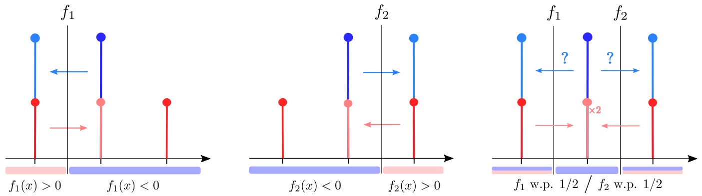
Mixed Nash Equilibria in the Adversarial Examples Game

Laurent Meunier[1,2,*], Meyer Scetbon[3,*], Rafael Pinot[4], Jamal Atif[2], Yann Chevaleyre[2]

[1] Facebook AI Research, Paris, France
[2] Miles Team, LAMSADE, Université Paris-Dauphine, Paris, France
[3] CREST, ENSAE, Paris, France
[4] Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

[*] Equal contribution

A Motivating Example



- **On left and in the middle:** the classifier is deterministic. The adversarial risk is 3/4.
- **On right:** The classifier is randomized. The adversarial risk is 1/2.

The best attack is also randomized: if the attacker takes a deterministic decision, the classifier can play a deterministic strategy to counter him.

There exists a <u>Mixed Nash equilibrium in this game.</u> Can we generalize it?

Setting

- Classification problem on $X \times Y$. P is a Borel probability distribution over $X \times Y$.
- Θ is a set of classifiers.
- A loss $l : \Theta \times (X \times Y) \to \mathbb{R}$ (possibly the 0/1 loss).

<u>Adversarial deterministic risk:</u>

$$R^{\varepsilon}_{adv}(\theta) := \mathbb{E}_{(x,y)\sim P}\left[\sup_{x' \in \mathcal{X},\ d(x,x') \leq \varepsilon} l(\theta, (x', y))\right].$$

<u>Adversarial randomized risk</u>:

$$R^{\varepsilon}_{adv}(\mu) := \mathbb{E}_{(x,y)\sim P}\left[\sup_{x' \in \mathcal{X},\ d(x,x') \leq \varepsilon} \mathbb{E}_{\theta\sim\mu}\left[l(\theta, (x', y))\right]\right].$$

<u>Risk minimization problems:</u>

$$V^{\varepsilon}_{rand} := \inf_{\mu \in \mathcal{M}^1_+(\Theta)} R^{\varepsilon}_{adv}(\mu),\ V^{\varepsilon}_{det} := \inf_{\theta\in\Theta} R^{\varepsilon}_{adv}(\theta)$$

<u>Remark:</u> $V^0_{rand} = V^0_{det} \leq V^{\varepsilon}_{rand} \leq V^{\varepsilon}_{det}$

<u>Adversarial distributions/randomized adversaries:</u>

$$\mathcal{A}_{\varepsilon}(P) := \left\{ Q \mid \exists \gamma, d(x, x') \leq \varepsilon,\ y = y' \quad \gamma\text{-a.s.},\ \Pi_{1\sharp}\gamma = P,\ \Pi_{2\sharp}\gamma = Q \right\}$$

Such an attacker can map any point randomly in the ball of radius ε. It is a Wasserstein ball for a well chosen cost.

<u>**Proposition:**</u> For a given classifier μ, the adversarial randomized risk equals:

$$R^{\varepsilon}_{adv}(\mu) = \sup_{Q \in \mathcal{A}_{\varepsilon}(P)} \mathbb{E}_{(x',y')\sim Q, \theta\sim\mu}\left[l(\theta, (x', y'))\right].$$

The supremum is attained and the optimum might be attained with a deterministic mapping.

Adversarial Examples Game

<u>Primal game:</u> $\inf_{\mu \in \mathcal{M}^1_+(\Theta)} \sup_{Q \in \mathcal{A}_{\varepsilon}(P)} \mathbb{E}_{Q,\mu}\left[l(\theta, (x, y))\right]$. **Classifier goal**: be robust to every attacks.

<u>Dual game:</u> $\sup_{Q \in \mathcal{A}_{\varepsilon}(P)} \inf_{\mu \in \mathcal{M}^1_+(\Theta)} \mathbb{E}_{(x,y)\sim Q, \theta\sim\mu}\left[l(\theta, (x, y))\right]$. **Attacker goal:** find an attack to fool any classifier.

Denoting by D^{ε} the value of the dual formulation, we have:

$$D^{\varepsilon} \leq V^{\varepsilon}_{rand} \leq V^{\varepsilon}_{det}.$$

Is there always equality of the left terms? Does there exist Nash equilibria in this game?

<u>**Theorem:**</u> Strong duality always holds in the randomized setting

$$\inf_{\mu \in \mathcal{M}^{\dagger}_1(\Theta)} \max_{Q \in \mathcal{A}_{\varepsilon}(P)} \mathbb{E}_{\theta\sim\mu, (x,y)\sim Q}\left[l(\theta, (x, y))\right] = \max_{Q \in \mathcal{A}_{\varepsilon}(P)} \inf_{\mu \in \mathcal{M}^{\dagger}_1(\Theta)} \mathbb{E}_{\theta\sim\mu, (x,y)\sim Q}\left[l(\theta, (x, y))\right]$$

<u>Interpretations:</u>

- Always exist approximate Mixed Nash Equilibria in the adversarial examples game.
- If the infimum is attained, there exist Mixed Nash Equilibria.

Entropic Regularization and Algorithms

Given empirical distribution $P_n = \sum_{i=1}^{n} \delta_{(x_i,y_i)}$ and a finite set of classifiers $\{\theta_1, \ldots, \theta_L\}$. **Can we learn the optimal randomized classifier over this finite set of classifiers, i.e. optimize the weights of the probability that θ_i appears?**

<u>Entropic Relaxation:</u>

$$\inf_{\mu \in \mathcal{M}^{\dagger}_1(\Theta)} \sum_{i=1}^{N} \sup_{Q_i \in \Gamma_{i,\varepsilon}} \mathbb{E}_{Q_i,\mu}\left[l(\theta, (x, y))\right] - \alpha_i \mathbf{KL}\left(Q_i \middle|\middle| \frac{1}{N}\mathbb{U}_{(x_i,y_i)}\right) = \inf_{\mu \in \mathcal{M}^{\dagger}_1(\Theta)} \sum_{i=1}^{N} \frac{\alpha_i}{N} \log\left(\int \exp \frac{\mathbb{E}_{\mu}\left[l(\theta, (x, y))\right]}{\alpha_i} d\mathbb{U}_{(x_i,y_i)}\right).$$

- Approximates well the adversarial risk.
- Convex and smooth objective: Algorithm with rate of convergence in $O(T^{-2})$.

<u>Oracle-Based Algorithm:</u>

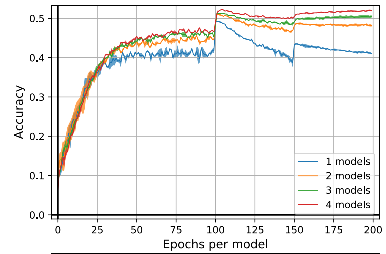
- Inspired by robust optimization and subgradient methods (Danskin Theorem).
- Rate of convergence of order $O(\delta + T^{-1/2})$ where δ denotes the quality of the gradient estimation.

Experiments

Algorithm 2 Adversarial Training for Mixtures

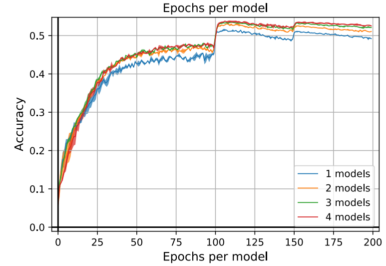
L: number of models, T: number of iterations,
T_θ: number of updates for the models $\boldsymbol{\theta}$,
T_λ: number of updates for the mixture $\boldsymbol{\lambda}$,
$\boldsymbol{\lambda}_0 = (\lambda^1_0, \ldots \lambda^L_0)$, $\boldsymbol{\theta}_0 = (\theta^1_0, \ldots \theta^L_0)$
for $t = 1, \ldots, T$ **do**
  Let B_t be a batch of data.
  if $t \mod (T_\theta L + 1) \neq 0$ **then**
    k sampled uniformly in $\{1, \ldots, L\}$
    $\tilde{B}_t \leftarrow$ Attack of images in B_t for the model $(\boldsymbol{\lambda}_t, \boldsymbol{\theta}_t)$
    $\theta^t_k \leftarrow$ Update θ^{t-1}_k with $\tilde{B}_t$ for fixed $\boldsymbol{\lambda}_t$ with a SGD step
  else
    $\boldsymbol{\lambda}_t \leftarrow$Update $\boldsymbol{\lambda}_{t-1}$ on B_t for fixed $\boldsymbol{\theta}_t$ with oracle-based or regularized algorithm with T_λ iterations.
  end
end

Proposed heuristic algorithm for deep learning



Models	Acc.	APGD_{CE}	APGD_{DLR}	Rob. Acc.
1	81.9%	47.6%	47.7%	45.6%
2	81.9%	49.0%	49.6%	47.0%
3	81.7%	49.0%	49.3%	46.9%
4	**82.6%**	**49.7%**	**49.8%**	**47.2%**

Accuracy under PGD attack on a ResNet18 model for CIFAR10 dataset using Adversarial Training loss



Models	Acc.	APGD_{CE}	APGD_{DLR}	Rob. Acc.
1	79.6%	50.9%	48.9%	48.3%
2	80.3%	52.3%	51.2%	50.2%
3	80.7%	52.8%	51.7%	50.7%
4	**80.9%**	**53.0%**	**51.8%**	**50.8%**

Accuracy under PGD attack on a ResNet18 model for CIFAR10 dataset using TRADES loss

Take a photo to learn more:

FACEBOOK AI   Dauphine | PSL   ENSAE   EPFL