# Adversarial robustness in classification via the lens of optimal transport

Jakwang Kim
**Joint works with: Nicolas García Trillos(UW-Madison), Matt Jacobs(UC Santa Barbara), Matt Werenski(Tufts)**

PIMS Kantotrovich Initiative, Department of Mathematics
University of British Columbia

KIAS Center for AI and Natural Sciences Fall Workshop
November 5, 2024

# Table of Contents

# Table of Contents

# Overview

In the series of papers[1]:

- provide geometric understanding of the multiclass adversarial training model by generalized Wasserstein barycenter problem.
- prove the existence of adversarial robust classifiers, and unify variants of adversarial training models.
- propose a new numerical scheme to approximate a lower bound the adversarial risk.

[1] supported by the IFDS at UW-Madison and NSF through TRIPODS grant 2023239, and PIMS postdoctoral fellowship through the Kantorovich Initiative PIMS Research Network (PRN) as well as National Science Foundation grant NSF-DMS 2133244

# Table of Contents



Pacific Institute for the Mathematical Sciences

# Better than human?: ImageNet

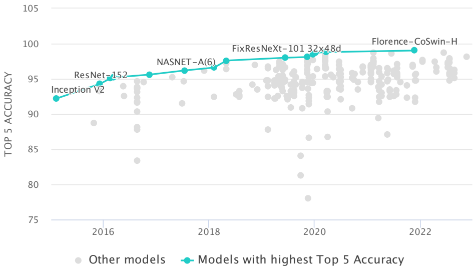



Figure: Image Classification on ImageNet: top 5 accuracy (Yuan et al.[2])

According to Dodge, Karam[3], human top-5 classification accuracy on the large scale ImageNet dataset has been reported to be 94.9%, while 2023 best performance show 99% accuracy.

[2] Yuan et al., "Florence: A new foundation model for computer vision".

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

# Instability of neural networks: adversarial attack

Neural networks are sometimes very sensitive to a small noise, the *adversarial attack*: $x \to x + \xi$ by choosing well-designed $\xi$ with $\|\xi\| \le \varepsilon$. It sabotages the performance of neural networks.

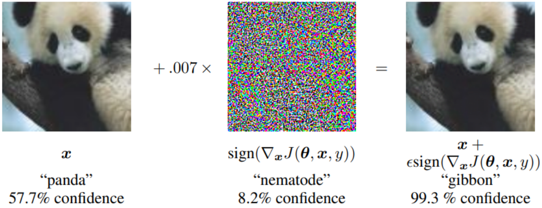



Figure: Adversarial examples generated for GoogLeNet (Goodfellow, Shlens, Szegedy[4]).

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

# Questions are

- How to understand this phenomenon? What is the meaning of adversarial attack?
- How to compute the risk of this model? How to obtain an optimal adversarial attack?
- How to train a classifier to make it optimal and robust against such all noise?

# Classification problem

- $(\mathcal{X}, d)$ : Feature space, $\mathcal{Y} := \{1, \ldots, K\}$ : Class space.
- $\Delta_{\mathcal{Y}} := \{(u_1, \ldots, u_K) : 0 \leq u_i \leq 1, \sum_{i=1}^{K} u_i \leq 1\}$ : the set of distributions over $\mathcal{Y}$.
- $\mu = (\mu_1, \ldots, \mu_K)$ : a data distribution; $\mu_i$ is a distribution over $\mathcal{X}$ given $Y = i$.
- $f = (f_1, \ldots, f_K) : \mathcal{X} \to \Delta_{\mathcal{Y}}$, a measurable probabilistic classifier.
- $\ell(f(x), i) := 1 - f_i(x)$ : 0-1 loss function.

A learning problem aims at solving

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

# Table of Contents

# Optimal transport

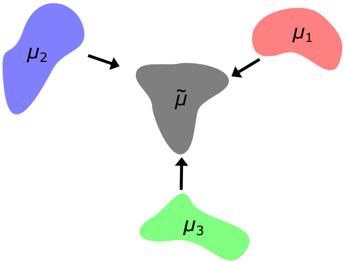
Given probability measures $\mu$, $\nu$ on spaces $\mathcal{S}_1$ and $\mathcal{S}_2$, respectively, and a cost function $c : \mathcal{S}_1 \times \mathcal{S}_2 \to [-\infty, \infty]$, optimal transport(OT) is defined as

$$C(\mu, \nu) := \min_{\pi \in \Pi(\mu,\nu)} \int_{\mathcal{S}_1 \times \mathcal{S}_2} c(s_1, s_2) d\pi(s_1, s_2)$$

where $\Pi(\mu, \nu)$ is the set of joint distributions whose marginals are $\mu$ and $\nu$.

- $\Pi(\mu, \nu)$ is convex and weakly compact.
- Under general conditions, there is a solution.

# Barycenter problem




- Introduced by Ekeland[5], Chiappori, McCann, Nesheim[6], Agueh, Carlier[7].
- $\tilde{\mu} \in \arg\min_\nu \sum_{i=1}^K C(\mu_i, \nu)$.

[5] Ekeland, "An optimal matching problem".
[6] Chiappori, McCann, and Nesheim, "Hedonic price equilibria, stable matching, and optimal transport: equivalence, topology, and uniqueness".
[7] Agueh and Carlier, "Barycenters in the Wasserstein space".

# Multimarginal optimal transport(MOT)

The multimarginal optimal transport(MOT) problem is the generalization of OT to $K$-marginal constraints:

$$\inf_{\pi \in \Pi(\mu_1,\ldots,\mu_K)} \int_{\mathcal{S}_1 \times \cdots \times \mathcal{S}_K} \boldsymbol{c}(s_1,\ldots,s_K) d\pi(s_1,\ldots,s_K).$$

- Applications in physics (density function theory).
- Deep connection to barycenter problems (by taking $\boldsymbol{c}(x_1,\ldots,x_K) = \inf_x \sum_{i=1}^K c(x,x_i)$).
- Machine learning, statistics and etc.

# DRO adversarial model

The adversary perturbs the distribution $\mu$:

$$\mu \longmapsto \tilde{\mu} \in \arg\max_{\nu} \{R(f,\nu) - C(\mu,\nu)\}$$

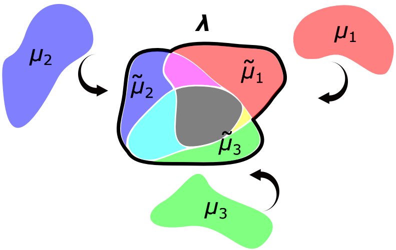
where $C(\mu,\tilde{\mu})$ is a transport cost defined as

$$C(\mu,\tilde{\mu}) := \sum_{i\in\mathcal{Y}} \inf_{\pi_i \in \Pi(\mu_i,\tilde{\mu}_i)} \int c(x,\tilde{x})d\pi(x,\tilde{x}).$$

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

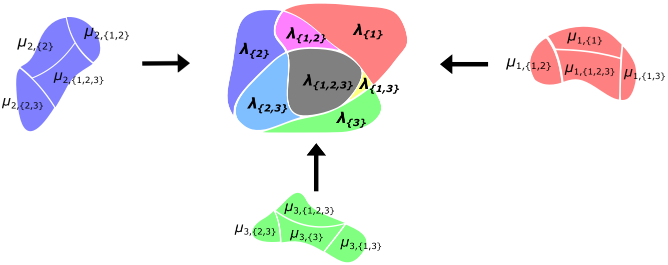
$$\inf_f \sup_{\tilde{\mu}} \{R(f,\tilde{\mu}) - C(\mu,\tilde{\mu})\}.$$

# Optimal adversarial attack




- A classification problem becomes harder as $\mu_i$'s are similar.
- The optimal adversarial attacks will be $\tilde{\mu}_i \approx \tilde{\mu}_j$, or a barycenter $\lambda$ of $\mu_i$'s such that $\lambda \approx \tilde{\mu}_i$ in some sense.

# Generalized barycenter problem




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

# Equivalence

## Theorem (**K.**, García Trillos, Jacobs[8](JMLR))

*DRO model is equivalent to generalized barycenter problem. Also, generalized barycenter problem has a solution, and MOT formulation.*

- First geometric understanding of the adversarial training model.
- Connect it to MOT, so computable explicitly.
- Extend previous literature of the binary setting (Bhagoji, Cullina, Mittal[9], Pydi, Jog[10], García Trillos, Murray[11]).

---

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

[9] Bhagoji, Cullina, and Mittal, "Lower Bounds on Adversarial Robustness from Optimal Transport".

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

[11] García Trillos and Murray, "Adversarial Classification: Necessary Conditions and Geometric Flows".

# Existence of robust classifier

**Theorem (K., García Trillos, Jacobs[12] (accepted by EJAM))**
*DRO model has a (Nash) equilibrium, a pair of optimal classifiers and adversarial attacks. Also, variants of adversarial training models are equivalent.*

- Rigorous proof of the existence results beyond the binary setting (Awasthi, Frank, Mohri[13], Frank, Niles-Weed[14]).
- Unify variant adversarial training models and total-variation regularization problem (Bungert, García Trillos, Murray[15]).

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

[13] Awasthi, Frank, and Mohri, "On the existence of the adversarial bayes classifier".

[14] Frank and Niles-Weed, "Existence and minimax theorems for adversarial surrogate risks in binary classification".

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

# Efficient Algorithm

Theorem (**K.**, García Trillos, Jacobs, Werenski[16](accepted by JMLR))

*With truncation level $L < K$, there is an algorithm to compute a lower bound of the adversarial risk within $\tilde{O}(n^L)$.*

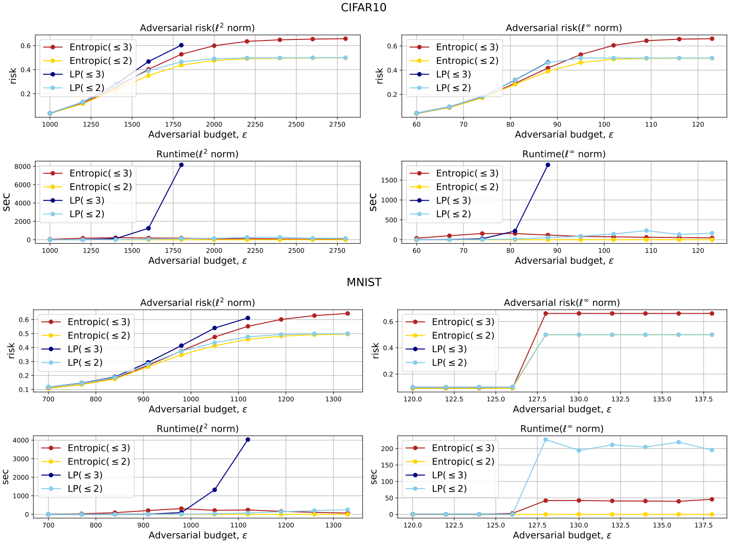
- Crucially depends on the special structure of this problem.
- Use entropic regularization (Lin et al.[17] shows for MOT with $K$ marginals, its complexity is $\tilde{O}(N^K)$).
- MOT is NP-hard in the worst case (Altschuler, Boix-Adserà[18]).

[16] García Trillos et al., *An Optimal Transport Approach for Computing Adversarial Training Lower Bounds in Multiclass Classification*.

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

[18] Altschuler and Boix-Adserà, "Wasserstein barycenters are NP-hard to compute".

# Synthetic data analysis



# Table of Contents

# Some future works

- Return to neural networks: how to apply this technique for them?
- PAC learnability of the multiclass adversarial learning: adversarial learning of the binary setting (Montasser, Hanneke, Srebro[19]), vanilla multiclass learning (Brukhim et al.[20]).
- Quantify the regularity of robust classifier (Bungert, García Trillos, Murray[21]).
- Sample complexity; unlike $W_2$, a popular cost function in adversarial training models is very singular.

[19] Montasser, Hanneke, and Srebro, "VC classes are adversarially robustly learnable, but only improperly".

[20] Brukhim et al., "A characterization of multiclass learnability".

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

# Thank you for your attention!