An analytical and geometric perspective on adversarial learning.

Nicolás García Trillos

UNIVERSITY OF WISCONSIN - MADISON

• Adversarial training / learning.

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

- μ distributions on $z = (x, y)$ $x \in \mathbb{R}^d$, $y \in \{0,1\}$
- θ parameter of statistical model
- $\ell(\cdot, \cdot)$
- $D(\cdot, \cdot)$

- Regularized Risk minimization.

$$(R) \quad \inf_{\theta \in \Theta} \mathbb{E}_{z \sim \mu}[\ell(z, \theta)] + \lambda R(\theta)$$

- Classical statistics.
- Inverse Problems
- Graph-Based Learning.

Q: What is the relationship between adversarial training and regularization?

Sometimes a very direct one.
Consider the following setup:

- $\Theta = \mathbb{R}^d$

- $\ell(z, \theta) = (\langle \theta, x \rangle - y)^2$

- $D(\mu, \tilde{\mu}) = W_{c_p}(\mu, \tilde{\mu})$

$$:= \inf_{\pi \in \Gamma(\mu, \tilde{\mu})} \iint c_p(z, \tilde{z}) \, d\pi(z, \tilde{z})$$

where $c_p(z, \tilde{z}) = \begin{cases} \|x - \tilde{x}\|_p & \text{if } y = \tilde{y} \\ +\infty & \text{if } y \neq \tilde{y}. \end{cases}$

Then:

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

$$=$$

$$\inf_{\theta \in \Theta} \left(\sqrt{\mathbb{E}_{(x,y) \sim \mu} [\ell(z, \theta)]} + \sqrt{\varepsilon} \, \|\theta\|_q \right)^2$$

$$\frac{1}{p} + \frac{1}{q} = 1$$

[Chen et al 20'], [Blanchet et al 19'], [Kuhn et al 19'], ...

Today's talk Based on:

- [NGT, MURRAY 20']:
  "Adversarial classification: necessary conditions and geometric flows".
- [C. GARCIA TRILLOS, NGT 21']
  "On the regularized risk of distributionally robust learning over deep neural networks".
- [BUNGERT, NGT, MURRAY, 21']
  "The geometry of adversarial learning in binary classification"

Setup:

- Θ : Borel Subsets of $\mathbb{R}^d$

From now on use $A \subseteq \mathbb{R}^d$ instead of Θ

- $(x,y) \in \mathbb{R}^d \times \{0,1\}$.

- $D(\mu,\tilde{\mu}) = W_\infty(\mu,\tilde{\mu})$

$$W_\infty(\mu,\tilde{\mu}) = \inf_{\pi \in \Gamma(\mu,\tilde{\mu})} \sup_{(z,\tilde{z}) \in \mathrm{Supp}(\pi)} \hat{d}(z,\tilde{z})$$

$$\hat{d}(z,\tilde{z}) := \begin{cases} d(x,\tilde{x}) & \text{if } y=\tilde{y} \\ +\infty & \text{if } y \neq \tilde{y} \end{cases}$$

- $$\ell(z,A) = \begin{cases} 1 & \text{if } \mathbb{1}_A(x) \neq y \\ 0 & \text{otherwise.} \end{cases}$$

Here:

$$\inf_{A} \underbrace{\sup_{\tilde{\mu}: W_\infty(\mu,\tilde{\mu}) \le \varepsilon} \mathbb{E}_{\tilde{z}\sim\tilde{\mu}} [\ell(\tilde{z},\theta)]}_{L_\varepsilon(A)}$$

$\varepsilon = 0$: $\quad A_0^* := \underset{A}{\operatorname{argmin}}\ L_0(A)$

$$A_0^* := \{ x \in \mathbb{R}^d : \omega_1 \rho_1(x) \ge \omega_0 \rho_0(x) \}.$$

Here:

$$\omega_0 \Rightarrow \mu(\mathbb{R}^d \times \{0\})$$
$$\omega_1 \Rightarrow \mu(\mathbb{R}^d \times \{1\})$$
$$\rho_0 \Rightarrow \mu(dx \mid y=0)$$
$$\rho_1 \Rightarrow \mu(dx \mid y=1)$$
$$\rho \longrightarrow \mu(dx)$$

Q: How should the boundary of A_0^* change to track solutions to $\inf_{A} L_\varepsilon(A)$ as ε grows?

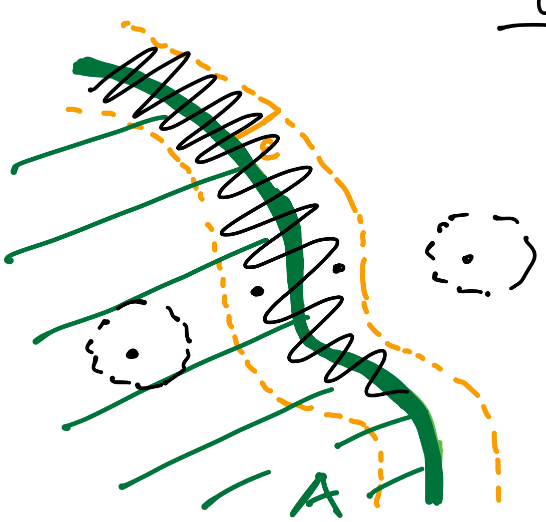
$$\begin{cases} \varepsilon \in (0,\varepsilon_0) \longmapsto A_\varepsilon \\ A_0 = A_0^* \end{cases}$$

s.t. $A_\varepsilon \in \underset{A}{\arg\min}\, L_\varepsilon(A)$

Here it is convenient to notice that:

$$L_\varepsilon(A) = \underline{\sup_{\tilde\mu : W_\infty(\mu,\tilde\mu)\le\varepsilon} \mathbb{E}_{\tilde z\sim\tilde\mu}\left[\ell(\tilde z, A)\right]}$$

$$= \underline{\mathbb{E}_{(x,y)\sim\mu}}\Big[\underline{\sup_{\tilde x\in B_\varepsilon(x)} \ell((\tilde x, y), A)}\Big]$$



$$= \int_{\partial_\varepsilon A} \rho(x)\,dx + \int_{A^{-\varepsilon}} \omega_0 \rho_0\,dx + \int_{(A^\varepsilon)^c} \omega_1 \rho_1\,dx$$

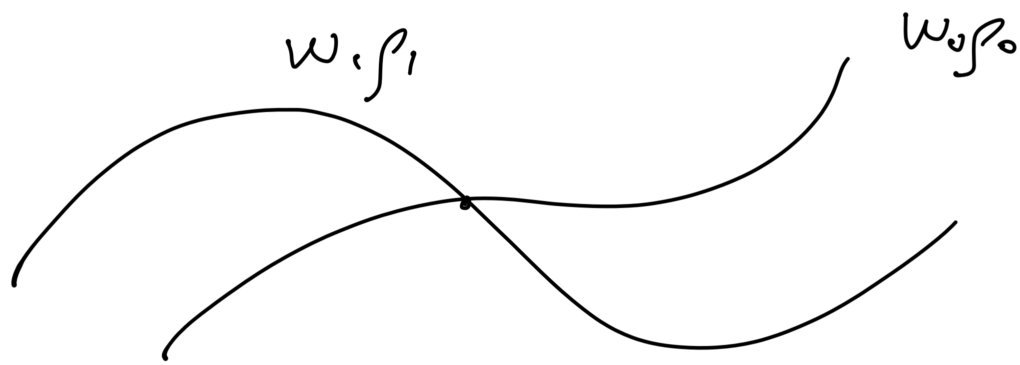
[NGT, MURRAY, 20']

In 1D first: Suppose Bayes classifier has the form $\boxed{A_0 = \bigcup_{k=1}^{K} [a_k(0), b_k(0)]}$

Under a "strict crossing" condition for $w_0 \rho_0$ and $w_1 \rho_1$, the following system of ODEs tracks solutions for all small enough ε:

$$\begin{cases} \dfrac{d b_k(\varepsilon)}{d\varepsilon} = -\left(\dfrac{w_0 \rho_0'(b_k+\varepsilon) + w_1 \rho_1'(b_k-\varepsilon)}{w_0 \rho_0'(b_k+\varepsilon) - w_1 \rho_1'(b_k-\varepsilon)} \right) \\ \\ b_k(0) \end{cases}$$

and similar eqns for a_k.



W_1, ρ_1
W_0, ρ_0

Comments:

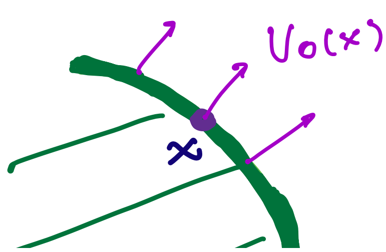
① Connection to Optimal Transport problem:

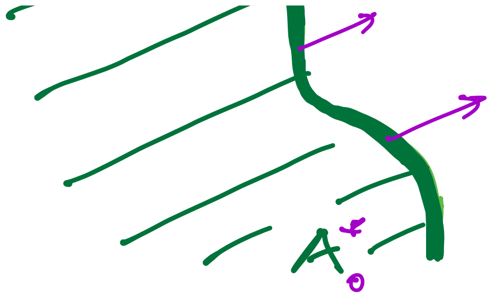
- [Bhagoji et al 19'] $\Big\}$ $w_0 = w_1$
- [Pydi + Jog 19']

② 1D setting does not reveal the geometric structure of the general problem...

• In $d > 1$:

Messy equations (existence?) but at $\varepsilon = 0$ we can try to answer how the boundary of A_0 changes infinitesimally:





$$V_0(x) = \left(- \frac{\nabla\rho \cdot \vec{n} + \rho \sum_i \varkappa_i}{(\omega_1 \nabla\rho_1 - \omega_0 \nabla\rho_0)\cdot\vec{n}} \right) \vec{n}$$

SAME infinitesimal change
as if we were tracking

solutions of: (R')

$$\boxed{\inf_A \left\{ \mathbb{E}_{z\sim\mu}\left[\ell(z,A)\right] + \varepsilon \, \mathrm{Per}(A) \right\}}$$

where:

$$\mathrm{Per}(A) = \int_{\partial A} \rho(x)\, d\mathcal{H}^{d-1}(x)$$

Comments:

1. So Perimeter is connected to the regularization induced by (AT).

2. Now, is (R') equivalent to (AT)?

NO, BUT:

Take :

$$\mathbb{E}_{z\sim\mu}\Big[\sup_{\tilde{x}\in B_\varepsilon(x)} \ell((\tilde{x},y),A)\Big]$$

$$= \mathbb{E}_{z\sim\mu}\big[\ell((x,y),A)\big] + \varepsilon\underbrace{\left(\frac{\mathbb{E}_{z\sim\mu}\Big[\sup_{\tilde{x}\in B_\varepsilon(x)} \ell((\tilde{x},y),A)\Big] - \mathbb{E}_{z\sim\mu}\big[\ell((x,y),A)\big]}{\varepsilon}\right)}_{Per_\varepsilon(A).}$$

Theorem : [Bungert, NGT, MURRAY, 21]

$$(AT) = \inf_{A}\left\{\mathbb{E}_{z\sim\mu}\big[\ell((x,y),A)\big] + \varepsilon\, Per_\varepsilon(A)\right\}$$

where :

(1) $Per_\varepsilon(A) \geq 0 \quad \forall A.$

(2) $Per_\varepsilon(\cdot)$ is submodular:

$$A, B \subseteq \mathbb{R}^d$$

$$Per_\varepsilon(A \cup B) + Per_\varepsilon(A \cap B) \leq Per_\varepsilon(A) + Per_\varepsilon(B)$$

(3) let

$$TV_\varepsilon(u) := \int_{-\infty}^{\infty} Per_\varepsilon(\{u \geq t\})\, dt$$

$u : \mathbb{R}^d \to \mathbb{R}$.

Then: TV_ε is convex, 1-homogeneous, and l.s.c w.r.t appropriate topology.

(4) The problem:

$$\min_{f : \mathbb{R}^d \to [0,1]} \left\{ \underline{\mathbb{E}_{(x,y) \sim \mu}\left[|f(x) - y| \right] + \varepsilon\, TV_\varepsilon(f)} \right\}$$

is an exact convex relaxation of (AL).

Remark:

- $TV_\varepsilon(u) =$

$$\frac{\omega_1}{\varepsilon}\int_{\mathbb{R}^d}\Big(u(x) - \inf_{\tilde{x}\in B_\varepsilon(x)} u(\tilde{x})\Big)\rho_1(dx)$$

$$+ \frac{\omega_0}{\varepsilon}\int_{\mathbb{R}^d}\Big(\sup_{\tilde{x}\in B_\varepsilon(x)} u(x) - u(x)\Big)\rho_0(dx)$$

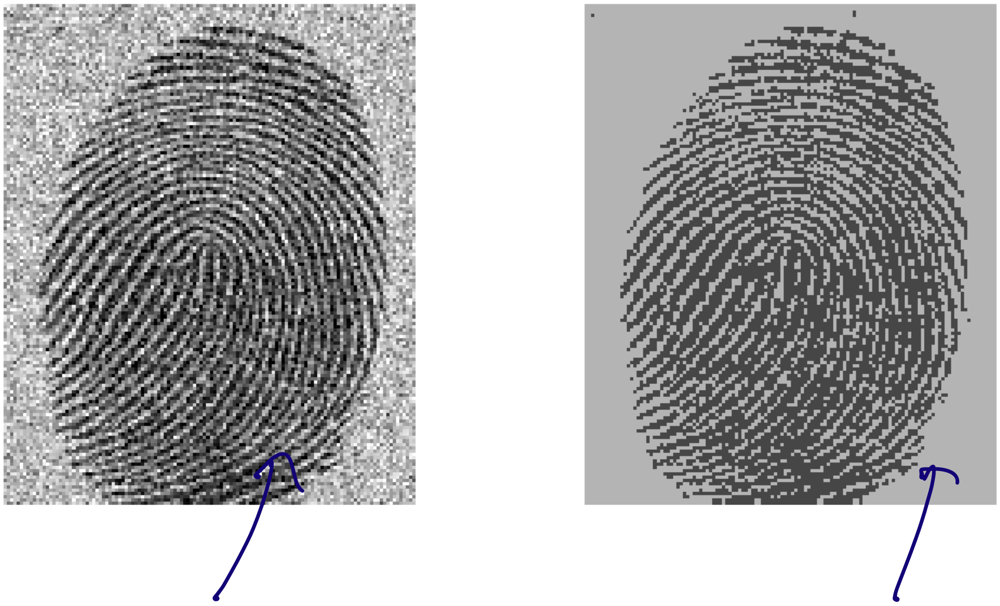
- Notice that TV_ε depends on the full μ!

($Per_\varepsilon(A) = TV_\varepsilon(\mathbb{1}_A)$ too)

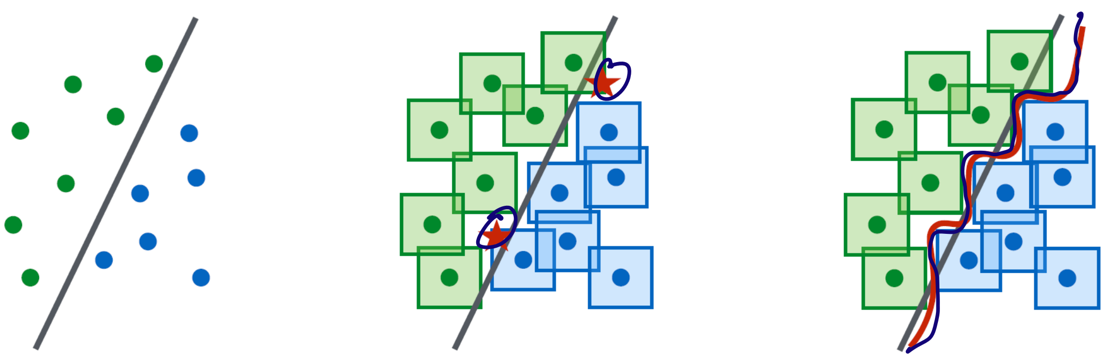
- take $w_0 = w_1 = \frac{1}{2}$ and
  $\rho_0 = \rho_1 = \rho$ in the formula
  for TV_2. We recover:

Related to:

- [Barchiesi et al 10']
- [Chambolle et al 12']



Per vs Per_{ε}



From MADRY et al 19'.

Q: What is the connection with

regularizers used in graph Based learning?

Different adversarial model:

- Nature chooses $\tilde{x} \sim \mathcal{U}_{B_\varepsilon(x)}$
  Adversary decides to accept / reject $\tilde{x}$ (with the goal of maximizing their payoff).

(AL')

$$= \inf_{u:\mathbb{R}^d \to [0,1]} \left\{ \mathbb{E}_{z\sim\mu}[|u(x)-y|] + \varepsilon \widetilde{TV}_\varepsilon(u) \right\}$$

where

$\widetilde{TV}_\varepsilon(u) =$

$$\frac{\omega_1}{\varepsilon} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \frac{\eta_\varepsilon(|x-\tilde{x}|)}{\rho(B_\varepsilon(x))} \left(u(x) - u(\tilde{x}) \right)_+ \rho(d\tilde{x}) \rho_1(dx)$$

$$+ \frac{\omega_0}{\varepsilon} \int_{\mathbb{R}^d} \int_{\mathbb{R}^d} \frac{\eta_\varepsilon(|x-\tilde{x}|)}{\rho(B_\varepsilon(x))} \left(u(x) - u(\tilde{x}) \right)_- \rho(d\tilde{x}) \rho_0(dx)$$

Remark: $\omega_0 = \omega_1 = \frac{1}{2}$

$\rho_0 = \rho_1 = \rho$:

$$\boxed{\frac{1}{2\varepsilon} \iint \frac{\eta_\varepsilon(|x-\tilde{x}|)}{\rho(B_\varepsilon(x))} |u(x) - u(\tilde{x})| \, \rho(d\tilde{x}) \rho(dx)}$$

When $\rho(dx) = \frac{1}{n} \sum_{i=1}^{n} \delta_{x_i}$

$-\rho(B_\varepsilon(x))$ = degree of geometric graph

$$\boxed{\frac{1}{2n^2\varepsilon}\sum_i\sum_j \frac{\eta_\varepsilon(|x_i-x_j|)}{d_\varepsilon(x_i)}\,|u(x_i)-u(x_j)|}$$

Thank you for your attention!

Special thanks to:

-NSF Grant: DMS-2005797
-All my collaborators.