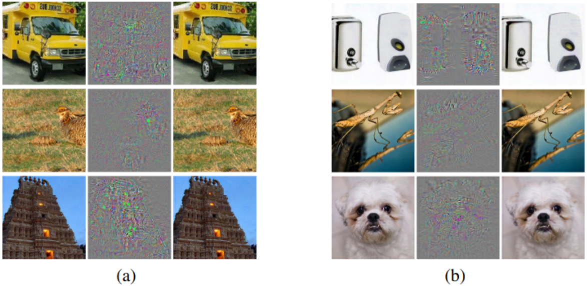


Figure: Adversarial examples generated for AlexNet [9].(Left) is a correctly predicted sample, (center) difference between correct image, and image predicted incorrectly magnified by 10x (values shifted by 128 and clamped), (right) adversarial example[4].

[4] Szegedy et al., "Intriguing properties of neural networks".

Adversarial attack(cont.)

We emphasize that the size of adversarial attacks is usually small so that they are imperceptible to humans. Also, not every noise is adversarial: adversarial attack is delicately designed by maximizing empirical risk.

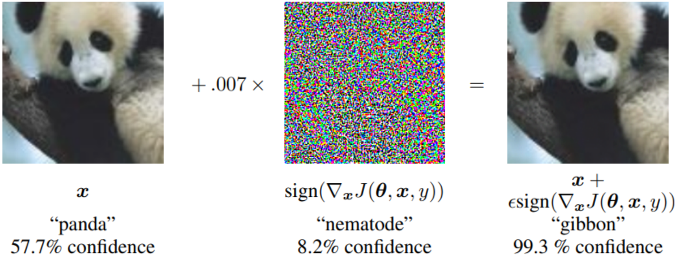

$+ .007 \times$
$=$
$\boldsymbol{x}$
"panda"
57.7% confidence
$\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"nematode"
8.2% confidence
$\boldsymbol{x} + \epsilon\text{sign}(\nabla_{\boldsymbol{x}} J(\boldsymbol{\theta}, \boldsymbol{x}, y))$
"gibbon"
99.3 % confidence

Figure: Adversarial examples generated for GoogLeNet. (Left) is a correctly predicted sample, (center) difference between correct image and (right) adversarial example(Goodfellow et al.[5]).

[5] Goodfellow, Shlens, and Szegedy, "Explaining and harnessing adversarial examples".

Adversarial attack is not artificial

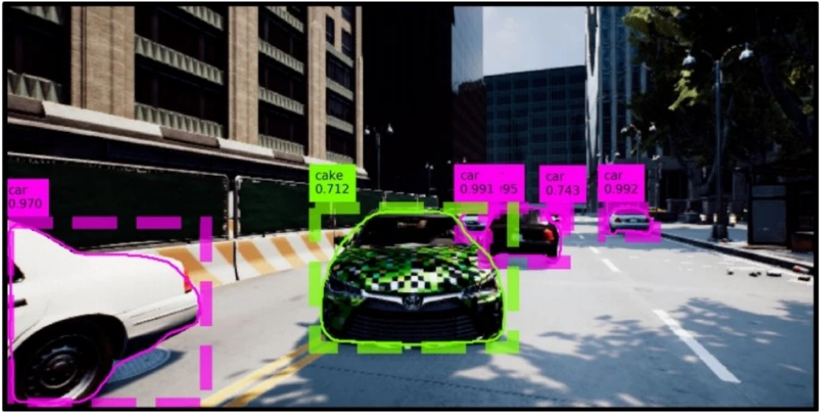
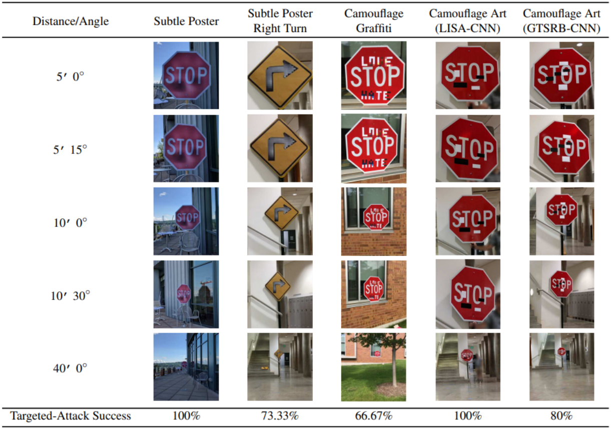


Figure: The car with a camouflage pattern is misdetected as a “cake” (Zhang et al.[6]).

[6] Zhang et al., “CAMOU: Learning Physical Vehicle Camouflages to Adversarially Attack Detectors in the Wild”.

Adversarial attack is not artificial(cont.)



Distance/Angle	Subtle Poster	Subtle Poster Right Turn	Camouflage Graffiti	Camouflage Art (LISA-CNN)	Camouflage Art (GTSRB-CNN)
5′ 0°					
5′ 15°					
10′ 0°					
10′ 30°					
40′ 0°					
Targeted-Attack Success	100%	73.33%	66.67%	100%	80%

Figure: Sample of physical adversarial examples against LISA-CNN and GTSRB-CNN[7].

[7] Eykholt et al., "Robust Physical-World Attacks on Deep Learning Visual Classification".

Adversarial attack is not artificial(cont.)

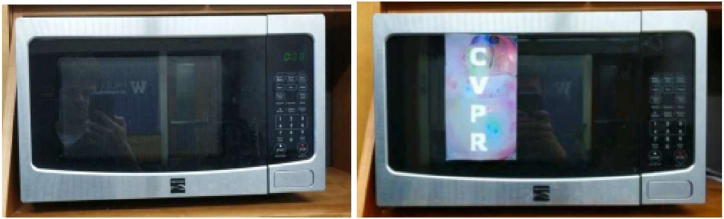


Figure: Physical adversarial example against the Inception-v3 classifier. The left shows the original cropped image identified as microwave (85.2%) while the right shows the cropped physical adversarial example identified as phone(77.8%)[8]).

[8] Eykholt et al., "Robust Physical-World Attacks on Deep Learning Visual Classification".

Connection between adversarial training and optimal transport

Classical classification problem : Bayes classifier

- $(\mathcal{X}, d)$: a feature space with metric d, $\mathcal{Y} := \{1, \ldots, K\}$: a class space, $\mathcal{Z} := \mathcal{X} \times \mathcal{Y}$.
- $\mu = (\mu_1, \ldots, \mu_K)$: a (Borel probability) data distribution; (after a normalization) each μ_i is a conditional distribution over $\mathcal{X}$ given $Y = i$.
- Θ : a parametric family of functions from $\mathcal{X}$ to $\mathcal{Y}$: e.g. neural networks.
- $\ell(x) := 1 - f_i(x)$: loss function.

A formal problem is

$$\inf_{\theta \in \Theta} R(f, \mu) := \inf_{\theta \in \Theta} \sum_{i \in \mathcal{Y}} \int_{\mathcal{X}} (1 - f_i^{\theta}(x)) d\mu_i(x).$$

There is always an optimal *Bayes hard* classifier: $f_i^*(x) = 1$ if $d\mu_i(x) \geq d\mu_j(x)$ for all $j \neq i$.

Data-perturbing adversarial model

Fix $\varepsilon > 0$ be the *adversarial budget*. The adversary only attacks a feature, i.e., for each $x \in \mathcal{X}$

$$x \longmapsto \tilde{x} \in \arg\max\{1 - f_i^\theta(x') : x' \in \overline{B}_\varepsilon(x)\}.$$

The usual adversarial training model is defined as

$$\inf_{\theta\in\Theta} \left\{ \sum_{i\in\mathcal{Y}} \int_{\mathcal{X}} \sup_{\tilde{x}\in\overline{B}_\varepsilon(x)} \{1 - f_i^\theta(\tilde{x})\} d\mu_i(x) \right\}.$$

Two problems of (AT)

- Neural networks in practice are huge and complex, so hard to understand $\sup_{\tilde{x}\in B_\varepsilon(x)}\{1-f_i^\theta(\tilde{x})\}$: consider the largest possible space(agnostic learner)

$$\mathcal{F} := \{(f_1,\dots,f_K) : 0 \le f_i \le 1, \sum f_i = 1, \text{Borel measurable}\}.$$

- $\sup_{\tilde{x}\in B_\varepsilon(x)}\{1-f_i^\theta(\tilde{x})\}$ might not be Borel measurable, hence the problem is not well-defined: replace μ by its universal completion $\overline{\mu}$. The *data-perturbing adversarial model* is

$$\inf_{f\in\mathcal{F}}\left\{\sum_{i\in\mathcal{Y}}\int_{\mathcal{X}}\sup_{\tilde{x}\in\overline{B}_\varepsilon(x)}\{1-f_i(\tilde{x})\}d\overline{\mu}_i(x)\right\}. \tag{AT}$$

DRO adversarial model

The adversary attacks the distribution μ, i.e., given unknown μ (the adversary is assumed to know)

$$\mu \longmapsto \tilde{\mu} \in \arg\max\{R(f,\nu) : \nu \in \mathcal{P}(\mathcal{Z}), W_p(\mu,\nu) \leq \varepsilon\}.$$

The *distributionally robust optimization(DRO) adversarial model* is

$$R_{\varepsilon}^{DRO} := \inf_{f \in \mathcal{F}} \sup_{\tilde{\mu} \in \mathcal{P}(\mathcal{Z})} \{R(f,\tilde{\mu}) - C(\mu,\tilde{\mu})\} \tag{DRO}$$

where $C : \mathcal{P}(\mathcal{Z}) \times \mathcal{P}(\mathcal{Z}) \to [0,\infty]$ is a transport cost defined as

$$C(\mu,\tilde{\mu}) := \inf_{\pi \in \Gamma(\mu,\tilde{\mu})} \int c_{\mathcal{Z}}(z,\tilde{z}) d\pi(z,\tilde{z}).$$

Cost function for (DRO)

We always assume that a cost function is

$$c_Z(z,\tilde{z}) := \begin{cases} c_\varepsilon(x,\tilde{x}) & \text{if } y=\tilde{y}, \\ \infty & \text{otherwise} \end{cases}.$$

Then,

$$C(\mu,\tilde{\mu}) = \sum_{i\in\mathcal{Y}} C(\mu_i,\tilde{\mu}_i).$$

A typical choice of $c_\varepsilon(x,\tilde{x})$ is

$$c_\varepsilon(x,\tilde{x}) := \begin{cases} 0 & \text{if } d(x,\tilde{x}) \le \varepsilon, \\ \infty & \text{if } d(x,\tilde{x}) > \varepsilon \end{cases}.$$

Connection between (AT) and (DRO)

Theorem

(Pydi and Jog[a]*, García Trillos, Jacobs and* ***K.***[b]*) With $c_\varepsilon(x, \tilde{x})$ as above,*

$$(\mathrm{AT}) = (\mathrm{DRO}).$$

[a] Pydi and Jog, "The Many Faces of Adversarial Risk".

[b] García Trillos, Jacobs, and Kim, *On the existence of solutions to adversarial training in multiclass classification*.

Question : (DRO) is a minimax problem: hard to compute in general. Can we get a better formulation to be computationally tractable?

Geometry of adversarial learning: generalized barycenter problem.

Toy example: three points distribution

Recall

$$c_\varepsilon(x,\tilde{x}) := \begin{cases} 0 & \text{if } d(x,\tilde{x}) \leq \varepsilon, \\ \infty & \text{if } d(x,\tilde{x}) > \varepsilon \end{cases}.$$

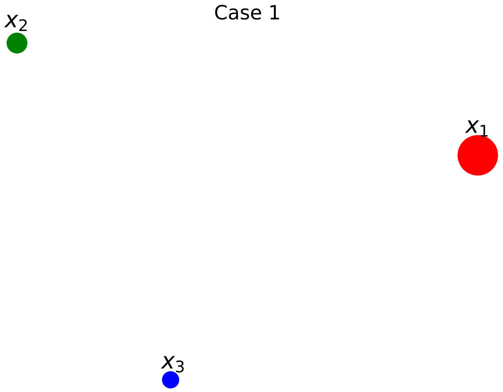
Consider a simple data distribution

$$\mu = (\omega_1 \delta_{(x_1,1)}, \omega_2 \delta_{(x_2,2)}, \omega_3 \delta_{(x_3,3)}), \quad \omega_1 \geq \omega_2 \geq \omega_3.$$

Q. What is the optimal attack for the adversary?
A. Use (possible) barycenters of $\omega_1 \delta_{(x_1,1)}, \omega_2 \delta_{(x_2,2)}, \omega_3 \delta_{(x_3,3)}$.

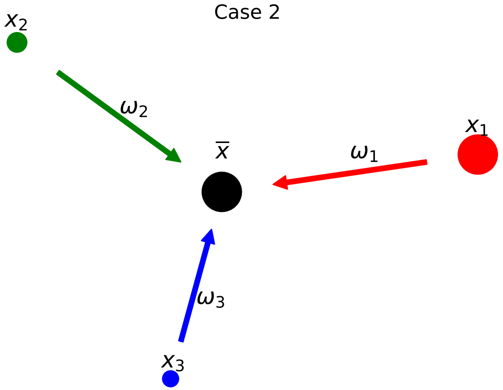
Toy example: three points distribution(cont.)

Case 1 : $d(x_i, x_j) > 2\varepsilon$ for all $1 \leq i \neq j \leq 3$.



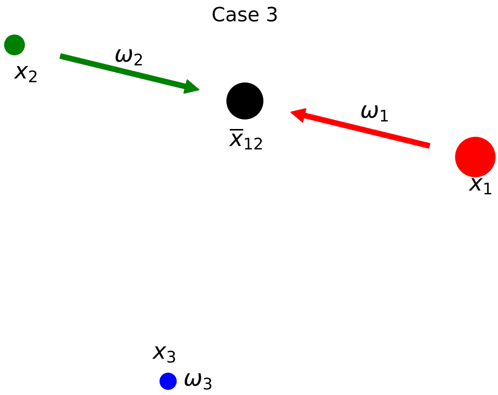
Toy example: three points distribution(cont.)

Case 2 : There is some $\overline{x}$ such that $d(\overline{x}, x_i) \leq \varepsilon$ for all $1 \leq i \leq 3$.



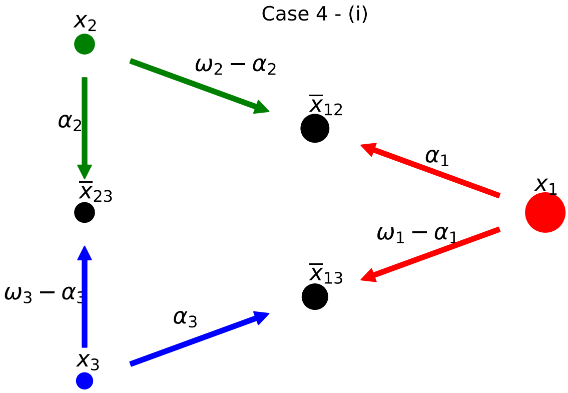
Toy example: three points distribution(cont.)

Case 3 : $d(x_1, x_2) \leq 2\varepsilon$, $d(x_1, x_3) > 2\varepsilon$ and $d(x_2, x_3) > 2\varepsilon$.



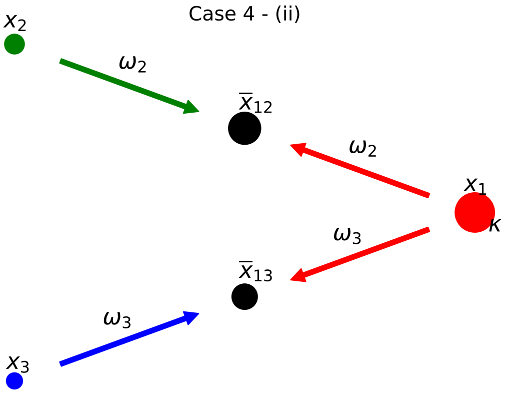
Toy example: three points distribution(cont.)

Case 4-(i) : $d(x_i, x_j) \le 2\varepsilon$ for any x_i, x_j but $\overline{B}(x_1, \varepsilon) \cap \overline{B}(x_2, \varepsilon) \cap \overline{B}(x_3, \varepsilon) = \emptyset$ and $\omega_1 < \omega_2 + \omega_3$.



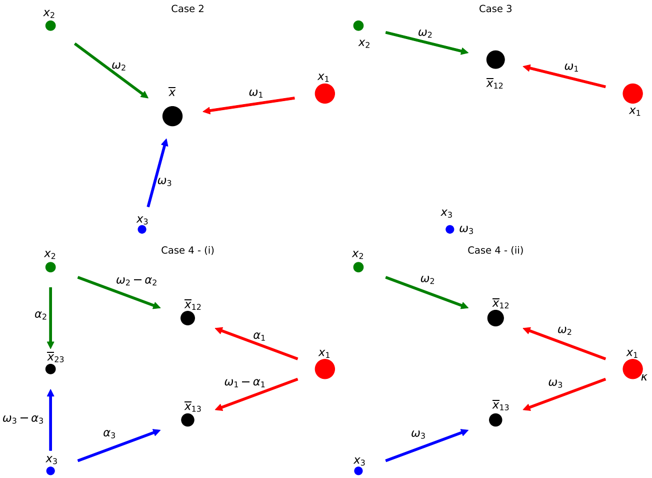
Toy example: three points distribution(cont.)

Case 4-(ii) : $d(x_i, x_j) \leq 2\varepsilon$ for any x_i, x_j but $\overline{B}(x_1,\varepsilon) \cap \overline{B}(x_2,\varepsilon) \cap \overline{B}(x_3,\varepsilon) = \emptyset$ and $\omega_1 \geq \omega_2 + \omega_3$.



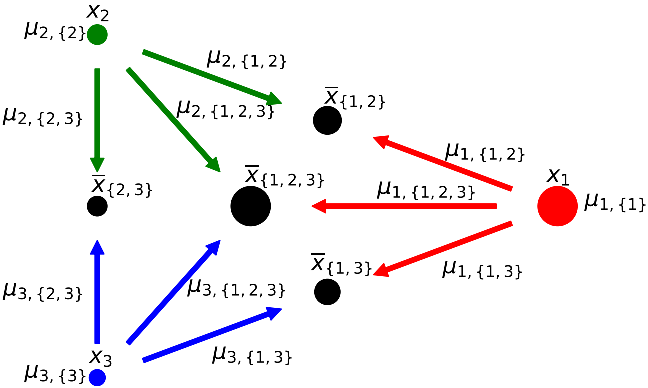
Toy example: three points distribution(cont.)

The optimal adversarial attack uses barycenters of $\{x_1, x_2, x_3\}$. Also, it depends on not only the geometry of the support of μ but also the magnitudes of its marginals, $(\omega_1, \omega_2, \omega_3)$.



Toy example: three points distribution(cont.)

For general cost function c_ε,



Generalized barycenter problem

Recall (DRO):

$$\begin{aligned}
&\inf_{f\in\mathcal{F}} \sup_{\tilde{\mu}\in\mathcal{P}(\mathcal{Z})} \{R(f,\tilde{\mu}) - C(\mu,\tilde{\mu})\}\\
&= \inf_{f\in\mathcal{F}} \sup_{\tilde{\mu}\in\mathcal{P}(\mathcal{Z})} \left\{\sum_{i\in\mathcal{Y}} \int_{\mathcal{X}} (1-f_i(\tilde{x}))d\tilde{\mu}_i(\tilde{x}) + \sum_{i\in\mathcal{Y}} C(\mu_i,\tilde{\mu}_i)\right\}\\
&= 1 - \sup_{f\in\mathcal{F}} \inf_{\tilde{\mu}\in\mathcal{P}(\mathcal{Z})} \left\{\sum_{i\in\mathcal{Y}} \int_{\mathcal{X}} f_i(\tilde{x})d\tilde{\mu}_i(\tilde{x}) + \sum_{i\in\mathcal{Y}} C(\mu_i,\tilde{\mu}_i)\right\}.
\end{aligned}$$

Generalized barycenter problem(cont.)

For the $\sup_{f\in\mathcal{F}} \inf_{\tilde{\mu}\in\mathcal{P}(\mathcal{Z})}$ term, if we can swap them,

$$\inf_{\tilde{\mu}\in\mathcal{P}(\mathcal{Z})} \sup_{f\in\mathcal{F}} \left\{ \sum_{i\in\mathcal{Y}} \int_{\mathcal{X}} f_i(\tilde{x}) d\tilde{\mu}_i(\tilde{x}) + \sum_{i\in\mathcal{Y}} C(\mu_i, \tilde{\mu}_i) \right\}$$
$$= \inf_{\tilde{\mu}\in\mathcal{P}(\mathcal{Z})} \left\{ \sup_{f\in\mathcal{F}} \sum_{i\in\mathcal{Y}} \int_{\mathcal{X}} f_i(\tilde{x}) d\tilde{\mu}_i(\tilde{x}) + \sum_{i\in\mathcal{Y}} C(\mu_i, \tilde{\mu}_i) \right\}.$$

Notice that

$$\sup_{f\in\mathcal{F}} \sum_{i\in\mathcal{Y}} \int_{\mathcal{X}} f_i(\tilde{x}) d\tilde{\mu}_i(\tilde{x}) = \inf_{\lambda} \left\{ \lambda(\mathcal{X}) : \lambda \geq \tilde{\mu}_i \text{ for all } i \in \mathcal{Y} \right\}.$$

Generalized barycenter problem(cont.)

The *generalized barycenter problem* is

$$\inf_{\lambda,\tilde{\mu}_1,\dots,\tilde{\mu}_K} \lambda(\mathcal{X}) + \sum_{i\in\mathcal{Y}} C(\mu_i,\tilde{\mu}_i) \quad \text{s.t. } \lambda \geq \tilde{\mu}_i \; \forall i \in \mathcal{Y}. \tag{GB}$$

Theorem

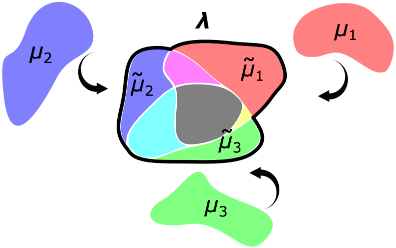
*(Garcıa Trillos, Jacobs and **K.**[a]) Under some assumptions on c_ε,*

$$(\mathrm{DRO}) = 1 - (\mathrm{GB}).$$

Furthermore, the infimum of (GB) *is achieved.*

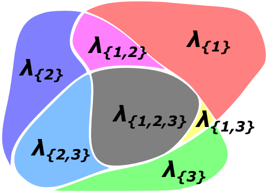
[a] Garcıa Trillos, Jacobs, and Kim, "The multimarginal optimal transport formulation of adversarial multiclass classification".

Generalized barycenter problem(cont.)



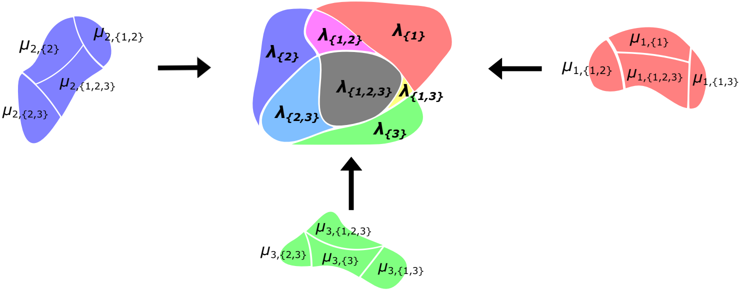
The adversary transports each μ_i to $\tilde{\mu}_i$ by paying $C(\mu_i, \tilde{\mu}_i)$. λ is chosen to be the minimum positive measure covering all $\tilde{\mu}_i$'s.

Generalized barycenter problem(cont.)



Partition λ indexed by non-empty interactions $A \subset \mathcal{Y}$. λ_A is in fact (Wasserstein) barycenter of $\mu_{i,A}$'s for $i \in A$.

MOT formulation


$\mu_{2,\{2\}}$
$\mu_{2,\{1,2\}}$
$\mu_{2,\{1,2,3\}}$
$\mu_{2,\{2,3\}}$
$\boldsymbol{\lambda}_{\{1,2\}}$
$\boldsymbol{\lambda}_{\{1\}}$
$\boldsymbol{\lambda}_{\{2\}}$
$\boldsymbol{\lambda}_{\{1,2,3\}}$
$\boldsymbol{\lambda}_{\{1,3\}}$
$\boldsymbol{\lambda}_{\{2,3\}}$
$\boldsymbol{\lambda}_{\{3\}}$
$\mu_{1,\{1\}}$
$\mu_{1,\{1,2\}}$
$\mu_{1,\{1,2,3\}}$
$\mu_{1,\{1,3\}}$
$\mu_{3,\{1,2,3\}}$
$\mu_{3,\{2,3\}}$
$\mu_{3,\{3\}}$
$\mu_{3,\{1,3\}}$

Using decompositions, the problem can be written in terms of $\mu_{i,A}$'s and λ_A's. Furthermore, $\lambda_A \in \arg\min_{\lambda'_A} \sum_{i \in A} C(\mu_{i,A}, \lambda'_A)$ is a solution to a classical (Wasserstein) barycenter problem.

Stratified MOT

Theorem

Let $S_K = \{A \subseteq \mathcal{Y} : A \neq \emptyset\}, S_k(i) = \{A \in S_K : i \in A\}$ *and*

$$c_{\varepsilon,A}(x_1, \dots, x_K) := \inf_{x' \in \mathcal{X}} \sum_{i \in A} c_\varepsilon(x_i, x').$$

Consider

$$\begin{aligned} &\inf_{\{\pi_A : A \in S_K\}} \sum_{A \in S_K} \int_{\mathcal{X}^K} \big(c_{\varepsilon,A}(x_1, \dots, x_K) + 1\big) d\pi_A(x_1, \dots, x_K) \\ &\text{s.t.} \sum_{A \in S_K(i)} \mathcal{P}_{i\#}\pi_A = \mu_i \text{ for all } i \in \mathcal{Y}. \end{aligned} \tag{MOT}$$

Then, (GB) $=$ (MOT)*.*

Duality

Theorem

Consider

$$\sup_{g_1,\dots,g_K \in \mathcal{C}_b(\mathcal{X})} \sum_{i \in \mathcal{Y}} \int_{\mathcal{X}} g_i(x_i) d\mu_i(x_i)$$

$$\text{s.t.} \sum_{i \in A} g_i(x_i) \leq 1 + c_{\varepsilon,A}(x_i : i \in A) \text{ for all } (x_i : i \in A) \in \mathcal{X}^A, A \in S_K. \tag{Dual}$$

Then, (MOT) $=$ (Dual)*.*

If c_ε is bounded Lipschitz, then (Dual) *is achieved by $g \in \mathcal{C}_b(\mathcal{X})^K$.*

Full MOT

Theorem

Under some assumptions on c, there exists $\boldsymbol{c} : \mathcal{X}_*^K \to \mathbb{R}$ which defines

$$\inf_{\pi \in \Pi(\hat{\mu}_1, \dots, \hat{\mu}_K)} \left\{ \int_{\mathcal{X}_*^K} \boldsymbol{c}(x_1, \dots, x_K) d\pi(x_1, \dots, x_K) \right\} \tag{MOT'}$$

where $\mathcal{X}_* = \mathcal{X} \cup \{\text{👻}\}$ and $d\hat{\mu}_i = d\mu_i + \left(\sum_{j \neq i} ||\mu_j|| \right) \delta_{\text{👻}}$. Then,

$$(\text{MOT}) = (\text{MOT'})$$

Implication : universal lower bound

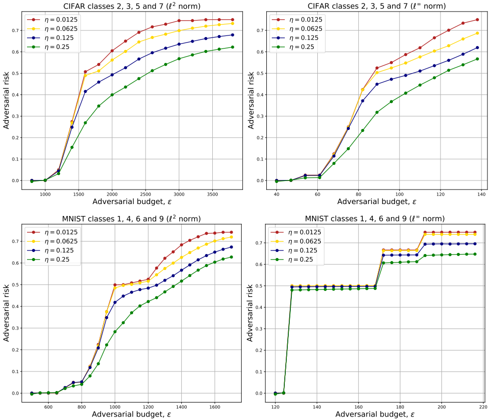
Suppose we were interested in

$$\inf_{f\in\mathcal{G}} \sup_{\tilde{\mu}\in\mathcal{P}(\mathcal{Z})} \{R(f,\tilde{\mu}) - C(\mu,\tilde{\mu})\}$$

where $\mathcal{G}$ can be any family of classifiers: e.g. neural networks. It is always the case that

$$\begin{aligned} &\inf_{f\in\mathcal{F}} \sup_{\tilde{\mu}\in\mathcal{P}(\mathcal{Z})} \{R(f,\tilde{\mu}) - C(\mu,\tilde{\mu})\} \\ &\leq \inf_{f\in\mathcal{G}} \sup_{\tilde{\mu}\in\mathcal{P}(\mathcal{Z})} \{R(f,\tilde{\mu}) - C(\mu,\tilde{\mu})\}. \end{aligned}$$

Real data: MNIST and CIFAR-10



Use off-the-shelf MOT solvers: Lin et al.[9], Tupitsa et al.[10].

[9] Lin et al., "On the complexity of approximating multimarginal optimal transport".

[10] Tupitsa et al., "Multimarginal optimal transport by accelerated alternating minimization".

The existence of optimal robust classifiers

For the learner?

Notice that (GB) and (MOT) are problems for the adversary. Solutions from these problems are optimal adversarial attacks.
Q: What about the learner? How do we compute an optimal robust classifier?
A: Consider the dual of (GB) and (MOT).

Existence of Borel measurable robust classifiers

Corollary

*(García Trillos, Jacobs, **K.**[a]) Let* $0 \leq g \leq 1$ *be a solution to* (Dual)*. Define*

$$f_i^*(\tilde{x}) := \sup_{x \in spt(\mu_i)} \{g_i(x) - c_\varepsilon(x, \tilde{x})\} \vee 0.$$

Suppose f^* is Borel measurable. Then, f^* is a Borel robust classifier.

[a] Garcıa Trillos, Jacobs, and Kim, "The multimarginal optimal transport formulation of adversarial multiclass classification".

The issue is that *a priori* we do not know f^* is Borel measurable unless c is continuous. A measurability issue is common in distributional robust optimization literature.

Existence of Borel measurable robust classifiers(cont.)

Theorem

*(García Trillos, Jacobs, **K.**[a])There exists a (Borel) solution f^* of* (DRO)*. Furthermore, there exists $\tilde{\mu}^* \in \mathcal{P}(\mathcal{Z})$ such that $(f^*, \tilde{\mu}^*)$ is a saddle point for* (DRO)*. In other words, the following holds: for any $\tilde{\mu} \in \mathcal{P}(\mathcal{Z})$ and any $f \in \mathcal{F}$ we have*

$$R(f^*, \tilde{\mu}) - C(\mu, \tilde{\mu}) \leq R(f^*, \tilde{\mu}^*) - C(\mu, \tilde{\mu}^*) \leq R(f, \tilde{\mu}^*) - C(\mu, \tilde{\mu}^*).$$

[a] Garcıa Trillos, Jacobs, and Kim, *On the existence of solutions to adversarial training in multiclass classification*.

Existence of Borel measurable robust classifiers(cont.)

Let c_ε^n be a bounded and Lipschitz cost function converging to c_ε pointwise. Let g^n be optimal for (Dual) with each c_ε^n. By Corollary, for each n

$$f_i^n := \sup_{x \in \mathrm{spt}(\mu_i)} \{g_i^n(x) - c_\varepsilon^n(x, \tilde{x})\} \vee 0$$

is a lower semi-continuous robust classifier. A candidate for robust classifier is

$$f_i^* = \limsup_{n \to \infty} f_i^n.$$

Existence of Borel measurable robust classifiers(cont.)

Proposition

Let g be the weak* limit of the g^n, and let f^* be defined as before. Then, for every $i \in \mathcal{Y}$,

$$g_i^*(x) = \inf_{\tilde{x} \in \mathcal{X}} \{f_i^*(\tilde{x}) + c_\varepsilon(x, \tilde{x})\}$$

for μ_i-a.e. $x \in \mathcal{X}$.

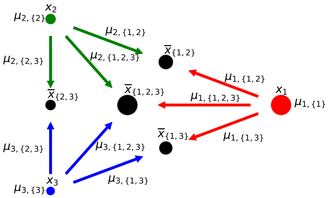
It holds that

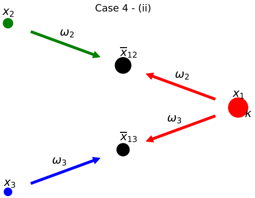
$$\int_{\mathcal{X}} f_i^*(\widetilde{x}) d\widetilde{\mu}_i^*(\widetilde{x}) + \int_{\mathcal{X}\times\mathcal{X}} c_\varepsilon(x, \widetilde{x}) d\pi_i^*(x, \widetilde{x}) = \int_{\mathcal{X}} g_i(x) d\mu_i(x).$$

In words, given optimal adversarial attack $\tilde{\mu}^*$, $(g, -f^*)$ is an optimal dual pair.

Efficient numerics

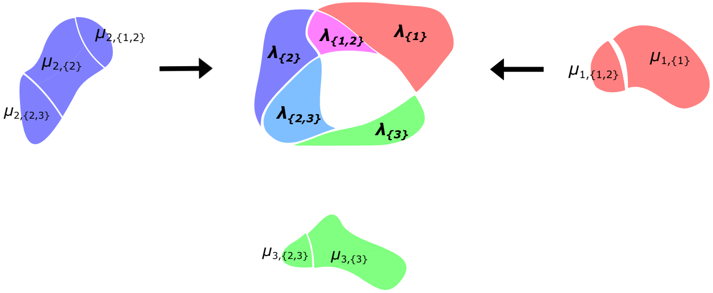
What happens in reality





In reality, different classes are mostly far from each other compared to the adversarial budget.

(GB) in reality



For example, since μ_1 and μ_3 are separated enough, you can't use $\lambda_{\{1,2,3\}}$: 3rd order interaction does not occur.

Truncation

From OT perspective, it suffices to solve

$$\inf_{\{\pi_A : A \in S_K, |A| \leq L\}} \sum_{A \in S_K} \int_{\mathcal{X}^K} \left(c_{\varepsilon, A}(x_1, \dots, x_K) + 1 \right) d\pi_A(x_1, \dots, x_K)$$

$$\text{s.t.} \sum_{A \in S_K(i)} \mathcal{P}_{i \#} \pi_A = \mu_i \text{ for all } i \in \mathcal{Y}.$$

for some L which is the maximum order of interactions. Notice that (MOT') cannot capture the truncation. It always needs to compute cost tensor $\boldsymbol{c}$ of order K.

Two approaches

Theorem (Informal)

(García Trillos et al[a]) Assume that classes are separated well, and let $L \ll K$ be the truncation level. Then, there are algorithms whose computational complexity is $\widetilde{O}(n^L)$.

[a] Garcıa Trillos et al., *Two approaches for computing adversarial training lower bounds based on optimal transport frameworks*.

Note that the complexity of off-the-shelf MOT solver is $\widetilde{O}(n^K)$.

Conclusions and future works

Conclusions

In a series of works,

- Connect (DRO) to (GB), generalized barycenter problem, and rewrite it in terms of (MOT) via mutlimarginal optimal transport.
- Prove the existence of Borel measurable robust classifiers of (DRO), and (AT) by using (Dual) and c-transform formula.
- In our recent work, we develop efficient algorithms based on (GB) and (MOT), respectively. The idea is that in real data, higher order interactions are rare so that sufficient to focus on lower order ones.

Future works

- Restrict $\mathcal{F}$ to a parameter family, e.g. neural nets, what can we say about it?
- Sample complexity of (DRO): how many samples do we need to achieve the approximate of the value of (DRO)?
- The asymptotic behavior of dual potentials/robust classifiers.
- How to choose a certain saddle point? There are possibly many equilibria...
- Regularity of robust classifiers: see Bungert, García Trillos and Murray[11]
- Different loss function: cross entropy, quadratic loss etc...
- Divergence-type barycenter problems: KL-divergence, χ^2-divergence, Rényi-divergence etc...
- A condition to characterize the optimal truncation level.

[11] Bungert, García Trillos, and Murray, "The geometry of adversarial training in binary classification".

Thank you for your attention!

This work is partially based on NSF Grant(TRIPODS grant 2023239).

Pacific Institute *for the*
Mathematical Sciences

- Bungert, Leon, Nicolás García Trillos, and Ryan Murray. “The geometry of adversarial training in binary classification”. In: *Information and Inference: A Journal of the IMA* 12.2 (2023), pp. 921–968.
- Dodge, Samuel and Lina Karam. “A study and comparison of human and deep learning recognition performance under visual distortions”. In: *2017 26th international conference on computer communication and networks (ICCCN)*. IEEE. 2017, pp. 1–7.
- Eykholt, Kevin et al. “Robust Physical-World Attacks on Deep Learning Visual Classification”. In: *2018 IEEE/CVF Conference on Computer Vision and Pattern Recognition* (2018), pp. 1625–1634. URL: `https://api.semanticscholar.org/CorpusID:29162614`.
- Garcıa Trillos, Nicolás, Matt Jacobs, and Jakwang Kim. *On the existence of solutions to adversarial training in multiclass classification*. 2023. arXiv: `2305.00075 [cs.LG]`.
- — .“The multimarginal optimal transport formulation of adversarial multiclass classification”. In: *Journal of Machine Learning Research* 24.45 (2023), pp. 1–56.

- Garcıa Trillos, Nicolás et al. *Two approaches for computing adversarial training lower bounds based on optimal transport frameworks*. 2023.
- Goodfellow, Ian J, Jonathon Shlens, and Christian Szegedy. "Explaining and harnessing adversarial examples". In: *arXiv preprint arXiv:1412.6572* (2014).
- Lin, Tianyi et al. "On the complexity of approximating multimarginal optimal transport". In: *Journal of Machine Learning Research* 23.65 (2022), pp. 1–43.
- Pydi, Muni Sreenivas and Varun Jog. "The Many Faces of Adversarial Risk". In: *Thirty-Fifth Conference on Neural Information Processing Systems*. 2021. URL: https://openreview.net/forum?id=-8QSntMuqBV.
- Szegedy, Christian et al. "Intriguing properties of neural networks". In: *CoRR* abs/1312.6199 (2014).
- Tupitsa, Nazarii et al. "Multimarginal optimal transport by accelerated alternating minimization". In: *2020 59th IEEE Conference on Decision and Control (CDC)*. IEEE. 2020, pp. 6132–6137.
- Yuan, Lu et al. "Florence: A new foundation model for computer vision". In: *arXiv preprint arXiv:2111.11432* (2021).

Zhang, Yang et al. “CAMOU: Learning Physical Vehicle Camouflages to Adversarially Attack Detectors in the Wild”. In: *International Conference on Learning Representations*. 2019. URL: https://openreview.net/forum?id=SJgEl3A5tm.