Supplementary Material

1. Omitted proofs and Additional results

Notations. Let us suppose that $(\mathcal{X},\|.\|)$ is a normed vector space. $B_{\|.\|}(x,\epsilon)=\{z\in\mathcal{X}\mid \|x-z\|\leq\epsilon\}$ is the closed ball of center x and radius ϵ for the norm $\|.\|$. Note that $\mathcal{H}:=\{h : x\mapsto \operatorname{sgn} g(x)\mid g:\mathcal{X}\to\mathbb{R}\text{ continuous}\}$, with sgn the function that outputs 1 if $g(x)>0$, -1 if $g(x)<0$, and 0 otherwise. Hence for any $(x,y)\sim D$, and $h\in\mathcal{H}$ one has $\mathbb{1}\{h(x)\neq y\}=\mathbb{1}\{g(x)y\leq 0\}$. Finally, we denote ν_1 and $\nu_{\text{-}1}$ respectively the probabilities of class 1 and -1.

Introducing remarks. Let us first note that in the paper, the penalties are defined with an ℓ_2 norm. However, Lemma 1 and 2 hold as long as $\mathcal{X}$ is an Hilbert space with dot product $<|>$ and associated norm $||.||=\sqrt{<.\,|\,.>}$. We first demonstrate Lemma 2 with these general notations. Then we present the proof of Lemma 1 that follows the same schema. Note that, for Lemma 1, we do not even need the norm to be Hilbertian, since the core argument rely on separation property of the norm, *i.e.* on the property $\|x-y\|=0 \iff x=y$.

Lemma 2. *Let $h\in\mathcal{H}$ and $\boldsymbol{\phi}\in\mathfrak{BR}_{\Omega_{norm}}(h)$. Then the following assertion holds:*

$$\phi_1(x)=\begin{cases}\pi(x) & \textit{if } x\in P_h(\epsilon_2)\\ x & \textit{otherwise.}\end{cases}$$

Where π is the orthogonal projection on $(P_h)^{\complement}$. $\phi_{\text{-}1}$ is characterized symmetrically.

Proof. Let us first simplify the worst case adversarial risk for h. Recall that $h=\operatorname{sgn}(g)$ with g continuous. From the definition of adversarial risk we have:

$$\sup_{\boldsymbol{\phi}\in\left(\mathcal{F}_{\mathcal{X}|\epsilon_2}\right)^2}\mathcal{R}_{\text{adv}}^{\Omega_{\text{norm}}}(h,\boldsymbol{\phi}) \tag{1}$$

$$=\sup_{\boldsymbol{\phi}\in(\mathcal{F}_{\mathcal{X}})^2}\sum_{y=\pm1}\nu_y\underset{X\sim\mu_y}{\mathbb{E}}\left[\mathbb{1}\left\{h\left(\phi_y(X)\right)\neq y\right\}-\lambda\|X-\phi_y(X)\|-\infty\mathbb{1}\left\{\|X-\phi_y(X)\|>\epsilon_2\right\}\right] \tag{2}$$

$$=\sup_{\boldsymbol{\phi}\in(\mathcal{F}_{\mathcal{X}})^2}\sum_{y=\pm1}\nu_y\underset{X\sim\mu_y}{\mathbb{E}}\left[\mathbb{1}\left\{g\left(\phi_y(X)\right)y\leq 0\right\}-\lambda\|X-\phi_y(X)\|-\infty\mathbb{1}\left\{\|X-\phi_y(X)\|>\epsilon_2\right\}\right] \tag{3}$$

$$=\sum_{y=\pm1}\nu_y\sup_{\phi_y\in\mathcal{F}_{\mathcal{X}}}\underset{X\sim\mu_y}{\mathbb{E}}\left[\mathbb{1}\left\{g\left(\phi_y(X)\right)y\leq 0\right\}-\lambda\|X-\phi_y(X)\|-\infty\mathbb{1}\left\{\|X-\phi_y(X)\|>\epsilon_2\right\}\right] \tag{4}$$

Finding ϕ_1 and ϕ_1 are two independent optimization problems, hence, we focus on characterizing ϕ_1 (*i.e.* $y=1$).

$$\sup_{\phi_1\in\mathcal{F}_{\mathcal{X}}}\underset{X\sim\mu_1}{\mathbb{E}}\left[\mathbb{1}\left\{g\left(\phi_1(X)\right)\leq 0\right\}-\lambda\|X-\phi_1(X)\|-\infty\mathbb{1}\left\{\|X-\phi_1(X)\|>\epsilon_2\right\}\right] \tag{5}$$

$$=\underset{X\sim\mu_1}{\mathbb{E}}\left[\underset{z\in B_{\|.\|}(X,\epsilon_2)}{\text{essup}}\;\mathbb{1}(g(z)\leq 0)-\lambda\|X-z\|\right] \tag{6}$$

$$=\int_{\mathcal{X}}\underset{z\in B_{\|.\|}(x,\epsilon_2)}{\text{essup}}\;\mathbb{1}\left\{g(z)\leq 0\right\}-\lambda\|x-z\|\;\;d\mu_1(x). \tag{7}$$

Let us now consider $(H_j)_{j\in J}$ a partition of $\mathcal{X}$, we can write.

$$\sup_{\phi_1\in\mathcal{F}_{\mathcal{X}}}\underset{X\sim\mu_1}{\mathbb{E}}\left[\mathbb{1}\left\{g\left(\phi_1(X)\right)\leq 0\right\}-\lambda\|X-\phi_1(X)\|-\infty\mathbb{1}\left\{\|X-\phi_1(X)\|>\epsilon_2\right\}\right] \tag{8}$$

$$= \sum_{j\in J} \int_{H_j} \operatorname*{essup}_{z\in B_{\|.\|}(x,\epsilon_2)} \mathbb{1}\left\{g(z)\leq 0\right\} - \lambda\|x-z\| \;\, d\mu_1(x) \tag{9}$$

In particular, we consider here $H_0 = P_h^{\complement}$, $H_1 = P_h \setminus P_h(\epsilon_2)$, and $H_2 = P_h(\epsilon_2)$.

For $x \in H_0 = P_h^{\complement}$. Taking $z = x$ we get $\mathbb{1}\left\{g(z)\leq 0\right\} - \lambda\|x-z\| = 1$. Since for any $z \in \mathcal{X}$ we have $\mathbb{1}\left\{g(z)\leq 0\right\} - \lambda\|x-z\| \leq 1$, this strategy is optimal. Furthermore, for any other optimal strategy z', we would have $\|x-z'\| = 0$, hence $z' = x$, and an optimal attack will never move the points of $H_0 = P_h^{\complement}$.

For $x \in H_1 = P_h \setminus P_h(\epsilon_2)$. We have $B_{\|.\|}(x,\epsilon_2) \subset P_h$ by definition of $P_h(\epsilon_2)$. Hence, for any $z \in B_{\|.\|}(x,\epsilon_2)$, one gets $g(z) > 0$. Then $\mathbb{1}\left\{g(z)\leq 0\right\} - \lambda\|x-z\| \leq 0$. The only optimal z will thus be $z = x$, giving value 0.

Let us now consider $x \in H_2 = P_h(\epsilon_2)$ which is the interesting case where an attack is possible. We know that $B_{\|.\|}(x,\epsilon_2) \cap P_h^{\complement} \neq \emptyset$, and for any z in this intersection, $\mathbb{1}(g(z)\leq 0) = 1$. Hence :

$$\operatorname*{essup}_{z\in B_{\|.\|}(x,\epsilon_2)} \mathbb{1}\left\{g(z)\leq 0\right\} - \lambda\|x-z\| = \max(1 - \lambda \operatorname*{essinf}_{z\in B_{\|.\|}(x,\epsilon_2)\cap P_h^{\complement}} \|x-z\|, 0) \tag{10}$$

$$= \max(1 - \lambda\pi_{B_{\|.\|}(x,\epsilon_2)\cap P_h^{\complement}}(x), 0) \tag{11}$$

Where $\pi_{B_{\|.\|}(x,\epsilon_2)\cap P_h^{\complement}}$ is the projection on the closure of $B_{\|.\|}(x,\epsilon_2) \cap P_h^{\complement}$. Note that $\pi_{B_{\|.\|}(x,\epsilon_2)\cap P_h^{\complement}}$ exists: g is continuous, so $B_{\|.\|}(x,\epsilon_2) \cap P_h^{\complement}$ is a closed set, bounded, and thus compact, since we are in finite dimension. The projection is however not guaranteed to be unique since we have no evidence on the convexity of the set. Finally, let us remark that, since $\lambda \in (0,1)$, and $\epsilon_2 \leq 1$, one has $1 - \lambda\pi_{B_{\|.\|}(x,\epsilon_2)\cap P_h^{\complement}}(x) \geq 0$ for any $x \in H_2$. Hence, on $P_h(\epsilon_2)$, the optimal attack projects all the points on the decision boundary. For simplicity, and since there is no ambiguity, we write the projection π.

Finally. Since $H_0 \cup H_1 \cup H_2 = \mathcal{X}$, Lemma 2 holds. Furthermore, the score for this optimal attack is:

$$\sup_{\phi\in\left(\mathcal{F}_{\mathcal{X}|\epsilon_2}\right)^2} \mathcal{R}_{\text{adv}}^{\Omega_{\text{norm}}}(h,\phi) \tag{12}$$

$$= \sum_{y=\pm 1} \nu_y \sum_{j\in J} \int_{H_j} \operatorname*{essup}_{z\in B_{\|.\|}(x,\epsilon_2)} \mathbb{1}\left\{g(z)y\leq 0\right\} - \lambda\|x-z\| \;\, d\mu_y(x) \tag{13}$$

Since the value is 0 on $P_h \setminus P_h(\epsilon_2)$ (resp. on $N_h \setminus N_h(\epsilon_2)$) for ϕ_1 (resp. $\phi_{\text{-}1}$), one gets:

$$= \nu_1 \left[\int_{P_h(\epsilon_2)} \big(1 - \lambda\|x-\pi(x)\|\big) d\mu_1(x) + \int_{P_h^{\complement}} 1 d\mu_1(x) \right] + \nu_{\text{-}1} \left[\int_{N_h(\epsilon_2)} \big(1 - \lambda\|x-\pi(x)\|\big) d\mu_{\text{-}1}(x) + \int_{N_h^{\complement}} 1 d\mu_{\text{-}1}(x) \right] \tag{14}$$

$$= \nu_1 \left[\int_{P_h(\epsilon_2)} \big(1 - \lambda\|x-\pi(x)\|\big) d\mu_1(x) + \mu_1(P_h^{\complement}) \right] + \nu_{\text{-}1} \left[\int_{N_h(\epsilon_2)} \big(1 - \lambda\|x-\pi(x)\|\big) d\mu_{\text{-}1}(x) + \mu_{\text{-}1}(N_h^{\complement}) \right] \tag{15}$$

$$= \mathcal{R}(h) + \nu_1 \int_{P_h(\epsilon_2)} \big(1 - \lambda\|x-\pi(x)\|\big) d\mu_1(x) + \nu_{\text{-}1} \int_{N_h(\epsilon_2)} \big(1 - \lambda\|x-\pi(x)\|\big) d\mu_{\text{-}1}(x) \tag{16}$$

(16) holds since $\mathcal{R}(h) = \mathbb{P}(h(X) \neq Y)\mathbb{P}(g(X)Y \leq 0) = \nu_1\mu_1(P_h^{\complement}) + \nu_{\text{-}1}\mu_{\text{-}1}(N_h^{\complement})$. This provides an interesting decomposition of the adversarial risk into the risk without attack and the loss on the attack zone.

□

Lemma 1. *Let $h \in \mathcal{H}$ and $\boldsymbol{\phi} \in \mathfrak{BR}_{\Omega_{mass}}(h)$. Then the following assertion holds:*

$$\begin{cases} \phi_1(x) \in (P_h)^{\complement} & \text{if } x \in P_h(\epsilon_2) \\ \phi_1(x) = x & \text{otherwise.} \end{cases}$$

Where $(P_h)^{\complement}$, the complement of P_h in $\mathcal{X}$. $\phi_{\text{-}1}$ is characterized symmetrically.

Proof. Following the same proof schema as before the adversarial risk writes as follows:

$$\sup_{\boldsymbol{\phi}\in\left(\mathcal{F}_{\mathcal{X}|\epsilon_2}\right)^2} \mathcal{R}^{\Omega_{\text{mass}}}_{\text{adv}}(h,\phi) \tag{17}$$

$$= \sup_{\boldsymbol{\phi}\in(\mathcal{F}_{\mathcal{X}})^2} \sum_{y=\pm1} \nu_y \mathop{\mathbb{E}}_{X\sim\mu_y} \left[\mathbb{1}\left\{h\left(\phi_y(X)\right) \neq y\right\} - \lambda\mathbb{1}\left\{X \neq \phi_y(X)\right\} - \infty\mathbb{1}\left\{\|X - \phi_y(X)\| > \epsilon_2\right\}\right] \tag{18}$$

$$= \sup_{\boldsymbol{\phi}\in(\mathcal{F}_{\mathcal{X}})^2} \sum_{y=\pm1} \nu_y \mathop{\mathbb{E}}_{X\sim\mu_y} \left[\mathbb{1}\left\{g\left(\phi_y(X)\right) y \leq 0\right\} - \lambda\mathbb{1}\left\{X \neq \phi_y(X)\right\} - \infty\mathbb{1}\left\{\|X - \phi_y(X)\| > \epsilon_2\right\}\right] \tag{19}$$

$$= \sum_{y=\pm1} \nu_y \sup_{\phi_y\in\mathcal{F}_{\mathcal{X}}} \mathop{\mathbb{E}}_{X\sim\mu_y} \left[\mathbb{1}\left\{g\left(\phi_y(X)\right) y \leq 0\right\} - \lambda\mathbb{1}\left\{X \neq \phi_y(X)\right\} - \infty\mathbb{1}\left\{\|X - \phi_y(X)\| > \epsilon_2\right\}\right] \tag{20}$$

Finding ϕ_1 and ϕ_1 are two independent optimization problem, hence we focus on characterizing ϕ_1 (*i.e.* $y = 1$).

$$\sup_{\phi_1\in\mathcal{F}_{\mathcal{X}}} \mathop{\mathbb{E}}_{X\sim\mu_1} \left[\mathbb{1}\left\{g\left(\phi_1(X)\right) \leq 0\right\} - \lambda\mathbb{1}\left\{X \neq \phi_1(X)\right\} - \infty\mathbb{1}\left\{\|X - \phi_1(X)\| > \epsilon_2\right\}\right] \tag{21}$$

$$= \mathop{\mathbb{E}}_{X\sim\mu_1} \left[\operatorname*{essup}_{z\in B_{\|.\|}(X,\epsilon_2)} \mathbb{1}\left\{g(z) \leq 0\right\} - \lambda\mathbb{1}\left\{X \neq z\right\}\right] \tag{22}$$

$$= \int_{\mathcal{X}} \operatorname*{essup}_{z\in B_{\|.\|}(x,\epsilon_2)} \mathbb{1}\left\{g(z) \leq 0\right\} - \lambda\mathbb{1}\left\{x \neq z\right\} \; d\mu_1(x). \tag{23}$$

Let us now consider $(H_j)_{j\in J}$ a partition of $\mathcal{X}$, we can write.

$$\sup_{\phi_1\in\mathcal{F}_{\mathcal{X}}} \mathop{\mathbb{E}}_{X\sim\mu_1} \left[\mathbb{1}\left\{g\left(\phi_1(X)\right) \leq 0\right\} - \lambda\mathbb{1}\left\{X \neq \phi_1(X)\right\} - \infty\mathbb{1}\left\{\|X - \phi_1(X)\| > \epsilon_2\right\}\right] \tag{24}$$

$$= \sum_{j\in J} \int_{H_j} \operatorname*{essup}_{z\in B_{\|.\|}(x,\epsilon_2)} \mathbb{1}\left\{g(z) \leq 0\right\} - \lambda\mathbb{1}\left\{x \neq z\right\} \; d\mu_1(x) \tag{25}$$

In particular, we can take $H_0 = P_h^{\complement}$, $H_1 = P_h \setminus P_h(\epsilon_2)$, and $H_2 = P_h(\epsilon_2)$.

For $x \in H_0 = P_h^{\complement}$ or $x \in H_1 = P_h \setminus P_h(\epsilon_2)$. With the same reasoning as before, any optimal attack will choose $\phi_1(x) = x$.

Let $x \in H_2 = P_h(\epsilon_2)$. We know that $B_{\|.\|}(x,\epsilon_2) \cap P_h^{\complement} \neq \emptyset$, and for any z in this intersection, one has $g(z) \leq 0$ and $z \neq x$. Hence $\operatorname*{essup}_{z\in B_{\|.\|}(x,\epsilon_2)} \mathbb{1}\left\{g(z) \leq 0\right\} - \lambda\mathbb{1}\left\{z \neq x\right\} = \max(1-\lambda, 0)$. Since $\lambda \in (0,1)$ one has $\mathbb{1}\left\{g(z) \leq 0\right\} - \lambda\mathbb{1}\left\{z \neq x\right\} = 1 - \lambda$ for any $z \in B_{\|.\|}(x,\epsilon_2) \cap P_h^{\complement}$. Then any function that given a $x \in \mathcal{X}$ outputs $\phi_1(x) \in B_{\|.\|}(x,\epsilon_2) \cap P_h^{\complement}$ is optimal on H_2.

Finally. Since $H_0 \cup H_1 \cup H_2 = \mathcal{X}$, Lemma 1 holds.

□

Lemma 3. *Let us consider $\boldsymbol{\phi} \in \left(\mathcal{F}_{\mathcal{X}|\epsilon_2}\right)^2$. If we take $h \in \mathfrak{BR}(\boldsymbol{\phi})$, then for $y = 1$ (resp. $y =$ -1), and for any $B \subset P_h$ (resp. $B \subset N_h$) one has*

$$\mathbb{P}(Y = y|X \in B) \geq \mathbb{P}(Y = -y|X \in B)$$

with $Y \sim \nu$ and for all $y \in \mathcal{Y}$, $X|(Y = y) \sim \phi_y\#\mu_y$.

Proof. We reason ad absurdum. Let us consider $y = 1$, the proof for $y = -1$ is symmetrical. Let us suppose that there exists $C \subset P_h$ such that $\nu_{-1}\phi_{-1}\#\mu_{-1}(C) > \nu_1\phi_1\#\mu_1(C)$. We can then construct h_1 as follows:

$$h_1(x) = \begin{cases} h(x) & \text{if } x \notin C \\ -1 & \text{otherwise.} \end{cases}$$

Since h and h_1 are identical outside C, the difference between the adversarial risks of h and h_1 writes as follows:

$$\mathcal{R}_{\text{adv}}^{\Omega_{\text{mass}}}(h, \phi) - \mathcal{R}_{\text{adv}}^{\Omega_{\text{mass}}}(h_1, \phi) \tag{26}$$

$$= \sum_{y=\pm1} \nu_y \int\limits_C \big(\mathbb{1}\left\{h(x) \neq y\right\} - \mathbb{1}\left\{h_1(x) \neq y\right\}\big) \;\; d(\phi_y\#\mu_y)(x) \tag{27}$$

$$=\nu_{-1}\mathbb{1}\left\{h(x) = 1\right\}\phi_{-1}\#\mu_{-1}(C) - \nu_1\mathbb{1}\left\{h_1(x) \neq 1\right\}\phi_1\#\mu_1(C) \tag{28}$$

$$=\nu_{-1}\phi_{-1}\#\mu_{-1}(C) - \nu_1\phi_1\#\mu_1(C) \tag{29}$$

Since by hypothesis $\nu_{-1}\phi_{-1}\#\mu_{-1}(C) > \nu_1\phi_1\#\mu_1(C)$ the difference between the adversarial risks of h and h_1 is strictly positive. This means that h_1 gives strictly better adversarial risk than the best response h. Since, by definition h is supposed to be optimal, this leads to a contradiction. Hence Lemma 3 holds. □

Additional Result. *Let us assume that there is a probability measure ζ that dominates both $\phi_1\#\mu_1$ and $\phi_{-1}\#\mu_{-1}$. Let us consider $\boldsymbol{\phi} \in \left(\mathcal{F}_{\mathcal{X}|\epsilon_2}\right)^2$. If we take $h \in \mathfrak{BR}(\boldsymbol{\phi})$, then h is the Bayes Optimal Classifier for the distribution characterized by $(\nu, \phi_1\#\mu_1, \phi_{-1}\#\mu_{-1})$.*

Proof. For simplicity, we denote $f_1 = \frac{(d\phi_1\#\mu_1)}{d\zeta}$ and $f_{-1} = \frac{d(\phi_{-1}\#\mu_{-1})}{d\zeta}$ the Radon-Nikodym derivatives of $\phi_1\#\mu_1$ and $\phi_{-1}\#\mu_{-1}$ w.r.t. ζ. The best response h minimizes adversarial risk under attack $\boldsymbol{\phi}$. This minimal risk writes:

$$\inf_{h\in\mathcal{H}} \mathcal{R}_{\text{adv}}^{\Omega_{\text{mass}}}(h, \boldsymbol{\phi}) \tag{30}$$

$$= \inf_{h\in\mathcal{H}} \sum_{y=\pm1} \nu_y \mathop{\mathbb{E}}_{x\sim\mu_y}\left[\mathbb{1}\left\{h(\phi_y(x)) \neq y\right\}\right] - \lambda\,\Omega\left(\boldsymbol{\phi}\right). \tag{31}$$

Since the the penalty function does not depend on h, it suffices to seek $\inf\limits_{h\in\mathcal{H}} \sum\limits_{y=\pm1} \nu_y \int\limits_{\mathcal{X}} \mathbb{1}\left\{h(x) \neq y\right\} \quad d(\phi_y\#\mu_y)(x)$. Moreover thanks to the transfer theorem, one gets the following:

$$\inf_{h\in\mathcal{H}} \sum_{y=\pm1} \nu_y \int\limits_{\mathcal{X}} \mathbb{1}\left\{h(x) \neq y\right\} \;\; d(\phi_y\#\mu_y)(x) \tag{32}$$

$$= \inf_{h\in\mathcal{H}} \sum_{y=\pm1} \nu_y \int\limits_{\mathcal{X}} \mathbb{1}\left\{h(x) \neq y\right\} f_y(x) \;\; d\zeta(x) \tag{33}$$

$$= \inf_{h\in\mathcal{H}} \int\limits_{\mathcal{X}} \sum_{y=\pm1} \nu_y \mathbb{1}\left\{h(x) \neq y\right\} f_y(x) \;\; d\zeta(x). \tag{34}$$

Finally, since the integral is bounded we get:

$$\inf_{h\in\mathcal{H}} \int\limits_{\mathcal{X}} \sum_{y=\pm1} \nu_y \mathbb{1}\left\{h(x) \neq y\right\} f_y(x) \;\; d\zeta(x) \tag{35}$$

$$= \int\limits_{\mathcal{X}} \left[\inf_{h\in\mathcal{H}} \sum_{y=\pm1} \nu_y \mathbb{1}\left\{h(x) \neq y\right\} f_y(x)\right] d\zeta(x). \tag{36}$$

Hence, the best response h is such that for every $x \in \mathcal{X}$, and $y \in \mathcal{Y}$, one has $h(x) = y$ if and only if $f_y(x) \leq f_{-y}(x)$. Thus, h is the optimal Bayes classifier for the distribution $(\nu, \phi_1\#\mu_1, \phi_{\text{-}1}\#\mu_{\text{-}1})$. Furthermore, for $y = 1$ (resp. $y =$ -1), and for any $B \subset P_h$ (resp. $B \subset N_h$) one has:

$$\mathbb{P}(Y = y|X \in B) \geq \mathbb{P}(Y = -y|X \in B)$$

with $Y \sim \nu$ and for all $y \in \mathcal{Y}$, $X|(Y = y) \sim \phi_y\#\mu_y$.

□

Theorem 1 (Non-existence of a pure Nash equilibrium)**.** *In our zero-sum game with* $\lambda \in (0, 1)$ *and penalty* $\Omega \in \{\Omega_{mass}, \Omega_{norm}\}$*, there is no Pure Nash Equilibrium.*

Proof. Let h be a classifier, $\phi \in \mathfrak{BR}_\Omega(h)$ an optimal attack against h. We will show that $h \notin \mathfrak{BR}(\boldsymbol{\phi})$, i.e. that h does not satisfy the condition from Lemma 3. This suffices for Theorem 1 to hold since it implies that there is no $(h, \boldsymbol{\phi}) \in \mathcal{H} \times \left(\mathcal{F}_{\mathcal{X}|\epsilon_2}\right)^2$ such that $h \in \mathfrak{BR}(\boldsymbol{\phi})$ *and* $\boldsymbol{\phi} \in \mathfrak{BR}_\Omega(h)$.

According to Lemmas 1 and 2, whatever penalty we use, there exists $\delta > 0$ such that $\phi_1\#\mu_1\left(P_h(\delta)\right) = 0$ or $\phi_{-1}\#\mu_{\text{-}1}\left(N_h(\delta)\right) = 0$. Both cases are symmetrical, so let us assume that $P_h(\delta)$ is of null measure for the transported distribution conditioned by $y = 1$. Furthermore we have $\phi_{-1}\#\mu_{\text{-}1}\left(P_h(\delta)\right) = \mu_{\text{-}1}\left(P_h(\delta)\right) > 0$ since ϕ_{-1} is the identity function on $P_h(\delta)$, and since $\mu_{\text{-}1}$ is of full support on $\mathcal{X}$. Hence we get the following:

$$\phi_{-1}\#\mu_{\text{-}1}\left(P_h(\delta)\right) > \phi_1\#\mu_1\left(P_h(\delta)\right). \tag{37}$$

Since the right side of the inequality is null, we also get:

$$\phi_{-1}\#\mu_{\text{-}1}\left(P_h(\delta)\right)\nu_{\text{-}1} > \phi_1\#\mu_1\left(P_h(\delta)\right)\nu_1. \tag{38}$$

This inequality is incompatible with the characterization of best response for the Defender of Lemma 3. Hence $h \notin \mathfrak{BR}(\boldsymbol{\phi})$.

□

Theorem 2. *(Randomization matters) Let us consider* $h_1 \in \mathcal{H}$*,* $\lambda \in (0, 1)$*,* $\Omega = \Omega_{mass}$*,* $\boldsymbol{\phi} \in \mathfrak{BR}_\Omega(h_1)$ *and* $h_2 \in \mathfrak{BR}(\boldsymbol{\phi})$*. Then for any* $\alpha \in (\max(\lambda, 1 - \lambda), 1)$ *and for any* $\boldsymbol{\phi}' \in \mathfrak{BR}_\Omega(m^{\mathbf{q}}_{\mathbf{h}})$ *one has*

$$\mathcal{R}^{\Omega_{mass}}_{\text{adv}}(m^{\mathbf{q}}_{\mathbf{h}}, \boldsymbol{\phi}') < \mathcal{R}^{\Omega_{mass}}_{\text{adv}}(h_1, \boldsymbol{\phi}).$$

Where $\mathbf{h} = (h_1, h_2)$*,* $\mathbf{q} = (\alpha, 1 - \alpha)$*, and* $m^{\mathbf{q}}_{\mathbf{h}}$ *is the mixture of* $\mathbf{h}$ *by* $\mathbf{q}$.

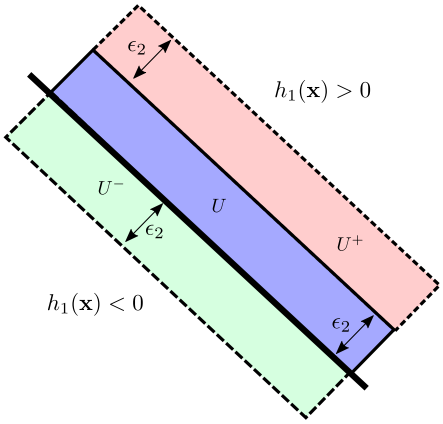

Figure 1. Illustration of the notations U, U^+, and U^- for proof of Theorem 2.

Proof. To demonstrate Theorem 2, let us denote $U = P_{h_1}(\epsilon_2)$ and define the ϵ_2*-dilation of* U as $U \oplus \epsilon_2 := \Big\{ u + v \mid (u, v) \in U \times \mathcal{X} \text{ and } \|v\|_p \leq \epsilon_2 \Big\}$. We can construct h_2 as follows

$$h_2(x) = \begin{cases} -h_1(x) & \text{if } x \in U \\ h_1(x) & \text{otherwise.} \end{cases}$$

This means that h_2 changes the class of all points in U, and do not change the rest, compared to h_1. Then taking $\alpha \in (0, 1)$, we can define $m^{\boldsymbol{q}}_{\boldsymbol{h}}$, and $\boldsymbol{\phi}' \in \mathfrak{BR}_\Omega(m^{\boldsymbol{q}}_{\boldsymbol{h}})$. We aim to find a condition on α so that the score of $m^{\boldsymbol{q}}_{\boldsymbol{h}}$ is lower than the score of h_1. Finally, let us recall that

$$\begin{aligned}
&\mathcal{R}^{\Omega_{\text{mass}}}_{\text{adv}}(m^{\boldsymbol{q}}_{\boldsymbol{h}}, \boldsymbol{\phi}') \\
= \ &\nu_1 \int_{\mathcal{X}} \operatorname*{essup}_{z \in B_{\|.\|}(x,\epsilon_2)} \alpha \mathbb{1}\left\{h_1(z) = \text{-}1\right\} + (1-\alpha)\mathbb{1}\left\{h_2(z) = \text{-}1\right\} - \lambda \mathbb{1}\left\{x \neq z\right\} \ d\mu_1(x) \\
+ \ &\nu_{\text{-}1} \int_{\mathcal{X}} \operatorname*{essup}_{z \in B_{\|.\|}(x,\epsilon_2)} \alpha \mathbb{1}\left\{h_1(z) = 10\right\} + (1-\alpha)\mathbb{1}\left\{h_2(z) = 1\right\} - \lambda \mathbb{1}\left\{x \neq z\right\} \ d\mu_{\text{-}1}(x).
\end{aligned}$$

The only terms that may vary between the score of h_1 and the score of $m^{\boldsymbol{q}}_{\boldsymbol{h}}$ are the integrals on U, $U \oplus \epsilon_2 \cap P_{h_1}$ and $\phi^{-1}_{\text{-}1}(U)$ – inverse image of U by $\phi_{\text{-}1}$. These sets represent respectively the points we mix on, the points that may become attacked – when changing from h_1 to $m^{\boldsymbol{q}}_{\boldsymbol{h}}$ – by moving them on U, and the ones that were – for h_1 – attacked before by moving them on U. Hence, for simplicity, we only write those terms. Furthermore, we denote

$$U^+ := U \oplus \epsilon_2 \cap P_{h_1} \setminus U, \ U^- := \phi^{-1}_{\text{-}1}(U) \text{ and recall } U := P_{h_1}(\epsilon_2).$$

One can refer to Figure 1 for visual interpretation of this sets. We can now evaluate the worst case adversarial score for h_1 restricted to the above sets. Thanks to Lemma 1 that characterizes $\boldsymbol{\phi}$, we can write

$$\begin{aligned}
&\mathcal{R}^{\Omega_{\text{mass}}}_{\text{adv}}(h_1, \boldsymbol{\phi})_{|U,\, U^+,\, U^-} \\
= \ &(1-\lambda) \times \nu_1 \mu_1\left(U\right) + \nu_{\text{-}1}\mu_{\text{-}1}(U) \\
&+ 0 \times \nu_1 \mu_1\left(U^+\right) + \nu_{\text{-}1}\mu_{\text{-}1}\left(U^+\right) \\
&+ \nu_1 \mu_1\left(U^-\right) + (1-\lambda) \times \nu_{\text{-}1}\mu_{\text{-}1}\left(U^-\right).
\end{aligned}$$

Similarly, we can write the worst case adversarial score of the mixture on the sets we consider. Note that the max operator comes from the fact that the adversary has to make a choice between attacking the zone or just take advantage of the error due to randomization.

$$\begin{aligned}
&\mathcal{R}^{\Omega_{\text{mass}}}_{\text{adv}}(m^{\boldsymbol{q}}_{\boldsymbol{h}}, \boldsymbol{\phi}')_{|U,\, U^+,\, U^-} \\
= \ &\max\left(1-\alpha, 1-\lambda\right) \times \nu_1\mu_1\left(U\right) + \max\left(\alpha, 1-\lambda\right) \times \nu_{\text{-}1}\mu_{\text{-}1}(U) \\
&+ \ \max\left(0, 1-\alpha-\lambda\right) \times \nu_1\mu_1\left(U^+\right) + \nu_{\text{-}1}\mu_{\text{-}1}\left(U^+\right) \\
&+ \nu_1\mu_1\left(U^-\right) + \max\left(0, \alpha-\lambda\right) \times \nu_{\text{-}1}\mu_{\text{-}1}\left(U^-\right).
\end{aligned}$$

Computing the difference between these two terms, we get the following

$$\begin{aligned}
&\mathcal{R}^{\Omega_{\text{mass}}}_{\text{adv}}(h_1, \boldsymbol{\phi}) - \mathcal{R}^{\Omega_{\text{mass}}}_{\text{adv}}(m^{\boldsymbol{q}}_{\boldsymbol{h}}, \boldsymbol{\phi}') && (39) \\
= \ &(1-\lambda-\max\left(1-\alpha, 1-\lambda\right)) \times \nu_1\mu_1\left(U\right) && (40) \\
&+ \ (1-\max\left(\alpha, 1-\lambda\right)) \times \nu_{\text{-}1}\mu_{\text{-}1}\left(U\right) && (41) \\
&- \ \max\left(0, 1-\alpha-\lambda\right) \times \nu_1\mu_1\left(U^+\right) && (42) \\
&+ (1-\lambda-\max\left(0, \alpha-\lambda\right)) \times \nu_{\text{-}1}\mu_{\text{-}1}\left(U^-\right) && (43)
\end{aligned}$$

Let us now simplify Equation (39) using additional assumptions.

- First, we have that Equation (41) is equal to

$$\min\left(1-\alpha,\lambda\right)\mu_{\text{-}1}(U)\nu_{\text{-}1}>0.$$

  Thus, a sufficient condition for the difference between the adversarial scores to be positive is to have the other terms greater or equal to 0.

- To have Equation (40) ≥ 0 we can always set $\max\left(1-\alpha,1-\lambda\right)=1-\lambda$. This gives us $\alpha\geq\lambda$.
- Also note that to get (42) ≥ 0, we can force $\max\left(1-\alpha-\lambda,0\right)=0$. This gives us $\alpha\geq 1-\lambda$.
- Finally, since $\alpha\geq\lambda$, we have that $1-\lambda-\max\left(0,\alpha-\lambda\right)=1-\alpha$ thus Equations (43) >0.

With the above simplifications, we have (39) >0 for any $\alpha>\max(\lambda,1-\lambda)$ which concludes the proof. □

Theorem 3. *(Randomization matters) Let us consider $h_1\in\mathcal{H}$, $\lambda\in(0,1)$, $\Omega=\Omega_{norm}$, $\boldsymbol{\phi}\in\mathfrak{BR}_\Omega(h_1)$ and $h_2\in\mathfrak{BR}(\boldsymbol{\phi})$. Let us take $\delta\in(0,\epsilon_2)$, then for any $\alpha\in(\max(1-\lambda\delta,\lambda(\epsilon_2-\delta)),1)$ and for any $\boldsymbol{\phi}'\in\mathfrak{BR}_\Omega(m_{\mathbf{h}}^{\mathbf{q}})$ one has*

$$\mathcal{R}_{\text{adv}}^{\Omega_{norm}}(m_{\mathbf{h}}^{\mathbf{q}},\boldsymbol{\phi}')<\mathcal{R}_{\text{adv}}^{\Omega_{norm}}(h_1,\boldsymbol{\phi}).$$

Where $\mathbf{h}=(h_1,h_2)$, $\mathbf{q}=(\alpha,1-\alpha)$, and $m_{\mathbf{h}}^{\mathbf{q}}$ is the mixture of $\mathbf{h}$ by $\mathbf{q}$.

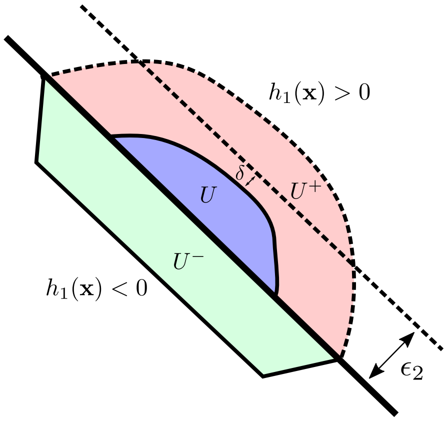

Figure 2. Illustration of the notations U, U^+, U^- and δ for proof of Theorem 3.

Proof. Let us take $U\subset P_{h_1}(\epsilon_2)$ such that

$$\min_{x\in U}\|x-\pi_{P_h\setminus P_h(\epsilon_2)}(x)\|=\delta\in(0,\epsilon_2)$$

. We construct h_2 as follows.

$$h_2(x)=\begin{cases}-h_1(x) & \text{if } x\in U\\ h_1(x) & \text{otherwise.}\end{cases}$$

This means that h_2 changes the class of all points in U, and do not change the rest. Let $\alpha\in(0,1)$, the corresponding mixture $m_{\boldsymbol{h}}^{\boldsymbol{q}}$, and $\boldsymbol{\phi}'\in\mathfrak{BR}_\Omega(m_{\boldsymbol{h}}^{\boldsymbol{q}})$. We will find a condition on α so that the score of $m_{\boldsymbol{h}}^{\boldsymbol{q}}$ is lower than the score of h_1. Recall that

$$\begin{aligned}&\mathcal{R}_{\text{adv}}^{\Omega_{\text{norm}}}(m_{\boldsymbol{h}}^{\boldsymbol{q}},\boldsymbol{\phi}')\\ =\;&\nu_1\int_{\mathcal{X}}\operatorname*{essup}_{z\in B_{\|.\|}(x,\epsilon_2)}\alpha\mathbb{1}\left\{h_1(z)=\text{-}1\right\}+(1-\alpha)\mathbb{1}\left\{h_2(z)=\text{-}1\right\}-\lambda\|x-z\|\,d\mu_1(x)\\ &+\nu_{\text{-}1}\int_{\mathcal{X}}\operatorname*{essup}_{z\in B_{\|.\|}(x,\epsilon_2)}\alpha\mathbb{1}\left\{h_1(z)=1\right\}+(1-\alpha)\mathbb{1}\left\{h_2(z)=1\right\}-\lambda\|x-z\|\,d\mu_{\text{-}1}(x).\end{aligned}$$

As we discussed in proof of Theorem 2, the only terms that may vary between the score of h_1 and the score of $m_{\boldsymbol{h}}^{\boldsymbol{q}}$ are the integrals on U, $U \oplus \epsilon_2 \cap P_{h_1}$ and $\phi_{-1}^{-1}(U)$. Hence, for simplicity, we only write those terms. Furthermore, we denote

$$U^+ := U \oplus \epsilon_2 \cap P_{h_1} \setminus U,\; U^- := \phi_{-1}^{-1}(U) \text{ and } P_{\epsilon_2} := P_{h_1}(\epsilon_2).$$

One can refer to Figure 2 for a visual interpretation of this ensembles. We can now evaluate the worst case adversarial score for h_1 restricted to the above sets. Thanks to Lemma 2 that characterizes $\boldsymbol{\phi}$, we can write

$$\begin{aligned}
&\mathcal{R}_{\text{adv}}^{\Omega_{\text{norm}}}(h_1, \boldsymbol{\phi}) \\
&= \nu_1 \int\limits_U \Big(1 - \lambda \|x - \pi_{P_{h_1}^{\complement}}(x)\|\Big)\, d\mu_1(x) + \nu_{-1}\mu_{-1}(U) \\
&+ \nu_1 \int\limits_{U^+ \setminus P_{\epsilon_2}} 0\, d\mu_1(x) + \nu_{-1}\mu_{-1}\left(U^+ \setminus P_{\epsilon_2}\right) \\
&+ \nu_1 \int\limits_{U^+ \cap P_{\epsilon_2}} \Big(1 - \lambda \|x - \pi_{P_{h_1}^{\complement}}(x)\|\Big)\, d\mu_1(x) + \nu_{-1}\mu_{-1}\left(U^+ \cap P_{\epsilon_2}\right) \\
&+ \nu_1\mu_1\left(U^-\right) + \nu_{-1} \int\limits_{U^-} \Big(1 - \lambda \|x - \pi_U(x)\|\Big) d\mu_{-1}(x).
\end{aligned}$$

Similarly we can evaluate the worst case adversarial score for the mixture,

$$\begin{aligned}
&\mathcal{R}_{\text{adv}}^{\Omega_{\text{norm}}}(m_{\boldsymbol{h}}^{\boldsymbol{q}}, \boldsymbol{\phi}') \\
&= \nu_1 \int\limits_U \max\Big(1-\alpha, 1 - \lambda \|x - \pi_{P_{h_1}^{\complement}}(x)\|\Big) \;\; d\mu_1(x) \\
&+ \nu_{-1} \int\limits_U \max\left(\alpha, 1 - \lambda\|x - \pi_{U^+}(x)\|\right) \;\; d\mu_{-1}(x) \\
&+ \nu_1 \int\limits_{U^+ \setminus P_{\epsilon_2}} \max\left(0, 1 - \alpha - \lambda\|x - \pi_U(x)\|\right) \;\; d\mu_1(x) + \nu_{-1}\mu_{-1}\left(U^+ \setminus P_{\epsilon_2}\right) \\
&+ \nu_1 \int\limits_{U^+ \cap P_{\epsilon_2}} \max\Big(1 - \alpha - \lambda\|x - \pi_U(x)\|, 1 - \lambda\|x - \pi_{P_{h_1}^{\complement}}(x)\|\Big) \;\; d\mu_1(x) \\
&+ \nu_{-1}\mu_{-1}\left(U^+ \cap P_{\epsilon_2}\right) + \nu_1\mu_1\left(U^-\right) \\
&+ \nu_{-1} \int\limits_{U^-} \max\Big(0, 1 - \lambda\|x - \pi_{N_{h_1}^{\complement} \setminus U}(x)\|, \alpha - \lambda\|x - \pi_U(x)\|\Big)\, d\mu_{-1}(x).
\end{aligned}$$

Note that we need to take into account the special case of the points in the dilation that were already in the attacked zone before, and that can now be attacked in two ways, either by projecting on U – but that works with probability α, since the classification on U is now randomized – or by projecting on $P_{h_1}^{\complement}$, which works with probability 1 but may use more distance and so pay more penalty. We can now compute the difference between both scores.

$$\mathcal{R}_{\text{adv}}^{\Omega_{\text{norm}}}(h_1, \boldsymbol{\phi}) - \mathcal{R}_{\text{adv}}^{\Omega_{\text{norm}}}(m_{\boldsymbol{h}}^{\boldsymbol{q}}, \boldsymbol{\phi}') \tag{44}$$

$$= \nu_1 \int\limits_U 1 - \lambda\|x - \pi_{P_{h_1}^{\complement}}(x)\| - \max\Big(1 - \alpha, 1 - \lambda\|x - \pi_{P_{h_1}^{\complement}}(x)\|\Big)\, d\mu_1(x) \tag{45}$$

$$+ \nu_{-1} \int\limits_U 1 - \max\left(\alpha, 1 - \lambda\|x - \pi_{U^+}(x)\|\right) d\mu_{-1}(x) \tag{46}$$

$$- \nu_1 \int\limits_{U^+ \setminus P_{\epsilon_2}} \max\left(1 - \alpha - \lambda\|x - \pi_U(x)\|, 0\right) d\mu_1(x) \tag{47}$$

$$\begin{aligned} &+ \nu_1 \int\limits_{U^+ \cap P_{\epsilon_2}} 1 - \lambda\|x - \pi_{P^{\complement}_{h_1}}(x)\| \\ &- \max\left(1 - \alpha - \lambda\|x - \pi_U(x)\|, 1 - \lambda\|x - \pi_{P^{\complement}_{h_1}}(x)\|\right) d\mu_1(x) \end{aligned} \tag{48}$$

$$\begin{aligned} &+ \nu_{\text{-}1} \int\limits_{U^-} 1 - \lambda\|x - \pi_U(x)\| \\ &- \max\left(0, 1 - \lambda\|x - \pi_{N^{\complement}_{h_1} \setminus U}(x)\|, \alpha - \lambda\|x - \pi_U(x)\|\right) d\mu_{\text{-}1}(x). \end{aligned} \tag{49}$$

Let us simplify Equation (44) using using additional hypothesis:

- First, note that Equation (46)> 0. Then a sufficient condition for the difference to be strictly positive is to ensure that other lines are ≥ 0.
- In particular to have (45) ≥ 0 it is sufficient to have for all $x \in U$

  $$\max\left(1 - \alpha, 1 - \lambda\|x - \pi_{P^{\complement}_{h_1}}(x)\|\right) = 1 - \lambda\|x - \pi_{P^{\complement}_{h_1}}(x)\|.$$

  This gives us $\alpha \geq \lambda(\epsilon_2 - \delta) \geq \lambda \max\limits_{x \in U} \|x - \pi_{P^{\complement}_{h_1}}(x)\|$.
- Similarly, to have (47) ≥ 0, we should set for all $x \in U^+ \setminus P_{\epsilon_2}$

  $$\alpha \geq 1 - \lambda\|x - \pi_U(x)\|.$$

  Since $\min\limits_{x \in U^+ \setminus P_{\epsilon_2}} \|x - \pi_U(x)\| = \delta$, we get the condition $\alpha \geq 1 - \lambda\delta$.
- Finally (49) ≥ 0, since by definition of U^-, for any $x \in U^-$ we have

  $$\|x - \pi_{N^{\complement}_{h_1} \setminus U}(x)\| \geq \|x - \pi_U(x)\|.$$

Finally, by summing all these simplifications, we have (44) > 0. Hence the result hold for any $\alpha > \max(1 - \lambda\delta, \lambda(\epsilon_2 - \delta))$ □

2. Experimental results

In the experimental section, we consider $\mathcal{X} = [0, 1]^{3\times32\times32}$ to be the set of images, and $\mathcal{Y} = \{1, ..., 10\}$ or $\mathcal{Y} = \{1, ..., 100\}$ according to the dataset at hand.

2.1. Adversarial attacks

Let $(x, y) \sim D$ and $h \in \mathcal{H}$. We consider the following attacks:

(i) ℓ_∞-PGD attack. In this scenario, the Adversary maximizes the loss objective function, under the constraint that the ℓ_∞ norm of the perturbation remains bounded by some value ϵ_∞. To do so, it recursively computes:

$$x^{t+1} = \Pi_{B_{\|\cdot\|}(x, \epsilon_\infty)}\left[x^t + \beta \operatorname{sgn}\left(\nabla_x \mathcal{L}\left(h\left(x^t\right), y\right)\right)\right] \tag{50}$$

where $\mathcal{L}$ is some differentiable loss (such as the cross-entropy), β is a gradient step size, and Π_S is the projection operator on S. One can refer to (Madry et al., 2018) for implementation details.

(ii) ℓ_2-C&W attack. In this attack, the Adversary optimizes the following objective:

$$\underset{\tau \in \mathcal{X}}{\operatorname{argmin}} \|\tau\|_2 + \lambda \times \operatorname{cost}(x + \tau) \tag{51}$$

where $\operatorname{cost}(x+\tau) < 0$ if and only if $h(x+\tau) \neq y$. The authors use a change of variable $\tau = \frac{1}{2}(\tanh(w) - x + 1)$ to ensure that $x + \tau \in \mathcal{X}$, a binary search to optimize the constant λ, and Adam or SGD to compute an approximated solution. One should refer to (Carlini & Wagner, 2017) for implementation details.

2.2. Experimental setup

Datasets. To illustrate our theoretical results we did experiments on the **CIFAR10** and **CIFAR100** datasets. See (Krizhevsky et al., 2009) for more details.

Classifiers. All the classifiers we use are WideResNets (see (Zagoruyko & Komodakis, 2016)) with 28 layers, a widen factor of 10, a dropout factor of 0.3 and LeakyRelu activations with a 0.1 slope.

Natural Training. To train an undefended classifier we use the following hyperparameters.

- **Number of Epochs:** 200
- **Batch size:** 128
- **Loss function:** Cross Entropy Loss
- **Optimizer :** SGD algorithm with momentum 0.9, weight decay of 2×10^{-4} and a learning rate that decreases during the training as follows:

$$lr = \begin{cases} 0.1 & \text{if} \quad 0 \leq \text{epoch} < 60 \\ 0.02 & \text{if} \quad 60 \leq \text{epoch} < 120 \\ 0.004 & \text{if} \quad 120 \leq \text{epoch} < 160 \\ 0.0008 & \text{if} \quad 160 \leq \text{epoch} < 200 \end{cases}$$

Adversarial Training. To adversarially train a classifier we use the same hyperparameters as above, and generate adversarial examples using the ℓ_∞**-PGD** attack with 20 iterations. When considering that the input space is $[0, 255]^{3\times32\times32}$, on **CIFAR10** and **CIFAR100**, a perturbation is considered to be imperceptible for $\epsilon_\infty = 8$. Here, we consider $\mathcal{X} = [0, 1]^{3\times32\times32}$ which is the normalization of the pixel space $[0.255]^{3\times32\times32}$. Hence, we choose $\epsilon_2 = 0.031$ ($\approx 8/255$) for each attack. Moreover, the step size we use for ℓ_∞**-PGD** is 0.008 ($\approx 2/255$), we use a random initialization for the gradient descent and we repeat the procedure three times to take the best perturbation over all the iterations *i.e* the one that maximises the loss. For the ℓ_∞**-PGD** attack against the mixture $m_{\mathbf{h}}^{\mathbf{q}}$, we use the same parameters as above, but compute the gradient over the loss of the expected logits (as explained in the main paper).

Evaluation Under Attack. At evaluation time, we use 100 iterations instead of 20 for **Adaptive-**ℓ_∞**-PGD**, and the same remaining hyperparameters as before. For the **Adaptive-**ℓ_2**-C&W** attack, we use 100 iterations, a learning rate equal to 0.01, 9 binary search steps, and an initial constant of 0.001. We give results for several different values of the rejection threshold: $\epsilon_2 \in \{0.4, 0.6, 0.8\}$.

Computing Adaptive-ℓ_2**-C&W on a mixture** To attack a randomized model, it is advised in the literature (Tramer et al., 2020) to compute the expected logits returned by this model. However this advice holds for randomized models that return logits in the same range for a same example (*e.g.* classifier with noise injection). Our randomized model is a mixture and returns logits that depend on selected classifier. Hence, for a same example, the logits can be very different. This phenomenon made us notice that for some example in the dataset, computing the expected loss over the classifier (instead of the expected logits) performs better to find a good perturbation (it can be seen as computing the expectation of the logits normalized thanks to the loss). To ensure a fair evaluation of our model, in addition of using EOT with the expected logits, we compute in parallel EOT with the expected loss and take the perturbation that maximizes the expected error of the mixture. See the submitted code for more details.

Library used. We used the Pytorch and Advertorch libraries for all implementations.

Machine used. 6 Tesla V100-SXM2-32GB GPUs

2.3. Experimental details

Sanity checks for Adaptive attacks In (Tramer et al., 2020), the authors give a lot of sanity checks and good practices to design an Adaptive attacks. We follow them and here are the information for **Adaptive-ℓ_∞-PGD** :

- We compute the gradient of the loss by doing the expected logits over the mixture.
- The attack is repeated 3 times with random start and we take the best perturbation over all the iterations.
- When adding a constant to the logits, it doesn't change anything to the attack
- When doing 200 iterations instead of 100 iterations, it doesn't change the performance of the attack
- When increasing the budget ϵ_∞, the accuracy goes to 0, which ensures that there is no gradient masking. Here are some values to back this statement:

Epsilon	0.015	0.031	0.125	0.250
Accuracy	0.638	0.546	0.027	0.000

Table 1. Evolution of the accuracy under **Adaptive-ℓ_∞-PGD** attack depending on the budget ϵ_∞

- The loss doesn't fluctuate at the end of the optimization process.

Selecting the first element of the mixture. Our algorithm creates classifiers in a boosting fashion, starting with an adversarially trained classifier. There are several ways of selecting this first element of the mixture: use the classifier with the best accuracy under attack (option 1, called bestAUA), or rather the one with the best natural accuracy (option 2). Table 2 compares both options.

Beside the fact that any of the two mixtures outperforms the first classifier, we see that the fisrt option always outperforms the second. In fact, when taking option 1 (bestAUA = True) the accuracy under ℓ_∞**-PGD** attack of the mixture is 3% better than with option 2 (bestAUA = False). One can also note that both mixtures have the same natural accuracy (0.80), which makes the choice of option 1 natural.

Training method	NA of the 1^{st} clf	AUA of the 1^{st} clf	NA of the mixture	AUA of the mixture
BAT (bestAUA=True)	0.77	**0.46**	**0.80**	**0.55**
BAT (bestAUA=False)	**0.83**	0.42	**0.80**	0.52

Table 2. Comparison of the mixture that has as first classifier the best one in term of natural accuracy and the mixture that has as first classifier the best one in term of Accuracy under attack. The accuracy under attack is computed with the ℓ_∞**-PGD** attack. NA means matural accuracy, and AUA means accuracy under attack.

2.4. Extension to more than two classifiers

As we mention in the main part of the paper, a mixture of more than two classifiers can be constructed by adding at each step t a new classifier trained naturally on the dataset $\tilde{D}$ that contains adversarial examples against the mixture at step $t-1$. Since $\tilde{D}$ has to be constructed from a mixture, one would have to use an adaptive attack as **Adaptive-ℓ_∞-PGD**. Here is the algorithm for the extented version :

Algorithm 1 Boosted Adversarial Training

```
Input : $n$ the number of classifiers, $D$ the training data set and $\alpha$ the weight update parameter.

Create and adversarially train $h_1$ on $D$
$\mathbf{h} = (h_1)$ ; $\mathbf{q} = (1)$
for $i = 2, \dots, n$ do
    Generate the adversarial data set $\tilde{D}$ against $m_{\mathbf{h}}^{\mathbf{q}}$.
    Create and naturally train $h_i$ on $\tilde{D}$

    $q_k \leftarrow (1 - \alpha) q_k \quad \forall k \in [i - 1]$
    $q_i \leftarrow \alpha$

    $\mathbf{q} \leftarrow (\alpha, \dots, q_i)$
    $\mathbf{h} \leftarrow (h_1, \dots, h_i)$
end
return $m_{\mathbf{h}}^{\mathbf{q}}$
```

Here to find the parameter α, the grid search is more costly. In fact in the two-classifier version we only need to train the first and second classifier without taking care of α, and then test all the values of α using the same two classifier we trained. For the extended version, the third classifier (and all the other ones added after) depends on the first classifier, the second one and their weights $1 - \alpha$ and α. Hence the third classifier for a certain value of α can't be use for another one and, to conduct the grid search, one have to retrain all the classifiers from the third one. Naturally the parameters α depends on the number of classifiers n in the mixtures.