Supplementary material

A. Notations

Let $(\mathcal{Z}, d)$ be a Polish metric space (i.e. complete and separable). We say that $(\mathcal{Z}, d)$ is proper if for all $z_0 \in \mathcal{Z}$ and $R > 0$, $B(z_0, R) := \{z \mid d(z, z_0) \leq R\}$ is compact. For $(\mathcal{Z}, d)$ a Polish space, we denote $\mathcal{M}_+^1(\mathcal{Z})$ the set of Borel probability measures on $\mathcal{Z}$ endowed with $\|\cdot\|_{TV}$ strong topology. We recall the notion of weak topology: we say that a sequence $(\mu_n)_n$ of $\mathcal{M}_+^1(\mathcal{Z})$ converges weakly to $\mu \in \mathcal{M}_+^1(\mathcal{Z})$ if and only if for every bounded continuous function f on $\mathcal{Z}$, $\int f d\mu_n \to_{n\to\infty} \int f d\mu$. Endowed with its weak topology, $\mathcal{M}_+^1(\mathcal{Z})$ is a Polish space. For $\mu \in \mathcal{M}_+^1(\mathcal{Z})$, we define $L^1(\mu)$ the set of integrable functions with respect to μ. We denote $\Pi_1 : (z, z') \in \mathcal{Z}^2 \mapsto z$ and $\Pi_2 : (z, z') \in \mathcal{Z}^2 \mapsto z'$ respectively the projections on the first and second component, which are continuous applications. For a measure μ and a measurable mapping μ, we denote $g_\sharp \mu$ the pushforward measure of μ by g. Let $L \geq 1$ be an integer and denote $\Delta_L := \{\lambda \in \mathbb{R}_+^L \text{ s.t. } \sum_{k=1}^L \lambda_k = 1\}$, the probability simplex of $\mathbb{R}^L$.

B. Useful Lemmas

Lemma 1 (Fubini's theorem)**.** *Let $l : \Theta \times (\mathcal{X} \times \mathcal{Y}) \to [0, \infty)$ satisfying Assumption 1. Then for all $\mu \in \mathcal{M}_+^1(\Theta)$, $\int l(\theta, \cdot) d\mu(\theta)$ is Borel measurable; for $\mathbb{Q} \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})$, $\int l(\cdot, (x, y)) d\mathbb{Q}(x, y)$ is Borel measurable. Moreover: $\int l(\theta, (x, y)) d\mu(\theta) d\mathbb{Q}(x, y) = \int l(\theta, (x, y)) d\mathbb{Q}(x, y) d\mu(\theta)$*

Lemma 2. *Let $l : \Theta \times (\mathcal{X} \times \mathcal{Y}) \to [0, \infty)$ satisfying Assumption 1. Then for all $\mu \in \mathcal{M}_+^1(\Theta)$, $(x, y) \mapsto \int l(\theta, (x, y)) d\mu(\theta)$ is upper semi-continuous and hence Borel measurable.*

Proof. Let $(x_n, y_n)_n$ be a sequence of $\mathcal{X} \times \mathcal{Y}$ converging to $(x, y) \in \mathcal{X} \times \mathcal{Y}$. For all $\theta \in \Theta$, $M - l(\theta, \cdot)$ is non negative and lower semi-continuous. Then by Fatou's Lemma applied:

$$\begin{aligned}\int M - l(\theta, (x, y)) d\mu(\theta) &\leq \int \liminf_{n\to\infty} M - l(\theta, (x_n, y_n)) d\mu(\theta) \\ &\leq \liminf_{n\to\infty} \int M - l(\theta, (x_n, y_n)) d\mu(\theta)\end{aligned}$$

Then we deduce that: $\int M - l(\theta, \cdot) d\mu(\theta)$ is lower semi-continuous and then $\int l(\theta, \cdot) d\mu(\theta)$ is upper-semi continuous. □

Lemma 3. *Let $l : \Theta \times (\mathcal{X} \times \mathcal{Y}) \to [0, \infty)$ satisfying Assumption 1 Then for all $\mu \in \mathcal{M}_+^1(\Theta)$, $\mathbb{Q} \mapsto \int l(\theta, (x, y)) d\mu(\theta) d\mathbb{Q}(x, y)$ is upper semi-continuous for weak topology of measures.*

Proof. $-\int l(\theta, \cdot) d\mu(\theta)$ is lower semi-continuous from Lemma 2. Then $M - \int l(\theta, \cdot) d\mu(\theta)$ is lower semi-continuous and non negative. Let denote v this function. Let $(v_n)_n$ be a non-decreasing sequence of continuous bounded functions such that $v_n \to v$. Let $(\mathbb{Q}_k)_k$ converging weakly towards $\mathbb{Q}$. Then by monotone convergence:

$$\int v d\mathbb{Q} = \lim_n \int v_n d\mathbb{Q} = \lim_n \lim_k \int v_n d\mathbb{Q}_k \leq \liminf_k \int v d\mathbb{Q}_k$$

Then $\mathbb{Q} \mapsto \int v d\mathbb{Q}$ is lower semi-continuous and then $\mathbb{Q} \mapsto \int l(\theta, (x, y)) d\mu(\theta) d\mathbb{Q}(x, y)$ is upper semi-continuous for weak topology of measures. □

Lemma 4. *Let $l : \Theta \times (\mathcal{X} \times \mathcal{Y}) \to [0, \infty)$ satisfying Assumption 1. Then for all $\mu \in \mathcal{M}_+^1(\Theta)$, $(x, y) \mapsto \sup_{(x', y'), d(x, x') \leq \varepsilon, y = y'} \int l(\theta, (x', y')) d\mu(\theta)$ is universally measurable (i.e. measurable for all Borel probability measures). And hence the adversarial risk is well defined.*

Proof. Let $\phi : (x, y) \mapsto \sup_{(x', y'), d(x, x') \leq \varepsilon, y = y'} \int l(\theta, (x', y')) d\mu(\theta)$. Then for $u \in \bar{\mathbb{R}}$:

$$\{\phi(x, y) > u\} = \text{Proj}_1 \left\{ ((x, y), (x', y')) \mid \int l(\theta, (x', y')) d\mu(\theta) - c_\varepsilon((x, y), (x', y')) > u \right\}$$

By Lemma 3: $((x,y),(x',y')) \mapsto \int l(\theta,(x',y'))d\mu(\theta) - c_\varepsilon((x,y),(x',y'))$ is upper-semicontinuous hence Borel measurable. So its level sets are Borel sets, and by (Bertsekas & Shreve, 2004, Proposition 7.39), the projection of a Borel set is analytic. And then $\{\phi(x,y) > u\}$ universally measurable thanks to (Bertsekas & Shreve, 2004, Corollary 7.42.1). We deduce that ϕ is universally measurable. □

C. Proofs

C.1. Proof of Proposition 1

Proof. Let $\eta > 0$. Let $\mathbb{Q} \in \mathcal{A}_\varepsilon(\mathbb{P})$. There exists $\gamma \in \mathcal{M}_1^+\left((\mathcal{X}\times\mathcal{Y})^2\right)$ such that, $d(x,x') \leq \varepsilon$, $y = y'$ γ-almost surely, and $\Pi_{1\sharp}\gamma = \mathbb{P}$, and $\Pi_{2\sharp}\gamma = \mathbb{Q}$. Then $\int c_\varepsilon d\gamma = 0 \leq \eta$. Then, we deduce that $W_{c_\varepsilon}(\mathbb{P},\mathbb{Q}) \leq \eta$, and $\mathbb{Q} \in \mathcal{B}_{c_\varepsilon}(\mathbb{P},\eta)$. Reciprocally, let $\mathbb{Q} \in \mathcal{B}_{c_\varepsilon}(\mathbb{P},\eta)$. Then, since the infimum is attained in the Wasserstein definition, there exists $\gamma \in \mathcal{M}_1^+\left((\mathcal{X}\times\mathcal{Y})^2\right)$ such that $\int c_\varepsilon d\gamma \leq \eta$. Since $c_\varepsilon((x,x'),(y,y')) = +\infty$ when $d(x,x') > \varepsilon$ and $y \neq y'$, we deduce that, $d(x,x') \leq \varepsilon$ and $y = y'$, γ-almost surely. Then $\mathbb{Q} \in \mathcal{A}_\varepsilon(\mathbb{P})$. We have then shown that: $\mathcal{A}_\varepsilon(\mathbb{P}) = \mathcal{B}_{c_\varepsilon}(\mathbb{P},\eta)$.

The convexity of $\mathcal{A}_\varepsilon(\mathbb{P})$ is then immediate from the relation with the Wasserstein uncertainty set.

Let us show first that $\mathcal{A}_\varepsilon(\mathbb{P})$ is relatively compact for weak topology. To do so we will show that $\mathcal{A}_\varepsilon(\mathbb{P})$ is tight and apply Prokhorov's theorem. Let $\delta > 0$, $(\mathcal{X}\times\mathcal{Y}, d\oplus d')$ being a Polish space, $\{\mathbb{P}\}$ is tight then there exists K_δ compact such that $\mathbb{P}(K_\delta) \geq 1-\delta$. Let $\tilde{K}_\delta := \{(x',y') \mid \exists (x,y) \in K_\delta,\ d(x',x) \leq \varepsilon,\ y = y'\}$. Recalling that $(\mathcal{X},d)$ is proper (i.e. the closed balls are compact), so $\tilde{K}_\delta$ is compact. Moreover for $\mathbb{Q} \in \mathcal{A}_\varepsilon(\mathbb{P})$, $\mathbb{Q}(\tilde{K}_\delta) \geq \mathbb{P}(K_\delta) \geq 1-\delta$. And then, Prokhorov's theorem holds, and $\mathcal{A}_\varepsilon(\mathbb{P})$ is relatively compact for weak topology.

Let us now prove that $\mathcal{A}_\varepsilon(\mathbb{P})$ is closed to conclude. Let $(\mathbb{Q}_n)_n$ be a sequence of $\mathcal{A}_\varepsilon(\mathbb{P})$ converging towards some $\mathbb{Q}$ for weak topology. For each n, there exists $\gamma_n \in \mathcal{M}_+^1(\mathcal{X}\times\mathcal{Y})$ such that $d(x,x') \leq \varepsilon$ and $y = y'$ γ_n-almost surely and $\Pi_{1\sharp}\gamma_n = \mathbb{P}$, $\Pi_{2\sharp}\gamma_n = \mathbb{Q}_n$. $\{\mathbb{Q}_n, n \geq 0\}$ is relatively compact, then tight, then $\bigcup_n \Gamma_{\mathbb{P},\mathbb{Q}_n}$ is tight, then relatively compact by Prokhorov's theorem. $(\gamma_n)_n \in \bigcup_n \Gamma_{\mathbb{P},\mathbb{Q}_n}$, then up to an extraction, $\gamma_n \to \gamma$. Then $d(x,x') \leq \varepsilon$ and $y = y'$ γ-almost surely, and by continuity, $\Pi_{1\sharp}\gamma = \mathbb{P}$ and by continuity, $\Pi_{2\sharp}\gamma = \mathbb{Q}$. And hence $\mathcal{A}_\varepsilon(\mathbb{P})$ is closed.

Finally $\mathcal{A}_\varepsilon(\mathbb{P})$ is a convex compact set for the weak topology. □

C.2. Proof of Proposition 2

Proof. Let $\mu \in \mathcal{M}_+^1(\Theta)$. Let $\tilde{f} : ((x,y),(x',y')) \mapsto \mathbb{E}_{\theta\sim\mu}\left[l(\theta,(x,y))\right] - c_\varepsilon((x,y),(x',y'))$. $\tilde{f}$ is upper-semi continuous, hence upper semi-analytic. Then, by upper semi continuity of $\mathbb{E}_{\theta\sim\mu}\left[l(\theta,\cdot)\right]$ on the compact $\{(x',y') \mid d(x,x') \leq \varepsilon, y = y'\}$ and (Bertsekas & Shreve, 2004, Proposition 7.50), there exists a universally measurable mapping T such that $\mathbb{E}_{\theta\sim\mu}\left[l(\theta,T(x,y))\right] = \sup_{(x',y'),\ d(x,x')\leq\varepsilon, y=y'} \mathbb{E}_{\theta\sim\mu}\left[l(\theta,(x,y))\right]$. Let $\mathbb{Q} = T_\sharp\mathbb{P}$, then $\mathbb{Q} \in \mathcal{A}_\varepsilon(\mathbb{P})$. And then $\mathbb{E}_{(x,y)\sim\mathbb{P}}\left[\sup_{(x',y'),\ d(x,x')\leq\varepsilon, y=y'} \mathbb{E}_{\theta\sim\mu}\left[l(\theta,(x',y'))\right]\right] \leq \sup_{\mathbb{Q}\in\mathcal{A}_\varepsilon(\mathbb{P})} \mathbb{E}_{(x,y)\sim\mathbb{Q}}\left[\mathbb{E}_{\theta\sim\mu}\left[l(\theta,(x,y))\right]\right]$.

Reciprocally, let $\mathbb{Q} \in \mathcal{A}_\varepsilon(\mathbb{P})$. There exists $\gamma \in \mathcal{M}_+^1((\mathcal{X}\times\mathcal{Y})^2)$, such that $d(x,x') \leq \varepsilon$ and $y = y'$ γ-almost surely, and, $\Pi_{1\sharp}\gamma = \mathbb{P}$ and $\Pi_{2\sharp}\gamma = \mathbb{Q}$. Then: $\mathbb{E}_{\theta\sim\mu}\left[l(\theta,(x',y'))\right] \leq \sup_{(u,v),\ d(x,u)\leq\varepsilon, y=v} \mathbb{E}_{\theta\sim\mu}\left[l(\theta,(u,v))\right]$ γ-almost surely. Then, we deduce that:

$$\begin{aligned}
\mathbb{E}_{(x',y')\sim\mathbb{Q}}\left[\mathbb{E}_{\theta\sim\mu}\left[l(\theta,(x',y'))\right]\right] &= \mathbb{E}_{(x,y,x',y')\sim\gamma}\left[\mathbb{E}_{\theta\sim\mu}\left[l(\theta,(x',y'))\right]\right] \\
&\leq \mathbb{E}_{(x,y,x',y')\sim\gamma}\left[\sup_{(u,v),\ d(x,u)\leq\varepsilon, y=v} \mathbb{E}_{\theta\sim\mu}\left[l(\theta,(u,v))\right]\right] \\
&\leq \mathbb{E}_{(x,y)\sim\mathbb{P}}\left[\sup_{(u,v),\ d(x,u)\leq\varepsilon, y=v} \mathbb{E}_{\theta\sim\mu}\left[l(\theta,(u,v))\right]\right]
\end{aligned}$$

Then we deduce the expected result:

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

Let us show that the optimum is attained. $\mathbb{Q} \mapsto \mathbb{E}_{(x,y)\sim\mathbb{Q}}\left[\mathbb{E}_{\theta\sim\mu}\left[l(\theta,(x,y))\right]\right]$ is upper semi continuous by Lemma 3 for the weak topology of measures, and $\mathcal{A}_\varepsilon(\mathbb{P})$ is compact by Proposition 1, then by (Bertsekas & Shreve, 2004, Proposition 7.32), the supremum is attained for a certain $\mathbb{Q}^* \in \mathcal{A}_\varepsilon(\mathbb{P})$.

□

C.3. Proof of Theorem 1

Let us first recall the Fan's Theorem.

Theorem 2. *Let U be a compact convex Haussdorff space and V be convex space (not necessarily topological). Let $\psi : U \times V \to \mathbb{R}$ be a concave-convex function such that for all $v \in V$, $\psi(\cdot, v)$ is upper semi-continuous then:*

$$\inf_{v \in V} \max_{u \in U} \psi(u, v) = \max_{u \in U} \inf_{v \in V} \psi(u, v)$$

We are now set to prove Theorem 1.

Proof. $\mathcal{A}_\varepsilon(\mathbb{P})$, endowed with the weak topology of measures, is a Hausdorff compact convex space, thanks to Proposition 1. Moreover, $\mathcal{M}_+^1(\Theta)$ is clearly convex and $(\mathbb{Q}, \mu) \mapsto \int l d\mu d\mathbb{Q}$ is bilinear, hence concave-convex. Moreover thanks to Lemma 3, for all μ, $\mathbb{Q} \mapsto \int l d\mu d\mathbb{Q}$ is upper semi-continuous. Then Fan's theorem applies and strong duality holds. □

In the related work (Section 6), we mentioned a particular form of Theorem 1 for convex cases. As mentioned, this result has limited impact in the adversarial classification setting. It is still a direct corollary of Fan's theorem. This theorem can be stated as follows:

Theorem 3. *Let $\mathbb{P} \in \mathcal{M}_+^1(\mathcal{X} \times \mathcal{Y})$, $\varepsilon > 0$ and Θ a convex set. Let l be a loss satisfying Assumption 1, and also, $(x, y) \in \mathcal{X} \times \mathcal{Y}$, $l(\cdot, (x, y))$ is a convex function, then we have the following:*

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

The supremum is always attained. If Θ is a compact set then, the infimum is also attained.

C.4. Proof of Proposition 3

Proof. Let us first show that for $\alpha \geq 0$, $\sup_{\mathbb{Q}_i \in \Gamma_{i,\varepsilon}} \mathbb{E}_{\mathbb{Q}_i, \mu} \left[l(\theta, (x, y)) \right] - \alpha \mathrm{KL} \left(\mathbb{Q}_i \middle|\middle| \frac{1}{N} \mathbb{U}_{(x_i, y_i)} \right)$ admits a solution. Let $\alpha \geq 0$, $(\mathbb{Q}_{\alpha,i}^n)_{n \geq 0}$ a sequence such that

$$\mathbb{E}_{\mathbb{Q}_{\alpha,i}^n, \mu} \left[l(\theta, (x, y)) \right] - \alpha \mathrm{KL} \left(\mathbb{Q}_{\alpha,i}^n \middle|\middle| \frac{1}{N} \mathbb{U}_{(x_i, y_i)} \right) \xrightarrow[n \to +\infty]{} \sup_{\mathbb{Q}_i \in \Gamma_{i,\varepsilon}} \mathbb{E}_{\mathbb{Q}_i, \mu} \left[l(\theta, (x, y)) \right] - \alpha \mathrm{KL} \left(\mathbb{Q}_i \middle|\middle| \frac{1}{N} \mathbb{U}_{(x_i, y_i)} \right).$$

As $\Gamma_{i,\varepsilon}$ is tight ($(\mathcal{X}, d)$ is a proper metric space therefore all the closed ball are compact) and by Prokhorov's theorem, we can extract a subsequence which converges toward $\mathbb{Q}_{\alpha,i}^*$. Moreover, l is upper semi-continuous (u.s.c), thus $\mathbb{Q} \to \mathbb{E}_{\mathbb{Q}, \mu} \left[l(\theta, (x, y)) \right]$ is also u.s.c.[8] Moreover $\mathbb{Q} \to -\alpha \mathrm{KL} \left(\mathbb{Q} \middle|\middle| \frac{1}{N} \mathbb{U}_{(x_i, y_i)} \right)$ is also u.s.c. [9], therefore, by considering the limit superior as n goes to infinity we obtain that

$$\begin{aligned}
&\limsup_{n \to +\infty} \mathbb{E}_{\mathbb{Q}_{\alpha,i}^n, \mu} \left[l(\theta, (x, y)) \right] - \alpha \mathrm{KL} \left(\mathbb{Q}_{\alpha,i}^n \middle|\middle| \frac{1}{N} \mathbb{U}_{(x_i, y_i)} \right) \\
&= \sup_{\mathbb{Q}_i \in \Gamma_{i,\varepsilon}} \mathbb{E}_{\mathbb{Q}_i, \mu} \left[l(\theta, (x, y)) \right] - \alpha \mathrm{KL} \left(\mathbb{Q}_i \middle|\middle| \frac{1}{N} \mathbb{U}_{(x_i, y_i)} \right) \\
&\leq \mathbb{E}_{\mathbb{Q}_{\alpha,i}^*, \mu} \left[l(\theta, (x, y)) \right] - \alpha \mathrm{KL} \left(\mathbb{Q}_{\alpha,i}^* \middle|\middle| \frac{1}{N} \mathbb{U}_{(x_i, y_i)} \right)
\end{aligned}$$

from which we deduce that $\mathbb{Q}_{\alpha,i}^*$ is optimal.

[8] Indeed by considering a decreasing sequence of continuous and bounded functions which converge towards $\mathbb{E}_\mu \left[l(\theta, (x, y)) \right]$ and by definition of the weak convergence the result follows.

[9] for $\alpha = 0$ the result is clear, and if $\alpha > 0$, note that $\mathrm{KL} \left(\cdot \middle|\middle| \frac{1}{N} \mathbb{U}_{(x_i, y_i)} \right)$ is lower semi-continuous

Let us now show the result. We consider a positive sequence of $(\alpha_i^{(\ell)})_{\ell\geq 0}$ such that $\alpha_i^{(\ell)} \to 0$. Let us denote $\mathbb{Q}^*_{\alpha_i^{(\ell)},i}$ and $\mathbb{Q}_i^*$ the solutions of $\max_{\mathbb{Q}_i\in\Gamma_{i,\varepsilon}} \mathbb{E}_{\mathbb{Q}_i,\mu}\left[l(\theta,(x,y))\right] - \alpha_i^{(\ell)}\mathrm{KL}\left(\mathbb{Q}_i \middle|\middle| \frac{1}{N}\mathbb{U}_{(x_i,y_i)}\right)$ and $\max_{\mathbb{Q}_i\in\Gamma_{i,\varepsilon}} \mathbb{E}_{\mathbb{Q}_i,\mu}\left[l(\theta,(x,y))\right]$ respectively. Since $\Gamma_{i,\varepsilon}$ is tight, $(\mathbb{Q}^*_{\alpha_i^{(\ell)},i})_{\ell\geq 0}$ is also tight and we can extract by Prokhorov's theorem a subsequence which converges towards $\mathbb{Q}^*$. Moreover we have

$$\mathbb{E}_{\mathbb{Q}_i^*,\mu}\left[l(\theta,(x,y))\right] - \alpha_i^{(\ell)}\mathrm{KL}\left(\mathbb{Q}_i^* \middle|\middle| \frac{1}{N}\mathbb{U}_{(x_i,y_i)}\right) \leq \mathbb{E}_{\mathbb{Q}^*_{\alpha_i^{(\ell)},i},\mu}\left[l(\theta,(x,y))\right] - \alpha_i^{(\ell)}\mathrm{KL}\left(\mathbb{Q}^*_{\alpha_i^{(\ell)},i} \middle|\middle| \frac{1}{N}\mathbb{U}_{(x_i,y_i)}\right)$$

from which follows that

$$0 \leq \mathbb{E}_{\mathbb{Q}_i^*,\mu}\left[l(\theta,(x,y))\right] - \mathbb{E}_{\mathbb{Q}^*_{\alpha_i^{(\ell)},i},\mu}\left[l(\theta,(x,y))\right] \leq \alpha_i^{(\ell)}\left(\mathrm{KL}\left(\mathbb{Q}_i^* \middle|\middle| \frac{1}{N}\mathbb{U}_{(x_i,y_i)}\right) - \mathrm{KL}\left(\mathbb{Q}^*_{\alpha_i^{(\ell)},i} \middle|\middle| \frac{1}{N}\mathbb{U}_{(x_i,y_i)}\right)\right)$$

Then by considering the limit superior we obtain that

$$\limsup_{\ell\to+\infty} \mathbb{E}_{\mathbb{Q}^*_{\alpha_i^{(\ell)},i},\mu}\left[l(\theta,(x,y))\right] = \mathbb{E}_{\mathbb{Q}_i^*,\mu}\left[l(\theta,(x,y))\right].$$

from which follows that

$$\mathbb{E}_{\mathbb{Q}_i^*,\mu}\left[l(\theta,(x,y))\right] \leq \mathbb{E}_{\mathbb{Q}^*,\mu}\left[l(\theta,(x,y))\right]$$

and by optimality of $\mathbb{Q}_i^*$ we obtain the desired result. □

C.5. Proof of Proposition 4

Proof. Let us denote for all $\mu \in \mathcal{M}_1^+(\Theta)$,

$$\widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}}(\mu) := \sum_{i=1}^N \frac{\alpha_i}{N}\log\left(\frac{1}{m_i}\sum_{j=1}^{m_i}\exp\frac{\mathbb{E}_\mu\left[l(\theta,u_j^{(i)})\right]}{\alpha_i}\right).$$

Let also consider $(\mu_n^{(\mathbf{m})})_{n\geq 0}$ and $(\mu_n)_{n\geq 0}$ two sequences such that

$$\widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}}(\mu_n^{(\mathbf{m})}) \xrightarrow[n\to+\infty]{} \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}}, \quad \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon}}_{adv,\boldsymbol{\alpha}}(\mu_n) \xrightarrow[n\to+\infty]{} \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},*}_{adv,\boldsymbol{\alpha}}.$$

We first remarks that

$$\begin{aligned}\widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}} - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},*}_{adv,\boldsymbol{\alpha}} &\leq \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}} - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}}(\mu_n) + \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}}(\mu_n) - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon}}_{adv,\boldsymbol{\alpha}}(\mu_n) + \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon}}_{adv,\boldsymbol{\alpha}}(\mu_n) - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},*}_{adv,\boldsymbol{\alpha}} \\ &\leq \sup_{\mu\in\mathcal{M}_1^+(\Theta)}\left|\widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}}(\mu) - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon}}_{adv,\boldsymbol{\alpha}}(\mu)\right| + \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon}}_{adv,\boldsymbol{\alpha}}(\mu_n) - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},*}_{adv,\boldsymbol{\alpha}},\end{aligned}$$

and by considering the limit, we obtain that

$$\widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}} - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},*}_{adv,\boldsymbol{\alpha}} \leq \sup_{\mu\in\mathcal{M}_1^+(\Theta)}\left|\widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}}(\mu) - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon}}_{adv,\boldsymbol{\alpha}}(\mu)\right|$$

Simarly we have that

$$\widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},*}_{adv,\boldsymbol{\alpha}} - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}} \leq \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},*}_{adv,\boldsymbol{\alpha}} - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon}}_{adv,\boldsymbol{\alpha}}(\mu_n^{(\boldsymbol{m})}) + \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon}}_{adv,\boldsymbol{\alpha}}(\mu_n^{(\boldsymbol{m})}) - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}}(\mu_n^{(\boldsymbol{m})}) + \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}}(\mu_n^{(\boldsymbol{m})}) - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}}$$

from which follows that

$$\widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},*}_{adv,\boldsymbol{\alpha}} - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}} \leq \sup_{\mu\in\mathcal{M}_1^+(\Theta)}\left|\widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}}(\mu) - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon}}_{adv,\boldsymbol{\alpha}}(\mu)\right|$$

Therefore we obtain that

$$\left|\widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},*}_{adv,\boldsymbol{\alpha}} - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}}\right| \leq \sum_{i=1}^{N} \frac{\alpha}{N} \sup_{\mu\in\mathcal{M}_1^+(\Theta)} \Big| \log\left(\frac{1}{m_i}\sum_{j=1}^{m_i} \exp\left(\frac{\mathbb{E}_{\theta\sim\mu}\left[l(\theta,u_j^{(i)}))\right]}{\alpha}\right)\right) - \log\left(\int_{\mathcal{X}\times\mathcal{Y}} \exp\left(\frac{\mathbb{E}_{\theta\sim\mu}\left[l(\theta,(x,y))\right]}{\alpha}\right) d\mathbb{U}_{(x_i,y_i)}\right)\Big|.$$

Observe that $l \geq 0$, therefore because the $\log$ function is 1-Lipschitz on $[1,+\infty)$, we obtain that

$$\left|\widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},*}_{adv,\boldsymbol{\alpha}} - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}}\right| \leq \sum_{i=1}^{N} \frac{\alpha}{N} \sup_{\mu\in\mathcal{M}_1^+(\Theta)} \Big|\frac{1}{m_i}\sum_{j=1}^{m_i} \exp\left(\frac{\mathbb{E}_{\theta\sim\mu}\left[l(\theta,u_j^{(i)}))\right]}{\alpha}\right) - \int_{\mathcal{X}\times\mathcal{Y}} \exp\left(\frac{\mathbb{E}_{\theta\sim\mu}\left[l(\theta,(x,y))\right]}{\alpha}\right) d\mathbb{U}_{(x_i,y_i)}\Big|.$$

Let us now denote for all $i = 1,\dots,N$,

$$\widehat{R}_i(\mu,\boldsymbol{u}^{(i)}) := \sum_{j=1}^{m_i} \exp\left(\frac{\mathbb{E}_{\theta\sim\mu}\left[l(\theta,u_j^{(i)}))\right]}{\alpha}\right)$$
$$R_i(\mu) := \int_{\mathcal{X}\times\mathcal{Y}} \exp\left(\frac{\mathbb{E}_{\theta\sim\mu}\left[l(\theta,(x,y))\right]}{\alpha}\right) d\mathbb{U}_{(x_i,y_i)}.$$

and let us define

$$f(\boldsymbol{u}^{(1)},\dots,\boldsymbol{u}^{(N)}) := \sum_{i=1}^{N} \frac{\alpha}{N} \sup_{\mu\in\mathcal{M}_1^+(\Theta)} \left|\widehat{R}_i(\mu) - R_i(\mu)\right|$$

where $\boldsymbol{u}^{(i)} := (u_1^{(i)},\dots,u_1^{(m)})$. By denoting $z^{(i)} = (u_1^{(i)},\dots,u_{k-1}^{(i)},z,u_{k+1}^{(i)},\dots,u_m^{(i)})$, we have that

$$\begin{aligned}
|f(\boldsymbol{u}^{(1)},\dots,\boldsymbol{u}^{(N)}) - f(\boldsymbol{u}^{(1)},\dots,\boldsymbol{u}^{(i-1)},\boldsymbol{z}^{(i)},\boldsymbol{u}^{(i+1)},\dots,\boldsymbol{u}^{(N)})| &\leq \frac{\alpha}{N}\Big| \sup_{\mu\in\mathcal{M}_1^+(\Theta)} \left|\widehat{R}_i(\mu,\boldsymbol{u}^{(i)}) - R_i(\mu)\right| \\
&\quad - \sup_{\mu\in\mathcal{M}_1^+(\Theta)} \left|\widehat{R}_i(\mu,\boldsymbol{z}^{(i)}) - R_i(\mu)\right|\Big| \\
&\leq \frac{\alpha}{N}\Big|\frac{1}{m}\left[\exp\left(\frac{\mathbb{E}_{\theta\sim\mu}\left[l(\theta,u_k^{(i)}))\right]}{\alpha}\right) - \exp\left(\frac{\mathbb{E}_{\theta\sim\mu}\left[l(\theta,z^{(i)}))\right]}{\alpha}\right)\right]\Big| \\
&\leq \frac{2\exp(M/\alpha)}{Nm}
\end{aligned}$$

where the last inequality comes from the fact that the loss is upper bounded by $l \leq M$. Then by appling the McDiarmid's Inequality, we obtain that with a probability of at least $1-\delta$,

$$\left|\widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},*}_{adv,\boldsymbol{\alpha}} - \widehat{\mathcal{R}}^{\boldsymbol{\varepsilon},\mathbf{m}}_{adv,\boldsymbol{\alpha}}\right| \leq \mathbb{E}(f(\boldsymbol{u}^{(1)},\dots,\boldsymbol{u}^{(N)})) + \frac{2\exp(M/\alpha)}{\sqrt{mN}}\sqrt{\frac{\log(2/\delta)}{2}}.$$

Thanks to (Shalev-Shwartz & Ben-David, 2014, Lemma 26.2), we have for all $i \in \{1,\dots,N\}$

$$\mathbb{E}(f(\boldsymbol{u}^{(1)},\dots,\boldsymbol{u}^{(N)})) \leq 2\mathbb{E}(\mathrm{Rad}(\mathcal{F}_i \circ \mathbf{u^{(i)}}))$$

where for any class of function $\mathcal{F}$ defined on $\mathcal{Z}$ and point $\boldsymbol{z} : (z_1,\dots,z_q) \in \mathcal{Z}^q$

$$\mathcal{F}\circ\boldsymbol{z} := \Big\{(f(z_1),\dots,f(z_q)),\ f\in\mathcal{F}\Big\} \quad , \quad \mathrm{Rad}(\mathcal{F}\circ\boldsymbol{z}) := \frac{1}{q}\mathbb{E}_{\boldsymbol{\sigma}\sim\{\pm1\}}\left[\sup_{f\in\mathcal{F}}\sum_{i=1}^{q}\sigma_i f(z_i)\right]$$
$$\mathcal{F}_i := \Big\{u \to \exp\left(\frac{\mathbb{E}_{\theta\sim\mu}\left[l(\theta,u))\right]}{\alpha}\right),\ \mu\in\mathcal{M}_1^+(\Theta)\Big\}.$$

Moreover as $x \to \exp(x/\alpha)$ is $\frac{\exp(M/\alpha)}{\alpha}$-Lipstchitz on $(-\infty, M]$, by (Shalev-Shwartz & Ben-David, 2014, Lemma 26.9), we have

$$\text{Rad}(\mathcal{F}_i \circ \mathbf{u^{(i)}}) \leq \frac{\exp(M/\alpha)}{\alpha} \text{Rad}(\mathcal{H}_i \circ \mathbf{u^{(i)}})$$

where

$$\mathcal{H}_i := \left\{ u \to \mathbb{E}_{\theta \sim \mu} \left[l(\theta, u) \right], \ \mu \in \mathcal{M}_1^+(\Theta) \right\}.$$

Let us now define

$$g(\boldsymbol{u}^{(1)}, \dots, \boldsymbol{u}^{(N)}) := \sum_{j=1}^{N} \frac{2\exp(M/\alpha)}{N} \text{Rad}(\mathcal{H}_j \circ \mathbf{u^{(j)}}).$$

We observe that

$$\begin{aligned} |g(\boldsymbol{u}^{(1)}, \dots, \boldsymbol{u}^{(N)}) - g(\boldsymbol{u}^{(1)}, \dots, \boldsymbol{u}^{(i-1)}, \boldsymbol{z}^{(i)}, \boldsymbol{u}^{(i+1)}, \dots, \boldsymbol{u}^{(N)})| &\leq \frac{2\exp(M/\alpha)}{N} |\text{Rad}(\mathcal{H}_i \circ \mathbf{u^{(i)}}) - \text{Rad}(\mathcal{H}_i \circ \mathbf{z^{(i)}})| \\ &\leq \frac{2\exp(M/\alpha)}{N} \frac{2M}{m}. \end{aligned}$$

By Applying the McDiarmid's Inequality, we have that with a probability of at least $1 - \delta$

$$\mathbb{E}(g(\boldsymbol{u}^{(1)}, \dots, \boldsymbol{u}^{(N)})) \leq g(\boldsymbol{u}^{(1)}, \dots, \boldsymbol{u}^{(N)}) + \frac{4\exp(M/\alpha)M}{\sqrt{mN}} \sqrt{\frac{\log(2/\delta)}{2}}.$$

Remarks also that

$$\begin{aligned} \text{Rad}(\mathcal{H}_i \circ \mathbf{u^{(i)}}) &= \frac{1}{m} \mathbb{E}_{\boldsymbol{\sigma} \sim \{\pm 1\}} \left[\sup_{\mu \in \mathcal{M}_1^+(\Theta)} \sum_{j=1}^{m} \sigma_i \mathbb{E}_{\mu}(l(\theta, u_j^{(i)})) \right] \\ &= \frac{1}{m} \mathbb{E}_{\boldsymbol{\sigma} \sim \{\pm 1\}} \left[\sup_{\theta \in \Theta} \sum_{j=1}^{m} \sigma_i l(\theta, u_j^{(i)}) \right] \end{aligned}$$

Finally, applying a union bound leads to the desired result.

□

C.6. Proof of Proposition 5

Proof. Following the same steps than the proof of Proposition 4, let $(\mu_n^{\varepsilon})_{n \geq 0}$ and $(\mu_n)_{n \geq 0}$ two sequences such that

$$\widehat{\mathcal{R}}_{adv,\boldsymbol{\alpha}}^{\varepsilon}(\mu_n^{\varepsilon}) \xrightarrow[n \to +\infty]{} \widehat{\mathcal{R}}_{adv,\boldsymbol{\alpha}}^{\varepsilon,*}, \quad \widehat{\mathcal{R}}_{adv}^{\varepsilon}(\mu_n) \xrightarrow[n \to +\infty]{} \widehat{\mathcal{R}}_{adv}^{\varepsilon,*}.$$

Remarks that

$$\begin{aligned} \widehat{\mathcal{R}}_{adv,\boldsymbol{\alpha}}^{\varepsilon,*} - \widehat{\mathcal{R}}_{adv}^{\varepsilon,*} &\leq \widehat{\mathcal{R}}_{adv,\boldsymbol{\alpha}}^{\varepsilon,*} - \widehat{\mathcal{R}}_{adv,\boldsymbol{\alpha}}^{\varepsilon}(\mu_n) + \widehat{\mathcal{R}}_{adv,\boldsymbol{\alpha}}^{\varepsilon}(\mu_n) - \widehat{\mathcal{R}}_{adv}^{\varepsilon}(\mu_n) + \widehat{\mathcal{R}}_{adv}^{\varepsilon}(\mu_n) - \widehat{\mathcal{R}}_{adv}^{\varepsilon,*} \\ &\leq \sup_{\mu \in \mathcal{M}_1^+(\Theta)} \left| \widehat{\mathcal{R}}_{adv,\boldsymbol{\alpha}}^{\varepsilon}(\mu) - \widehat{\mathcal{R}}_{adv}^{\varepsilon}(\mu) \right| + \widehat{\mathcal{R}}_{adv}^{\varepsilon}(\mu_n) - \widehat{\mathcal{R}}_{adv}^{\varepsilon,*} \end{aligned}$$

Then by considering the limit we obtain that

$$\widehat{\mathcal{R}}_{adv,\boldsymbol{\alpha}}^{\varepsilon,*} - \widehat{\mathcal{R}}_{adv}^{\varepsilon,*} \leq \sup_{\mu \in \mathcal{M}_1^+(\Theta)} \left| \widehat{\mathcal{R}}_{adv,\boldsymbol{\alpha}}^{\varepsilon}(\mu) - \widehat{\mathcal{R}}_{adv}^{\varepsilon}(\mu) \right|.$$

Similarly, we obtain that

$$\widehat{\mathcal{R}}^{\varepsilon,*}_{adv} - \widehat{\mathcal{R}}^{\varepsilon,*}_{adv,\boldsymbol{\alpha}} \leq \sup_{\mu \in \mathcal{M}_1^+(\Theta)} \left| \widehat{\mathcal{R}}^{\varepsilon}_{adv,\boldsymbol{\alpha}}(\mu) - \widehat{\mathcal{R}}^{\varepsilon}_{adv}(\mu) \right|,$$

from which follows that

$$\left| \widehat{\mathcal{R}}^{\varepsilon,*}_{adv,\boldsymbol{\alpha}} - \widehat{\mathcal{R}}^{\varepsilon,*}_{adv} \right| \leq \frac{1}{N} \sum_{i=1}^{N} \sup_{\mu \in \mathcal{M}_1^+(\Theta)} \left| \alpha \log \left(\int_{\mathcal{X}\times\mathcal{Y}} \exp\left(\frac{\mathbb{E}_\mu[l(\theta,(x,y))]}{\alpha} \right) d\mathbb{U}_{(x_i,y_i)} \right) - \sup_{u \in S^{\varepsilon}_{(x_i,y_i)}} \mathbb{E}_\mu[l(\theta,u)] \right|.$$

Let $\mu \in \mathcal{M}_1^+(\Theta)$ and $i \in \{1,\dots,N\}$, then we have

$$\begin{aligned}
&\left| \alpha \log \left(\int_{\mathcal{X}\times\mathcal{Y}} \exp\left(\frac{\mathbb{E}_\mu[l(\theta,(x,y))]}{\alpha} \right) d\mathbb{U}_{(x_i,y_i)} \right) - \sup_{u \in S^{\varepsilon}_{(x_i,y_i)}} \mathbb{E}_\mu[l(\theta,u)] \right| \\
&= \left| \alpha \log \left(\int_{\mathcal{X}\times\mathcal{Y}} \exp\left(\frac{\mathbb{E}_\mu[l(\theta,(x,y))] - \sup_{u \in S^{\varepsilon}_{(x_i,y_i)}} \mathbb{E}_\mu[l(\theta,u)]}{\alpha} \right) d\mathbb{U}_{(x_i,y_i)} \right) \right| \\
&= \alpha \Bigg| \log \left(\int_{A^{(x_i,y_i)}_{\beta,\mu}} \exp\left(\frac{\mathbb{E}_\mu[l(\theta,(x,y))] - \sup_{u \in S^{\varepsilon}_{(x_i,y_i)}} \mathbb{E}_\mu[l(\theta,u)]}{\alpha} \right) d\mathbb{U}_{(x_i,y_i)} \right. \\
&\quad \left. + \int_{(A^{(x_i,y_i)}_{\beta,\mu})^c} \exp\left(\frac{\mathbb{E}_\mu[l(\theta,(x,y))] - \sup_{u \in S^{\varepsilon}_{(x_i,y_i)}} \mathbb{E}_\mu[l(\theta,u)]}{\alpha} \right) d\mathbb{U}_{(x_i,y_i)} \right) \Bigg| \\
&\leq \alpha \left| \log \left(\exp(-\beta/\alpha) \mathbb{U}_{(x_i,y_i)} \left(A^{(x_i,y_i)}_{\beta,\mu} \right) \right) \right| \\
&\quad + \alpha \left| \log \left(1 + \frac{\exp(\beta/\alpha)}{\mathbb{U}_{(x_i,y_i)} \left(A^{(x_i,y_i)}_{\beta,\mu} \right)} \int_{(A^{(x_i,y_i)}_{\beta,\mu})^c} \exp\left(\frac{\mathbb{E}_\mu[l(\theta,(x,y))] - \sup_{u \in S^{\varepsilon}_{(x_i,y_i)}} \mathbb{E}_\mu[l(\theta,u)]}{\alpha} \right) d\mathbb{U}_{(x_i,y_i)} \right) \right| \\
&\leq \alpha \log(1/C_\beta) + \beta + \frac{\alpha}{C_\beta} \\
&\leq 2\alpha \log(1/C_\beta) + \beta
\end{aligned}$$

□

C.7. Proof of Proposition 6

Proof. Thanks to Danskin theorem, if $\mathbb{Q}^*$ is a best response to $\boldsymbol{\lambda}$, then $\boldsymbol{g}^* := \left(\mathbb{E}_{\mathbb{Q}^*}\left[l(\theta_1,(x,y)) \right], \dots, \mathbb{E}_{\mathbb{Q}^*}\left[l(\theta_L,(x,y)) \right] \right)^T$ is a subgradient of $\boldsymbol{\lambda} \to \mathcal{R}^{\varepsilon}_{adv}(\boldsymbol{\lambda})$. Let $\eta \geq 0$ be the learning rate. Then we have for all $t \geq 1$:

$$\begin{aligned}
\|\boldsymbol{\lambda}_t - \boldsymbol{\lambda}^*\|^2 &\leq \|\boldsymbol{\lambda}_{t-1} - \eta \boldsymbol{g}_t - \boldsymbol{\lambda}^*\|^2 \\
&= \|\boldsymbol{\lambda}_{t-1} - \boldsymbol{\lambda}^*\|^2 - 2\eta \langle \boldsymbol{g}_t, \boldsymbol{\lambda}_{t-1} - \boldsymbol{\lambda}^* \rangle + \eta^2 \|\boldsymbol{g}_t\|^2 \\
&\leq \|\boldsymbol{\lambda}_{t-1} - \boldsymbol{\lambda}^*\|^2 - 2\eta \langle \boldsymbol{g}^*_t, \boldsymbol{\lambda}_{t-1} - \boldsymbol{\lambda}^* \rangle + 2\eta \langle \boldsymbol{g}^*_t - \boldsymbol{g}_t, \boldsymbol{\lambda}_{t-1} - \boldsymbol{\lambda}^* \rangle + \eta^2 M^2 L \\
&\leq \|\boldsymbol{\lambda}_{t-1} - \boldsymbol{\lambda}^*\|^2 - 2\eta \left(\mathcal{R}^{\varepsilon}_{adv}(\boldsymbol{\lambda}_t) - \mathcal{R}^{\varepsilon}_{adv}(\boldsymbol{\lambda}^*) \right) + 4\eta\delta + \eta^2 M^2 L
\end{aligned}$$

We then deduce by summing:

$$2\eta \sum_{t=1}^{T} \mathcal{R}^{\varepsilon}_{adv}(\boldsymbol{\lambda}_t) - \mathcal{R}^{\varepsilon}_{adv}(\boldsymbol{\lambda}^*) \leq 4\delta\eta T + \|\boldsymbol{\lambda}_0 - \boldsymbol{\lambda}^*\|^2 + \eta^2 M^2 L T$$

Then we have:

$$\min_{t \in [T]} \mathcal{R}^{\varepsilon}_{adv}(\boldsymbol{\lambda}_t) - \mathcal{R}^{\varepsilon}_{adv}(\boldsymbol{\lambda}^*) \leq 2\delta + \frac{4}{\eta T} + M^2 L \eta$$

The left-hand term is minimal for $\eta = \frac{2}{M\sqrt{LT}}$, and for this value:

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

□

.

D. Additional Experimental Results

D.1. Experimental setting.

Optimizer. For each of our models, The optimizer we used in all our implementations is SGD with learning rate set to 0.4 at epoch 0 and is divided by 10 at half training then by 10 at the three quarters of training. The momentum is set to 0.9 and the weight decay to 5×10^{-4}. The batch size is set to 1024.

Adaptation of Attacks. Since our classifier is randomized, we need to adapt the attack accordingly. To do so we used the expected loss:

$$\tilde{l}\left((\boldsymbol{\lambda}, \boldsymbol{\theta}), (x, y)\right) = \sum_{k=1}^{L} \lambda_k l(\theta_k, (x, y))$$

to compute the gradient in the attacks, regardless the loss (DLR or cross-entropy). For the inner maximization at training time, we used a PGD attack on the cross-entropy loss with $\varepsilon = 0.03$. For the final evaluation, we used the untargeted DLR attack with default parameters.

Regularization in Practice. The entropic regularization in higher dimensional setting need to be adapted to be more likely to find adversaries. To do so, we computed PGD attacks with only 3 iterations with 5 different restarts instead of sampling uniformly 5 points in the ℓ_∞-ball. In our experiments in the main paper, we use a regularization parameter $\alpha = 0.001$. The learning rate for the minimization on $\boldsymbol{\lambda}$ is always fixed to 0.001.

Alternate Minimization Parameters. Algorithm 2 implies an alternate minimization algorithm. We set the number of updates of $\boldsymbol{\theta}$ to $T_\theta = 50$ and, the update of $\boldsymbol{\lambda}$ to $T_\lambda = 25$.

D.2. Effect of the Regularization

In this subsection, we experimentally investigate the effect of the regularization. In Figure 4, we notice, that the regularization has the effect of stabilizing, reducing the variance and improving the level of the robust accuracy for adversarial training for mixtures (Algorithm 2). The standard accuracy curves are very similar in both cases.

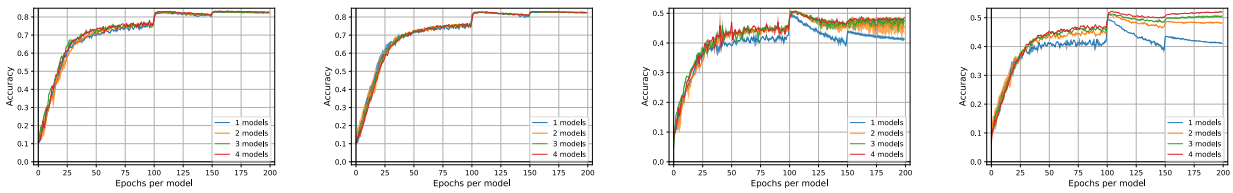

Figure 4. On left and middle-left: Standard accuracies over epochs with respectively no regularization and regularization set to $\alpha = 0.001$. On middle right and right: Robust accuracies for the same parameters against PGD attack with 20 iterations and $\varepsilon = 0.03$.

D.3. Additional Experiments on WideResNet28x10

We now evaluate our algorithm on WideResNet28x10 (Zagoruyko & Komodakis, 2016) architecture. Due to computation costs, we limit ourselves to 1 and 2 models, with regularization parameter set to 0.001 as in the paper experiments section.

Results are reported in Figure 5. We remark this architecture can lead to more robust models, corroborating the results from (Gowal et al., 2020).

Models	Acc.	APGD_{CE}	APGD_{DLR}	Rob. Acc.
1	85.2%	49.9%	50.2%	48.5%
2	**86.0%**	**51.5%**	**52.1%**	**49.6%**

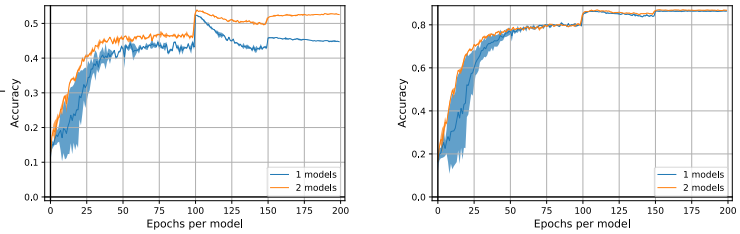

Figure 5. On left: Comparison of our algorithm with a standard adversarial training (one model) on WideResNet28x10. We reported the results for the model with the best robust accuracy obtained over two independent runs because adversarial training might be unstable. Standard and Robust accuracy (respectively in the middle and on right) on CIFAR-10 test images in function of the number of epochs per classifier with 1 and 2 WideResNet28x10 models. The performed attack is PGD with 20 iterations and $\varepsilon = 8/255$.

D.4. Overfitting in Adversarial Robustness

We further investigate the overfitting of our heuristic algorithm. We plotted in Figure 6 the robust accuracy on ResNet18 with 1 to 5 models. The most robust mixture of 5 models against PGD with 20 iterations arrives at epoch 198, *i.e.* at the end of the training, contrary to 1 to 4 models, where the most robust mixture occurs around epoch 101. However, the accuracy against AGPD with 100 iterations in lower than the one at epoch 101 with global robust accuracy of 47.6% at epoch 101 and 45.3% at epoch 198. This strange phenomenon would suggest that the more powerful the attacks are, the more the models are subject to overfitting. We leave this question to further works.

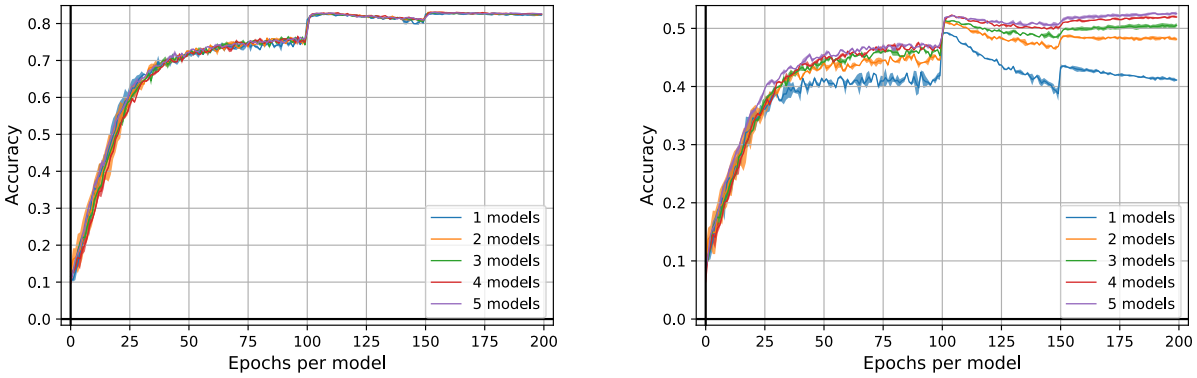

Figure 6. Standard and Robust accuracy (respectively on left and on right) on CIFAR-10 test images in function of the number of epochs per classifier with 1 to 5 ResNet18 models. The performed attack is PGD with 20 iterations and $\varepsilon = 8/255$. The best mixture for 5 models occurs at the end of training (epoch 198).

E. Additional Results

E.1. Equality of Standard Randomized and Deterministic Minimal Risks

Proposition 7. *Let $\mathbb{P}$ be a Borel probability distribution on $\mathcal{X} \times \mathcal{Y}$, and l a loss satisfying Assumption 1, then:*

$$\inf_{\mu \in \mathcal{M}_+^1(\Theta)} \mathcal{R}(\mu) = \inf_{\theta \in \Theta} \mathcal{R}(\theta)$$

Proof. It is clear that: $\inf_{\mu\in\mathcal{M}_+^1(\Theta)}\mathcal{R}(\mu)\leq\inf_{\theta\in\Theta}\mathcal{R}(\theta)$. Now, let $\mu\in\mathcal{M}_+^1(\Theta)$, then:

$$\begin{aligned}\mathcal{R}(\mu)=\mathbb{E}_{\theta\sim\mu}(\mathcal{R}(\theta))&\geq\operatorname*{essinf}_{\mu}\mathbb{E}_{\theta\sim\mu}\left(\mathcal{R}(\theta)\right)\\&\geq\inf_{\theta\in\Theta}\mathcal{R}(\theta).\end{aligned}$$

where essinf denotes the essential infimum. □

We can deduce an immediate corollary.

Corollary 2. *Under Assumption 1, the dual for randomized and deterministic classifiers are equal.*

E.2. Decomposition of the Empirical Risk for Entropic Regularization

Proposition 8. *Let $\hat{\mathbb{P}}:=\frac{1}{N}\sum_{i=1}^N\delta_{(x_i,y_i)}$. Let l be a loss satisfying Assumption 1. Then we have:*

$$\frac{1}{N}\sum_{i=1}^N\sup_{x,\ d(x,x_i)\leq\varepsilon}\mathbb{E}_{\theta\sim\mu}\left[l(\theta,(x,y))\right]=\sum_{i=1}^N\sup_{\mathbb{Q}_i\in\Gamma_{i,\varepsilon}}\mathbb{E}_{(x,y)\sim\mathbb{Q}_i,\theta\sim\mu}\left[l(\theta,(x,y))\right]$$

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

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

Proof. This proposition is a direct application of Proposition 2 for diracs $\delta_{(x_i,y_i)}$. □

E.3. On the NP-Hardness of Attacking a Mixture of Classifiers

In general, the problem of finding a best response to a mixture of classifiers is in general NP-hard. Let us justify it on a mixture of linear classifiers in binary classification: $f_{\theta_k}(x)=\langle\theta,x\rangle$ for $k\in[L]$ and $\boldsymbol{\lambda}=\mathbf{1}_L/L$. Let us consider the ℓ_2 norm and $x=0$ and $y=1$. Then the problem of attacking x is the following:

$$\sup_{\tau,\ \|\tau\|\leq\varepsilon}\frac{1}{L}\sum_{k=1}^L\mathbf{1}_{\langle\theta_k,\tau\rangle\leq0}$$

This problem is equivalent to a linear binary classification problem on τ, which is known to be NP-hard.

E.4. Case of Separated Conditional Distribtions

Proposition 9. *Let $\mathcal{Y}=\{-1,+1\}$. Let $\mathbb{P}\in\mathcal{M}_+^1(\mathcal{X}\times\mathcal{Y})$. Let $\varepsilon>0$. For $i\in\mathcal{Y}$, let us denote $\mathbb{P}_i$ the distribution of $\mathbb{P}$ conditionally to $y=i$. Let us assume that $d_{\mathcal{X}}(\operatorname{supp}(\mathbb{P}_{1+1}),\operatorname{supp}(\mathbb{P}_{-1}))>2\varepsilon$. Let us consider the nearest neighbor deterministic classifier : $f(x)=d(x,\operatorname{supp}(\mathbb{P}_{+1}))-d(x,\operatorname{supp}(\mathbb{P}_{-1}))$ and the $0/1$ loss $l(f,(x,y))=\mathbf{1}_{yf(x)\leq0}$. Then f satisfies both optimal standard and adversarial risks: $\mathcal{R}(f)=0$ and $\mathcal{R}_{adv}^\varepsilon(f)=0$.*

Proof. Let Let denote $p_i=\mathbb{P}(y=i)$. Then we have

$$\mathcal{R}_{adv}^\varepsilon(f)=p_{+1}\mathbb{E}_{\mathbb{P}_{+1}}\left[\sup_{x',\ d(x,x')\leq\varepsilon}\mathbf{1}_{f(x')\leq0}\right]+p_{-1}\mathbb{E}_{\mathbb{P}_{-1}}\left[\sup_{x',\ d(x,x')\leq\varepsilon}\mathbf{1}_{f(x')\geq0}\right]$$

For $x\in\operatorname{supp}(\mathbb{P}_{+1})$, we have, for all x' such that $d(x,x')\neq0$, $f(x')>0$, then: $\mathbb{E}_{\mathbb{P}_{+1}}\left[\sup_{x',\ d(x,x')\leq\varepsilon}\mathbf{1}_{f(x')\leq0}\right]=0$. Similarly, we have $\mathbb{E}_{\mathbb{P}_{-1}}\left[\sup_{x',\ d(x,x')\leq\varepsilon}\mathbf{1}_{f(x')\geq0}\right]=0$. We then deduce the result. □