# Mixed Nash Equilibria in the Adversarial Examples Game

Laurent Meunier, Meyer Scetbon, Rafaël Pinot, Jamal Atif, Yann Chevaleyre

# Mixed Nash Equilibria in the Adversarial Examples Game

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

## Abstract

This paper tackles the problem of adversarial examples from a game theoretic point of view. We study the open question of the existence of mixed Nash equilibria in the zero-sum game formed by the attacker and the classifier. While previous works usually allow only one player to use randomized strategies, we show the necessity of considering randomization for both the classifier and the attacker. We demonstrate that this game has no duality gap, meaning that it always admits approximate Nash equilibria. We also provide the first optimization algorithms to learn a mixture of a finite number of classifiers that approximately realizes the value of this game, *i.e.* procedures to build an optimally robust randomized classifier.

## 1. Introduction

Adversarial examples (Biggio et al., 2013; Szegedy et al., 2014) are one of the most dizzling problems in machine learning: state of the art classifiers are sensitive to imperceptible perturbations of their inputs that make them fail. Last years, research have concentrated on proposing new defense methods (Cohen et al.; Madry et al., 2018; Moosavi-Dezfooli et al., 2019) and building more and more sophisticated attacks (Carlini & Wagner, 2017; Croce & Hein, 2020; Goodfellow et al., 2015; Kurakin et al., 2016). So far, most defense strategies proved to be vulnerable to these new attacks or are computationally intractable. This asks the following question: can we build classifiers that are robust against any adversarial attack?

A recent line of research argued that randomized classifiers could help countering adversarial attacks (Dhillon et al., 2018; Pinot et al., 2019; Wang et al., 2019; Xie et al., 2018). Along this line, (Pinot et al., 2020) demonstrated, using game theory, that randomized classifiers are indeed more robust than deterministic ones against regularized adversaries. However, the findings of these previous works depends on the definition of considered adversary. In particular, they did not investigate scenarios where the adversary also uses randomized strategies, which is essential to account for if we want to give a principled answer to the above question. Previous works studying adversarial examples from the scope of game theory investigated the randomized framework (for both the classifier and the adversary) in restricted settings where the adversary is either parametric or has a finite number of strategies (Bose et al., 2021; Perdomo & Singer, 2019; Rota Bulò et al., 2017). Our framework does not assume any constraint on the definition of the adversary, making our conclusions independent on the adversary the classifiers are facing. More precisely, we answer the following questions.

**Q1:** Is it always possible to reach a Mixed Nash equilibrium in the adversarial example game when both the adversary and the classifier can use randomized strategies?

**A1:** We answer positively to this question. First we motivate in Section 2 the necessity for using randomized strategies both with the attacker and the classifier. Then, we extend the work of (Pydi & Jog, 2020), by rigorously reformulating the adversarial risk as a linear optimization problem over distributions. In fact, we cast the adversarial risk minimization problem as a Distributionally Robust Optimization (DRO) (Blanchet & Murthy, 2019) problem for a well suited cost function. This formulation naturally leads us, in Section 3, to analyze adversarial risk minimization as a zero-sum game. We demonstrate that, in this game, the duality gap always equals $0$, meaning that it always admits approximate mixed Nash equilibria.

**Q2:** Can we design efficient algorithms to learn an optimally robust randomized classifier?

**A2:** To answer this question, we focus on learning a finite mixture of classifiers. Taking inspiration from robust optimization (Sinha et al., 2017) and subgradient methods (Boyd, 2003), we derive in Section 4 a first oracle algorithm to optimize a finite mixture. Then, following the line of work of (Cuturi, 2013), we introduce an entropic regularization to effectively compute an approximation of the optimal mixture. We validate our findings with experi-

[*]Equal contribution [1] Miles Team, LAMSADE, Université Paris-Dauphine, Paris, France [2] Facebook AI Research, Paris, France [3] CREST, ENSAE, Paris, France [4]Ecole Polytechnique Fédérale de Lausanne (EPFL), Switzerland. Correspondence to: Laurent Meunier <laurentmeunier@fb.com>, Meyer Scetbon <meyer.scetbon@ensae.fr>, Rafael Pinot <rafael.pinot@epfl.ch>, Jamal Atif <jamal.atif@dauphine.fr>, Yann Chevaleyre <yann.chevaleyre@dauphine.fr>.

*Proceedings of the $38^{th}$ International Conference on Machine Learning*, PMLR 139, 2021. 

ments on simulated and real datasets, namely CIFAR-10 an CIFAR-100 (Krizhevsky & Hinton, 2009).

## 2. The Adversarial Attack Problem

### 2.1. A Motivating Example

Consider the binary classification task illustrated in Figure 1. We assume that all input-output pairs $(X, Y)$ are sampled from a distribution $\mathbb{P}$ defined as follows

$$\mathbb{P}(Y = \pm 1) = 1/2 \text{ and } \begin{cases} \mathbb{P}(X = 0 \mid Y = -1) = 1 \\ \mathbb{P}(X = \pm 1 \mid Y = 1) = 1/2 \end{cases}$$

Given access to $\mathbb{P}$, the adversary aims to maximize the expected risk, but can only move each point by at most 1 on the real line. In this context, we study two classifiers: $f_1(x) = -x - 1/2$ and $f_2(x) = x - 1/2$[1]. Both $f_1$ and $f_2$ have a standard risk of $1/4$. In the presence of an adversary, the risk (*a.k.a.* the adversarial risk) increases to 1. Here, using a randomized classifier can make the system more robust. Consider $f$ where $f = f_1$ w.p. $1/2$ and $f_2$ otherwise. The standard risk of $f$ remains $1/4$ but its adversarial risk is $3/4 < 1$. Indeed, when attacking $f$, any adversary will have to choose between moving points from 0 to 1 or to $-1$. Either way, the attack only works half of the time; hence an overall adversarial risk of $3/4$. Furthermore, if $f$ knows the strategy the adversary uses, it can always update the probability it gives to $f_1$ and $f_2$ to get a better (possibly deterministic) defense. For example, if the adversary chooses to always move 0 to 1, the classifier can set $f = f_1$ w.p. 1 to retrieve an adversarial risk of $1/2$ instead of $3/4$.

Now, what happens if the adversary can use randomized strategies, meaning that for each point it can flip a coin before deciding where to move? In this case, the adversary could decide to move points from 0 to 1 w.p. $1/2$ and to $-1$ otherwise. This strategy is still optimal with an adversarial risk of $3/4$ but now the classifier cannot use its knowledge of the adversary's strategy to lower the risk. We are in a state where neither the adversary nor the classifier can benefit from unilaterally changing its strategy. In the game theory terminology, this state is called a Mixed Nash equilibrium.

### 2.2. General setting

Let us consider a classification task with input space $\mathcal{X}$ and output space $\mathcal{Y}$. Let $(\mathcal{X}, d)$ be a proper (i.e. closed balls are compact) Polish (i.e. completely separable) metric space representing the inputs space[2]. Let $\mathcal{Y} = \{1, \dots, K\}$ be the labels set, endowed with the trivial metric $d'(y, y') = \mathbf{1}_{y \neq y'}$. Then the space $(\mathcal{X} \times \mathcal{Y}, d \oplus d')$ is a proper Polish space. For any Polish space $\mathcal{Z}$, we denote $\mathcal{M}_+^1(\mathcal{Z})$ the Polish space of Borel probability measures on $\mathcal{Z}$. Let us assume the data is drawn from $\mathbb{P} \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})$. Let $(\Theta, d_\Theta)$ be a Polish space (not necessarily proper) representing the set of classifier parameters (for instance neural networks). We also define a loss function: $l : \Theta \times (\mathcal{X} \times \mathcal{Y}) \to [0, \infty)$ satisfying the following set of assumptions.

**Assumption 1** (Loss function)**.** *1) The loss function $l$ is a non negative Borel measurable function. 2) For all $\theta \in \Theta$, $l(\theta, \cdot)$ is upper-semi continuous. 3) There exists $M > 0$ such that for all $\theta \in \Theta$, $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $0 \leq l(\theta, (x, y)) \leq M$.*

It is usual to assume upper-semi continuity when studying optimization over distributions (Blanchet & Murthy, 2019; Villani, 2003). Furthermore, considering bounded (and positive) loss functions is also very common in learning theory (Bartlett & Mendelson, 2002) and is not restrictive.

In the adversarial examples framework, the loss of interest is the $0/1$ loss, for whose surrogates are misunderstood (Bao et al., 2020; Cranko et al., 2019); hence it is essential that the $0/1$ loss satisfies Assumption 1. In the binary classification setting (*i.e.* $\mathcal{Y} = \{-1, +1\}$) the $0/1$ loss writes $l_{0/1}(\theta, (x, y)) = \mathbf{1}_{y f_\theta(x) \leq 0}$. Then, assuming that for all $\theta$, $f_\theta(\cdot)$ is continuous and for all $x$, $f_\cdot(x)$ is continuous, the $0/1$ loss satisfies Assumption 1. In particular, it is the case for neural networks with continuous activation functions.

### 2.3. Adversarial Risk Minimization

The standard risk for a single classifier $\theta$ associated with the loss $l$ satisfying Assumption 1 writes: $\mathcal{R}(\theta) := \mathbb{E}_{(x,y)\sim\mathbb{P}}[l(\theta, (x, y))]$. Similarly, the adversarial risk of $\theta$ at level $\varepsilon$ associated with the loss $l$ is defined as[3]

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

It is clear that $\mathcal{R}_{adv}^{0}(\theta) = \mathcal{R}(\theta)$ for all $\theta$. We can generalize these notions with distributions of classifiers. In other terms the classifier is then randomized according to some distribution $\mu \in \mathcal{M}_+^1(\Theta)$. A classifier is randomized if for a given input, the output of the classifier is a probability distribution. The standard risk of a randomized classifier $\mu$ writes $\mathcal{R}(\mu) = \mathbb{E}_{\theta\sim\mu}[\mathcal{R}(\theta)]$. Similarly, the adversarial risk of the randomized classifier $\mu$ at level $\varepsilon$ is[4]

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

For instance, for the $0/1$ loss, the inner maximization problem, consists in maximizing the probability of misclassification for a given couple $(x, y)$. Note that $\mathcal{R}(\delta_\theta) = \mathcal{R}(\theta)$ and $\mathcal{R}_{adv}^{\varepsilon}(\delta_\theta) = \mathcal{R}_{adv}^{\varepsilon}(\theta)$. In the remainder of the paper,

[1] $(X, Y) \sim \mathbb{P}$ is misclassified by $f_i$ if and only if $f_i(X)Y \leq 0$

[2] For instance, for any norm $\|\cdot\|$, $(\mathbb{R}^d, \|\cdot\|)$ is a proper Polish metric space.

[3] For the well-posedness, see Lemma 4 in Appendix.

[4] This risk is also well posed (see Lemma 4 in the Appendix).

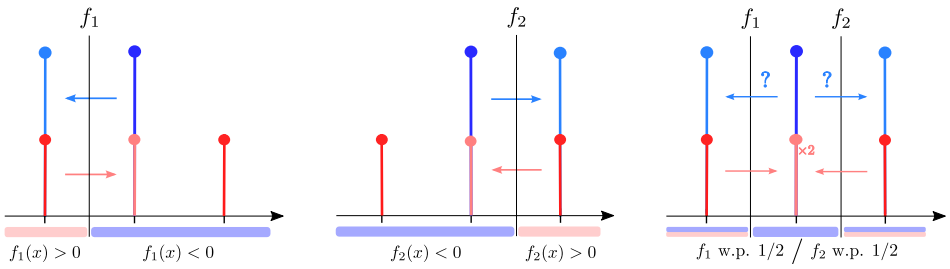



*Figure 1.* Motivating example: blue distribution represents label $-1$ and the red one, label $+1$. The height of columns represents their mass. The red and blue arrows represent the attack on the given classifier. On left: deterministic classifiers ($f_1$ on the left, $f_2$ in the middle) for whose, the blue point can always be attacked. On right: a randomized classifier, where the attacker has a probability $1/2$ of failing, regardless of the attack it selects.

we study the adversarial risk minimization problems with randomized and deterministic classifiers and denote

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

**Remark 1.** *We can show (see Appendix E) that the standard risk infima are equal : $\mathcal{V}^0_{rand} = \mathcal{V}^0_{det}$. Hence, no randomization is needed for minimizing the standard risk. Denoting $\mathcal{V}$ this common value, we also have the following inequalities for any $\varepsilon > 0$, $\mathcal{V} \leq \mathcal{V}^{\varepsilon}_{rand} \leq \mathcal{V}^{\varepsilon}_{det}$.*

### 2.4. Distributional Formulation of the Adversarial Risk

To account for the possible randomness of the adversary, we rewrite the adversarial attack problem as a convex optimization problem over distributions. Let us first introduce the set of adversarial distributions.

**Definition 1** (Set of adversarial distributions)**.** *Let $\mathbb{P}$ be a Borel probability distribution on $\mathcal{X} \times \mathcal{Y}$ and $\varepsilon > 0$. We define the set of adversarial distributions as*

$$\mathcal{A}_{\varepsilon}(\mathbb{P}) := \big\{\mathbb{Q} \in \mathcal{M}^+_1(\mathcal{X} \times \mathcal{Y}) \mid \exists \gamma \in \mathcal{M}^+_1\left((\mathcal{X} \times \mathcal{Y})^2\right), \\ d(x, x') \leq \varepsilon,\ y = y'\ \ \gamma\text{-a.s.},\ \Pi_{1\sharp}\gamma = \mathbb{P},\ \Pi_{2\sharp}\gamma = \mathbb{Q}\big\}$$

*where $\Pi_i$ denotes the projection on the $i$-th component, and $g_\sharp$ the push-forward measure by a measurable function $g$.*

An attacker that can move the initial distribution $\mathbb{P}$ anywhere in $\mathcal{A}_{\varepsilon}(\mathbb{P})$ is not applying a point-wise deterministic perturbation as considered in the standard adversarial risk. In other words, for a point $(x, y) \sim \mathbb{P}$, the attacker could choose a distribution $q(\cdot \mid (x, y))$ whose support is included in $\{(x', y') \mid d(x, x') \leq \epsilon,\ y = y'\}$ from which he will sample the adversarial attack. In this sense, we say the attacker is allowed to be randomized.

**Link with DRO.** Adversarial examples have been studied in the light of DRO by former works (Sinha et al., 2017; Tu et al., 2018), but an exact reformulation of the adversarial risk as a DRO problem has not been made yet. When $(\mathcal{Z}, d)$ is a Polish space and $c : \mathcal{Z}^2 \to \mathbb{R}^+ \cup \{+\infty\}$ is a lower semi-continuous function, for $\mathbb{P}, \mathbb{Q} \in \mathcal{M}^+_1(\mathcal{Z})$ , the primal Optimal Transport problem is defined as

$$W_c(\mathbb{P}, \mathbb{Q}) := \inf_{\gamma \in \Gamma_{\mathbb{P},\mathbb{Q}}} \int_{\mathcal{Z}^2} c(z, z') d\gamma(z, z')$$

with $\Gamma_{\mathbb{P},\mathbb{Q}} := \big\{\gamma \in \mathcal{M}^+_1(\mathcal{Z}^2) \mid \Pi_{1\sharp}\gamma = \mathbb{P},\ \Pi_{2\sharp}\gamma = \mathbb{Q}\big\}$. When $\eta > 0$ and for $\mathbb{P} \in \mathcal{M}^+_1(\mathcal{Z})$, the associated Wasserstein uncertainty set is defined as:

$$\mathcal{B}_c(\mathbb{P}, \eta) := \big\{\mathbb{Q} \in \mathcal{M}^+_1(\mathcal{Z}) \mid W_c(\mathbb{P}, \mathbb{Q}) \leq \eta\big\}$$

A DRO problem is a linear optimization problem over Wasserstein uncertainty sets $\sup_{\mathbb{Q} \in \mathcal{B}_c(\mathbb{P},\eta)} \int g(z) d\mathbb{Q}(z)$ for some upper semi-continuous function $g$ (Yue et al., 2020). For an arbitrary $\varepsilon > 0$, we define the cost $c_\varepsilon$ as follows

$$c_\varepsilon((x, y), (x', y')) := \begin{cases} 0 & \text{if } d(x, x') \leq \varepsilon \text{ and } y = y' \\ +\infty & \text{otherwise.} \end{cases}$$

This cost is lower semi-continuous and penalizes to infinity perturbations that change the label or move the input by a distance greater than $\varepsilon$. As Proposition 1 shows, the Wasserstein ball associated with $c_\varepsilon$ is equal to $\mathcal{A}_{\varepsilon}(\mathbb{P})$.

**Proposition 1.** *Let $\mathbb{P}$ be a Borel probability distribution on $\mathcal{X} \times \mathcal{Y}$ and $\varepsilon > 0$ and $\eta \geq 0$, then $\mathcal{B}_{c_\varepsilon}(\mathbb{P}, \eta) = \mathcal{A}_{\varepsilon}(\mathbb{P})$. Moreover, $\mathcal{A}_{\varepsilon}(\mathbb{P})$ is convex and compact for the weak topology of $\mathcal{M}^+_1(\mathcal{X} \times \mathcal{Y})$.*

Thanks to this result, we can reformulate the adversarial risk as the value of a convex problem over $\mathcal{A}_{\varepsilon}(\mathbb{P})$.

**Proposition 2.** *Let $\mathbb{P}$ be a Borel probability distribution on $\mathcal{X} \times \mathcal{Y}$ and $\mu$ a Borel probability distribution on $\Theta$. Let $l : \Theta \times (\mathcal{X} \times \mathcal{Y}) \to [0, \infty)$ satisfying Assumption 1. Let $\varepsilon > 0$. Then:*

$$\mathcal{R}^{\varepsilon}_{adv}(\mu) = \sup_{\mathbb{Q} \in \mathcal{A}_{\varepsilon}(\mathbb{P})} \mathbb{E}_{(x', y') \sim \mathbb{Q}, \theta \sim \mu}\left[l(\theta, (x', y'))\right]. \quad (2)$$

*The supremum is attained. Moreover $\mathbb{Q}^* \in \mathcal{A}_{\varepsilon}(\mathbb{P})$ is an optimum of Problem (2) if and only if there exists $\gamma^* \in \mathcal{M}^+_1\left((\mathcal{X} \times \mathcal{Y})^2\right)$ such that: $\Pi_{1\sharp}\gamma^* = \mathbb{P}$, $\Pi_{2\sharp}\gamma^* = \mathbb{Q}^*$, $d(x, x') \leq \varepsilon$, $y = y'$ and $l(x', y') = \sup_{u \in \mathcal{X}, d(x,u) \leq \varepsilon} l(u, y)$ $\gamma^*$-almost surely.*

The adversarial attack problem is a DRO problem for the cost $c_\varepsilon$. Proposition 2 means that, against a fixed classifier $\mu$, the randomized attacker that can move the distribution in $\mathcal{A}_\varepsilon(\mathbb{P})$ has exactly the same power as an attacker that moves every single point $x$ in the ball of radius $\varepsilon$. By Proposition 2, we also deduce that the adversarial risk can be casted as a linear optimization problem over distributions.

**Remark 2.** *In a recent work, (Pydi & Jog, 2020) proposed a similar adversary using Markov kernels but left as an open question the link with the classical adversarial risk, due to measurability issues. Proposition 2 solves these issues. The result is similar to (Blanchet & Murthy, 2019). Although we believe its proof might be extended for infinite valued costs, (Blanchet & Murthy, 2019) did not treat that case. We provide an alternative proof in this special case.*

## 3. Nash Equilibria in the Adversarial Game

### 3.1. Adversarial Attacks as a Zero-Sum Game

Thanks to Proposition 2, the adversarial risk minimization problem can be seen as a two-player zero-sum game that writes as follows,

$$\inf_{\mu\in\mathcal{M}_+^1(\Theta)} \sup_{\mathbb{Q}\in\mathcal{A}_\varepsilon(\mathbb{P})} \mathbb{E}_{(x,y)\sim\mathbb{Q},\theta\sim\mu}\left[l(\theta,(x,y))\right]. \tag{3}$$

In this game the classifier objective is to find the best distribution $\mu\in\mathcal{M}_1^+(\Theta)$ while the adversary is manipulating the data distribution. For the classifier, solving the infimum problem in Equation (3) simply amounts to solving the adversarial risk minimization problem – Problem (1), whether the classifier is randomized or not. Then, given a randomized classifier $\mu\in\mathcal{M}_1^+(\Theta)$, the goal of the attacker is to find a new data-set distribution $\mathbb{Q}$ in the set of adversarial distributions $\mathcal{A}_\varepsilon(\mathbb{P})$ that maximizes the risk of $\mu$. More formally, the adversary looks for

$$\mathbb{Q}\in\operatorname*{argmax}_{\mathbb{Q}\in\mathcal{A}_\varepsilon(\mathbb{P})} \mathbb{E}_{(x,y)\sim\mathbb{Q},\theta\sim\mu}\left[l(\theta,(x,y))\right].$$

In the game theoretic terminology, $\mathbb{Q}$ is also called the best response of the attacker to the classifier $\mu$.

**Remark 3.** *Note that for a given classifier $\mu$ there always exists a "deterministic" best response, i.e. every single point $(x,y)$ is mapped to another single point $T(x,y)$. Let $T:\mathcal{X}\times\mathcal{Y}\to\mathcal{X}$ be defined such that for all $(x,y)\in\mathcal{X}\times\mathcal{Y}$, $l(T(x,y),y)=\sup_{x',\,d(x,x')\leq\varepsilon} l(x',y)$. Thanks to (Bertsekas & Shreve, 2004, Proposition 7.50), $(T,id)$ is $\mathbb{P}$-measurable. Moreover, we get that $\mathbb{Q}=(T,id)_\sharp\mathbb{P}$ belongs to the best response to $\mu$. Therefore, $T$ is the optimal "deterministic" attack against the classifier $\mu$.*

### 3.2. Dual Formulation of the Game

Every zero sum game has a dual formulation that allows a deeper understanding of the framework. Here, from Proposition 2, we can define the dual problem of adversarial risk minimization for randomized classifiers. This dual problem also characterizes a two-player zero-sum game that writes as follows,

$$\sup_{\mathbb{Q}\in\mathcal{A}_\varepsilon(\mathbb{P})} \inf_{\mu\in\mathcal{M}_+^1(\Theta)} \mathbb{E}_{(x,y)\sim\mathbb{Q},\theta\sim\mu}\left[l(\theta,(x,y))\right]. \tag{4}$$

In this dual game problem, the adversary plays first and seeks an adversarial distribution that has the highest possible risk when faced with an arbitrary classifier. This means that it has to select an adversarial perturbation for every input $x$, without seeing the classifier first. In this case, as pointed out by the motivating example in Section 2.1, the attack can (and should) be randomized to ensure maximal harm against several classifiers. Then, given an adversarial distribution, the classifier objective is to find the best possible classifier on this distribution. Let us denote $\mathcal{D}^\varepsilon$ the value of the dual problem. Since the weak duality is always satisfied, we get

$$\mathcal{D}^\varepsilon \leq \mathcal{V}^\varepsilon_{rand} \leq \mathcal{V}^\varepsilon_{det}. \tag{5}$$

Inequalities in Equation (5) mean that the lowest risk the classifier can get (regardless of the game we look at) is $\mathcal{D}^\varepsilon$. In particular, this means that the primal version of the game, *i.e.* the adversarial risk minimization problem, will always have a value greater or equal to $\mathcal{D}^\varepsilon$. As we discussed in Section 2.1, this lower bound may not be attained by a deterministic classifier. As we will demonstrate in the next section, optimizing over randomized classifiers allows to approach $\mathcal{D}^\varepsilon$ arbitrary closely.

**Remark 4.** *Note that, we can always define the dual problem when the classifier is deterministic,*

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

*Furthermore, we can demonstrate that the dual problems for deterministic and randomized classifiers have the same value [5]; hence the inequalities in Equation (5).*

### 3.3. Nash Equilibria for Randomized Strategies

In the adversarial examples game, a Nash equilibrium is a couple $(\mu^*,\mathbb{Q}^*)\in\mathcal{M}_+^1(\Theta)\times\mathcal{A}_\varepsilon(\mathbb{P})$ where both the classifier and the attacker have no incentive to deviate unilaterally from their strategies $\mu^*$ and $\mathbb{Q}^*$. More formally, $(\mu^*,\mathbb{Q}^*)$ is a Nash equilibrium of the adversarial examples game if $(\mu^*,\mathbb{Q}^*)$ is a saddle point of the objective function

$$(\mu,\mathbb{Q})\mapsto \mathbb{E}_{(x,y)\sim\mathbb{Q},\theta\sim\mu}\left[l(\theta,(x,y))\right].$$

Alternatively, we can say that $(\mu^*,\mathbb{Q}^*)$ is a Nash equilibrium if and only if $\mu^*$ solves the adversarial risk minimization problem – Problem (1), $\mathbb{Q}^*$ the dual problem – Problem (6),

[5] See Appendix E for more details

and $\mathcal{D}^\varepsilon = \mathcal{V}^\varepsilon_{rand}$. In our problem, $\mathbb{Q}^*$ always exists but it might not be the case for $\mu^*$. Then for any $\delta > 0$, we say that $(\mu_\delta, \mathbb{Q}^*)$ is a $\delta$-approximate Nash equilibrium if $\mathbb{Q}^*$ solves the dual problem and $\mu_\delta$ satisfies $\mathcal{D}^\varepsilon \geq \mathcal{R}^\varepsilon_{adv}(\mu_\delta) - \delta$.

We now state our main result: the existence of approximate Nash equilibria in the adversarial examples game when both the classifier and the adversary can use randomized strategies. More precisely, we demonstrate that the duality gap between the adversary and the classifier problems is zero, which gives as a corollary the existence of Nash equilibria.

**Theorem 1.** *Let $\mathbb{P} \in \mathcal{M}^1_+(\mathcal{X} \times \mathcal{Y})$. Let $\varepsilon > 0$. Let $l : \Theta \times (\mathcal{X} \times \mathcal{Y}) \to [0, \infty)$ satisfying Assumption 1. Then strong duality always holds in the randomized setting:*

$$\inf_{\mu \in \mathcal{M}^+_1(\Theta)} \max_{\mathbb{Q} \in \mathcal{A}_\varepsilon(\mathbb{P})} \mathbb{E}_{\theta \sim \mu, (x,y) \sim \mathbb{Q}} \left[ l(\theta, (x, y)) \right] \quad (6)$$
$$= \max_{\mathbb{Q} \in \mathcal{A}_\varepsilon(\mathbb{P})} \inf_{\mu \in \mathcal{M}^+_1(\Theta)} \mathbb{E}_{\theta \sim \mu, (x,y) \sim \mathbb{Q}} \left[ l(\theta, (x, y)) \right]$$

*The supremum is always attained. If $\Theta$ is a compact set, and for all $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $l(\cdot, (x, y))$ is lower semi-continuous, the infimum is also attained.*

**Corollary 1.** *Under Assumption 1, for any $\delta > 0$, there exists a $\delta$-approximate Nash-Equibilrium $(\mu_\delta, \mathbb{Q}^*)$. Moreover, if the infimum is attained, there exists a Nash equilibrium $(\mu^*, \mathbb{Q}^*)$ to the adversarial examples game.*

Theorem 1 shows that $\mathcal{D}^\varepsilon = \mathcal{V}^\varepsilon_{rand}$. From a game theoretic perspective, this means that the minimal adversarial risk for a randomized classifier against any attack (primal problem) is the same as the maximal risk an adversary can get by using an attack strategy that is oblivious to the classifier it faces (dual problem). This suggests that playing randomized strategies for the classifier could substantially improve robustness to adversarial examples. In the next section, we will design an algorithm that efficiently learn a randomized classifier and show improved adversarial robustness over classical deterministic defenses.

**Remark 5.** *Theorem 1 remains true if one replaces $\mathcal{A}_\varepsilon(\mathbb{P})$ with any other Wasserstein compact uncertainty sets (see (Yue et al., 2020) for conditions of compactness).*

## 4. Finding the Optimal Classifiers

### 4.1. An Entropic Regularization

Let $\{(x_i, y_i)\}_{i=1}^N$ samples independently drawn from $\mathbb{P}$ and denote $\widehat{\mathbb{P}} := \frac{1}{N} \sum_{i=1}^N \delta_{(x_i, y_i)}$ the associated empirical distribution. One can show the adversarial empirical risk minimization can be casted as:

$$\widehat{\mathcal{R}}^{\varepsilon,*}_{adv} := \inf_{\mu \in \mathcal{M}^+_1(\Theta)} \sum_{i=1}^N \sup_{\mathbb{Q}_i \in \Gamma_{i,\varepsilon}} \mathbb{E}_{(x,y) \sim \mathbb{Q}_i, \theta \sim \mu} \left[ l(\theta, (x, y)) \right]$$

where $\Gamma_{i,\varepsilon}$ is defined as :

$$\Gamma_{i,\varepsilon} := \Big\{ \mathbb{Q}_i \mid \int d\mathbb{Q}_i = \frac{1}{N}, \ \int c_\varepsilon((x_i, y_i), \cdot) d\mathbb{Q}_i = 0 \Big\}.$$

More details on this decomposition are given in Appendix E. In the following, we regularize the above objective by adding an entropic term to each inner supremum problem. Let $\boldsymbol{\alpha} := (\alpha_i)_{i=1}^N \in \mathbb{R}^N_+$ such that for all $i \in \{1, \dots, N\}$, and let us consider the following optimization problem:

$$\widehat{\mathcal{R}}^{\varepsilon,*}_{adv,\boldsymbol{\alpha}} := \inf_{\mu \in \mathcal{M}^+_1(\Theta)} \sum_{i=1}^N \sup_{\mathbb{Q}_i \in \Gamma_{i,\varepsilon}} \mathbb{E}_{\mathbb{Q}_i, \mu} \left[ l(\theta, (x, y)) \right] - \alpha_i \mathrm{KL} \left( \mathbb{Q}_i \Big|\Big| \frac{1}{N} \mathbb{U}_{(x_i, y_i)} \right)$$

where $\mathbb{U}_{(x,y)}$ is an arbitrary distribution of support equal to:

$$S^{(\varepsilon)}_{(x,y)} := \Big\{ (x', y') : \text{ s.t. } c_\varepsilon((x, y), (x', y')) = 0 \Big\},$$

and for all $\mathbb{Q}, \mathbb{U} \in \mathcal{M}_+(\mathcal{X} \times \mathcal{Y})$,

$$\mathrm{KL}(\mathbb{Q} || \mathbb{U}) := \begin{cases} \int \log(\frac{d\mathbb{Q}}{d\mathbb{U}}) d\mathbb{Q} + |\mathbb{U}| - |\mathbb{Q}| & \text{if } \mathbb{Q} \ll \mathbb{U} \\ +\infty & \text{otherwise.} \end{cases}$$

Note that when $\boldsymbol{\alpha} = 0$, we recover the problem of interest $\widehat{\mathcal{R}}^{\varepsilon,*}_{adv,\boldsymbol{\alpha}} = \widehat{\mathcal{R}}^{\varepsilon,*}_{adv}$. Moreover, we show the regularized supremum tends to the standard supremum when $\boldsymbol{\alpha} \to 0$.

**Proposition 3.** *For $\mu \in \mathcal{M}^+_1(\Theta)$, one has*

$$\lim_{\alpha_i \to 0} \sup_{\mathbb{Q}_i \in \Gamma_{i,\varepsilon}} \mathbb{E}_{\mathbb{Q}_i, \mu} \left[ l(\theta, (x, y)) \right] - \alpha_i KL \left( \mathbb{Q} \Big|\Big| \frac{1}{N} \mathbb{U}_{(x_i, y_i)} \right)$$
$$= \sup_{\mathbb{Q}_i \in \Gamma_{i,\varepsilon}} \mathbb{E}_{(x,y) \sim \mathbb{Q}_i, \theta \sim \mu} \left[ l(\theta, (x, y)) \right].$$

By adding an entropic term to the objective, we obtain an explicit formulation of the supremum involved in the sum: as soon as $\boldsymbol{\alpha} > 0$ (which means that each $\alpha_i > 0$), each sub-problem becomes just the Fenchel-Legendre transform of $\mathrm{KL}(\cdot | \mathbb{U}_{(x_i, y_i)}/N)$ which has the following closed form:

$$\sup_{\mathbb{Q}_i \in \Gamma_{i,\varepsilon}} \mathbb{E}_{\mathbb{Q}_i, \mu} \left[ l(\theta, (x, y)) \right] - \alpha_i \mathrm{KL} \left( \mathbb{Q}_i || \frac{1}{N} \mathbb{U}_{(x_i, y_i)} \right)$$
$$= \frac{\alpha_i}{N} \log \left( \int_{\mathcal{X} \times \mathcal{Y}} \exp \left( \frac{\mathbb{E}_{\theta \sim \mu} \left[ l(\theta, (x, y)) \right]}{\alpha_i} \right) d\mathbb{U}_{(x_i, y_i)} \right).$$

Finally, we end up with the following problem:

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

In order to solve the above problem, one needs to compute the integral involved in the objective. To do so, we estimate it by randomly sampling $m_i \geq 1$ samples $(u^{(i)}_1, \dots, u^{(i)}_{m_i}) \in$

$(\mathcal{X}\times\mathcal{Y})^{m_i}$ from $\mathbb{U}_{(x_i,y_i)}$ for all $i\in\{1,\dots,N\}$ which leads to the following optimization problem

$$\inf_{\mu\in\mathcal{M}_1^+(\Theta)}\sum_{i=1}^{N}\frac{\alpha_i}{N}\log\left(\frac{1}{m_i}\sum_{j=1}^{m_i}\exp\frac{\mathbb{E}_\mu\left[l(\theta,u_j^{(i)})\right]}{\alpha_i}\right) \quad (7)$$

denoted $\widehat{\mathcal{R}}^{\varepsilon,\boldsymbol{m}}_{adv,\boldsymbol{\alpha}}$ where $\boldsymbol{m}:=(m_i)_{i=1}^N$ in the following. Now we aim at controlling the error made with our approximations. We decompose the error into two terms

$$|\widehat{\mathcal{R}}^{\varepsilon,\boldsymbol{m}}_{adv,\boldsymbol{\alpha}}-\widehat{\mathcal{R}}^{\varepsilon,*}_{adv}|\leq|\widehat{\mathcal{R}}^{\varepsilon,*}_{adv,\boldsymbol{\alpha}}-\widehat{\mathcal{R}}^{\varepsilon,\boldsymbol{m}}_{adv,\boldsymbol{\alpha}}|+|\widehat{\mathcal{R}}^{\varepsilon,*}_{adv,\boldsymbol{\alpha}}-\widehat{\mathcal{R}}^{\varepsilon,*}_{adv}|$$

where the first one corresponds to the statistical error made by our estimation of the integral, and the second to the approximation error made by the entropic regularization of the objective. First, we show a control of the statistical error using Rademacher complexities (Bartlett & Mendelson, 2002).

**Proposition 4.** *Let $m\geq 1$ and $\alpha>0$ and denote $\boldsymbol{\alpha}:=(\alpha,\dots,\alpha)\in\mathbb{R}^N$ and $\boldsymbol{m}:=(m,\dots,m)\in\mathbb{R}^N$. Then by denoting $\tilde{M}=\max(M,1)$, we have with a probability of at least $1-\delta$*

$$|\widehat{\mathcal{R}}^{\varepsilon,*}_{adv,\boldsymbol{\alpha}}-\widehat{\mathcal{R}}^{\varepsilon,\boldsymbol{m}}_{adv,\boldsymbol{\alpha}}|\leq\frac{2e^{M/\alpha}}{N}\sum_{i=1}^{N}R_i+6\tilde{M}e^{M/\alpha}\sqrt{\frac{\log(\frac{4}{\delta})}{2mN}}$$

*where $R_i:=\frac{1}{m}\mathbb{E}_{\boldsymbol{\sigma}}\left[\sup_{\theta\in\Theta}\sum_{j=1}^m\sigma_j l(\theta,u_j^{(i)})\right]$ and $\boldsymbol{\sigma}:=(\sigma_1,\dots,\sigma_m)$ with $\sigma_i$ i.i.d. sampled as $\mathbb{P}[\sigma_i=\pm1]=1/2$.*

We deduce from the above Proposition that in the particular case where $\Theta$ is finite such that $|\Theta|=L$, with probability of at least $1-\delta$

$$|\widehat{\mathcal{R}}^{\varepsilon,*}_{adv,\boldsymbol{\alpha}}-\widehat{\mathcal{R}}^{\varepsilon,\boldsymbol{m}}_{adv,\boldsymbol{\alpha}}|\in\mathcal{O}\left(Me^{M/\alpha}\sqrt{\frac{\log(L)}{m}}\right).$$

This case is of particular interest when one wants to learn the optimal mixture of some given classifiers in order to minimize the adversarial risk. In the following proposition, we control the approximation error made by adding an entropic term to the objective.

**Proposition 5.** *Denote for $\beta>0$, $(x,y)\in\mathcal{X}\times\mathcal{Y}$ and $\mu\in\mathcal{M}_1^+(\Theta)$, $A^{(x,y)}_{\beta,\mu}:=\{u|\sup_{v\in S^{(\varepsilon)}_{(x,y)}}\mathbb{E}_\mu[l(\theta,v)]\leq\mathbb{E}_\mu[l(\theta,u)]+\beta\}$. If there exists $C_\beta$ such that for all $(x,y)\in\mathcal{X}\times\mathcal{Y}$ and $\mu\in\mathcal{M}_1^+(\Theta)$, $\mathbb{U}_{(x,y)}\left(A^{(x,y)}_{\beta,\mu}\right)\geq C_\beta$ then we have*

$$|\widehat{\mathcal{R}}^{\varepsilon,*}_{adv,\boldsymbol{\alpha}}-\widehat{\mathcal{R}}^{\varepsilon,*}_{adv}|\leq2\alpha|\log(C_\beta)|+\beta.$$

The assumption made in the above Proposition states that for any given random classifier $\mu$, and any given point $(x,y)$, the set of $\beta$-optimal attacks at this point has at least a certain amount of mass depending on the $\beta$ chosen. This assumption is always met when $\beta$ is sufficiently large. However in order to obtain a tight control of the error, a trade-off exists between $\beta$ and the smallest amount of mass $C_\beta$ of $\beta$-optimal attacks.

Now that we have shown that solving (7) allows to obtain an approximation of the true solution $\widehat{\mathcal{R}}^{\varepsilon,*}_{adv}$, we next aim at deriving an algorithm to compute it.

### 4.2. Proposed Algorithms

From now on, we focus on finite class of classifiers. Let $\Theta=\{\theta_1,\dots,\theta_L\}$, we aim to learn the optimal mixture of classifiers in this case. The adversarial empirical risk is therefore defined as:

$$\widehat{\mathcal{R}}^{\varepsilon}_{adv}(\boldsymbol{\lambda})=\sum_{i=1}^{N}\sup_{\mathbb{Q}_i\in\Gamma_{i,\varepsilon}}\mathbb{E}_{(x,y)\sim\mathbb{Q}_i}\left[\sum_{k=1}^{L}\lambda_k l(\theta_k,(x,y))\right]$$

for $\boldsymbol{\lambda}\in\Delta_L:=\{\boldsymbol{\lambda}\in\mathbb{R}_+^L\text{ s.t. }\sum_{i=1}^L\lambda_i=1\}$, the probability simplex of $\mathbb{R}^L$. One can notice that $\widehat{\mathcal{R}}^{\varepsilon}_{adv}(\cdot)$ is a continuous convex function, hence $\min_{\boldsymbol{\lambda}\in\Delta_L}\mathcal{R}^{\varepsilon}_{adv}(\boldsymbol{\lambda})$ is attained for a certain $\boldsymbol{\lambda}^*$. Then there exists a non-approximate Nash equilibrium $(\boldsymbol{\lambda}^*,\mathbb{Q}^*)$ in the adversarial game when $\Theta$ is finite. Here, we present two algorithms to learn the optimal mixture of the adversarial risk minimization problem.

**Algorithm 1** Oracle-based Algorithm

$\boldsymbol{\lambda}_0=\frac{\mathbf{1}_L}{L};T;\ \eta=\frac{2}{M\sqrt{LT}}$
**for** $t=1,\dots,T$ **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 $k\in[L]$, $|\mathbb{E}_{\tilde{\mathbb{Q}}}(l(\theta_k,(x,y)))-\mathbb{E}_{\mathbb{Q}^*}(l(\theta_k,(x,y)))|\leq\delta$
  $\boldsymbol{g}_t=\left(\mathbb{E}_{\tilde{\mathbb{Q}}}(l(\theta_1,(x,y)),\dots,\mathbb{E}_{\tilde{\mathbb{Q}}}(l(\theta_L,(x,y))\right)^T$
  $\boldsymbol{\lambda}_t=\Pi_{\Delta_L}\left(\boldsymbol{\lambda}_{t-1}-\eta\boldsymbol{g}_t\right)$
**end**

**An Entropic Relaxation.** Using the results from Section 4.1, adding an entropic term to the objective allows to have a simple reformulation of the problem, as follows:

$$\inf_{\boldsymbol{\lambda}\in\Delta_L}\sum_{i=1}^{N}\frac{\varepsilon_i}{N}\log\left(\frac{1}{m_i}\sum_{j=1}^{m_i}\exp\left(\frac{\sum_{k=1}^L\lambda_k l(\theta_k,u_j^{(i)})}{\varepsilon_i}\right)\right)$$

Note that in $\boldsymbol{\lambda}$, the objective is convex and smooth. One can apply the accelerated PGD (Beck & Teboulle, 2009; Tseng, 2008) which enjoys an optimal convergence rate for first order methods of $\mathcal{O}(T^{-2})$ for $T$ iterations.

**A First Oracle Algorithm.** Indepedently from the entropic regularization,we present an oracle-based algorithm inspired from (Sinha et al., 2017) and the convergence of projected sub-gradient methods (Boyd, 2003). The computation of

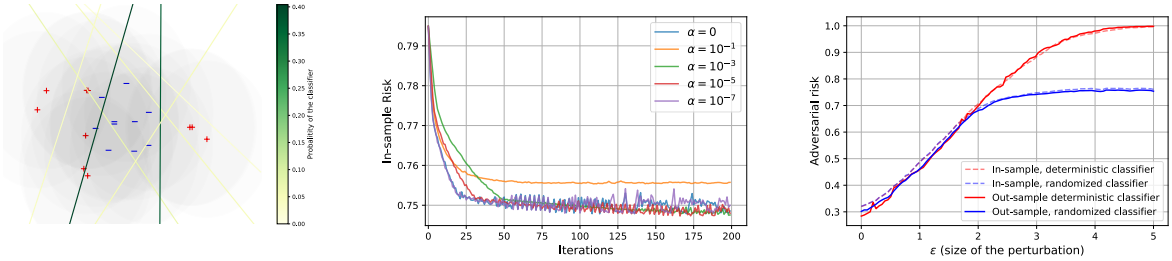



*Figure 2.* On left, 40 data samples with their set of possible attacks represented in shadow and the optimal randomized classifier, with a color gradient representing the probability of the classifier. In the middle, convergence of the oracle ($\alpha = 0$) and regularized algorithm for different values of regularization parameters. On right, in-sample and out-sample risk for randomized and deterministic minimum risk in function of the perturbation size $\varepsilon$. In the latter case, the randomized classifier is optimized with oracle Algorithm 1.

the inner supremum problem is usually NP-hard [6], but one may assume the existence of an approximate oracle to this supremum. The algorithm is presented in Algorithm 1. We get the following guarantee for this algorithm.

**Proposition 6.** *Let $l : \Theta \times (\mathcal{X} \times \mathcal{Y}) \to [0, \infty)$ satisfying Assumption 1. Then, Algorithm 1 satisfies:*

$$\min_{t \in [T]} \widehat{\mathcal{R}}^{\varepsilon}_{adv}(\boldsymbol{\lambda}_t) - \widehat{\mathcal{R}}^{\varepsilon,*}_{adv} \leq 2\delta + \frac{2M\sqrt{L}}{\sqrt{T}}$$

The main drawback of the above algorithm is that one needs to have access to an oracle to guarantee the convergence of the proposed algorithm. In the following we present its regularized version in order to approximate the solution and propose a simple algorithm to solve it.

### 4.3. A General Heuristic Algorithm

So far, our algorithms are not easily practicable in the case of deep learning. Adversarial examples are known to be easily transferrable from one model to another (Papernot et al., 2016; Tramèr et al., 2017). So we aim at learning diverse models. To this end, and support our theoretical claims, we propose an heuristic algorithm (see Algorithm 2) to train a robust mixture of $L$ classifiers. We alternatively train these classifiers with adversarial examples against the current mixture and update the probabilities of the mixture according to the algorithms we proposed in Section 4.2. More details on this algorithm are available in Appendix D.

**Algorithm 2** Adversarial Training for Mixtures

$L$: number of models, $T$: number of iterations,
$T_\theta$: number of updates for the models $\boldsymbol{\theta}$,
$T_\lambda$: 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 $\theta_k^{t-1}$ 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 $T_\lambda$ iterations.
  **end**
**end**

## 5. Experiments

### 5.1. Synthetic Dataset

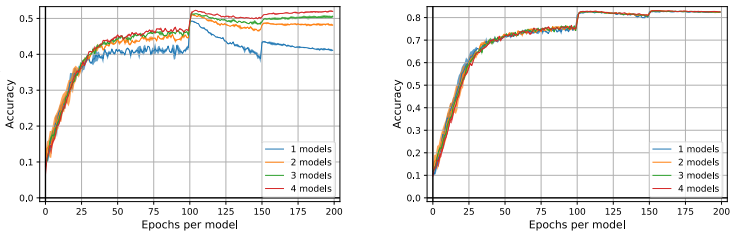
To illustrate our theoretical findings, we start by testing our learning algorithm on the following synthetic two-dimensional problem. Let us consider the distribution $\mathbb{P}$ defined as $\mathbb{P}(Y = \pm 1) = 1/2$, $\mathbb{P}(X \mid Y = -1) = \mathcal{N}(0, I_2)$ and $\mathbb{P}(X \mid Y = 1) = \frac{1}{2}[\mathcal{N}((-3, 0), I_2) + \mathcal{N}((3, 0), I_2)]$. We sample 1000 training points from this distribution and randomly generate 10 linear classifiers that achieves a standard training risk lower than 0.4. To simulate an adversary with budget $\varepsilon$ in $\ell_2$ norm, we proceed as follows. For every sample $(x, y) \sim \mathbb{P}$ we generate 1000 points uniformly at random in the ball of radius $\varepsilon$ and select the one maximizing the risk for the $0/1$ loss. Figure 2 (left) illustrates the type of mixture we get after convergence of our algorithms. Note that in this toy problem, we are likely to find the optimal adversary with this sampling strategy if we sample enough attack points.

To evaluate the convergence of our algorithms, we compute the adversarial risk of our mixture for each iteration of both the oracle and regularized algorithms. Figure 2 illustrates the convergence of the algorithms w.r.t the regularization parameter. We observe that the risk for both algorithms converge. Moreover, they converge towards the oracle minimizer when the regularization parameter $\alpha$ goes to 0.

Finally, to demonstrate the improvement randomized techniques offer against deterministic defenses, we plot in Figure 2 (right) the minimum adversarial risk for both random-

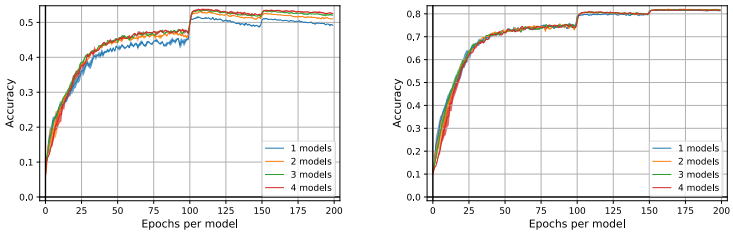
[6] See Appendix E for details.

| Models | Acc. | $\text{APGD}_{\text{CE}}$ | $\text{APGD}_{\text{DLR}}$ | 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%** |



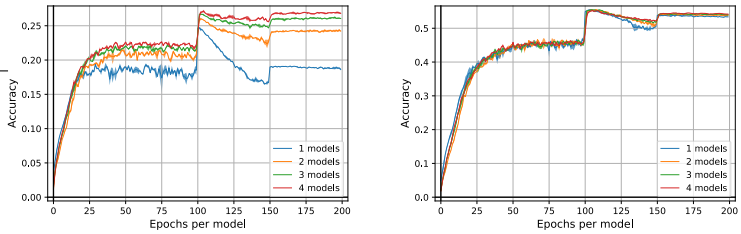

Adversarial Training, CIFAR-10 dataset results

| Models | Acc. | $\text{APGD}_{\text{CE}}$ | $\text{APGD}_{\text{DLR}}$ | 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%** |




TRADES, CIFAR-10 dataset results

| Models | Acc. | $\text{APGD}_{\text{CE}}$ | $\text{APGD}_{\text{DLR}}$ | Rob. Acc. |
|---|---|---|---|---|
| 1 | 55.2% | 24.1% | 23.8% | 22.5% |
| 2 | 55.2% | 25.3% | 26.1% | 23.6% |
| 3 | **55.4%** | 25.7% | 26.8% | 24.2% |
| 4 | 55.3% | **26.0%** | **27.5%** | **24.5%** |




Adversarial Training, CIFAR-100 dataset results

*Figure 3.* Upper plots: Adversarial Training, CIFAR-10 dataset results. Middle plots: TRADES, CIFAR-10 dataset results. Bottom plots: CIFAR-100 dataset results. On left: Comparison of our algorithm with a standard adversarial training (one model). We reported the results for the model with the best robust accuracy obtained over two independent runs because adversarial training might be unstable. Standard and Robust accuracy (respectively in the middle and on right) on CIFAR-10 test images in function of the number of epochs per classifier with 1 to 3 ResNet18 models. The performed attack is PGD with 20 iterations and $\varepsilon = 8/255$.

ized and deterministic classifiers w.r.t. $\varepsilon$. The adversarial risk is strictly better for randomized classifier whenever the adversarial budget $\varepsilon$ is bigger than 2. This illustration validates our analysis of Theorem 1, and motivates a in depth study of a more challenging framework, namely image classification with neural networks.

### 5.2. CIFAR Datasets

**Experimental Setup.** We now implement our heuristic algorithm (Alg. 2) on CIFAR-10 and CIFAR-100 datasets for both Adversarial Traning (Madry et al., 2018) and TRADES (Zhang et al., 2019) loss. To evaluate the performance of Algorithm 2, we trained from 1 to 4 ResNet18 (He et al., 2016) models on 200 epochs per model[7]. We study the robustness with regards to $\ell_\infty$ norm and fixed adversarial budget $\varepsilon = 8/255$. The attack we used in the inner maximization of the training is an adapted (adaptative) version of PGD for mixtures of classifiers with 10 steps. Note that for one single model, Algorithm 2 exactly corresponds to adversarial training (Madry et al., 2018) or TRADES. For each of our setups, we made two independent runs and select the best one. The training time of our algorithm is around four times longer than a standard Adversarial Training (with PGD 10 iter.) with two models, eight times with three models and twelve times with four models. We trained our models with a batch of size 1024 on 8 Nvidia V100 GPUs. We give more details on implementation in Appendix D.

**Evaluation Protocol.** At each epoch, we evaluate the current mixture on test data against PGD attack with 20 iterations. To select our model and avoid overfitting (Rice et al.,

[7] $L \times 200$ epochs in total, where $L$ is the number of models.

2020), we kept the most robust against this PGD attack. To make a final evaluation of our mixture of models, we used an adapted version of AutoPGD untargeted attacks (Croce & Hein, 2020) for randomized classifiers with both Cross-Entropy (CE) and Difference of Logits Ratio (DLR) loss. For both attacks, we made $100$ iterations and $5$ restarts.

**Results.** The results are presented in Figure 3. We remark our algorithm outperforms a standard adversarial training in all the cases by more $1\%$ on CIFAR-10 and CIFAR-100, without additional loss of standard accuracy as it is attested by the left figures. On TRADES, the gain is even more important by more than $2\%$ in robust accuracy. Moreover, it seems our algorithm, by adding more and more models, reduces the overfitting of adversarial training. It also appears that robustness increases as the number of models increases. So far, experiments are computationally very costful and it is difficult to raise precise conclusions. Further, hyperparameter tuning (Gowal et al., 2020) such as architecture, unlabeled data (Carmon et al., 2019) or activation function may still increase the results.

## 6. Related Work and Discussions

**Distributionally Robust Optimization.** Several recent works (Lee & Raginsky, 2018; Sinha et al., 2017; Tu et al., 2018) studied the problem of adversarial examples through the scope of distributionally robust optimization. In these frameworks, the set of adversarial distributions is defined using an $\ell_p$ Wasserstein ball (the adversary is allowed to have an *average* perturbation of at most $\varepsilon$ in $\ell_p$ norm). This however does not match the usual adversarial attack problem, where the adversary cannot move any point by more than $\varepsilon$. In the present work, we introduce a cost function allowing us to cast the adversarial example problem as a DRO one, without changing the adversary constraints.

**Optimal Transport (OT).** Bhagoji et al. (2019) and Pydi & Jog (2020) investigated classifier-agnostic lower bounds on the adversarial risk of any deterministic classifier using OT. These works only evaluate lower bounds on the primal deterministic formulation of the problem, while we study the existence of mixed Nash equilibria. Note that Pydi & Jog (2020) started to investigate a way to formalize the adversary using Markov kernels, but did not investigate the impact of randomized strategies on the game. We extended this work by rigorously reformulating the adversarial risk as a linear optimization problem over distributions and we study this problem from a game theoretic point of view.

**Game Theory.** Adversarial examples have been studied under the notions of Stackelberg game in (Brückner & Scheffer, 2011), and zero-sum game in (Bose et al., 2021; Perdomo & Singer, 2019; Rota Bulò et al., 2017). These works considered restricted settings (convex loss, parametric adversaries, etc.) that do not comply with the nature of the problem. Indeed, we prove in Appendix C.3 that when the loss is convex and the set $\Theta$ is convex, the duality gap is zero for deterministic classifiers. However, it has been proven that no convex loss can be a good surrogate for the $0/1$ loss in the adversarial setting (Bao et al., 2020; Cranko et al., 2019), narrowing the scope of this result. If one can show that for sufficiently separated conditional distributions, an optimal deterministic classifier always exists (see Appendix E for a clear statement), necessary and sufficient conditions for the need of randomization are still to be established. (Pinot et al., 2020) studied partly this question for regularized deterministic adversaries, leaving the general setting of randomized adversaries and mixed equilibria unanswered, which is the very scope of this paper.

## Acknowledgements

We thank anonymous reviewers for their valuable comments that helped us improve the readability of our paper. We also thank Clément Royer for fruitful discussions on the formulation of our problem as well as Alexandre Araujo for helping us in our expeiments.
This work was supported by a "Chaire d'excellence de l'IDEX Paris Saclay" and in part by Ecocloud, an EPFL research center (Postdoctoral Research Award).

## References

Bao, H., Scott, C., and Sugiyama, M. Calibrated surrogate losses for adversarially robust classification. In Abernethy, J. and Agarwal, S. (eds.), *Proceedings of Thirty Third Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pp. 408–451. PMLR, 09–12 Jul 2020. URL http://proceedings.mlr.press/v125/bao20a.html.

Bartlett, P. L. and Mendelson, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Beck, A. and Teboulle, M. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM journal on imaging sciences*, 2(1):183–202, 2009.

Bertsekas, D. P. and Shreve, S. *Stochastic optimal control: the discrete-time case*. 2004.

Bhagoji, A. N., Cullina, D., and Mittal, P. Lower bounds on adversarial robustness from optimal transport. In *Advances in Neural Information Processing Systems 32*, pp. 7496–7508. Curran Associates, Inc., 2019.

Biggio, B., Corona, I., Maiorca, D., Nelson, B., Šrndić, N., Laskov, P., Giacinto, G., and Roli, F. Evasion attacks

against machine learning at test time. In *Joint European conference on machine learning and knowledge discovery in databases*, pp. 387–402. Springer, 2013.

Blanchet, J. and Murthy, K. Quantifying distributional model risk via optimal transport. *Mathematics of Operations Research*, 44(2):565–600, 2019.

Bose, A. J., Gidel, G., Berard, H., Cianflone, A., Vincent, P., Lacoste-Julien, S., and Hamilton, W. L. Adversarial example games, 2021.

Boyd, S. Subgradient methods. 2003.

Brückner, M. and Scheffer, T. Stackelberg games for adversarial prediction problems. In *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '11, pp. 547–555, New York, NY, USA, 2011. Association for Computing Machinery. doi: 10.1145/2020408.2020495.

Carlini, N. and Wagner, D. Adversarial examples are not easily detected: Bypassing ten detection methods. In *Proceedings of the 10th ACM Workshop on Artificial Intelligence and Security*, pp. 3–14, 2017.

Carmon, Y., Raghunathan, A., Schmidt, L., Liang, P., and Duchi, J. C. Unlabeled data improves adversarial robustness. *arXiv preprint arXiv:1905.13736*, 2019.

Cohen, J. M., Rosenfeld, E., and Kolter, J. Z. Certified adversarial robustness via randomized smoothing. *arXiv preprint arXiv:1902.02918*.

Cranko, Z., Menon, A., Nock, R., Ong, C. S., Shi, Z., and Walder, C. Monge blunts bayes: Hardness results for adversarial training. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 1406–1415. PMLR, 09–15 Jun 2019. URL http://proceedings.mlr.press/v97/cranko19a.html.

Croce, F. and Hein, M. Reliable evaluation of adversarial robustness with an ensemble of diverse parameter-free attacks. In *International Conference on Machine Learning*, 2020.

Cuturi, M. Sinkhorn distances: Lightspeed computation of optimal transport. *Advances in neural information processing systems*, 26:2292–2300, 2013.

Dhillon, G. S., Azizzadenesheli, K., Bernstein, J. D., Kossaifi, J., Khanna, A., Lipton, Z. C., and Anandkumar, A. Stochastic activation pruning for robust adversarial defense. In *International Conference on Learning Representations*, 2018.

Goodfellow, I., Shlens, J., and Szegedy, C. Explaining and harnessing adversarial examples. In *International Conference on Learning Representations*, 2015.

Gowal, S., Qin, C., Uesato, J., Mann, T., and Kohli, P. Uncovering the limits of adversarial training against norm-bounded adversarial examples. *arXiv preprint arXiv:2010.03593*, 2020.

He, K., Zhang, X., Ren, S., and Sun, J. Deep residual learning for image recognition. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical report, Citeseer, 2009.

Kurakin, A., Goodfellow, I., and Bengio, S. Adversarial examples in the physical world. *arXiv preprint arXiv:1607.02533*, 2016.

Lee, J. and Raginsky, M. Minimax statistical learning with wasserstein distances. In *Advances in Neural Information Processing Systems 31*, pp. 2687–2696. Curran Associates, Inc., 2018.

Madry, A., Makelov, A., Schmidt, L., Tsipras, D., and Vladu, A. Towards deep learning models resistant to adversarial attacks. In *International Conference on Learning Representations*, 2018.

Moosavi-Dezfooli, S.-M., Fawzi, A., Uesato, J., and Frossard, P. Robustness via curvature regularization, and vice versa. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 9078–9086, 2019.

Papernot, N., McDaniel, P., and Goodfellow, I. Transferability in machine learning: from phenomena to black-box attacks using adversarial samples. *arXiv preprint arXiv:1605.07277*, 2016.

Perdomo, J. C. and Singer, Y. Robust attacks against multiple classifiers. *arXiv preprint arXiv:1906.02816*, 2019.

Pinot, R., Meunier, L., Araujo, A., Kashima, H., Yger, F., Gouy-Pailler, C., and Atif, J. Theoretical evidence for adversarial robustness through randomization. In *Advances in Neural Information Processing Systems*, pp. 11838–11848, 2019.

Pinot, R., Ettedgui, R., Rizk, G., Chevaleyre, Y., and Atif, J. Randomization matters. how to defend against strong adversarial attacks. *International Conference on Machine Learning*, 2020.

Pydi, M. S. and Jog, V. Adversarial risk via optimal transport and optimal couplings. In *International Conference on Machine Learning*. 2020.

Rice, L., Wong, E., and Kolter, Z. Overfitting in adversarially robust deep learning. In *International Conference on Machine Learning*, pp. 8093–8104. PMLR, 2020.

Rota Bulò, S., Biggio, B., Pillai, I., Pelillo, M., and Roli, F. Randomized prediction games for adversarial machine learning. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11):2466–2478, 2017.

Shalev-Shwartz, S. and Ben-David, S. *Understanding machine learning: From theory to algorithms*. Cambridge university press, 2014.

Sinha, A., Namkoong, H., and Duchi, J. Certifying some distributional robustness with principled adversarial training. *arXiv preprint arXiv:1710.10571*, 2017.

Szegedy, C., Zaremba, W., Sutskever, I., Bruna, J., Erhan, D., Goodfellow, I., and Fergus, R. Intriguing properties of neural networks. In *International Conference on Learning Representations*, 2014.

Tramèr, F., Papernot, N., Goodfellow, I., Boneh, D., and McDaniel, P. The space of transferable adversarial examples. *arXiv preprint arXiv:1704.03453*, 2017.

Tseng, P. On accelerated proximal gradient methods for convex-concave optimization. *submitted to SIAM Journal on Optimization*, 1, 2008.

Tu, Z., Zhang, J., and Tao, D. Theoretical analysis of adversarial learning: A minimax approach. *arXiv preprint arXiv:1811.05232*, 2018.

Villani, C. *Topics in optimal transportation*. Number 58. American Mathematical Soc., 2003.

Wang, B., Shi, Z., and Osher, S. Resnets ensemble via the feynman-kac formalism to improve natural and robust accuracies. In *Advances in Neural Information Processing Systems 32*, pp. 1655–1665. Curran Associates, Inc., 2019.

Xie, C., Wang, J., Zhang, Z., Ren, Z., and Yuille, A. Mitigating adversarial effects through randomization. In *International Conference on Learning Representations*, 2018.

Yue, M.-C., Kuhn, D., and Wiesemann, W. On linear optimization over wasserstein balls. *arXiv preprint arXiv:2004.07162*, 2020.

Zagoruyko, S. and Komodakis, N. Wide residual networks. In *Proceedings of the British Machine Vision Conference (BMVC)*, pp. 87.1–87.12. BMVA Press, 2016. ISBN 1-901725-59-6. doi: 10.5244/C.30.87.

Zhang, H., Yu, Y., Jiao, J., Xing, E. P., Ghaoui, L. E., and Jordan, M. I. Theoretically principled trade-off between robustness and accuracy. *International conference on Machine Learning*, 2019.