Adversarial Training Through the Lens of Optimal Transport

Nicolás García Trillos
University of Wisconsin-Madison

Kantorovich Initiative seminar series
February 2023

Based on joint works with:

- Jakwang Kim (Wisc)
- Camilo García Trillos (UCL)
- Matt Jacobs (Purdue)

Based on joint works with:

- Jakwang Kim (joining Kantorovich initiative this Fall!)
- Camilo García Trillos (UCL)
- Matt Jacobs (Purdue)

Neural networks, although accurate on clean data, are sensitive to adversarial attacks:

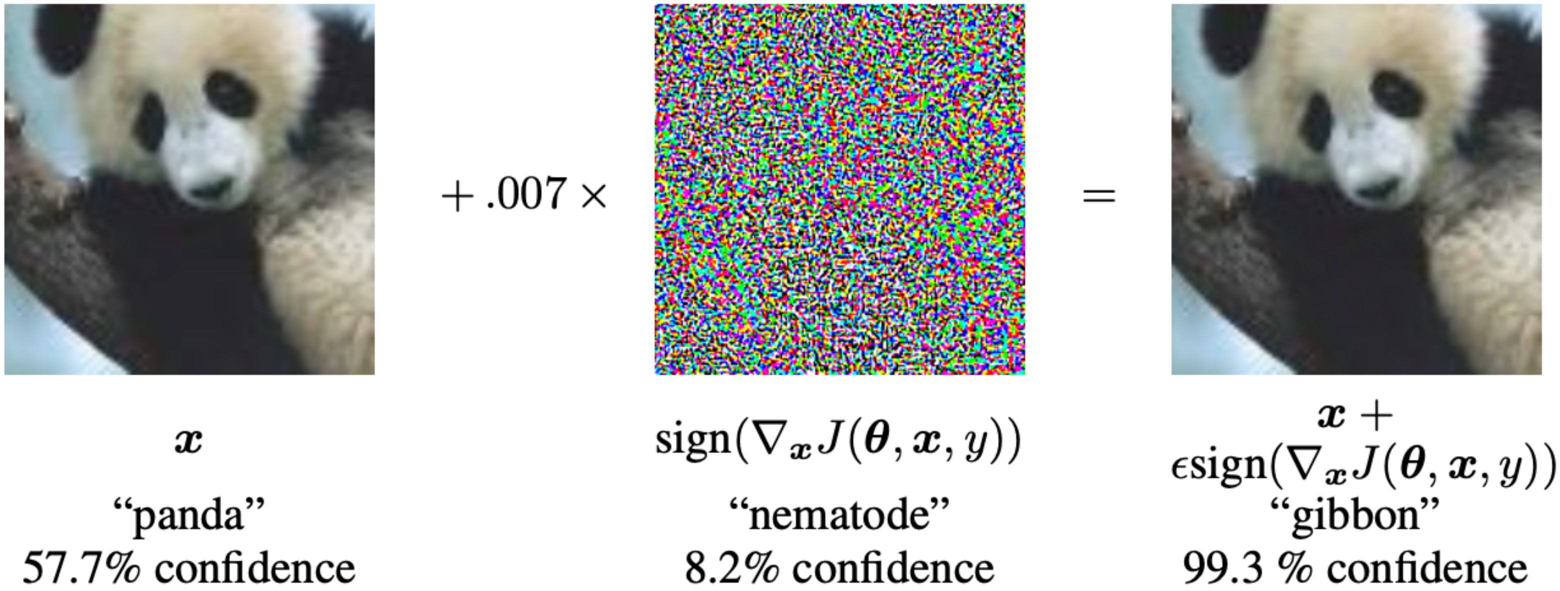


Figure: Picture taken from Goodfellow et al. (2015)

[Szegedy et al. 2014] , [Goodfellow et al. 2015]

Figure: An adversarial attack of a clean image in a safety-critical setting. Picture taken from Eykholt et al. (2018)

Formalization of adversarial training problem

How to train classifiers to be robust to (specific) adversarial attacks?:

$$\min_{\theta \in \Theta} \mathbb{E}_{(x,y)\sim\mu} \left[\sup_{\tilde{x} \in B_\varepsilon(x)} \ell(\theta, (\tilde{x}, y)) \right]. \qquad \text{(AT)}$$

[Madry et al 2017]

Compare to unrobust problem:

$$\min_{\theta \in \Theta} \mathbb{E}_{(x,y)\sim\mu} \left[\ell(\theta, (x, y)) \right].$$

Formalization of adversarial training problem

How to train classifiers to be robust to (specific) adversarial attacks?:

$$\min_{\theta \in \Theta} \mathbb{E}_{(x,y)\sim\mu} \left[\sup_{\tilde{x} \in B_\varepsilon(x)} \ell(\theta, (\tilde{x}, y)) \right],$$

or its **distributionally robust optimization (DRO)** version:

$$\inf_{\theta \in \Theta} \sup_{\tilde{\mu} : D(\mu, \tilde{\mu}) \leq \varepsilon} \mathbb{E}_{\tilde{z}\sim\tilde{\mu}} \left[\ell(\tilde{z}, \theta) \right].$$

Formalization of adversarial training problem

How to train classifiers to be robust to (specific) adversarial attacks?:

$$\min_{\theta \in \Theta} \mathbb{E}_{(x,y)\sim\mu} \left[\sup_{\tilde{x} \in B_\varepsilon(x)} \ell(\theta, (\tilde{x}, y)) \right],$$

or its distributionally robust optimization (DRO) version:

$$\inf_{\theta \in \Theta} \sup_{\tilde{\mu} : D(\mu, \tilde{\mu}) \leq \varepsilon} \mathbb{E}_{\tilde{z}\sim\tilde{\mu}} \left[\ell(\tilde{z}, \theta) \right],$$

or its explicit penalization version:

$$\inf_{\theta \in \Theta} \sup_{\tilde{\mu}} \mathbb{E}_{\tilde{z}\sim\tilde{\mu}} \left[\ell(\tilde{z}, \theta) \right] - C(\mu, \tilde{\mu}).$$

$$\inf_{\theta \in \Theta} \sup_{\tilde{\mu}} \mathbb{E}_{\tilde{z} \sim \tilde{\mu}} [\ell(\tilde{z}, \theta)] - C(\mu, \tilde{\mu}). \tag{AT}$$

- How do we find a solution to this problem?
- Can we find meaningful upper and lower bounds?

$$\inf_{\theta\in\Theta} \sup_{\tilde{\mu}: D(\mu,\tilde{\mu})\leq\varepsilon} \mathbb{E}_{\tilde{z}\sim\tilde{\mu}} \left[\ell(\tilde{z},\theta)\right]. \tag{AT}$$

- How do we find a solution to this problem?
- Can we find meaningful upper and lower bounds?
- How is a problem like (AT) related to regularization methods?

$$\inf_{\theta \in \Theta} \sup_{\tilde{\mu}} \mathbb{E}_{\tilde{z} \sim \tilde{\mu}} [\ell(\tilde{z}, \theta)] - C(\mu, \tilde{\mu}). \tag{AT}$$

- How do we find a solution to this problem?
- Can we find meaningful lower and upper bounds?
- How is a problem like (AT) related to regularization methods? **i.e. a problem like:**

$$\inf_{\theta \in \Theta} R(\mu, \theta) + \lambda F(\theta), \tag{Reg}$$

$$\inf_{\theta \in \Theta} \sup_{\tilde{\mu}: D(\mu, \tilde{\mu}) \leq \varepsilon} \mathbb{E}_{\tilde{z} \sim \tilde{\mu}} \left[\ell(\tilde{z}, \theta) \right]. \tag{AT}$$

- How do we find a solution to this problem?
- Can we find meaningful lower and upper bounds?
- How is a problem like (AT) related to regularization methods? **i.e. a problem like:**

$$\inf_{\theta \in \Theta} R(\mu, \theta) + \lambda F(\theta), \tag{Reg}$$

$$\inf_{\theta \in \Theta} \sup_{\tilde{\mu}} \mathbb{E}_{\tilde{z} \sim \tilde{\mu}} [\ell(\tilde{z}, \theta)] - C(\mu, \tilde{\mu}). \tag{AT}$$

What is the **geometry** of:

- Optimal robust classifiers.
- Optimal adversarial attacks.

Instead of the parametric problem

$$\inf_{\theta \in \Theta} \sup_{\tilde{\mu}} \mathbb{E}_{\tilde{z} \sim \tilde{\mu}} [\ell(\tilde{z}, \theta)] - C(\mu, \tilde{\mu}). \tag{1}$$

we'll consider non-parametric problems:

$$\inf_{f \in \mathcal{F}} \sup_{\tilde{\mu}} \mathbb{E}_{\tilde{z} \sim \tilde{\mu}} [\ell(\tilde{z}, f)] - C(\mu, \tilde{\mu}). \tag{2}$$

Outline

We'll consider two settings:

1. A multilabel classification problem with an agnostic learner.
2. A regression problem in a mean field regime.

Outline

We'll consider two settings:

1. A multilabel classification problem with an agnostic learner.
   - Lower bounds for general AT problems.
   - Connections to MOT and (generalized) barycenter problems.
2. A regression problem in a mean field regime.
   - How to find (approximate) Nash equilibria in mean-field learning settings.

Overarching goal: an invitation to look at (AT) from geometric and analytic perspectives.

1. A multilabel classification problem with an agnostic learner

A multilabel classification problem with an agnostic learner

- *Type of data:* $z = (x, y) \in \mathbb{R}^d \times \{1, \dots, k\}$, $\mathbf{k \geq 2}$.
- *Learner's actions:* measurable $f = (f_1, \dots, f_k)$ with: $f_l : \mathbb{R}^d \to [0, 1]$, and $\sum_{l=1}^{K} f_l = 1$ (Agnostic learner).
- *Loss function:* $\ell(z, f) = \ell((x, y), f) = 1 - f_y(x)$, i.e. 0-1 loss.

A multilabel classification problem with an agnostic learner

$$\inf_{f} \sup_{\tilde{\mu} \in \mathcal{P}(\mathcal{Z})} \left\{ \mathbb{E}_{(\tilde{x},\tilde{y})\sim\tilde{\mu}} \left(\ell(\tilde{z}, f) \right) - C(\mu, \tilde{\mu}) \right\},$$

where

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

for some cost function $c_{\mathcal{Z}} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ of the form:

$$c_{\mathcal{Z}}(z, \tilde{z}) = \begin{cases} c(x, \tilde{x}) & \text{if } y = \tilde{y} \\ \infty & \text{if } y \neq \tilde{y}, \end{cases} \qquad c : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty].$$

Lower bounds for more general AT problems:

$$\inf_{f \text{ measurable}} \sup_{\tilde{\mu} \in \mathcal{P}(\mathcal{Z})} \left\{ \mathbb{E}_{(\tilde{x},\tilde{y})\sim\tilde{\mu}} \left(\ell(\tilde{z}, f) \right) - C(\mu, \tilde{\mu}) \right\},$$

is smaller than

$$\inf_{f \in \mathcal{F}'} \sup_{\tilde{\mu} \in \mathcal{P}(\mathcal{Z})} \left\{ \mathbb{E}_{(\tilde{x},\tilde{y})\sim\tilde{\mu}} \left(\ell(\tilde{z}, f) \right) - C(\mu, \tilde{\mu}) \right\}.$$

Example of cost function:

$$\inf_f \sup_{\tilde{\mu} \in \mathcal{P}(\mathcal{Z})} \left\{ \mathbb{E}_{(\tilde{x},\tilde{y})\sim\tilde{\mu}} \left(\ell(\tilde{z}, f) \right) - C(\mu, \tilde{\mu}) \right\},$$

where

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

for some cost function $c_{\mathcal{Z}} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ of the form:

$$c_{\mathcal{Z}}(z, \tilde{z}) = \begin{cases} c(x, \tilde{x}) & \text{if } y = \tilde{y} \\ \infty & \text{if } y \neq \tilde{y}, \end{cases} \qquad c(x, \tilde{x}) = \begin{cases} 0 & \text{if } d(x, \tilde{x}) \leq \varepsilon \\ \infty & \text{if } d(x, \tilde{x}) > \varepsilon \end{cases}$$

Example of cost function:

When

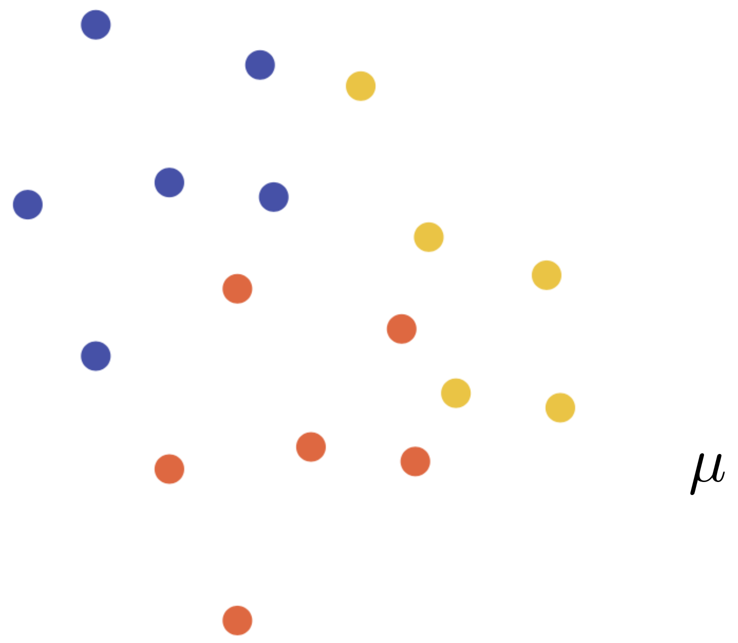
$$c(x,\tilde{x}) = \begin{cases} 0 & \text{if } d(x,\tilde{x}) \leq \varepsilon \\ \infty & \text{if } d(x,\tilde{x}) > \varepsilon \end{cases},$$

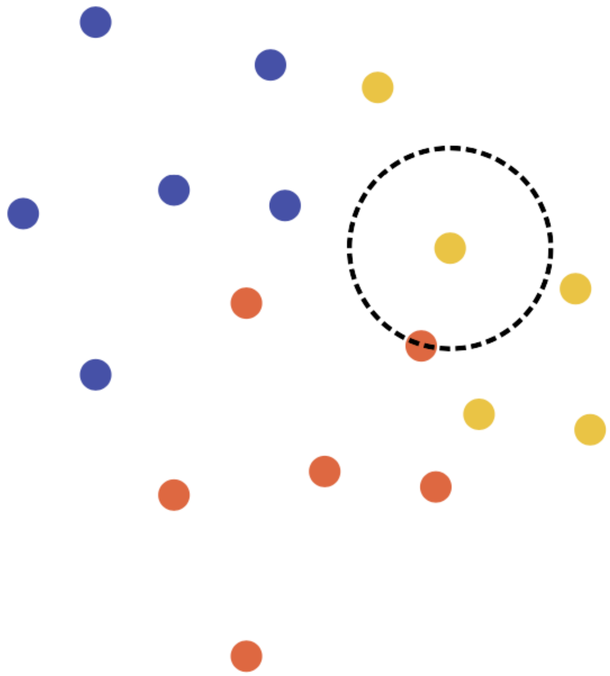
problem

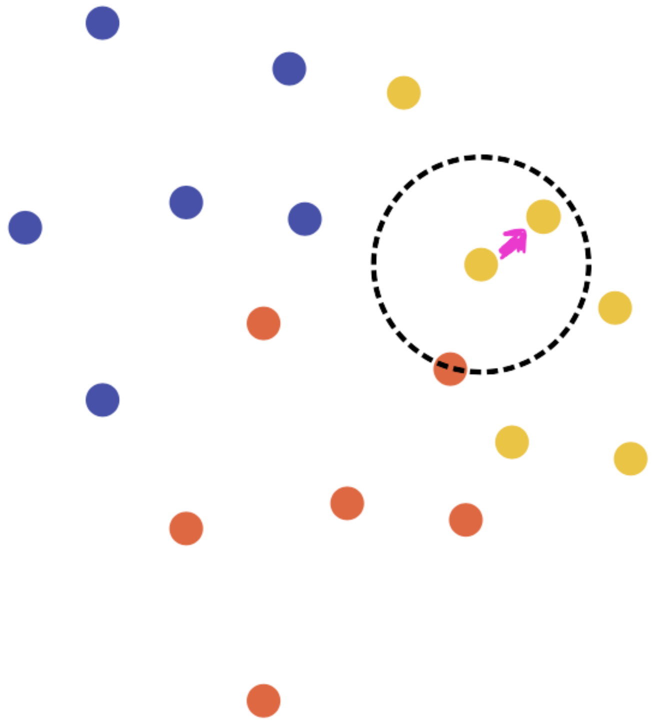
$$\inf_f \sup_{\tilde{\mu} \in \mathcal{P}(\mathcal{Z})} \left\{ \mathbb{E}_{(\tilde{x},\tilde{y})\sim\tilde{\mu}} \left(\ell(\tilde{z}, f)\right) - C(\mu,\tilde{\mu}) \right\}$$

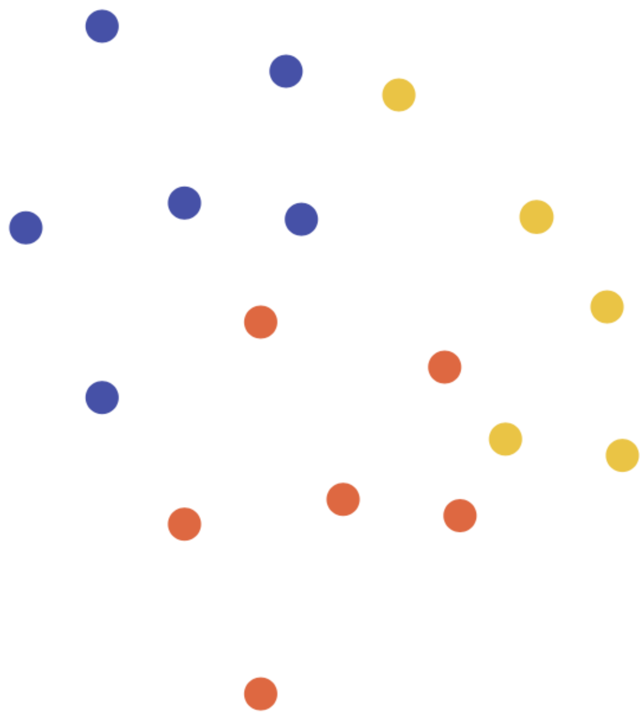
becomes:

$$\inf_f \mathbb{E}_{(x,y)\sim\mu} \left(\sup_{\tilde{x} \in B_\varepsilon(x)} \ell((\tilde{x}, y), f) \right).$$


μ







$\tilde{\mu}$

$$\inf_f \sup_{\tilde{\mu}\in\mathcal{P}(\mathcal{Z})} \left\{ \mathbb{E}_{(\tilde{x},\tilde{y})\sim\tilde{\mu}} \left(\ell(\tilde{z}, f)\right) - C(\mu, \tilde{\mu}) \right\},$$

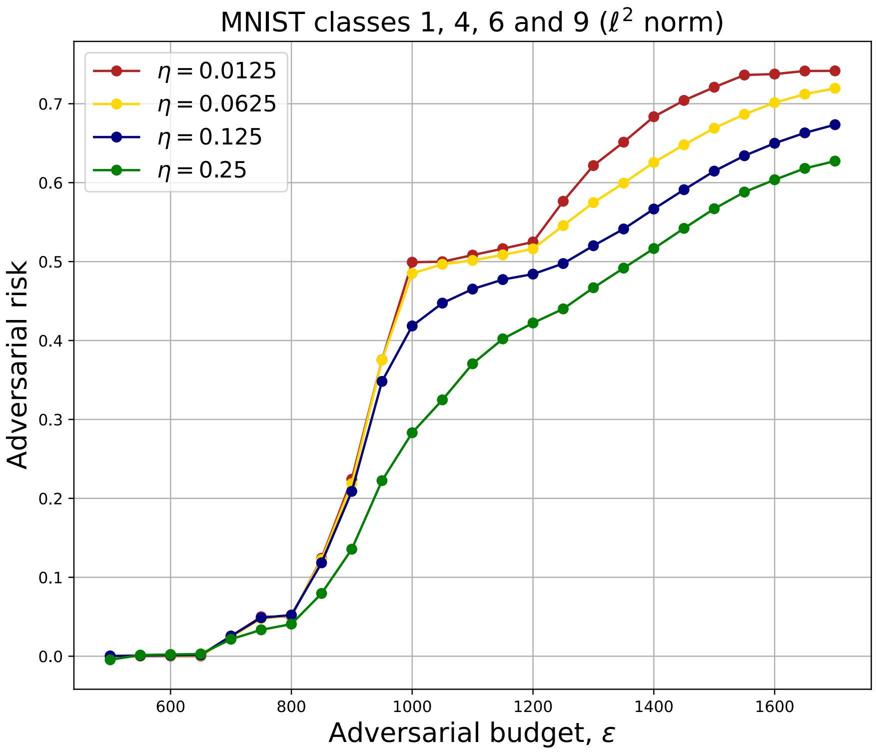
where

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

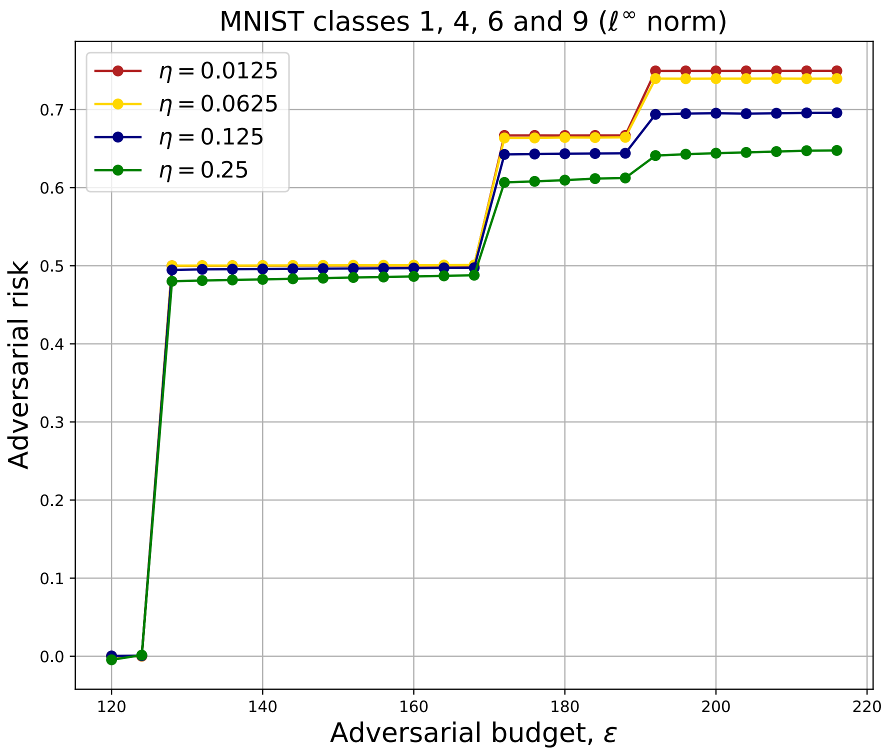
for some cost function $c_{\mathcal{Z}} : \mathcal{Z} \times \mathcal{Z} \to \mathbb{R}$ of the form:

$$c_{\mathcal{Z}}(z, \tilde{z}) = \begin{cases} c(x, \tilde{x}) & \text{if } y = \tilde{y} \\ \infty & \text{if } y \neq \tilde{y}, \end{cases} \qquad c : \mathbb{R}^d \times \mathbb{R}^d \to [0, \infty].$$

MNIST

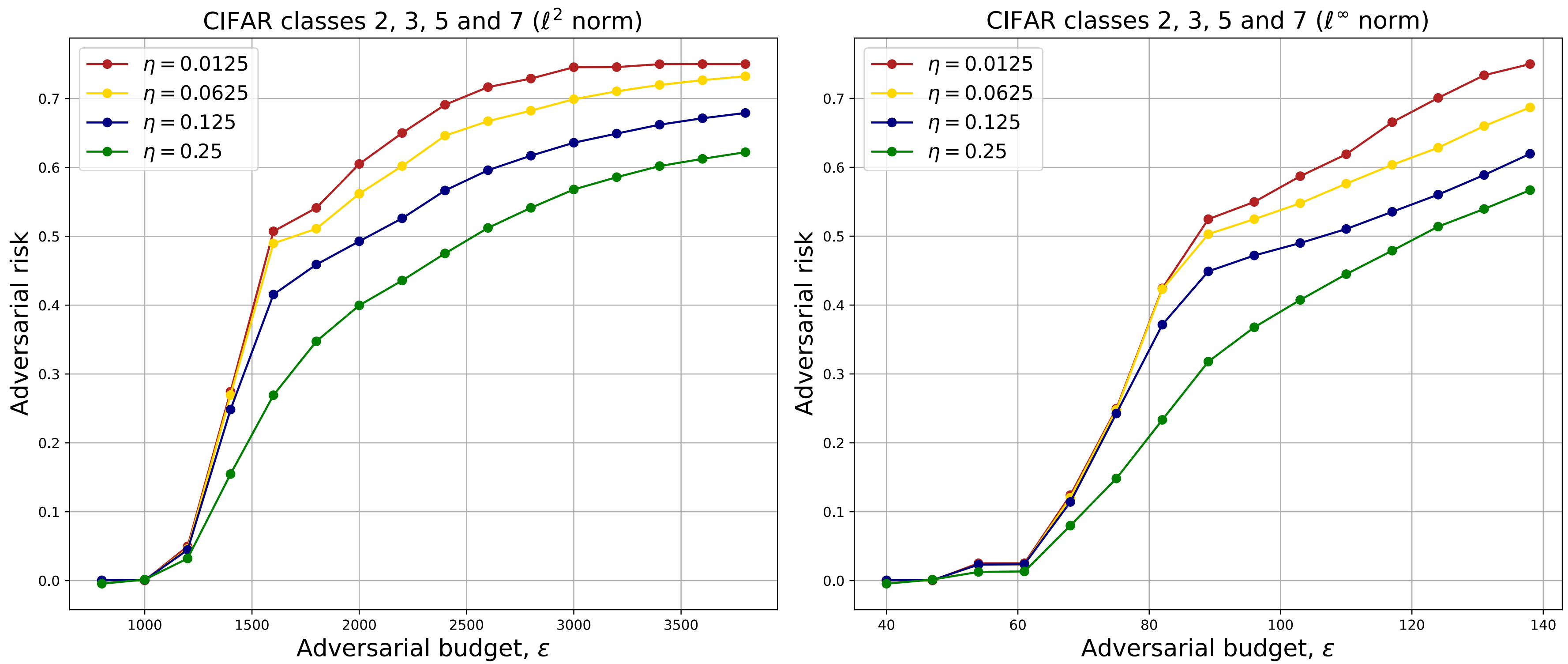


MNIST classes 1, 4, 6 and 9 (ℓ^2 norm)
$\eta = 0.0125$
$\eta = 0.0625$
$\eta = 0.125$
$\eta = 0.25$
Adversarial risk
Adversarial budget, ε
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
600
800
1000
1200
1400
1600



MNIST classes 1, 4, 6 and 9 (ℓ^∞ norm)
$\eta = 0.0125$
$\eta = 0.0625$
$\eta = 0.125$
$\eta = 0.25$
Adversarial risk
Adversarial budget, ε
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
120
140
160
180
200
220

CIFAR-10



We computed the above using off-the-shelf MOT solvers...

Multimarginal Optimal Transport (MOT)

$$\inf_{\Gamma(\rho_1,\dots,\rho_K)} \int \mathbf{c}(\xi_1,\dots,\xi_K) d\pi(\xi_1,\dots,\xi_K) \tag{MOT}$$

- Applications in Physics.
- Applications in Economics.
- Machine learning.

$$\inf_{\Gamma(\rho_1,\dots,\rho_K)} \int \mathbf{c}(\xi_1,\dots,\xi_K)\, d\pi(\xi_1,\dots,\xi_K)$$

where

$$\mathbf{c}(\xi_1,\dots,\xi_K) := \sum_{1\leq i<j\leq K} f(d(x_i,x_j)).$$

[Seidl 1999], [Gori-Giorgi et al. 2009].

Barycenter problems

$$\inf_{\Gamma(\rho_1,\dots,\rho_K)} \int \mathbf{c}(\xi_1,\dots,\xi_K)d\pi(\xi_1,\dots,\xi_K)$$

where

$$\mathbf{c}(\xi_1,\dots,\xi_K) := \inf_{\xi'\in\mathcal{X}} \sum_{i=1}^{K} c(\xi',\xi_i).$$

Barycenter problems

$$\inf_{\Gamma(\rho_1,\ldots,\rho_K)} \int c(\xi_1,\ldots,\xi_K) d\pi(\xi_1,\ldots,\xi_K)$$

where

$$c(\xi_1,\ldots,\xi_K) := \inf_{\xi' \in \mathcal{X}} \sum_{i=1}^{K} c(\xi', \xi_i).$$

Equivalent to:

$$\inf_{\rho} \sum_{i=1}^{K} C(\rho_i, \rho),$$

where

$$C(\rho_i, \rho) := \inf_{\pi \in \Gamma(\rho_i, \rho)} \int c(x, x') d\pi(x, x').$$

[Agueh and Carlier 2011].

What is the connection between (AT) and (MOT)?

What is the connection between (AT) and (MOT)?

- How to find a saddle $(\tilde{\mu}^*, f^*)$ for the (AT) problem?

What is the connection between (AT) and (MOT)?

- How to find a saddle $(\tilde{\mu}^*, f^*)$ for the (AT) problem?
  Answer: Solve a certain MOT problem and its dual.

Theorem [NGT, Jacobs, Kim 22']: For arbitrary $\mathbf{k} \geq \mathbf{2}$

$$(AT)(\mu) = 1 - \frac{1}{2} \inf_{\pi \in \Pi_k(\mu)} \int_{\mathcal{Z}_*^K} \mathbf{c}(z_1, \dots, z_K) d\pi(z_1, \dots, z_K),$$

for some cost function $\mathbf{c}$.

- From π^* can construct $\tilde{\mu}^*$.
- $\tilde{\mu}^*$ concentrates on barycenters (w.r.t. cost c) of groups of k or less points in the support of μ_x.
- From dual of (MOT) can construct f^*.

Theorem [NGT, Jacobs, Kim 22']: For arbitrary **k** $\geq$ **2**

$$(AT)(\mu) = 1 - \frac{1}{2} \inf_{\pi \in \Pi_k(\mu)} \int_{\mathcal{Z}_*^K} \mathbf{c}(z_1, \dots, z_K) d\pi(z_1, \dots, z_K),$$

for some cost function **c**. A given $c : \mathcal{X} \times \mathcal{X} \to [0, \infty]$ induces a **c**.

- From π^* can construct $\tilde{\mu}^*$.
- $\tilde{\mu}^*$ concentrates on barycenters (w.r.t. cost c) of groups of k or less points in the support of μ_x.
- From dual of (MOT) can construct f^*.

Precise MOT problem

Set $\mathcal{Z}_* := \mathcal{Z} \cup \{\mathghost\}$.

$$\inf_{\pi \in \Pi_k(\mu)} \int_{\mathcal{Z}_*^K} \mathbf{c}(z_1, \dots, z_K) d\pi(z_1, \dots, z_K).$$

- Couplings:

$$\Pi_K(\mu) := \left\{ \pi \in \mathcal{P}(\mathcal{Z}_*^K) \ : \ P_{i\sharp}\pi = \frac{1}{2\mu(\mathcal{Z})}\mu(\cdot \cap \mathcal{Z}) + \frac{1}{2}\delta_{\mathghost}, \quad \forall i \right\}.$$

- Cost:

$$\mathbf{c}(z_1, \dots, z_K) := \widehat{\mu}_{\vec{z}}(\mathcal{Z}) - AT(\widehat{\mu}_{\vec{z}}),$$

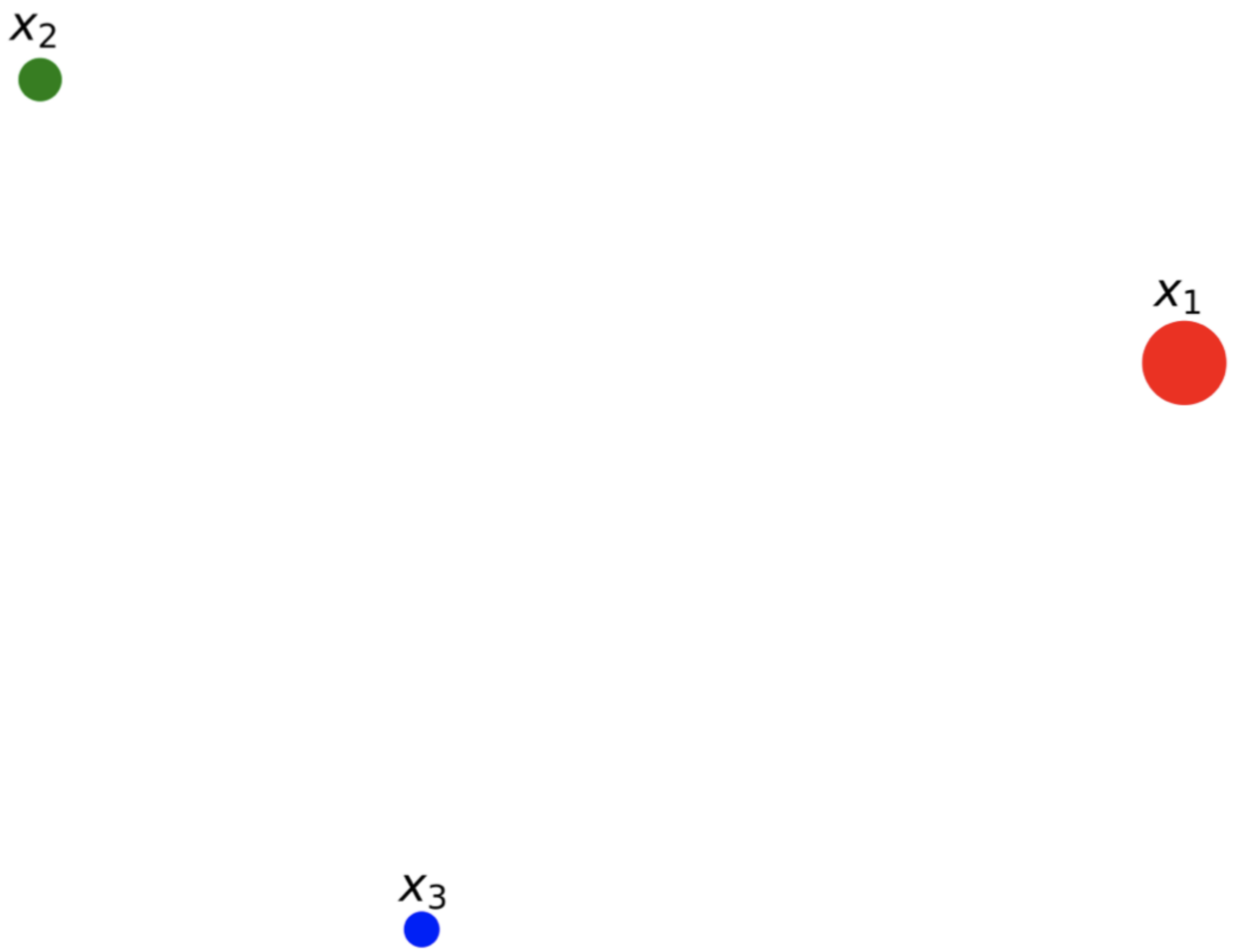
where $\widehat{\mu}_{\vec{z}}$ is the positive measure defined as:

$$\widehat{\mu}_{\vec{z}} := \frac{1}{K} \sum_{l \text{ s.t. } z_l \neq \mathghost}^{K} \delta_{z_l}.$$

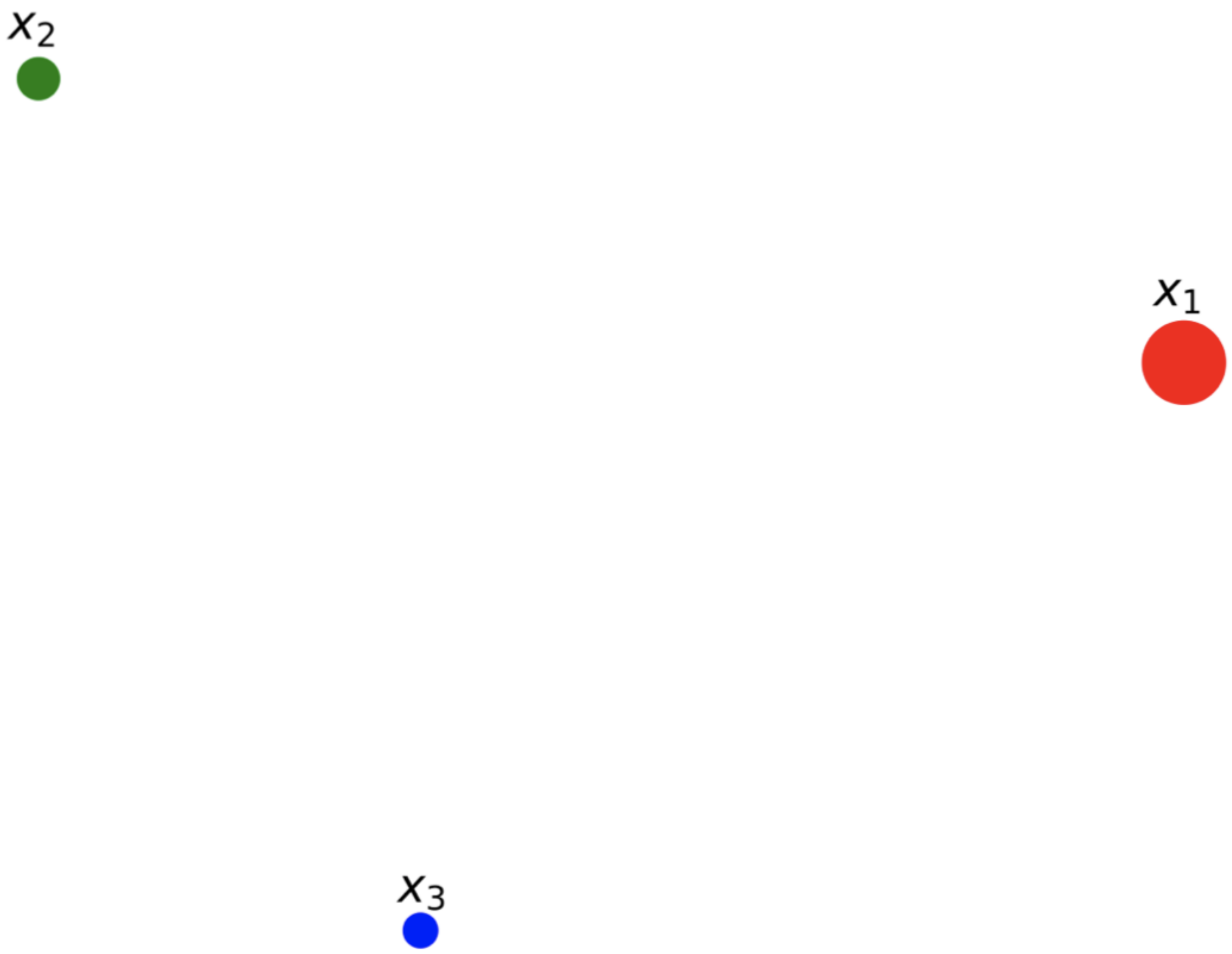
Toy example

Let $c(x,\tilde{x}) = c_\varepsilon(x,\tilde{x}) = \begin{cases} 0 & \text{if } d(x,\tilde{x}) \leq \varepsilon \\ +\infty & \text{otherwise} \end{cases}$

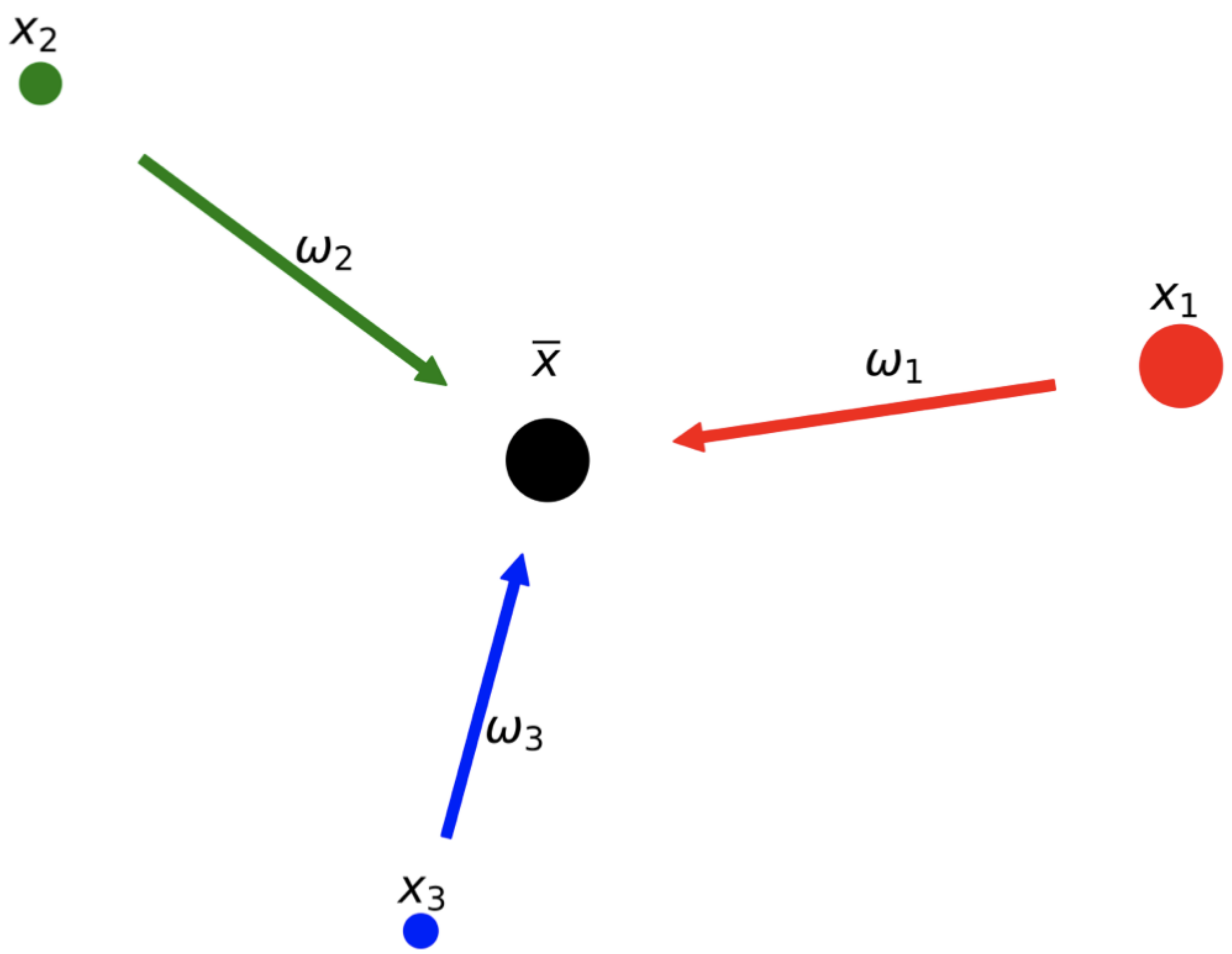
$$\mu = \omega_1 \delta_{(x_1,1)} + \omega_2 \delta_{(x_2,2)} + \omega_3 \delta_{(x_3,3)}$$



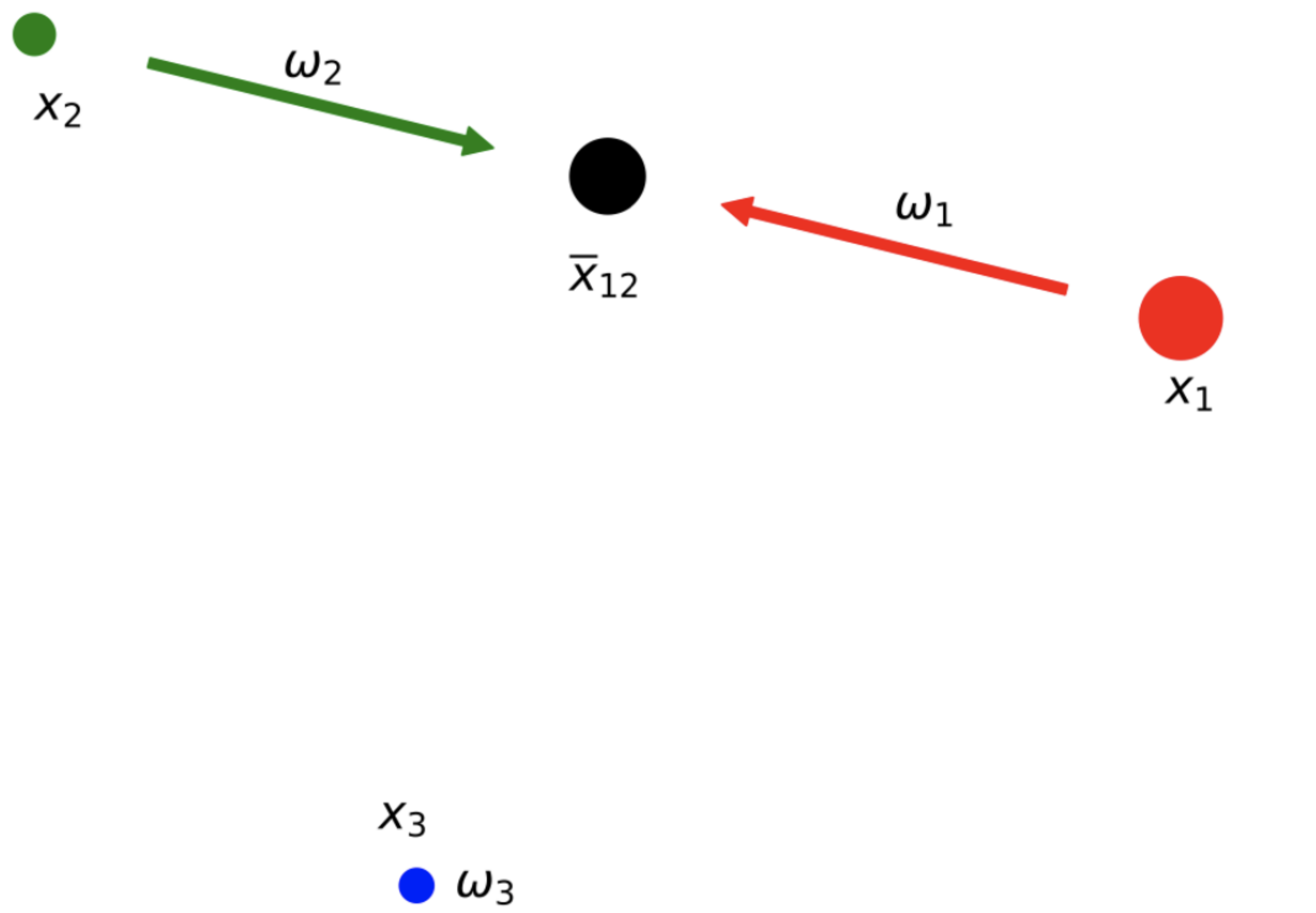
Case 1:



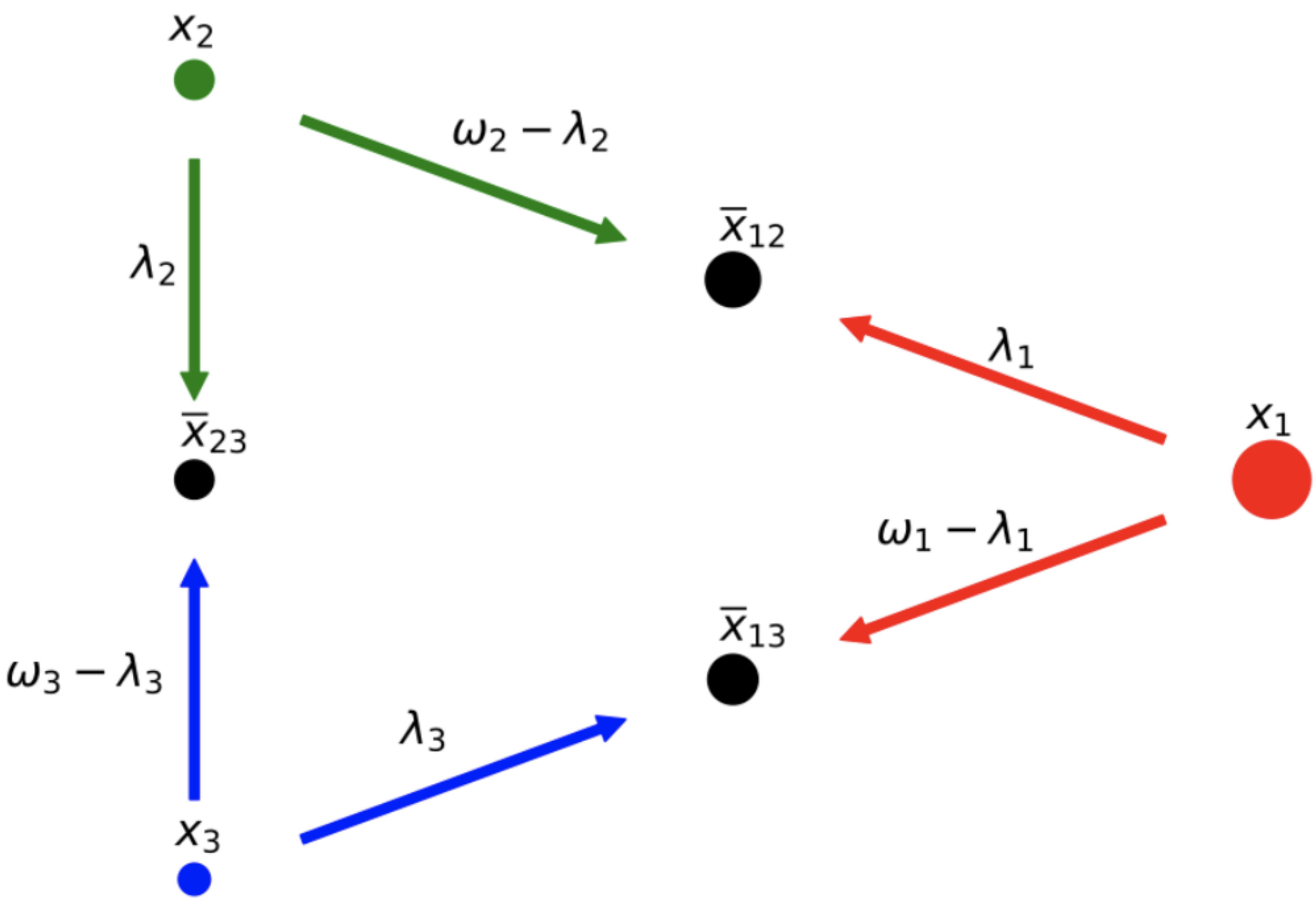
Case 2:



Case 3:

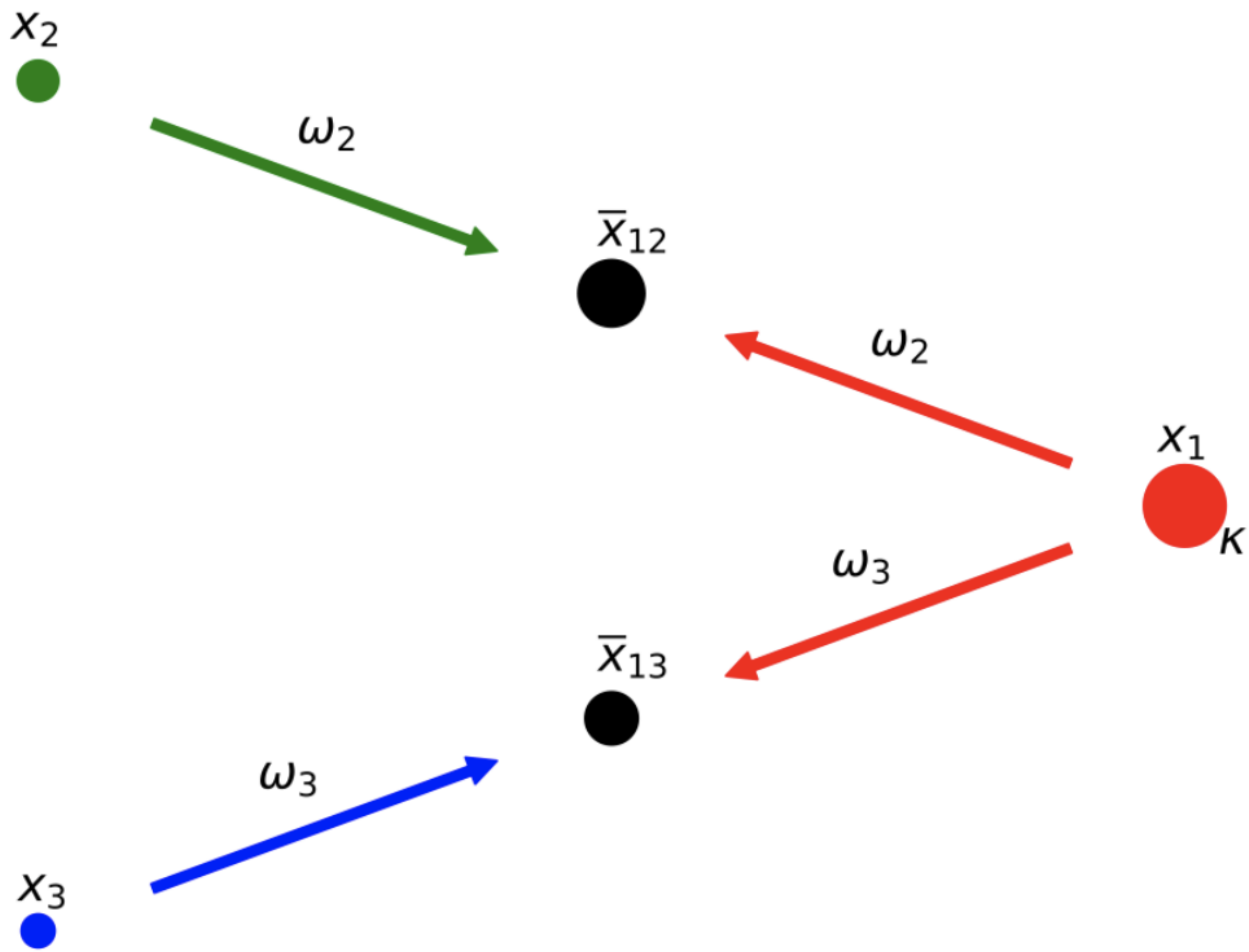


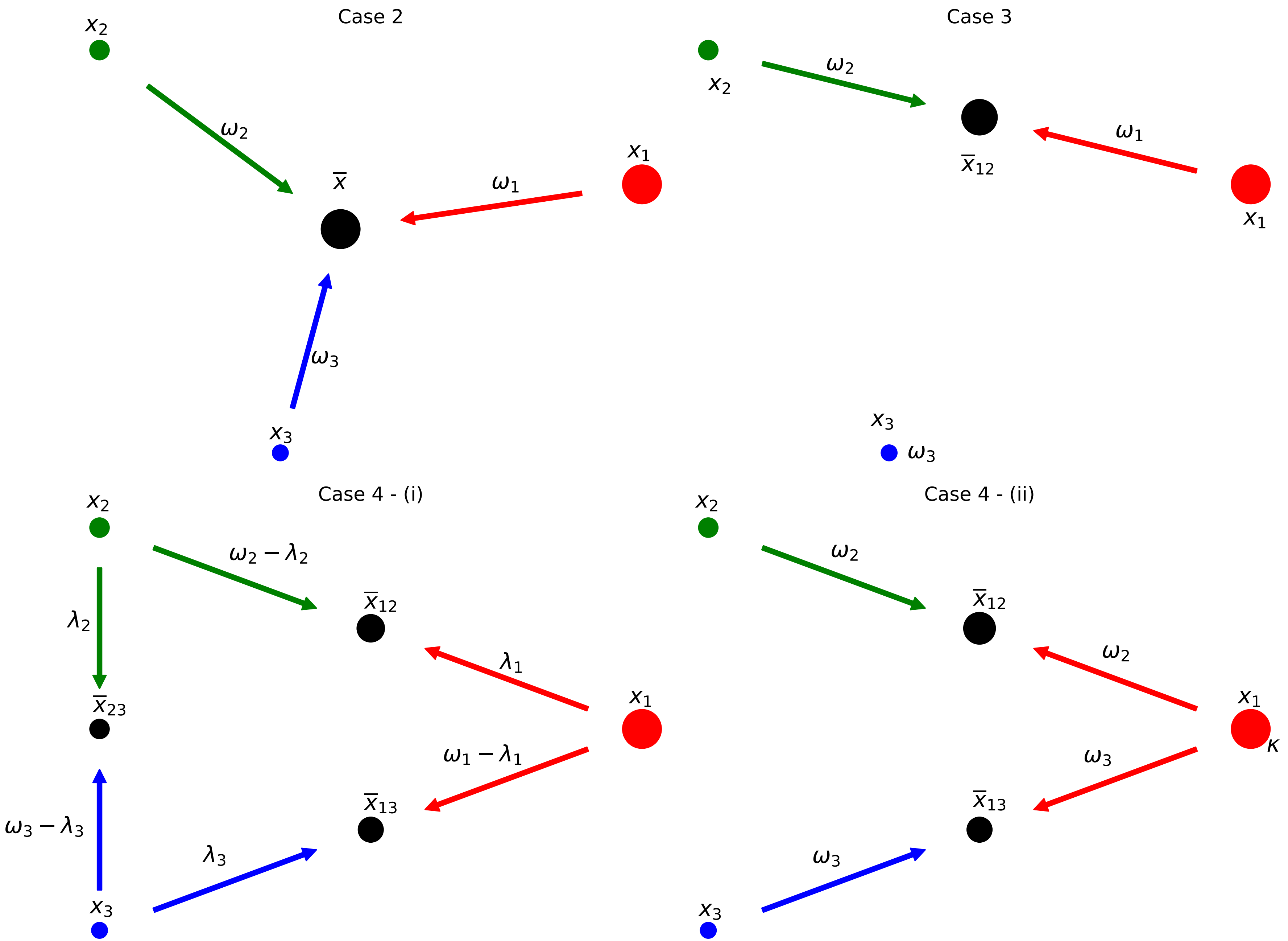
Case 4 i:



x_2
$\omega_2 - \lambda_2$
λ_2
$\bar{x}_{12}$
λ_1
$\bar{x}_{23}$
x_1
$\omega_1 - \lambda_1$
$\bar{x}_{13}$
$\omega_3 - \lambda_3$
λ_3
x_3

Case 4 ii:




Case 2
Case 3
Case 4 - (i)
Case 4 - (ii)

$$
\color{red}{(\mathrm{AT})(\mu)} \color{black}{= 1 - \frac{1}{2} \inf_{\pi \in \Pi_k(\mu)} \int_{\mathcal{Z}_*^K} \mathbf{c}(z_1, \dots, z_K) d\pi(z_1, \dots, z_K),}
$$

for cost function $\mathbf{c}$:

$$
\mathbf{c}(z_1, \dots, z_K) := \color{red}{\widehat{\mu}_{\vec{z}}(\mathcal{Z}) - AT(\widehat{\mu}_{\vec{z}})}\color{black}{,}
$$

where $\widehat{\mu}_{\vec{z}}$ is the positive measure defined as:

$$
\widehat{\mu}_{\vec{z}} := \frac{1}{K} \sum_{l \text{ s.t. } z_l \neq \text{👻}}^{K} \delta_{z_l}.
$$

Theorem (NGT, Jacobs, Kim, 2022)

Suppose that (π^*, ϕ^*) *is a solution pair for the MOT problem and its dual. Define* f^* *and* $\widetilde{\mu}^*$ *according to:*

$$f_i^* := \Big(\max \Big\{ \sum_{j=1}^{K} \phi_j^*(\cdot, i) + \sum_{j=1}^{K} \phi_j^*(\mathghost), 0 \Big\} \Big)^{\bar{c}}$$

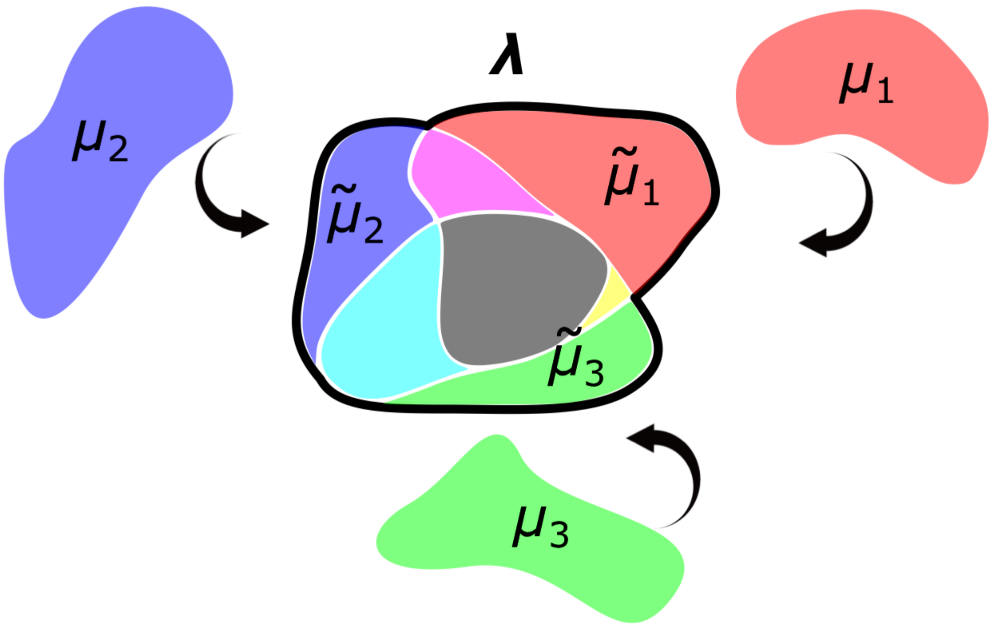
and for any test function h *on* $\mathcal{X}$*,*

$$\int_{\mathcal{X}} h(\widetilde{x}) d\widetilde{\mu}_i^*(\widetilde{x}) := \int_{\mathcal{Z}_*^K} \left\{ \int_{\mathcal{X}} h(\widetilde{x}) d\widetilde{\mu}_{\vec{z},i}^*(\widetilde{x}) \right\} d\pi^*(\vec{z}),$$

where $\widetilde{\mu}_{\vec{z},i}^*$ *is the i-th marginal of* $\widetilde{\mu}_{\vec{z}}^*$*, an optimal adversarial attack which achieves* $\mathbf{c}(z_1, \dots, z_K)$ *given* $\vec{z} = (z_1, \dots, z_K)$*. Then* $(f^*, \widetilde{\mu}^*)$ *is a saddle for problem* (AT)*.*

Generalized barycenter problems

$$\inf_{\lambda, \widetilde{\mu}_1, \ldots, \widetilde{\mu}_K} \lambda(\mathcal{X}) + \sum_{i=1}^{K} C(\mu_i, \widetilde{\mu}_i) \quad \text{s.t. } \lambda \geq \tilde{\mu}_i \; \forall i = 1, \ldots, K.$$



Generalized barycenter problems

$$\inf_{\lambda,\widetilde{\mu}_1,\dots,\widetilde{\mu}_K} \lambda(\mathcal{X}) + \sum_{i=1}^{K} C(\mu_i, \widetilde{\mu}_i) \quad \text{s.t. } \lambda \geq \tilde{\mu}_i \, \forall i = 1, \dots, K.$$

$$\Updownarrow$$

$$(\mathrm{AT})(\mu)$$

Generalized barycenter problems

$$\inf_{\lambda,\widetilde{\mu}_1,\dots,\widetilde{\mu}_K} \lambda(\mathcal{X}) + \sum_{i=1}^{K} C(\mu_i, \widetilde{\mu}_i) \quad \text{s.t. } \lambda \geq \tilde{\mu}_i \; \forall i = 1, \dots, K.$$

$\hookrightarrow$ **(MOT)**

[NGT, Jacobs, Kim 22']

$$\inf_{\rho} \sum_{i=1}^{K} C(\rho_i, \rho)$$

$\hookrightarrow$ **(MOT)**

[Agueh and Carlier 2011].

$$(\mathrm{AT})(\mu) = 1 - \frac{1}{2} \inf_{\pi \in \Pi_k(\mu)} \int_{\mathcal{Z}_*^K} \mathbf{c}(z_1, \dots, z_K) d\pi(z_1, \dots, z_K),$$

- Equivalence between (AT) and computational OT!
- Geometric description of optimal adversarial attacks!
- Specific OT algorithms for this problem?
- Generalizations to other loss functions?
- In binary case (i.e., $k = 2$): [Baghoji, Cullina, Mittal 19'], [Pydi, Jog 20'], [NGT and Murray 20'].

2. A regression problem in a mean field regime

A mean field model of NNs

- $z = (x, y)$, $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$.
- $f(x) = f_\nu(x) := \int_\Theta ah(b \cdot x) d\nu(a, b)$, where: $\theta = (a, b) \in \Theta$, $\nu \in \mathcal{P}(\Theta)$; h is non-linearity.
- $\ell(\tilde{z}, f_\nu) := (f_\nu(\tilde{x}) - \tilde{y})^2$.

- $z = (x, y)$, $x \in \mathbb{R}^d$ and $y \in \mathbb{R}$.
- $f(x) = f_\nu(x) := \int_\Theta a h(b \cdot x) d\nu(a, b)$, where: $\theta = (a, b) \in \Theta$, $\nu \in \mathcal{P}(\Theta)$; h is non-linearity.
- $\ell(\tilde{z}, f_\nu) := (f_\nu(\tilde{x}) - \tilde{y})^2$.
- *(AT) problem*:

$$\inf_{\nu \in \mathcal{P}(\Theta)} \sup_{\tilde{\mu} \in \mathcal{P}(\mathcal{Z})} \left\{ \mathbb{E}_{(\tilde{x}, \tilde{y}) \sim \tilde{\mu}} \left(\ell(\tilde{z}, f_\nu) \right) - c_a W_2^2(\mu, \tilde{\mu}) \right\}.$$

Now, problem:

$$\inf_{\nu \in \mathcal{P}(\Theta)} \sup_{\tilde{\mu} \in \mathcal{P}(\mathcal{Z})} \left\{ \mathbb{E}_{(\tilde{x},\tilde{y})\sim\tilde{\mu}} \left(\ell(\tilde{z}, f_\nu)\right) - c_a W_2^2(\mu, \tilde{\mu}) \right\}$$

is equivalent to

$$\min_{\nu \in \mathcal{P}(\Theta)} \max_{\pi \in \mathcal{P}(\mathcal{Z}\times\mathcal{Z}) \text{ s.t. } \pi_z = \mu} \mathcal{U}(\pi, \nu),$$

where

$$\mathcal{U}(\pi, \nu) := \int_{\mathcal{Z}} \int_{\mathcal{Z}} (f_\nu(\tilde{x}) - \tilde{y})^2 d\pi(z, \tilde{z}) - c_a \int_{\mathcal{Z}\times\mathcal{Z}} |z - \tilde{z}|^2 d\pi(z, \tilde{z}).$$

Target:

$$\min_{\nu\in\mathcal{P}(\Theta)} \max_{\pi\in\mathcal{P}(\mathcal{Z}\times\mathcal{Z}) \text{ s.t. } \pi_z=\mu} \mathcal{U}(\pi,\nu),$$

Ascent-Descent in spaces of measures:

$$\begin{cases} \partial_t \pi_t & = -\eta_t \mathrm{div}_{z,\tilde{z}}(\pi_t(0, \nabla_{\tilde{z}}\mathcal{U}_\pi)) \\ & \quad + \kappa_t \pi_t \left(\mathcal{U}_\pi(z,\tilde{z}) - \int \mathcal{U}_\pi(z,\tilde{z}')d\pi_t(\tilde{z}'|z)\right) \\ \partial_t \nu_t & = \eta_t \mathrm{div}_\theta(\nu_t \nabla_\theta \mathcal{U}_\nu(\theta)) - \kappa_t \nu_t \left(\mathcal{U}_\nu(\theta) - \int \mathcal{U}_\nu(\theta')d\nu_t(\theta')\right), \end{cases}$$

where $\mathcal{U}_\pi, \mathcal{U}_\nu$ first variations of $\mathcal{U}$ w.r.t. π, ν, respectively.

Target:

$$\min_{\nu\in\mathcal{P}(\Theta)} \max_{\pi\in\mathcal{P}(\mathcal{Z}\times\mathcal{Z}) \text{ s.t. } \pi_z=\mu} \mathcal{U}(\pi,\nu),$$

Ascent-Descent in spaces of measures (precisely, projected ascent-descent w.r.t. Wasserstein-Fisher-Rao metric):

$$\begin{cases}\partial_t \pi_t &= -\eta_t \text{div}_{z,\tilde{z}}(\pi_t(0, \nabla_{\tilde{z}}\mathcal{U}_\pi)) \\ & \quad +\kappa_t \pi_t \left(\mathcal{U}_\pi(z,\tilde{z}) - \int \mathcal{U}_\pi(z,\tilde{z}')d\pi_t(\tilde{z}'|z)\right) \\ \partial_t \nu_t &= \eta_t \text{div}_\theta(\nu_t \nabla_\theta \mathcal{U}_\nu(\theta)) - \kappa_t \nu_t \left(\mathcal{U}_\nu(\theta) - \int \mathcal{U}_\nu(\theta')d\nu_t(\theta')\right),\end{cases}$$

where $\mathcal{U}_\pi, \mathcal{U}_\nu$ first variations of $\mathcal{U}$ w.r.t. π, ν, respectively.

Particle system approximation:

$$\pi_t^N = \frac{1}{N}\sum_{i=1}^{N} \omega_t^i \delta_{(Z_t^i, \tilde{Z}_t^i)}, \quad \nu_t^N = \frac{1}{N}\sum_{i=1}^{N} \alpha_t^i \delta_{\theta_t^i},$$

where:

$$d_t(Z_t^i, \tilde{Z}_t^i) = (0, \eta_t \nabla_{\tilde{z}} \mathcal{U}_\pi(\pi_t^N, \nu_t^N; Z_t^i, \tilde{Z}_t^i))$$

$$d_t\omega_t^i = \kappa_t \omega_t^i \left(\mathcal{U}_\pi(\pi_t^N, \nu_t^N; Z_t^i, \tilde{Z}_t^i) - \int \mathcal{U}_\pi(\pi_t^N, \nu_t^N; Z_t^i, \tilde{z}')d\pi_t^N(\tilde{z}'|Z_t^i) \right)$$

$$d_t\theta_t^i = -\eta_t \nabla_\theta \mathcal{U}_\nu(\pi_t^N, \nu_t^N; \theta_t^i)$$

$$d_t\alpha_t^i = -\kappa_t \alpha_t^i \left(\mathcal{U}_\nu(\pi_N^N, \nu_t^N; \vartheta_t^i) - \int \mathcal{U}_\nu(\pi_t^N, \nu_t^N; \theta')d\nu_t^N(\theta') \right);$$

and given initial condition $(Z_0^i, \tilde{Z}_0^i, \omega_0^i, \vartheta_0^i, \alpha_0^i)$ (possibly random).

Part 1: Mean field limit of particle system

Theorem (C.A. García Trillos, NGT 23')

Suppose that:

- $\Theta, \mathcal{Z}$ *are bounded subsets of Euclidean space.*
- $\nabla U_\pi, \nabla \mathcal{U}_\nu$ *are Lipschitz.*
- *Initial conditions* $(Z_0^i, \tilde{Z}_0^i, \omega_0^i, \theta_0^i, \alpha_0^i)$ *are well prepared.*

Then, for every fixed $T > 0$, *we have:*

$$\sup_{t\in[0,T]} W_1(\pi_t^N, \pi_t) \to 0; \qquad \sup_{t\in[0,T]} W_1(\nu_t^N, \nu_t) \to 0,$$

as $N \to \infty$, *where* (π_t, ν_t) *solve Ascent-Descent dynamics PDE.*

Both (π_t^N, ν_t^N) and (π_t, ν_t) solve the same equation:

$$\begin{cases} \partial_t \pi_t & = -\eta_t \mathrm{div}_{z,\tilde{z}}(\pi_t(0, \nabla_{\tilde{z}} \mathcal{U}_\pi)) \\ & \quad + \kappa_t \pi_t \left(\mathcal{U}_\pi(z, \tilde{z}) - \int \mathcal{U}_\pi(z, \tilde{z}') d\pi_t(\tilde{z}' | z) \right) \\ \partial_t \nu_t & = \eta_t \mathrm{div}_\theta(\nu_t \nabla_\theta \mathcal{U}_\nu(\theta)) - \kappa_t \nu_t \left(\mathcal{U}_\nu(\theta) - \int \mathcal{U}_\nu(\theta') d\nu_t(\theta') \right), \end{cases}$$

but they differ in their initial conditions (π_0^N, ν_0^N) and (π_0, ν_0).

Mean field limit of particle system

Theorem (C.A. García Trillos, NGT 23')

Suppose that:

- $\Theta, \mathcal{Z}$ *are bounded subsets of Euclidean space.*
- $\nabla U_\pi, \nabla \mathcal{U}_\nu$ *are Lipschitz.*
- *Initial conditions* $(Z_0^i, \tilde{Z}_0^i, \omega_0^i, \theta_0^i, \alpha_0^i)$ *are well prepared.*

Then, for every fixed $T > 0$*, we have:*

$$\sup_{t\in[0,T]} W_1(\pi_t^N, \pi_t) \to 0; \qquad \sup_{t\in[0,T]} W_1(\nu_t^N, \nu_t) \to 0,$$

as $N \to \infty$*, where* (π_t, ν_t) *solve Ascent-Descent dynamics PDE.*

An example of well prepared initial conditions

Set $\omega_0^i = \alpha_0^i = 1$ and suppose that, as $N \to \infty$, we have:

$$W_1(\nu_0^N, \nu_0) \to 0,$$

as well as

$$\inf_{\upsilon \in \Gamma_{\text{Opt}}(\pi_{0,z}^N, \pi_{0,z})} \int W_1(\pi_0^N(\cdot|z_0'), \pi_0(\cdot|z_0)) d\upsilon(z_0', z_0) \to 0$$

(Knothe transport and reminescent to TLp metric).

An example of well prepared initial conditions

To satisfy:

$$\inf_{\upsilon \in \Gamma_{\mathrm{Opt}}(\pi^N_{0,z}, \pi_{0,z})} \int W_1(\pi^N_0(\cdot|z'_0), \pi_0(\cdot|z_0)) d\upsilon(z'_0, z_0) \to 0,$$

set, for example,

$$\pi^N_0 = \frac{1}{nm} \sum_{ij} \delta_{(Z^i_0, \tilde{Z}^{ij}_0)},$$

where

- $Z^i_0 \sim \pi_{0,z} = \mu$, $i = 1, \dots, n$,
- $\tilde{Z}^{ij}_0 \sim \pi_0(\cdot|Z^i_0)$, $j = 1, \dots, m$, $i = 1, \dots, n$.

An example of well prepared initial conditions

Lemma (C.A. García Trillos, NGT 23')

Let A, B be two bounded Borel subsets of $\mathbb{R}^d$ and $\mathbb{R}^{d'}$, respectively. Let $\mu \in \mathcal{P}(A)$, and let $u \in A \mapsto \mu_u(\cdot) \in \mathcal{P}(B)$ be a measurable map.
Then, for every sequence $\{\Upsilon_n\}_{n\in\mathbb{N}} \subseteq \Gamma(\mu, \mu)$ satisfying

$$\lim_{n\to\infty} \int_{A\times A} |u - u'| d\Upsilon_n(u, u') = 0,$$

we have

$$\lim_{n\to\infty} \int_{A\times A} W_1(\mu_u, \mu_{u'}) d\Upsilon_n(u, u') = 0.$$

Part 2: Long time behavior mean field system

Theorem (C.A. García Trillos, NGT 23')

Fix $\delta > 0$. Let π, ν the solution to descent-ascent dynamics for η_t, κ_t appropriately tuned. Define:

$$\overline{\nu}_t = \frac{1}{t}\int_0^t \nu_s ds, \quad \overline{\pi}_t = \frac{1}{t}\int_0^t \pi_s ds.$$

Then, for all large enough t,

$$\sup_{\tilde{\pi} \text{ s.t. } \tilde{\pi}_z = \mu} \mathcal{U}(\tilde{\pi}, \bar{\nu}(t)) - \inf_{\tilde{\nu}} \mathcal{U}(\bar{\pi}(t), \tilde{\nu}) \leq \delta.$$

Long time behavior mean field system

Theorem (C.A. García Trillos, NGT 23')

Fix $\delta > 0$. Let π, ν the solution to descent-ascent dynamics for η_t, κ_t appropriately tuned. Define:

$$\overline{\nu}_t = \frac{1}{t}\int_0^t \nu_s ds, \quad \overline{\pi}_t = \frac{1}{t}\int_0^t \pi_s ds.$$

Then, for all large enough t,

$$\sup_{\tilde{\pi} \text{ s.t. } \tilde{\pi}_z = \mu} \mathcal{U}(\tilde{\pi}, \bar{\nu}(t)) - \inf_{\tilde{\nu}} \mathcal{U}(\bar{\pi}(t), \tilde{\nu}) \leq \delta.$$

However, this is under very stringent conditions on initializations (both π_0, ν_0). On ν_0, these conditions are not so different to those in **Chizat and Bach 17'**, for example.

The “strongly concave" case

However, roles of π and ν are quite different. In the setting:

$$\mathcal{U}(\pi,\nu) = \int_{\mathcal{Z}}\int_{\mathcal{Z}} (f_\nu(\tilde{x}) - \tilde{y})^2 d\pi(z,\tilde{z}) - c_a \int_{\mathcal{Z}\times\mathcal{Z}} |z-\tilde{z}|^2 d\pi(z,\tilde{z}),$$

The “strongly concave" case

However, roles of π and ν are quite different. In the setting:

$$\mathcal{U}(\pi, \nu) = \int_{\mathcal{Z}} \int_{\mathcal{Z}} (f_\nu(\tilde{x}) - \tilde{y})^2 d\pi(z, \tilde{z}) - c_a \int_{\mathcal{Z} \times \mathcal{Z}} |z - \tilde{z}|^2 d\pi(z, \tilde{z}),$$

if c_a sufficiently large, then there exists $\lambda > 0$ such that
$\forall \nu \in \mathcal{P}(\Theta)$, $\forall \pi \in \mathcal{P}(\mathcal{Z}^2)$ with $\pi_z = \mu$:

$$\int |\nabla_{\tilde{z}} \mathcal{U}_\pi(\pi, \nu; z, \tilde{z})|^2 d\pi(z, \tilde{z}) \geq \lambda(m_\nu^* - \mathcal{U}(\pi, \nu)), \qquad \text{(PL)}$$

where $m_\nu^* := \sup_{\tilde{\pi} \text{ s.t. } \tilde{\pi}_z = \mu} \mathcal{U}(\tilde{\pi}, \nu)$.

Theorem (C.A. García Trillos, NGT 23')

Fix $\delta > 0$. Suppose PL assumption holds Let π, ν the solution to (slightly modified) descent-ascent dynamics for η_t, κ_t appropriately tuned, and with ν_0 appropriately initialized and π_0 arbitrary .
Define:

$$\overline{\nu}_t = \frac{1}{t}\int_0^t \nu_s ds, \quad \overline{\pi}_t = \frac{1}{t}\int_0^t \pi_s ds.$$

Then, for all large enough t,

$$\sup_{\tilde{\pi} \text{ s.t. } \tilde{\pi}_z = \mu} \mathcal{U}(\tilde{\pi}, \bar{\nu}(t)) - \inf_{\tilde{\nu}} \mathcal{U}(\bar{\pi}(t), \tilde{\nu}) \leq \delta.$$

Related work: "Certifying Some Distributional Robustness with Principled Adversarial Training" **Sinha, Namkoong, and Duchi 18'.**

$$\min_{\nu \in \mathcal{P}(\Theta)} \max_{\pi \in \mathcal{P}(\mathcal{Z} \times \mathcal{Z}) \text{ s.t. } \pi_z = \mu} \mathcal{U}(\pi, \nu),$$

- Ascent-descent algorithms.
- Some convergence results.
- Less stringent assumptions.
- Other geometries modeling adversarial costs?
- Other related work:
  1. “A mean-field analysis of two-player zero-sum games" **Domingo-Enrich et al 20'**.
  2. "An Exponentially Converging Particle Method for the Mixed Nash Equilibrium of Continuous Games" **Chizat and Wang 22'**.

An analyst's perspective on adversarial training:

- NGT and R. Murray "Adversarial classification: necessary conditions and geometric flows" *Journal of Machine Learning research (JMLR)* 22'.
- C. García Trillos, NGT "On the regularized risk of distributionally robust learning over deep neural networks" *Research in the Mathematical Sciences (RMS)* 22'.
- L. Bungert, NGT, R. Murray "The Geometry of Adversarial Training in Binary Classification" *To appear in Information and Inference: A Journal of the IMA*.
- NGT, M. Jacobs, J. Kim "The multimarginal optimal transport formulation of adversarial multiclass classification" *To appear in JMLR*.
- C. García Trillos, NGT "On adversarial robustness and the use of Wasserstein ascent-descent dynamics to enforce it" https://arxiv.org/abs/2301.03662 23'.

Thank you for your attention!

Special thanks to:

-NSF Grants: DMS-2005797 and DMS-2236447
-All my collaborators.