Mathematics of Adversarial Machine Learning
UT Winter School 2024

Nicolás García Trillos

January 2024

Introduction

Adversarial machine learning is an area of modern machine learning whose main goal is to study and develop methods for the design of learning models that are robust to adversarial perturbations of data. It became a prominent research field in machine learning less than a decade ago, not long after neural networks became the state of the art technology for tackling image processing and natural language processing tasks, when it was noticed that neural network models, as well as other learning models, although highly effective at making accurate predictions on clean data, were quite sensitive to adversarial attacks. This mini-course seeks an exploration of the mathematical underpinnings of this active and vibrant field. We will be particularly interested in exploring it from analytic and geometric perspectives and discussing connections with topics such as regularization theory, game theory, optimal transport, geometry, and distributionally robust optimization. The mini-course aims to present the topic of adversarial machine learning within the bigger objective of designing safe, secure, and trustworthy AI models.

1 Lecture 1

Scribes: Rachel Morris and Kevin Ren

Adversarial Learning/Training Problem:

$$\inf_{f\in\mathcal{F}} \mathbb{E}_{(x,y)\sim\mu}\left[\sup_{\tilde{x}\in B_\varepsilon(x)} \ell(f(\tilde{x}),y)\right] = \inf_{f\in\mathcal{F}} \int \sup_{\tilde{x}\in B_\varepsilon(x)} \ell(f(\tilde{x}),y)\ d\mu(x,y) \qquad \text{(ATP)}$$

- $(x,y)\in\mathcal{X}\times\mathcal{Y}$; $\mathcal{X}=\mathbb{R}^d$, $\mathcal{Y}=\begin{cases}\mathbb{R} & \text{regression}\\ \{1,\dots,k\} & \text{classification}\end{cases}$
- $\mu\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})$ is Borel probability measure, μ not necessarily empirical
- $f:\mathcal{X}\to\mathcal{Y}$ or $f:\mathcal{X}\to\mathcal{P}(\mathcal{Y})$; in the case that $\mathcal{Y}=\{1,\dots,k\}$ then $f(x)=(f_1(x),\dots,f_k(x))$ (f is a probability vector)
- $\mathcal{F}$ is a family of such functions f; e.g. linear functions, neural nets, all Borel functions
- Note that enlarging $\mathcal{F}$ makes the inf in (ATP) smaller
- $\ell(\cdot,\cdot)$ is a loss function; e.g. 0-1 loss $\ell(f(x),y)=1-f_y(x)$, cross entropy loss $\ell(f(x),y)=-\log(f_y(x))$
- We need to check integrability in (ATP)...
- $B_\varepsilon(x)$: should it be open or closed? If ℓ and f are continuous, then it doesn't make a difference. If f is Borel and ℓ is continuous, then $(x,y)\mapsto\sup_{\tilde{x}\in B_\varepsilon(x)}\ell(f(\tilde{x}),y)$ is Borel when $B_\varepsilon(x)$ is open (see exercises; not necessarily true if $B_\varepsilon(x)$ is closed).
- Universal σ-algebra of $\mathcal{X}\times\mathcal{Y}$:

  $$\mathcal{U}=\bigcap_{\gamma\in\mathcal{P}(\mathcal{X}\times\mathcal{Y})}\Sigma_\gamma$$

  where Σ_γ = completion of Borel σ-algebra with γ-null sets (sets with γ-measure zero). Then we can write (ATP) as

  $$\inf_{f\in\mathcal{F}}\int \sup_{\tilde{x}\in B_\varepsilon(x)} \ell(f(\tilde{x}),y)\ d\overline{\mu}(x,y) \qquad \text{(ATP*)}$$

  where $\overline{\mu}$ is the extension of μ to $\mathcal{U}$.
- **Remark.** For all but at most countably many $\varepsilon>0$, the two optimzation problems (ATP), (ATP*) are equivalent. (See exercises)

1. Distributionally Robust Optimzation version of AT (DRO)

$$\inf_{f\in\mathcal{F}} \sup_{\tilde{\mu}\in\rho(\mathcal{X}\times\mathcal{Y})} \mathbb{E}_{(\tilde{x},\tilde{y})\sim\tilde{\mu}}[\ell(f(\tilde{x}),\tilde{y})] - C(\mu,\tilde{\mu}) \tag{DRO}$$

   where $C(\mu,\tilde{\mu})$ is a positive function to penalize the attacker for changing the underlying distribution. A particular but interesting subcase:

$$\inf_{f\in\mathcal{F}} \sup_{\tilde{\mu} \text{ s.t.} D(\tilde{\mu},\mu)\leq\varepsilon} \mathbb{E}_{(\tilde{x},\tilde{y})\sim\tilde{\mu}}[\ell(f(\tilde{x}),\tilde{y})] \tag{DRO2}$$

   Here, D is a "distance function" between distributions. DRO generalizes (ATP). It can be used to find lower bounds for (ATP).

   Structure of $C(\mu,\tilde{\mu})$: an "optimal transport metric"

$$C(\mu,\tilde{\mu}) = \inf_{\pi\in\Gamma(\mu,\tilde{\mu})} \int C_z(z,\tilde{z})\ d\pi(z,\tilde{z}) \tag{OTM}$$

   where $\Gamma(\mu,\tilde{\mu}) = \{\pi\in\rho((\mathcal{X}\times\mathcal{Y})\times(\mathcal{X},\mathcal{Y})) : \pi_1=\mu, \pi_2=\tilde{\mu}\}$, i.e. Γ is the set of "couplings" or "transport plans" between μ and $\tilde{\mu}$. Notice that the constraints and objective function are linear. This problem is explicitly a min-max game. Lots of numerical algorithms for computing optimal transport between two distributions in last decade!

   To see the connection with ATP (i.e. how DRO is a larger class of problems that contains ATP):

$$C_z((x,y),(\tilde{x},\tilde{y})) = \begin{cases} 0 & \text{if } d(x,\tilde{x})\leq\varepsilon \text{ and if } y=\tilde{y} \\ \infty & \text{otherwise} \end{cases}$$

   In other words: the adversary is free to change data points in X by distance $\leq\varepsilon$, but infinite cost to change values in Y

2. Probabilistically Robust Learning - interpolate between best classification on clean data and robustness of the ATP

$$\inf_{f\in\mathcal{F}} \mathbb{E}_{(x,y)\sim\mu}\left[\max\left\{\ell(f(x),y), g_{p,\varepsilon}(\ell(\cdot,y))\right\}\right] \tag{PRL}$$

   where $g_{p,\varepsilon}$ is some softer sup operation. You can tune the parameter p to be closer to ATP or closer to best accuracy on clean data. This reveals a further connection between AT and regularization

3. Regularization:

$$\inf_{f\in\mathcal{F}} \mathbb{E}_{(x,y)\sim\mu}[\ell(f(x),y)] + R_{\varepsilon,p}(f)$$

   $R_{\varepsilon,p}(f)$ is the regularization term.

   Smoother decision boundaries are more robust to small perturbations (compare a smooth circular boundary vs. a rough squiggly boundary)

Outline for remaining days:

- Tuesday, Wednesday:
  - (DRO) Family
  - How to find universal lower bounds
  - Highlights connections with Optimal Transport
- Thursday, Friday:
  - (PRL) model
  - (ATP) as a perimeter minimization problem
  - Highlights some tools from calculus of variation

1.1 Further Remarks and References

1. The fact that $x \mapsto \sup_{\tilde{x} \in \overline{B}_\varepsilon(x)} g(\tilde{x})$ may not be Borel measurable if we only assume Borel measurability of g can be found in Lemma 4.1. in the paper *The Many Faces of Adversarial Risk,* by Muni Sreenivas Pydi and Varun Jog (whose ArXiv version you can find here `https://arxiv.org/pdf/2201.08956.pdf`). In that same paper, in Lemma 4.2 and Lemma 4.4 the authors discuss the universal measurability of this function.

2. Related to the second question in the first problem set, you can take a look at Theorem 8 in `https://arxiv.org/pdf/2006.09568.pdf`, where some results on estimating the (true) Bayes adversarial risk using finite data are discussed.

1.2 Exercises

1. (The Bayes classifier). In the classification setting $\mathcal{Y} = \{1, \dots, k\}$, let $\mathcal{F} = \mathcal{F}_{all}$ be the set of *all* Borel (weak or probabilistic) classifiers, i.e., all Borel functions $f : \mathcal{X} \to \mathcal{P}(Y)$. In other words, for every $x \in \mathcal{X}$ $f(x)$ is a probability vector in $\mathbb{R}^k$, each of whose coordinates represents the level of confidence in classifying x as that coordinate. Consider the (extended) $0-1$ loss function:
$$\ell(f(x), y) := 1 - f_y(x).$$
Prove that
$$\inf_{f \in \mathcal{F}} \mathbb{E}_{(x,y)\sim\mu}[\ell(f(x), y)]$$
admits solutions f^* such that for every x the vector $f^*(x)$ has a 1 in one coordinate and zeros in all other coordinates. Moreover, provide an expression for one such minimizer in terms of the distribution μ.

2. (On overfitting). Suppose that $(x_1, y_1), \dots, (x_n, y_n)$ are i.i.d. samples from a distribution μ. Let μ_n be the empirical measure associated to these samples. Let $\mathcal{F}_{all}$ be the set of all Borel measurable functions from $\mathcal{X}$ into $\mathcal{Y}$.

   Explain why solving the problem

$$\inf_{f \in \mathcal{F}_{all}} \mathbb{E}_{(x,y)\sim\mu_n}[\ell(f(x), y)]$$

   is in general useless for estimating solutions to the problem

$$\inf_{f \in \mathcal{F}_{all}} \mathbb{E}_{(x,y)\sim\mu}[\ell(f(x), y)].$$

   Does your answer change if we replace RM with AT for some value of adversarial budget ε? Discuss.

3. Let g be a Borel measurable function. Show that the function

$$x \in \mathbb{R}^d \mapsto \sup_{\tilde{x} \in B_\varepsilon(x)} g(\tilde{x})$$

   is Borel measurable. In the above, $B_\varepsilon(x)$ is an open ball.

4. (Equivalence between open and closed ball models) Let $\mu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ and let $\mathcal{F}$ be a family of regression functions/classifiers. Consider the following two problems:

$$\inf_{f \in \mathcal{F}} \int \sup_{\tilde{x} \in B_\varepsilon(x)} \{\ell(f(\tilde{x}), y)\} d\mu(x, y)$$

   and

$$\inf_{f \in \mathcal{F}} \int \sup_{\tilde{x} \in \overline{B}_\varepsilon(x)} \{\ell(f(\tilde{x}), y)\} d\overline{\mu}(x, y),$$

   where $B_\varepsilon(x)$ and $\overline{B}_\varepsilon(x)$ are the open and closed balls of radius ε centered around x, and $\overline{\mu}$ is the extension of μ to the universal σ-algebra (see definition below).

   Prove that for all but at most countably many values of ε the above problems satisfy:

   (a) The two infima are equal.

   (b) If $f^* \in \mathcal{F}$ is a solution to the closed ball problem, then f^* is also a solution to the open ball problem.

   Definition: For every $\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ let Σ_γ be the γ-completion of the Borel σ-algebra over $\mathcal{X} \times \mathcal{Y}$. That is, Σ_γ is the σ-algebra generated by all Borel sets and all γ-null sets. The universal σ-algebra $\mathcal{U}$ over $\mathcal{X} \times \mathcal{Y}$ is defined as

$$\mathcal{U} := \bigcap_{\gamma \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})} \Sigma_\gamma.$$

Notice that the Borel σ-algebra is contained in $\mathcal{U}$. Thus, any $\mu \in \mathcal{P}(\mathcal{X} \times \mathcal{Y})$ can be extended to a measure $\overline{\mu}$ over $\mathcal{U}$. As an extension, we should have

$$\int h(x) d\mu(x) = \int h(x) d\overline{\mu}(x)$$

for every Borel h.

5. (Regularization and AT Part 1) Consider the problem

$$\inf_{f \in \mathcal{F}} \mathbb{E}_{(x,y)\sim\mu}[\sup_{\tilde{x} \in \overline{B}_\varepsilon(x)} |f(\tilde{x}) - y|],$$

where $\mathcal{Y} = \mathbb{R}$, $\overline{B}_\varepsilon(\cdot)$ is the ball associated to some norm $\|\cdot\|$ in $\mathbb{R}^d$, and $\mathcal{F}$ is the family of linear functions of the form:

$$f(\cdot) = \langle \cdot, \theta \rangle, \quad \theta \in \mathbb{R}^d.$$

Prove that the above (AT) is equivalent to

$$\inf_{\theta \in \mathbb{R}^d} \mathbb{E}_{(x,y)\sim\mu}[|\langle x, \theta \rangle - y|] + \varepsilon ||\theta||_*,$$

where $\|\cdot\|_*$ is $\|\cdot\|$'s dual norm.

2 Lecture 2

Scribes: Kevin Ren

2.0.1 Optimal transport revisited from last time

Discrete setting: suppose we have sources $x_1, \cdots, x_n$ with weights $p_1, p_2, \cdots, p_n$, mapping into sinks $y_1, \cdots, y_m$ with weights $q_1, \cdots, q_m$, with $\sum_i p_i = \sum_j q_j = 1$. Then optimal transport becomes: over all matrices $\pi \in \mathbb{R}^{n\times m}$ with non-negative entries such that $\sum_j \pi_{ij} = p_i$ for all j and $\sum_i \pi_{ij} = q_j$ for all i, what is the minimum cost $\sum_{i,j} C_{ij}\pi_{ij}$, where C_{ij} is the cost of transporting i to j?

Continuous setting:

$$\inf_{\pi\in\Gamma(\mu,\tilde{\mu})} \int C_{\mathcal{Z}}(z,\tilde{z})\, d\pi(z,\tilde{z})$$

$\Gamma(\mu,\tilde{\mu}) = \{\pi \in \rho(\mathcal{Z}\times\mathcal{Z}) : P_{1\#}\pi = \mu, P_{2\#}\pi = \tilde{\mu}\}$

$P_{i\#}\pi$ is the i-th marginal of π.

Multi-marginal optimal transport will generalize the two-marginal optimal transport considered above. Uses higher-order tensors

Discrete setting: $\pi_{i_1\cdots i_d}$ is the amount of mass that is assigned at $x_{i_1}x_{i_2}\cdots x_{i_d}$.

Given marginals $\sum_{i_1\cdots i_{k-1}i_{k+1}\cdots i_d} \pi_{i_1\cdots i_d} = p^k_{i_k}$ for all $k, i_1, \cdots, i_d$, what is $\min_{\pi\times\mathbb{R}^{n_1\times\cdots\times n_d}} \sum_{i_1,\cdots,i_d} C_{i_1\cdots i_d}\pi_{i_1\cdots i_d}$?

Continuous setting: $\Gamma(\mu_1, \cdots, \mu_d) = \{\pi \in \rho(\mathcal{Z}^{\times d}) : P_{\ell\#}\pi = \mu_d\}$

2.0.2 Upper and lower bounds on AT and DRO (Universal lower bounds)

Upper bounds: $\mathcal{F}$

Suppose you can find $C(f,x,y)$ (the **certificate**) s.t.

$$\mathbb{E}_{(x,y)\sim\mu}[\sup_{\tilde{x}\in\overline{B_\varepsilon(x)}} \ell(f(x),y)] \le \mathbb{E}_{(x,y)\sim\mu}[\sup_{\tilde{x}\in\overline{B_\varepsilon(x)}} C(f,x,y)]$$

for all $f\in\mathcal{F}$. We can try to solve the surrogate problem

$$\inf_{f\in\mathcal{F}} \mathbb{E}_{(x,y)\sim\mu} C(f,x,y),$$

which will be an upper bound for the adversarial risk. Moreover, for the solution f_s^* of the surrogate (hopefully a much easier problem) we can “certify" how large is its adversarial risk: it is no larger than $\mathbb{E}_{(x,y)\sim\mu}[C(f_s^*,x,y)]$.

How to find certificates: $\sup_{\tilde{x}\in B_\varepsilon(x)} \ell(f(\tilde{x}),y)$. Treat $\ell(f(\tilde{x}),y)$ as a function $g(\tilde{x})$ of $\tilde{x}$. Then by fundamental theorem of calcuus one can get:

$$\sup_{\tilde{x}\in B_\varepsilon(x)} g(\tilde{x}) \le g(x) + \varepsilon \sup_{\tilde{x}\in B_\varepsilon(x)} \|\nabla g(\tilde{x})\|_*.$$

Student question: what value of ε to choose? Random networks?

A: If one has enough computational power, then try many different values of ε. But maybe a better approach is: we have a tradeoff between robustness and accuracy. For a given accuracy, tune ε to get the given accuracy and best possible robustness. For random networks, see references.

What about universal lower bounds? Notice that, regardless of $\mathcal{F}$ we always have:

$$\inf_{f\in\mathcal{F}} \sup_{\tilde{\mu}\in\rho(\mathcal{X}\times\mathcal{Y})} \mathbb{E}_{(\tilde{x},\tilde{y})\sim\tilde{\mu}}[\ell(f(\tilde{x}),\tilde{y})]-C(\mu,\tilde{\mu}) \geq \inf_{f\in\mathcal{F}_{\text{all}}} \sup_{\tilde{\mu}\in\rho(\mathcal{X}\times\mathcal{Y})} \mathbb{E}_{(\tilde{x},\tilde{y})\sim\tilde{\mu}}[\ell(f(\tilde{x}),\tilde{y})]-C(\mu,\tilde{\mu})$$

The larger the class $\mathcal{F}$, the closer these two quantities are. The second quantity does not suffer from overfitting and hence has an ok answer (see exercise from last time!) Indeed, the lecturer has a paper on comparing these two quantities.

2.0.3 Classification setting

$\mathcal{Y} = \{1, \cdots, K\}$

$f : X = \mathbb{R}^d \to \rho(Y)$

$0-1$ loss: $\ell(p, y) = 1 - f_y(x)$

Open question for tomorrow: if $\ell(p, y) = -\log f_y(x)$ (cross-entropy loss), what should be the correct answer?

$$C(\mu,\tilde{\mu}) := \inf_{\pi\in\Gamma(\mu,\tilde{\mu})} \int c_{\mathcal{Z}}(z,\tilde{z})\, d\pi(z,\tilde{z}), \qquad c_{\mathcal{Z}} = \begin{cases} c(x,\tilde{x}) & \text{if } y=\tilde{y}, \\ \infty & \text{otherwise.} \end{cases}$$

The adversary, to confuse the learner, would like to merge points with different colors if possible (the more different colors merged, the better for teh adversary). Thus, the adversary has a tradeoff between cost and reward. This tradeoff is interesting and will be explored soon.

Question: why is $C(\mu, \tilde{\mu})$ lower semi-continuous?

A: to guarantee that the optimization problem has a minimizer

2.0.4 Back to universal lower bounds

Claim:

$$\inf_{f\in\mathcal{F}_{\text{all}}} \sup_{\tilde{\mu}\in\rho(\mathcal{X}\times\mathcal{Y})} \mathbb{E}_{(\tilde{x},\tilde{y})\sim\tilde{\mu}}[\ell(f(\tilde{x}),\tilde{y})]-C(\mu,\tilde{\mu}) = \sup_{\tilde{\mu}\in\rho(\mathcal{X}\times\mathcal{Y})} \inf_{f\in\mathcal{F}_{\text{all}}} \mathbb{E}_{(\tilde{x},\tilde{y})\sim\tilde{\mu}}[\ell(f(\tilde{x}),\tilde{y})]-C(\mu,\tilde{\mu})$$

This needs justification, but we will do it *a posteriori*.

Assuming claim (and using the notation $\mu_i(\cdot) := \mu(\cdot \times \{i\})$, $C(\mu_i, \tilde{\mu}_i) = \inf_{\pi\in\Gamma(\mu_i,\tilde{\mu}_i)} \int c(x,\tilde{x})\, d\pi(x,\tilde{x})$), we can decouple by color:

$$\inf_{f\in\mathcal{F}_{\text{all}}} \mathbb{E}_{(\tilde{x},\tilde{y})\sim\tilde{\mu}}[\ell(f(\tilde{x}),\tilde{y})] - C(\mu,\tilde{\mu}) = \sum_{i=1}^{K} \int \ell(f(\tilde{x}),i)\, d\tilde{\mu}_i(\tilde{x}) - \sum_{i=1}^{K} C(\mu_i,\tilde{\mu}_i).$$

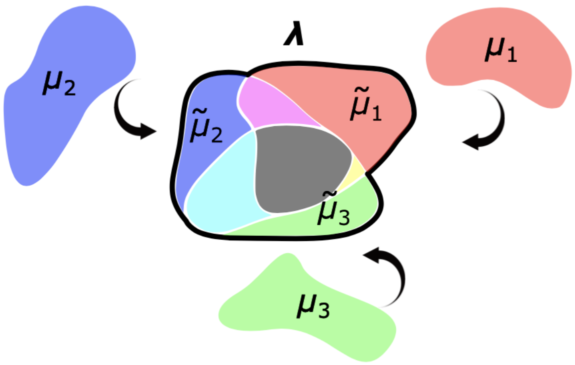

Figure 1: Illustration of generalized Wasserstein barycenter problem. The goal is to achieve the largest overlap (smallest $\lambda(\mathcal{X})$) between the classes at the cheapest cost.

We wish to minimize this over $f \in \mathcal{F}_{\text{all}}$. Indeed, consider pointwise for all $\tilde{x}$. The best strategy (see Exercise 1 from last session) is to select $p_i = 1$ when $\frac{d\tilde{\mu}_i}{d\tilde{\mu}}(\tilde{x})$ is largest. (Here, we abuse notation and let $\tilde{\mu} = \sum_{i=1}^{K} \tilde{\mu}_i$.) Thus,

$$\sup_{\tilde{\mu} \in \rho(\mathcal{X} \times \mathcal{Y})} \inf_{f \in \mathcal{F}_{\text{all}}} \mathbb{E}_{(\tilde{x},\tilde{y}) \sim \tilde{\mu}}[\ell(f(\tilde{x}), \tilde{y})] - C(\mu, \tilde{\mu}) = \sup_{\tilde{\mu}_1, \cdots, \tilde{\mu}_K} 1 - \int \max_{i=1,\cdots,K} \frac{d\tilde{\mu}_i(\tilde{x})}{d\tilde{\mu}}(\tilde{x}) d\tilde{\mu}(\tilde{x}) - \sum_{i=1}^{K} C(\mu_i, \tilde{\mu}_i).$$

The max looks scary, but we can give it a name. Rewrite as

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

From this last expression, we see our previous heuristic in quantitative form: the adversary wishes to minimize the cost while making the new measures $\tilde{\mu}_i$ overlap as much as possible. This is the **generalized Wasserstein barycenter problem.** Compare with the classical Wasserstein problem: given $\gamma_1(\mathcal{X}), \cdots, \gamma_k(\mathcal{X})$, find $\inf_\gamma \sum_{i=1}^{K} C(\gamma_i, \gamma)$. See exercise 3 today.

2.1 Further Remarks and References

1. The notion of generalized Wasserstein barycenter problem and its connection to AT were introduced in the paper *The multimarginal optimal transport formulation of adversarial multiclass classification*, which you can access here `https://www.jmlr.org/papers/v24/22-0698.html`.
2. One specific case where you can find useful upper bounds for AT is presented in the paper *Certified defenses against adversarial examples*, which you can

find here `https://arxiv.org/pdf/1801.09344.pdf`. There, the idea of constructing certificates for 1-hidden layer neural networks binary classifiers is discussed.

3. The connection between DRO and Lasso models discussed in problems 1 and 2 below appears in the work *Robust Wasserstein Profile Inference and Applications to Machine Learning*, whose ArXiv version you can find here `https://arxiv.org/pdf/1610.05627.pdf`.

4. The paper *Square-Root Lasso: Pivotal Recovery of Sparse Signals via Conic Programming* introduced the squared-root Lasso model (see `https://arxiv.org/abs/1009.5689`). In contrast to the standard Lasso model, squared-root Lasso is pivotal, a concept that is discussed in that paper and that has important consequences for the practical estimation of parameters from finite data.

2.2 Exercises

1. (Regularization and AT Part 2) Consider the DRO problem

$$\min_{f \in \mathcal{F}} \sup_{\tilde{\mu} \text{ s.t. } W_2^2(\mu,\tilde{\mu}) \leq \varepsilon} \mathbb{E}_{(\tilde{x},\tilde{y})\sim\tilde{\mu}}[\ell(f(\tilde{x}),\tilde{y})].$$

where $\mathcal{Y} = \mathbb{R}$, $\mathcal{F}$ is the set of all linear functions of the form $f(\cdot) = \langle \cdot, a \rangle + b$ ($a \in \mathbb{R}^d$ and $b \in \mathbb{R}$), and W_2^2 is the (squared) 2-OT distance:

$$W_2^2(\mu,\tilde{\mu}) = \inf_{\pi \in \Gamma(\mu,\tilde{\mu})} \int ||z - \tilde{z}||^2 d\pi(z,\tilde{z})$$

for some norm $\|\cdot\|$ in $\mathbb{R}^{d+1}$.

Prove that the above (DRO) is equivalent (same value and same minimizer) to the problem

$$\min_{a \in \mathbb{R}^d, b \in \mathbb{R}} \left(\sqrt{\mathbb{E}_{(x,y)\sim\mu}[(\langle a, x \rangle + b - y)^2]} + \sqrt{\varepsilon} \|(a,-1)\|_* \right)^2,$$

where $\|\cdot\|_*$ is the dual norm of $\|\cdot\|$. In particular, (DRO) has the same minimizers as

$$\min_{a \in \mathbb{R}^d, b \in \mathbb{R}} \sqrt{\mathbb{E}_{(x,y)\sim\mu}[(\langle a, x \rangle + b - y)^2]} + \sqrt{\varepsilon} \|(a,-1)\|_*.$$

2. (DRO and squared-root Lasso). Consider now a norm $\|\cdot\|$ in $\mathbb{R}^d$ and let $\|\cdot\|_*$ be its dual norm. Use the above problem to deduce that the DRO model

$$\min_{f \in \mathcal{F}} \sup_{\tilde{\mu} \text{ s.t. } C(\mu,\tilde{\mu}) \leq \varepsilon} \mathbb{E}_{(\tilde{x},\tilde{y})\sim\tilde{\mu}}[\ell(f(\tilde{x}),\tilde{y})],$$

with $\mathcal{F}$ as in problem 1 and

$$C(\mu, \tilde{\mu}) = \inf_{\pi \in \Gamma(\mu, \tilde{\mu})} \int c(z, \tilde{z}) d\pi(z, \tilde{z})$$

$$c(z, \tilde{z}) := \begin{cases} ||x - \tilde{x}||^2 & \text{if } y = \tilde{y} \\ \infty & \text{otherwise,} \end{cases}$$

has the same minimizers as the regularization problem

$$\min_{a \in \mathbb{R}^d, b \in \mathbb{R}} \sqrt{\mathbb{E}_{(x,y)\sim\mu}[(\langle a, x\rangle + b - y)^2]} + \sqrt{\varepsilon}\|a\|_*.$$

Note: When $\|\cdot\|$ is chosen to be the ℓ^∞ norm in $\mathbb{R}^d$, then $\|\cdot\|_*$ is the ℓ^1 norm and the resulting regularization problem is the so called *squared-root Lasso model*, a popular model in statistics.

3. Consider the generalized barycenter problem:

$$\min_{\lambda, \widetilde{\mu}_1, \dots, \widetilde{\mu}_K} \left\{ \lambda(\mathcal{X}) + \beta \sum_{i \in \mathcal{Y}} C(\mu_i, \widetilde{\mu}_i) : \lambda \succeq \widetilde{\mu}_i \text{ for all } i \in \mathcal{Y} \right\},$$

for a cost function c that, for simplicity, is continuous. In the above, β is a positive parameter. Suppose, in addition, that all classes are balanced, so that

$$\mu_1(\mathcal{X}) = \cdots = \mu_K(\mathcal{X}).$$

Prove that if $\{\beta_n\}_{n\in\mathbb{N}}$ is a sequence converging to zero, and if $(\lambda^{*,n}, \tilde{\mu}_1^{*,n}, \dots, \tilde{\mu}_K^{*,n})$ is a solution to the generalized barycenter problem for $\beta = \beta_n$, then, up to subsequence, all of $\lambda^{*,n}, \tilde{\mu}_1^{*,n}, \dots, \tilde{\mu}_K^{*,n}$ converge weakly (narrowly) toward the same positive measure $\overline{\lambda}$, which is a solution to the standard Wasserstein barycenter problem:

$$\inf_{\overline{\lambda}} \sum_{i=1}^{K} C(\mu_i, \overline{\lambda}).$$

3 Lecture 3

Scribes: Rachel Morris and Liane Xu

Recall the universal (DRO):

$$\inf_{f \in \mathcal{F}_{\text{all}}} \sup_{\tilde{\mu} \in \rho(\mathcal{X} \times \mathcal{Y})} \mathbb{E}_{(\tilde{x}, \tilde{y}) \sim \tilde{\mu}}[\ell(f(\tilde{x}), \tilde{y})] - C(\mu, \tilde{\mu}) \tag{DROuniv}$$

where $C(\cdot, \cdot) = \inf_{\pi \in \Gamma(\mu, \tilde{\mu})} \int C_z(z, \tilde{z}) \, d\pi(z, \tilde{z})$ where $C_z(z, \tilde{z}) = \begin{cases} c(x, \tilde{x}) & \text{if } y = \tilde{y} \\ \infty & \text{else} \end{cases}$

We are most interested in the classification problem. For classification, $\mathcal{Y} = \{1, \cdots, K\}$, where K is the number of classes (e.g. 10 or 100). Our $f : X \to \rho(Y)$, a probability vector whose i-th entry is the confidence assigned to class i.

Recall from last time that we reduced the universal problem (DROuniv) to the following generalized Wasserstein barycenter problem:

$$\inf_{\lambda, \tilde{\mu}_1, \ldots, \tilde{\mu}_k} \lambda(\mathcal{X}) + \sum_{i=1}^{K} C(\mu_i, \tilde{\mu}_i)$$

s..t. $\lambda \geq \tilde{\mu}_i$ for all $i = 1, \ldots, K$. This has inputs $\mu_1, \ldots, \mu_k$ which are measures that represent the data distribution of the colors and cost function C. Intuitively, the cost function c models the adversary/what it can do. The generalized Wasserstein barycenter problem is related to the standard Wasserstein barycenter problem:

$$\inf_{\gamma} \sum_{i=1}^{K} C(\gamma_i, \gamma)$$

where $\gamma_1(X) = ... = \gamma_k(X)$. In the standard problem, we think of γ as the barycenter whereas in the generalized problem we think of λ as the generalized barycenter.

As an analogue to the barycenter in $\mathbb{R}^d$, observe that the barycenter of some points $x_1, ..., x_k \in \mathbb{R}^d$ is

$$\frac{1}{k} \sum_i x_i = \arg \min_{x \in \mathbb{R}^d} \sum_i |X - X_i|^2$$

.

There are algorithms already for the standard Wasserstein barycenter problem. We can do something similar for the generalized Wasserstein barycenter problem. First, we will look at the standard Wasserstein barycenter problem.

Theorem 1 (Agick-Carlier '10)**.** *Let* $\mathbf{c}(x_1, ..., x_k) := \inf_{\overline{x}} \sum_{i=1}^{k} c(x_i, \overline{x})$. *The problem*

$$\inf_{\gamma} \sum_{i=1}^{k} C(\gamma_i, \gamma)$$

is equivalent to the following multi-marginal optimal transport (MOT) problem

$$\inf_{\pi \in \Gamma(\gamma_1, \ldots, \gamma_k)} \int \mathbf{C}(x_1, \ldots, x_k)\ d\pi(x_1, \ldots, x_k).$$

(Note that each x_i correspond to the γ_i)

We will now discuss algorithms for MOT, which will solve the standard Wasserstein barycenter problem. We want to follow a similar approach for the generalized Wasserstein barycenter problem as well: find a MOT analogue and then solve the MOT-analogue problem.

We will show this in the case where there are two measures ($K = 2$). The generalization follows.

Sinkhorn Algorithm for OT:

To solve:

$$\inf_{\pi \in \Gamma(\mu, \nu)} \int C(x, y)\ d\pi(x, y)$$

where $\mu, \nu \in \mathcal{P}(\mathcal{X})$ and $x, y \in \mathcal{X} = \mathbb{R}^d$.

Let $x_1, \ldots x_n$ be the support points of μ and $\mu_1, \ldots, \mu_i$ are the masses at each of the x_i.

Let $y_1, ..., y_m$ be the support points for ν, $\nu_1, ..., \nu_m$ the masses at each y_j.

We are now just solving the linear program

$$\inf_{\pi} \sum_{i,j} c_{ij}\pi_{ij} + \eta \sum_{i,j} (\log \pi_{ij} - 1)\pi_{ij}$$

where $c_{ij} = C(x_i, y_j)$ such that $\sum_{i=1}^{n} \pi_{ij} = \nu_j, \forall j = 1, \ldots, m$ and $\sum_{j=1}^{m} \pi_{ij} = \mu_i, \forall i = 1, \ldots, n$ (conservation of mass) and $\pi_{ij} \geq 0$.

The linear problem $\inf_{\pi} \sum_{i,j} c_{ij}\pi_{ij}$ is not strongly convex, so we add on a second term (seen above) that is the “entropy regularization" term, $\eta \sum_{i,j} (\log \pi_{ij} - 1)\pi_{ij}$. Now, the problem is strongly convex. Notice that in the strongly convex version, we no longer need the constraint $\pi_{ij} \geq 0$. We can control η (the original problem is when η is 0, but this method works better when $\eta \neq 0$). Some of the best results come from allowing η to be adjustable depending on how close we want to be to the solution of the original problem. There are other forms of regularization too, but this one is nice so we'll focus on it.

Trick: Look at the dual!

Let's denote the collection of Lagrange multipliers by $\phi = (\phi_1, ..., \phi_n)$ (Lagrange multipliers corresponding to μ_i) and $\psi = (\psi_1, ..., \psi_m)$ (Lagrange multipliers corresponding to ν_j).

$$\mathcal{L}(\pi; \phi, \psi) = \sum_{i,j} c_{ij}\pi_{ij} + \eta \sum_{i,j} (\log \pi_{ij} - 1)\pi_{ij} + \sum_{i=1}^{n} \phi_i(\mu_i - \sum_j \pi_{ij}) + \sum_{j=1}^{m} \psi_j(\nu_j - \sum_i \pi_{ij})$$

The dual objective is

$$g(\phi, \psi) = \inf_{\pi} \mathcal{L}(\pi, \phi, \psi) = \sum_i \phi_i \mu_i + \sum_j \psi_j \nu_j - \eta \sum_{i,j} \exp\left(\frac{1}{n}(\phi_i + \psi_j - C_{ij})\right)$$

where the second equality comes from differentiating and setting to 0. Then, the dual problem is,

$$\max_{\phi,\psi} \sum_i \phi_i \mu_i + \sum_j \psi_j \nu_j - \eta \sum_{i,j} \exp\left(\frac{1}{\eta}(\phi_i + \psi_j - C_{ij})\right).$$

For a given (ϕ, ψ), we can construct an associated π via

$$\pi(\phi,\psi)_{ij} = \exp\left(\frac{1}{\eta}(\phi_i + \psi_j - c_{ij})\right).$$

This is the optimal π that attains the minimum of $\inf_\pi \mathcal{L}(\pi, \phi, \psi)$. Moreover, if the above is feasible for the primal problem (feasible, i.e. it satisfies the constraints), then its is a solution to the primal problem. This comes from the KKT conditions.

Here's the idea for Sinkhorn's algorithm. Suppose at some timepoint t we have a guess (ϕ^t, ψ^t) of ϕ, ψ. To update our guesses, we first maximize

$$\psi^{t+1} = \arg\max_{\psi} \sum_i \phi_i^t \mu_i + \sum_j \psi_j \nu_j + \eta \sum \exp(\frac{1}{\eta}(\phi_i^t + \psi - C_{ij}))$$

Fix one variable, (above ϕ) and then optimize over the other variable to determine ψ^{t+1}. Then, you can do the other variable by maximizing over ϕ and setting $\psi = \psi^{t+1}$ in the equation above.

By differentiating our objective functions and setting them to 0, we obtain the Sinkhorn iterations:

- At $t = 0$, set $\phi^0 = 0, \psi^0 = 0$
- Update ϕ: $\phi^{t+1} = \phi^t + \eta \log(\mu) - \eta \log(P_{1\#}\phi(\phi^t, \psi^t))$ where $P_{1\#}\phi(\phi^t, \psi^t) = \sum_j \pi(\phi^t, \psi^t)_{ij}$
- Update ψ: $\psi^{t+1} = \psi^t + \eta \log(\nu) - \eta \log(P_{2\#}\pi(\phi^{t+1}, \psi^t)$
- $t = t + 1$
- Until: $||\mu - P_{1\#}\pi(\phi^t, \psi^t)||_1 + ||\nu - P_{2\#}\pi(\phi^t, \psi^t)||_1 \leq \delta$ where δ is some predefined tolerance

This algorithm has almost linear convergence. The cost for each iterate is about $O(n^2)$.

The stopping criteria intuitively says that ϕ^t, ψ^t are close to being feasible for the primal, and recall that KKT says that if they are feasible for the primal, then they are the solution to the primal. Therefore, this is an intuitive stopping criteria.

Notice that Sinkhorn is unstable for small η. Also potentially if μ or ν is close to zero. See the book *Computational Optimal Transport* by Peyre and Cuturi for more details.

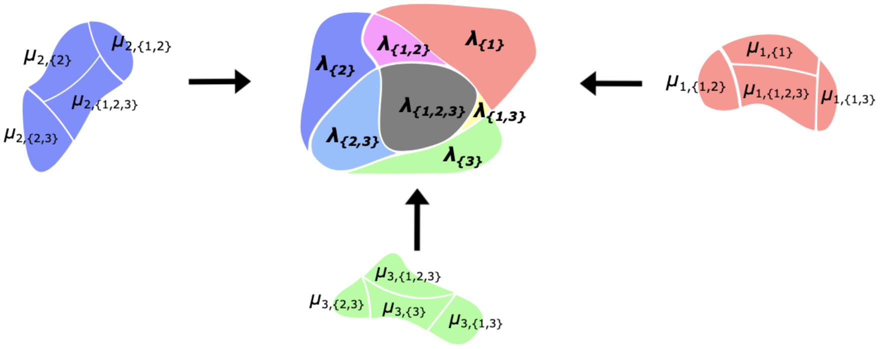

Figure 2: Illustration of splitting of masses.

Generalization to MLT Sinkhorn:

$$\sum_{\ell=i}^{k}\sum_{i_\ell}\phi^\ell_{i_\ell}\mu^\ell_{i_\ell} - \eta \sum_{i_1,\dots,i_k} \exp\left(\frac{1}{\eta}(\phi^1_{i_1}+\cdots+\phi^k_{i_k} - c_{i_1 i_1\cdots i_k}\right)$$

Initialize $\phi^{l,0} \equiv 0$ for all l. Then, find greedy coordinate (i.e. update the worst coordinate)

$$I = \arg\max_{\ell=1,\dots,k}\left\{D_{KL}(\mu^\ell || P_{\ell\#}\pi(\phi^{1,t},\dots,\phi^{k,t}))\right\}$$

where D_{KL} is the KL divergence.

We can view this as a version of gradient ascent, in a coordinate-wise fashion.

What do we do for the generalized barycenter problem?

A is a subset of $\{1,\dots,k\}$ s.t. $i \in A$. Recall our color example with colors red, blue and green. We will consider measures $\mu_{i,A}$ where color i can interact with the colors in A. For example, $\mu_{\text{red, \{blue, red\}}}$ are the red points that overlap with some blue points but not green points (pink points in Figure 2).

We can rewrite the problem as

$$\inf_{\{\lambda_A,\mu_{1,A},\dots,\mu_{k,A}\}_{A\in[k]}} \sum_A \lambda_A(X) + \sum_{i\in A} C(\mu_{i,A},\lambda_A)$$

This becomes the analog of MOT,

$$\inf_{\{\pi_A\}_A}\sum_A \int (1 + C_A(x_A))\, d\pi_A(x_A) \quad \text{s.t.} \sum_{A \text{ s.t. } i\in A} P_{i\#}\pi_A = \mu_i, \ \forall i = 1,\dots,k$$

where $C_A(x_A) = \inf_{\bar{x}} \sum_{i\in A} c(x_i,\bar{x})$ and $x_A = (X_i)_{i\in A}$.

In practice, we don't have to consider all $A \subseteq [k]$; we can truncate and only consider, e.g., A consisting of at most 4 colors/classes.

3.1 Further Remarks and References

1. The following are important references that study the computational complexity of Sinkhorn iterations for OT (or MOT) problems:

   - *Near-linear time approximation algorithms for optimal transport via Sinkhorn iteration* (`https://arxiv.org/abs/1705.09634`), which analyzes the computational complexity of Sinkhorn iterations, not only to approximate the entropy regularized OT problem, but also to approximate the original OT problem.
   - *On the Complexity of Approximating Multimarginal Optimal Transport* (`https://arxiv.org/abs/1910.00152`), which analyzes the multimarginal case.

2. The analog of the MOT problem for the adversarial training problem can be found in Equation (21) in the paper *The multimarginal optimal transport formulation of adversarial multiclass classification*, which you can access here `https://www.jmlr.org/papers/v24/22-0698.html`. Based on this reformulation for the DRO problem, the paper *An Optimal Transport Approach for Computing Adversarial Training Lower Bounds in Multiclass Classification* (`https://arxiv.org/abs/2401.09191`) adapts Sinkhorn iterations (after introducing an appropriate entropic regularization) to obtain a scalable algorithm for computing adversarial training lower bounds.

3.2 Exercises

1. (Lower bounds for some specific models) Consider the (AT) problem:

   $$\inf_{f\in\mathcal{F}} \mathbb{E}_{(x,y)\sim\mu}[\sup_{\tilde{x}\in\overline{B}_\varepsilon(x)} \ell(f(\tilde{x}),y)]$$

   for $\mathcal{F}$ the set of liner functions of the form:

   $$f(x)=\langle a,x\rangle+b$$

   for some $a\in\mathbb{R}^d$ and $b\in\mathbb{R}$. Suppose, in addition, that the function $\ell(\cdot,y)$ is convex (and differentiable) regardless of the value of the admissible y in the learning problem. Prove that the quantity

   $$\inf_{f\in\mathcal{F}} \mathbb{E}_{(x,y)\sim\mu}[\ell(f(x),y)+\varepsilon\|\nabla_x\ell(f(x),y)\|_*]$$

   is a lower bound for (AT) as above. Here, $\|\cdot\|_*$ is the dual to the norm that determines the ball $B_\varepsilon(x)$.

2. (Open problem) The discussion on universal lower bounds for DRO problems in the multiclass classification setting assumed that the loss function was the $0-1$ loss. How would you extend this discussion if the loss function

was more general? For example, what would happen if the loss function took the form:

$$\ell(f(x), y) = -\log(f_y(x)),$$

i.e., it was the cross entropy? Is it possible to come up with an efficient algorithm that computes the universal lower bound for this type of loss function?

4 Lecture 4

Scribes: Rachel Morris and Kevin Ren

4.0.1 Existence and regularity of solutions to (ATP)

Setting: Binary Classification

- Labels: $\mathcal{Y} = \{0, 1\}$
- Classifiers: $f : \mathcal{X} \to \mathcal{Y} = 1_A, A \in \mathcal{B}(\mathcal{X})$, indicators of sets

The question of regularity boils down to the regularity of the decision boundary. Rewrite (ATP) for the decision problem as:

$$\inf_{A \in \mathcal{B}(X)} \mathbb{E}_{(x,y)\sim\mu}[\sup_{\tilde{x} \in B_\varepsilon(x)} |1_A(\tilde{x}) - y|] \tag{ATPdecision}$$

We define μ_0, μ_1 as $\mu_0(\cdot) = \mu(\cdot \times \{0\}$ and $\mu_1(\cdot) = \mu(\cdot \times \{1\}$. Now, we reformulate ATPdecision by disintegrating the measure μ:

$$\begin{aligned}\mathbb{E}_{(x,y)\sim\mu}[\sup_{\tilde{x} \in B_\varepsilon(x)} |1_A(\tilde{x}) - y|] &= \int \sup_{\tilde{x} \in B_\varepsilon(x)} |1_A(\tilde{x}) - y| \; d\mu(x,y) \\ &= \int \sup_{\tilde{x} \in B_\varepsilon(x)} |1_A(\tilde{x}) - 0| \; d\mu_0(x) + \int \sup_{\tilde{x} \in B_\varepsilon(x)} |1_A(\tilde{x}) - 1| \; d\mu_1(x) \\ &= \int_{A^c} d\mu_1(x) + \int_A d\mu_1(x) \\ &+ \int \sup_{\tilde{x} \in B_\varepsilon(x)} 1_A(\tilde{x}) - 1_A(x) \; d\mu_0(x) + \int 1_A(x) - \inf_{\tilde{x} \in B_\varepsilon(x)} 1_A(\tilde{x}) \; d\mu_1(x)\end{aligned}$$

Above, the first two integrals represent the misclassification due to the choice in classifier. Then, the second two terms represent the misclassificaiton due to adversarial perturbation (i.e. they are correctly classified but they are too close to the decision boundary). Then, ATPdecision admits the following representation:

$$R_\varepsilon(A) = \text{Risk}(1_A) + \varepsilon \text{Per}_\varepsilon(A)$$

Alternatively, adversarial risk:

$$R_\varepsilon(A) = \int_{A^c} d\mu_1(x) + \int_A d\mu_1(x) + \int \sup_{\tilde{x} \in B_\varepsilon(x)} 1_A(\tilde{x}) - 1_A(x) \, d\mu_0(x) + \int 1_A(x) - \inf_{\tilde{x} \in B_\varepsilon(x)} 1_A(\tilde{x}) \, d\mu_1(x)$$

Reinterpret as the following: the set of x such that $B_\varepsilon(x)$ intersects both A and A^c. (The ε-neighborhood of the "boundary".)

4.0.2 Tools from calculus of variations

Q: Does there exist a solution to ATPdecision?

There are two approaches: (1) the direct method of calculus of variations and (2) relaxation.

For our problem, the direct method doesn't work. Instead, we construct a relaxed problem which does have a solution (via the direct method), and use the solution to the relaxed problem to construct a solution to the original problem.

Direct Method of Calculus of Variations: Topological space $(\mathcal{H}, \tau)$, trying to find solutions to $\inf_{h\in\mathcal{H}} \psi(h)$. We will assume

1. Sequential Lower Semi-continuity (l.s.c.): For a sequence $\{h_n\}_{n\in\mathbb{N}}$ s.t. $h_n \to_\tau h$, then $h \leq \liminf_{n\to\infty} \psi(h_n)$.
2. $\inf_{h\in\mathcal{H}} \psi(h) > -\infty$
3. Sequential Precompactness of level sets: If $\{h_n\}_{n\in\mathbb{N}}$ is a sequence with $\sup_{n\in\mathbb{N}} \psi(h_n) < \infty$, then there exists a convergent (in topology τ) subsequence of $\{h_n\}_{n\in\mathbb{N}}$ with
$$\exists \{n_k\}_k \text{ s.t. } h_{n_k} \to_\tau h \text{ for some } h.$$

One needs a "fine line" between compactness and lower semi-continuity. If your topology is too fine, then it isn't compact; if it is too coarse, then we don't have lower semi-continuity.

If the above (1) - (3) hold, then there exists h^* solution to the problem.

Proof. Let $\{h_n\}_n$ be a minimizing sequence for the problem. By the definition of infimum,
$$\lim_{n\to\infty} \psi(h_n) = \inf_{h\in\mathcal{H}} \psi(h)$$
Clearly $\sup_{n\in\mathbb{N}} \psi(h_n) < \infty$, so by compactness (3), we pass to a subsequence (which we keep denoting $\{h_n\}_n$) such that $h_n \to h$ for some h. By lower semi-continuity (1), we have
$$\psi(h) \leq \liminf_{n\to\infty} \psi(h_n) = \inf_{h\in\mathcal{H}} \psi(h).$$
□

Can we apply this direct method argument to ATPdecision? Implicit is that we have a defined topology. What is the topology on $\mathcal{B}(\mathcal{X})$, and can we expect l.s.c. or compactness properties? We will do some relaxations in order to get to a setting where we can apply the direct method.

Modified Adversarial Risk:

$$\begin{aligned}\tilde{R}_\varepsilon(A) = &\int_{A^c} d\mu_1(x) + \int_A d\mu_1(x) \\ &+ \int \nu\text{-ess }\sup_{\tilde{x}\in B_\varepsilon(x)} 1_A(\tilde{x}) - 1_A(x)\; d\mu_0(x) + \int 1_A(x) - \nu\text{-ess }\inf_{\tilde{x}\in B_\varepsilon(x)} 1_A(\tilde{x})\; d\mu_1(x)\end{aligned}$$

where ν is a positive finite Borel measure to be chosen shortly and

$$\nu\text{-ess }\sup_{\tilde{x}\in B_\varepsilon(x)} g(\tilde{x}) := \inf\{p : \nu(\{\tilde{x} \text{ s.t. } \tilde{x} \in B_\varepsilon(x) \text{ and } g(\tilde{x}) > p\}) = 0\}$$

We want $\tilde{R}$ to be defined on equivalence classes of ν.

Assumptions on ν:

- $\mu_0 + \mu_1$ is absolutely continuous (a.c.) wrt ν.

Observation: The relaxed problem is defined on a subset of $L^\infty(\mathcal{X}, \nu)$.

Let $\mathcal{T}$ be the weak* topology in $L^\infty(X, \nu)$.

Definition 1. $\{g_n\}_{n\in\mathbb{N}} \subset L^\infty(X, \nu)$ converges weak* towards $g \in L^\infty(X, \nu)$ if

$$\lim \int g_n \phi \, d\nu(x) = \int g\phi \, d\nu(x), \qquad \forall \phi \in L^1(X, \nu).$$

Essential property of weak* topology:

Theorem 2. *(Banach-Alaoglu) Suppose* $\{g_n\}_{n\in\mathbb{N}}$ *and* $\sup_{n\in\mathbb{N}} ||g_n||_\infty < \infty$*, then* $\{g_n\}_n$ *has a weak* convergent subsequence.*

An issue is that sequences of indicator functions don't converge to indicator functions.

Second relaxation to ensure that can stilld efine a related problem with a solution: Let $u(x) \in [0, 1]$ be a weak classifier. Then,

$$\begin{aligned} R_\varepsilon(u) = &\int 1 - u(x) \ d\mu_1(x) + \int u(x) \ d\mu_0(x) \\ &+ \int \nu\text{-ess }\sup_{\tilde{x}\in B_\varepsilon(x)} u(\tilde{x}) - u(x) \ d\mu_0(x) + \int u(x) - \nu\text{-ess }\inf_{\tilde{x}\in B_\varepsilon(x)} u(\tilde{x}) \ d\mu_1(x) \end{aligned}$$

The compactness property is better for the space of weak classifiers: sequences of functions mapping into $[0, 1]$ converge to functions mapping into $[0, 1]$.

After these relaxations, we can now apply the direct method.

Conclusion: There exist solutions to

$$\inf u \in L^\infty(\mathcal{X}, \nu), u \in [0, 1]\tilde{R}_\varepsilon(u). \tag{ATPrelax}$$

Coarea Formula:

$$\tilde{R}_\varepsilon(u) = \int_0^1 \tilde{R}_\varepsilon(1_{\{u\geq s\}}) \ ds$$

As a consequence of the coarea formula. Let u^* be the solution to ATPrelax, then

$$\inf_A \tilde{R}_\varepsilon(1_A) \leq \int_0^1 \tilde{R}_\varepsilon(1_{\{u^*>s\}}) \ ds = \tilde{R}_\varepsilon(u^*) \leq \inf_A \tilde{R}_\varepsilon(1_A).$$

So, for Lebesgue a.e. $s \in [0, 1]$, $\tilde{R}_\varepsilon(1_{\{u^*>s\}}) = \inf_A \tilde{R}_\varepsilon(1_A)$, which solves the binary classification problem. (i.e. weak and hard classifier problems are equivalent for binary classification in the adversarial setting) However, this does not extend to the K-classification problem for $K > 2$.

Finally, back to (ATPdecision).

Assume: if $\cup_{x \in \operatorname{supp}(\mu_0+\mu_1)} B_\varepsilon(x) \subseteq \operatorname{supp} \nu$, then we can show that for every $A \in \mathcal{B}(\mathcal{X})$, there exists a set $\tilde{A}$ that is ν-equivalent (i.e. $\nu(A \Delta \tilde{A}) = 0$) such that

$$\sup_{\tilde{x} \in B_\varepsilon(x)} 1_{\tilde{A}}(\tilde{x}) = \nu\text{-ess sup}_{\tilde{x} \in B_\varepsilon(x)} 1_A(\tilde{x})$$
$$\inf_{\tilde{x} \in B_\varepsilon(x)} 1_{\tilde{A}}(\tilde{x}) = \nu\text{-ess inf}_{\tilde{x} \in B_\varepsilon(x)} 1_A(\tilde{x})$$

for all $x \in \operatorname{supp}(\mu_0 + \mu_1)$.

In particular, a good choice is $\nu = \mu_0 + \mu_1 + \gamma$, where γ is the standard Gaussian measure.

4.1 Further Remarks and References

1. In class I mentioned that the AT problem in the classification setting for weak classifiers may not be equivalent to the problem for hard classifiers once the number of classes is greater than 2. You can find a nicely illustrated example of this situation in Appendix C.5. in the paper *On the Role of Randomization in Adversarially Robust Classification* (see `https://hal.science/hal-04312028/document`). Compare this with Exercise 1 from Lecture 1 for the standard risk minimization problem.

2. The exercises for this lecture are developed at different places in the paper *The geometry of adversarial training in binary classification* (`https://arxiv.org/abs/2111.13613`).

4.2 Exercises

1. Let ν be the uniform distribution over the interval $[0, 1]$.

   (a) Give an example of a collection $\{A_n\}_{n\in\mathbb{N}}$ of Borel subsets of the interval $[0, 1]$ such that f$\mathbb{1}_{A_n}$ converges weakly* toward some $u \in L^\infty([0, 1], \nu)$ as n goes to infinity, but u does not take the values 0 or 1 (in particular u is not an indicator function of a set).

   (b) (Lower and upper bounds are preserved by weak* convergence) Suppose that $\{u_n\}_{n\in\mathbb{N}}$ is a sequence in $L^\infty([0, 1], \nu)$ that satisfies:

   $$0 \leq u_n(x) \leq 1,$$

   for ν-a.e. $x \in [0, 1]$ and all n. Show that if the sequence converges in the weak* topology of $L^\infty(\mathcal{X}, \nu)$ toward some u, then u also satisfies

   $$0 \leq u(x) \leq 1,$$

   for ν-a.e. $x \in [0, 1]$.

2. (Sequential l.s.c. for relaxation of AT) In the context of what was discussed in class, suppose that ν is a (finite) positive measure that satisfies the following properties:

   (a) The measure $\mu_0 + \mu_1$ is absolutely continuous with respect to ν.

   (b) (Lebesgue differentiation theorem). For every $g \in L^1(\mathcal{X}, \nu)$ we have:

$$\lim_{r \to 0^+} \frac{1}{\nu(B_r(x))} \int_{B_r(x)} g(\tilde{x}) d\nu(\tilde{x}) = g(x), \quad \nu - \text{a.e.} \quad x \in \mathcal{X}.$$

   Prove that the function:

$$\begin{aligned} u \in L^\infty(\mathcal{X}, \nu) \mapsto \Psi(u) := & \int u(x) d\mu_0(x) + \int (1 - u(x)) d\mu_1(x) \\ & + \int (\nu\text{-ess sup}_{\tilde{x} \in B_\varepsilon(x)} u(\tilde{x}) - u(x)) d\mu_0(x) \\ & + \int (u(x) - \nu\text{-ess inf}_{\tilde{x} \in B_\varepsilon(x)} u(\tilde{x})) d\mu_1(x) \end{aligned}$$

   is sequentially lower semicontinuous with respect to the weak* topology.

3. (Coarea formula) Prove that Ψ from the previous exercise satisfies the coarea formula. Namely, if u takes values between 0 and 1, then

$$\Psi(u) = \int_0^1 \Psi(\mathbb{1}_{\{u > s\}}) ds.$$

4. Prove that if A^* is a solution to the (AT) problem we discussed in class today, then any Borel set B satisfying

$$\text{Op}_\varepsilon(A^*) \subseteq B \subseteq \text{Cl}_\varepsilon(A^*)$$

   is a solution to the (AT) problem.

5 Lecture 5

Scribes: Rachel Morris and Kevin Ren

Definition 2. (Morphological Operations) Let $A \subset \mathcal{X}$ and $\varepsilon > 0$. We define

1. Dilation as $A^{\varepsilon} = \{x \in \mathcal{X} : d(x, A) < \varepsilon\}$
2. Erosion as $A^{-\varepsilon} = \{x \in \mathcal{X} : d(a, A^c) \geq \varepsilon\}$
3. Closing as $cl_{\varepsilon}(A) = (A^{\varepsilon})^{-\varepsilon}$
4. Opening as $op_{\varepsilon}(A) = (A^{-\varepsilon})^{\varepsilon}$

Recall (ATPdecision):

$$\min_{A \in \mathcal{B}(\mathcal{X})} R(1_A) + \int \sup_{\tilde{x} \in B_{\varepsilon}(x)} 1_A(\tilde{x}) - 1_A(x)\, d\mu_0(x) + \int 1_A(x) - \inf_{\tilde{x} \in B_{\varepsilon}(x)} 1_A(\tilde{x})\, d\mu_1(x)$$

where $R(1_A)$ is the standard Bayes risk. We need to be careful with topology and relax the problem twice in order to show existence of solutions (pervious class for details).

Once we have shown $\exists A^*$ minimizer of (AT), then any B satisfying

$$op_{\varepsilon}(A^*) \subseteq B \subseteq cl_{\varepsilon}(A^*)$$

is a solution to (AT). This is one of the exercises from the previous day.

Using these operations, we can find solutions that have one-sided regularity. Specifically, the set $op_{\varepsilon}(A^*)$ is ε inner regular, i.e.e for every $x \in \partial op_{\varepsilon}(A^*)$, there exists $\tilde{x} \in op_{\varepsilon}(A^*)$ such that $x \in \partial B_{\varepsilon}(\tilde{x})$ and $B_{\varepsilon}(\tilde{x}) \subseteq op_{\varepsilon}(A^*)$. Similarly $cl_{\varepsilon}(A^*)$ is ε outer regular, i.e. for every $s \in \partial cl_{\varepsilon}(A^*)$, there exists $\tilde{x} \in cl_{\varepsilon}(A^*)^c$ such that $x \in \partial B_{\varepsilon}(\tilde{x})$ and $B_{\varepsilon}(\tilde{x}) \subseteq cl_{\varepsilon}(A^*)^c$.

To construct regular solutions, we want to interpolate between $op_{\varepsilon}(A^*)$ and $cl_{\varepsilon}(A^*)$. Due to the inner and outer regularity, this guarantees that the boundaries do not oscillate too much when $op_{\varepsilon}(A^*)$ and $cl_{\varepsilon}(A^*)$ touch. We will not go through the details of this proof. Usually, we would expect the regularity to be $C^{1,1}$ for these sorts of conditions, but this has not been proven yet.

Probabilistic version of (ATPdecision) (binary case): First, rewrite (AT) as

$$R(1_A) + \int_{A^c} 1_{\sup_{\tilde{x} \in B_{\varepsilon}(x)} 1_A(\tilde{x}) > 0}\, d\mu_0(x) + \int_A 1_{\sup_{\tilde{x} \in B_{\varepsilon}(x)} 1_{A^c}(\tilde{x}) > 0}\, d\mu_1(x).$$

In other words, we penalize the existence of an attack in $B_{\varepsilon}(x)$. For the probabilistic version, we will now need the probability of an attack to be greater than some $p \in [0, 1]$, that is for some measure $m_{x,\varepsilon}$, $m_{x,\varepsilon}(A) > 0$ for $p = 0$. Here, $\{m_{x,\varepsilon}\}_x$ is a family of probability measures localized around $B_{\varepsilon}(x)$. One example would be $m_{x,\varepsilon} = \mathrm{Unif}(B_{\varepsilon}(x))$ (although in higher dimensions this may concentrate too much along the boundary). A more quantitative probabilistic version of (ATPdecision) is the following probabilistically robust learning (PRL) problem:

$$\mathrm{PRL}_{\varepsilon,p}(1_A) := R(1_A) + \int_{A^c} 1_{m_{x,\varepsilon}(A)>p}\, d\mu_0(x) + \int_A 1_{m_{x,\varepsilon}(A)>p}\, d\mu_1(x). \quad \text{(PRL)}$$

Currently, there is no proof of existence for solutions of PRL. The relaxation technique may not work when redefining for weak classifiers $u \in [0,1]$, we cannot use a coarea formula argument like in proving existence of solutions for (ATPdecision).

We will further generalize PRL. Given a non-decreasing function $\Psi : [0,1] \to [0,\infty)$, we define the generalized PRL risk w.r.t. Ψ as

$$\mathrm{PRL}_{\varepsilon,\Psi} := R(1_A) + \int_{A^c} \Psi(m_{x,\varepsilon}(A))\, d\mu_0(x) + \int_A \Phi(m_{x,\varepsilon}(A^c))\, d\mu_1 \quad \text{(PRLgen)}$$

Note that if $\Psi(t) = 1_{t>p}$, then we recover $\mathrm{PRL}_{\varepsilon,p}$. It will be sufficient to assume that Ψ is concave to get desired existence of solutions. However, not that indicator functions do not satisfy this criteria.

Example. $\Psi(t) := \begin{cases} \frac{t}{p} & \text{if } t \le p, \\ 1 & \text{if } t > p. \end{cases}$

Note that the above example is the smallest concave function that lies above indicator functions.

Q: how to choose the measure $m_{x,\varepsilon}$?

A: usually uniform measure. In high dimensions, may need to be a little careful due to concentration of measure around the boundary. In the original formulation of PRL from 2022 (not the one above), you may get weird phenomenon where if you choose certain measures, then the adversary is "helping" instead of hurting you.

Proof for existence of solutions: as before, switch to weak classifiers.

$$R(u) + J_\Psi(u)$$

where R is the standard risk for weak classifiers (as seen previously) and

$$J_\Psi(u) = \int (1-u(x))\Psi\left(\int u(\tilde{x})\, dm_{x,\varepsilon}(\tilde{x})\right) d\mu_0(x) + \int u(x)\Psi\left(\int 1 - u(\tilde{x})\, dm_{x,\varepsilon}(\tilde{x})\right) d\mu_1(x)$$

Can we use the direct method of calculus of variation to find a solution to this relaxed problem? We don't have a coarea formula even if Ψ is concave. Fortunately, we can save this by relaxing further. Define

$$V_\Psi(u) := \int_0^1 J_\Psi(1_{\{u>s\}})\, ds. \quad (1)$$

Observe that $V_\Psi(1_A) = J_\Psi(1_A)$ for all $A \in \mathcal{B}(\mathcal{X})$. In particular, using this and (1), we see that V_Ψ satisfies the coarea formula

$$V_\Psi(u) = \int_0^1 V_\Psi(1_{\{u>s\}})\, ds.$$

Note that because J_Ψ does not satisfy the coarea formula in general, we cannot say $V_\Psi(u) = J_\Psi(u)$ for all weak classifiers u.

However, if Ψ is a concave function, the following properties hold for the triplet $(R + J_\Psi,\ R + V_\Psi,$ weak* in $L^\infty(\mathcal{X}, \nu))$:

1. $V_\Psi(1_A) = J_\Psi(1_A)$ for all $A \in \mathcal{B}(\mathcal{X})$
2. $J_\Psi(u) \geq V_\Psi(u)$ for all $u : \mathcal{X} \to [0, 1]$ (using Jensen's inequality and the fact the Ψ is concave)
3. J_Ψ is sequentially lower semi-continuous wrt weak* in $L^\infty(\mathcal{X}, \nu)$
4. V_Ψ satisfies the coarea formula, $V_\Psi(u) = \int_0^1 J_\Psi(1_{\{u>s\}})\, ds$

Note that we cannot say that V_Ψ is l.s.c. (or else we could just jump to the definition of V_Ψ without J_Ψ).

Proof. Let $\{A_n\_n \in \mathbb{N}\}$ be such that $1_{A_n} \to_* u$ and let it be a minimizing sequence of

$$\begin{aligned}
\inf_{A \in \mathcal{B}(\mathcal{X})} \{R(1_A) + J_\Psi(1_A)\} &= \lim_{n\to\infty} R(1_{A_n}) + J_\Psi(1_{A_n}) \\
&= \liminf_{n\to\infty} R(1_{A_n}) + J_\Psi(1_{A_n}) \\
&\geq R(u) + J_\Psi(u) \quad \text{(Property 3)} \\
&\geq R(u) + V_\Psi(u) \quad \text{(Property 2)} \\
&= \int_0^1 R(1_{\{u>s\}} + V_\Psi(1_{\{u>s\}})\ ds \quad \text{(Property 4)} \\
&= \int_0^1 R(1_{\{u>s\}}) + J_\Psi(1_{\{u>s\}})\ ds \quad \text{(Property 1)} \\
&\geq \inf_A \{R(1_A) + J_\Psi(1_A)\}
\end{aligned}$$

Yesterday, we used the coarea formula for J_Ψ to finish the proof. This is not true, but today we instead dominated $J_\Psi(u)$ by $V_\Psi(u)$, and used the coarea formula for V_Ψ. Thus, convexity of Ψ was crucial for the proof.

□

Q: can we use distributions (Young measures) instead of functions to simplify the argument?

A: sounds reasonable, haven't tried it yet

Further properties of $A \in \mathcal{B}(\mathcal{X}) \to J_\Psi(1_A)$

Assume Ψ is concave.

1. $J_\Psi(1_{A\cap B}) + J_\Psi(1_{A\cup B}) \leq J_\Psi(1_A) + J_\Psi(1_B)$ for all $A, B \in \mathcal{B}(\mathcal{X})$ (submodularity)
2. If $\Psi(0) = 0$, then $J_\Psi(1_\emptyset) = J_\Psi(1_X) = 0$

Proposition 1. *Let A be a set with "smooth enough" boundary.*

$$\lim_{\varepsilon\to 0}\frac{J_{\Psi,\varepsilon}(1_A)}{\varepsilon}=c_\Psi\int_{\partial A}(\rho_0(x)+\rho_1(x))\,d\mathcal{H}^{d-1}(x)$$

where $m_{x,\varepsilon}=Unif(B_\varepsilon(x))$, $d\mu_0=\rho_0\,dx$, and $d\mu_1=\rho_1\,dx$.

Conclusion: PRL for general learning models

There's an advantage to just having integrals rather than sup's, so we don't have to solve an optimization problem.

$$\inf_{f\in\mathcal{F}}\mathbb{E}_{(x,y)\sim\mu}\left[\max\{\ell(f(x),y),\mathrm{CVar}_p(\ell(f(\tilde{x}),y;\tilde{x}\sim m_{x,\varepsilon}\}\right]$$

where the CVar is

$$\mathrm{CVar}_p(\xi,\xi\sim\gamma)=\mathbb{E}_{\xi\sim\gamma}[\xi|\xi\geq\mathrm{Var}_p(\xi;\xi\sim\gamma)]$$

where $\mathrm{Var}_p(\xi;\xi\sim\gamma)=x$ if $\int_{-\infty}^{x}d\gamma=p$. (Not to be confused with variance!)

Then,

$$R(1_A)+J_\Psi(1_A)=\mathbb{E}_{(x,y)\sim}[\max\{\ell(1_A(x),y),\mathrm{CVar})\ell(1_A(\tilde{x},y);\tilde{x}\sim m_{x,\varepsilon}\}$$

where $\Psi=\begin{cases}\frac{t}{p} & \text{if } t\leq p\\ 1 & \text{if } t>p.\end{cases}$

Nice property: If you rescale loss function by a scalar, then solutions won't change because CVar is homogeneous.

Advertisement: AMS workshop in the summer!

Q: How to compute CVar?

A: Solution to $\inf_{\alpha\in\mathbb{R}}\alpha+\frac{\mathbb{E}[(\xi-\alpha)_+]}{p}$.

5.1 Exercises

1. Consider the standard AT problem in the binary classifications setting with $\mathcal{F}=\mathcal{F}_{all}$, the $0-1$ loss function, the Euclidean distance as metric for the attacks, and some $\varepsilon>0$. Give an example of a data distribution $\mu\in\mathcal{P}(\mathcal{X}\times\{0,1\})$ for which there is no solution to the corresponding AT problem that is ε pseudo-certifiable. You may consider simple distributions supported on finitely many points.

   Definition: A set is said to be ε pseudocertifiable if it is both ε inner regular and ε outer regular.

2. Prove that if $\Psi:[0,1]\mapsto[0,\infty)$ is a concave and non-decreasing function, then the functional
$$A\in\mathfrak{B}(\mathcal{X})\mapsto J_\Psi(\mathbb{1}_A)$$
introduced in class is submodular.

3. Prove that the functional

$$u \in L^\infty(\mathcal{X}, \nu) \mapsto J_\Psi(u)$$

introduced in class is sequentially l.s.c. with respect to the weak* convergence in $L^\infty(\mathcal{X}, \nu)$, provided that Ψ is continuous.

4. In the binary classification setting with the $0-1$ loss, we mentioned in class that the functional

$$R(\mathbb{1}_A) + J_{\Psi_p}(\mathbb{1}_A),$$

for Ψ_p the function

$$\Psi_p(t) := \begin{cases} \frac{t}{p} & \text{if } t \leq p \\ 1 & \text{if } t > p, \end{cases}$$

could be written as

$$\mathbb{E}_{(x,y)\sim\mu}[\max\{\ell(\mathbb{1}_A(x), y), \mathrm{CVar}_p(\ell(\mathbb{1}_A(\tilde{x}), y); \tilde{x} \sim m_{x,\varepsilon})\}].$$

Prove this fact.

5. (Open and not well formulated open problem) How could one "learn" a suitable cost function to implement AT in applications? It may be reasonable to assume that you are familiar with a very good adversarial attack in the setting of interest and that thus you can assume is optimal for some AT problem relative to some hidden cost/metric. One can then, perhaps, use an idea similar to the one discussed in this paper about “inverse optimal transport" `https://arxiv.org/pdf/1905.03950.pdf`, as presumably one can exploit the connection between AT and OT that we discussed in class to implement "inverse AT".