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

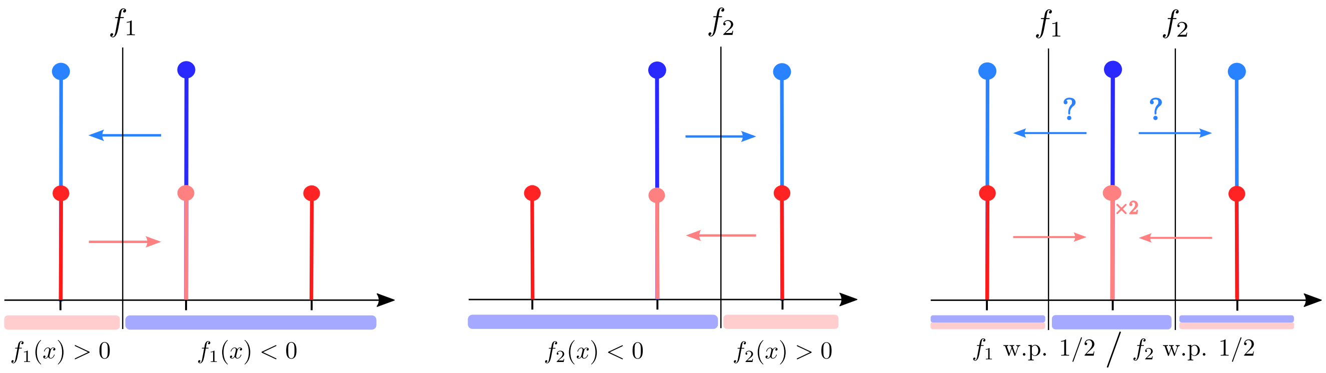
[3] CREST, ENSAE, Paris, France

[4] Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland

[*] Equal contribution

FACEBOOK AI

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 **Mixed Nash Equilibria in this game.** Can we generalize it?

General setting

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).

Adversarial deterministic risk:

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

Adversarial randomized risk :

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

Risk minimization problems:

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

Remark: $V^0_{rand} = V^0_{det} \leq V^{\varepsilon}_{rand} \leq V^{\varepsilon}_{det}$

Adversarial distributions/randomized adversaries:

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

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

Proposition: 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.

FACEBOOK AI | Dauphine Université Paris | PSL | ENSAE | EPFL

Adversarial Examples Game

- **Primal formulation:**

$$\inf_{\mu \in \mathcal{M}^1_+(\Theta)} \sup_{Q \in \mathcal{A}_\varepsilon(P)} \mathbb{E}_{Q,\mu} \left[l(\theta, (x, y)) \right].$$

Classifier objective: being robust to every attacks.

- **Dual Formulation:**

$$\sup_{Q \in \mathcal{A}_\varepsilon(P)} \inf_{\mu \in \mathcal{M}^1_+(\Theta)} \mathbb{E}_{(x,y)\sim Q, \theta \sim \mu} \left[l(\theta, (x, y)) \right].$$

Attacker objective: finding 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?

Theorem: Strong duality always holds in the randomized setting

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

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

Interpretations:

- 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 these class, i.e. optimize the weights of the probability that θ_i appears?**

Entropic Relaxation:

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

$$= \inf_{\mu \in \mathcal{M}_1^+(\Theta)} \sum_{i=1}^{N} \frac{\alpha_i}{N} \log\left(\int \exp \frac{\mathbb{E}_{\mu}\left[l(\theta,(x,y))\right]}{\alpha_i} d\mathbb{U}_{(x_i,y_i)}\right).$$

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

Oracle-Based Algorithm:

- Inspired by robust optimization and subgradient methods (Danskin Theorem)
- Rate of convergence of order $O(\delta + T^{-1/2})$

Algorithm 1 Oracle-based Algorithm

$\boldsymbol{\lambda}_0 = \frac{\mathbf{1}_L}{L}; T;\ \eta = \frac{2}{M\sqrt{LT}}$

for $t = 1, \ldots, T$ **do**

  $\tilde{\mathbb{Q}}$ s.t. $\exists \mathbb{Q}^* \in \mathcal{A}_\varepsilon(\mathbb{P})$ best response to $\boldsymbol{\lambda}_{t-1}$ and for all $k \in [L]$, $|\mathbb{E}_{\tilde{\mathbb{Q}}}(l(\theta_k,(x,y))) - \mathbb{E}_{\mathbb{Q}^*}(l(\theta_k,(x,y)))| \leq \delta$

  $\boldsymbol{g}_t = \left(\mathbb{E}_{\tilde{\mathbb{Q}}}(l(\theta_1,(x,y))), \ldots, \mathbb{E}_{\tilde{\mathbb{Q}}}(l(\theta_L,(x,y))\right)^T$

  $\boldsymbol{\lambda}_t = \Pi_{\Delta_L}\left(\boldsymbol{\lambda}_{t-1} - \eta \boldsymbol{g}_t\right)$

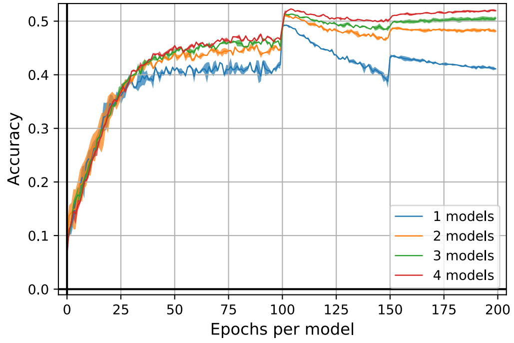
end

Experiments

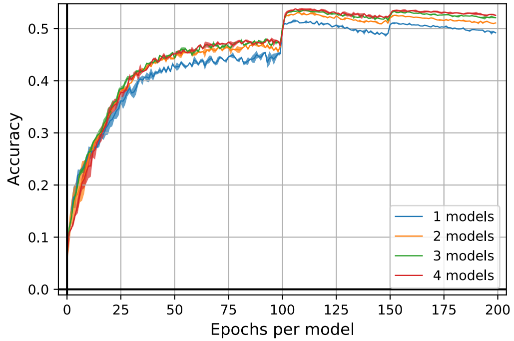
Algorithm 2 Adversarial Training for Mixtures

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 $\boldsymbol{\lambda}$,
$\boldsymbol{\lambda}_0 = (\lambda_0^1, \dots \lambda_0^L)$, $\boldsymbol{\theta}_0 = (\theta_0^1, \dots \theta_0^L)$
for $t = 1, \dots, T$ **do**
  Let B_t be a batch of data.
  if $t \mod (T_\theta L + 1) \neq 0$ **then**
    k sampled uniformly in $\{1, \dots, L\}$
    $\tilde{B}_t \leftarrow$ Attack of images in B_t for the model $(\boldsymbol{\lambda}_t, \boldsymbol{\theta}_t)$
    $\theta_k^t \leftarrow$ Update θ_k^{t-1} 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



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



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