(Supplementary material) Limitations of Adversarial Robustness: Strong No Free Lunch Theorem

Elvis Dohmatob [1]

A. Proofs

Proof of claims on the toy problem from section 1.1. These claims were already proved in (Tsipras et al., 2018). We provide a proof here just for completeness.

Now, one computes

$$\begin{aligned}
&\mathrm{acc}(h_{\text{avg}}) := \mathbb{P}_{(X,Y)}\left(h_{\text{avg}}(X) = Y\right) = \mathbb{P}\left(Yw^TX \geq 0\right)\\
&= \mathbb{P}_Y\left((Y/(p-1))\sum_{j\geq 2}\mathcal{N}(\eta Y, 1) \geq 0\right)\\
&= \mathbb{P}\left(\mathcal{N}(\eta, 1/(p-1)) \geq 0\right) = \mathbb{P}\left(\mathcal{N}(0, 1/(p-1)) \geq -\eta\right)\\
&= \mathbb{P}\left(\mathcal{N}(0, 1/(p-1)) \leq \eta\right) \geq 1 - e^{-(p-1)\eta^2/2},
\end{aligned}$$

which is $\geq 1-\delta$ if $\eta \geq \sqrt{2\log(1/\delta)/(p-1)}$. Likewise, for $\epsilon \geq \eta$, it was shown in (Tsipras et al., 2018) that the adversarial robustness accuracy of h_{avg} writes

$$\begin{aligned}
&\mathrm{acc}_\epsilon(h_{\text{avg}}) := \mathbb{P}_{(X,Y)}\left(Yh_{\text{avg}}(X+\Delta x) \geq 0\ \forall \|\Delta x\|_\infty \leq \epsilon\right)\\
&= \mathbb{P}_{(X,Y)}\left(\inf_{\|\Delta x\|_\infty \leq \epsilon} Yw^T(X+\Delta x) \geq 0\right)\\
&= \mathbb{P}_{(X,Y)}\left(Yw^TX - \epsilon\|Yw\|_1 \geq 0\right)\\
&= \mathbb{P}_{(X,Y)}\left(Yw^TX - \epsilon \geq 0\right)\\
&= \mathbb{P}(\mathcal{N}(0, 1/(p-1)) \geq \epsilon - \eta) \leq e^{-(p-1)(\epsilon-\eta)^2/2}.
\end{aligned}$$

Thus $\mathrm{acc}_\epsilon(h_{\text{avg}}) \leq \delta$ for $\epsilon \geq \eta + \sqrt{2\log(1/\delta)/(p-1)}$, which completes the proof. □

Proof of Theorem 2. Let $h : \mathcal{X} \to \{1, \ldots, K\}$ be a classifier, and for a fixed class label $k \in \{1, 2, \ldots, K\}$, define the set $B(h,k) := \{x \in \mathcal{X} | h(x) \neq k\}$. Because we only consider $P_{X|Y}$-a.e continuous classifiers, each $B(h,k)$ is Borel. Conditioned on the event "$y = k$", the probability of $B(h,k)$ is precisely the average error made by the classifier h on the class label k. That is, $\mathrm{acc}(h|k) = 1 - P_{X|k}(B(h,k))$. Now, the assumptions imply by virtue of Lemma 1, that $P_{X|k}$ has the BLOWUP(c) property. Thus, if $\epsilon \geq \sigma_k\sqrt{2\log(1/(P_{X|Y}(B(h,k))} = \sigma_k\sqrt{2\log(1/\mathrm{err}(h|k)} =: \epsilon(h|k)$, then one has

$$\begin{aligned}
&\mathrm{acc}_\epsilon(h|k) = 1 - P_{X|k}(B(h,k)^\epsilon_{d_{\text{geo}}})\\
&\leq e^{-\frac{1}{2\sigma_k^2}(\epsilon - \sigma_k\sqrt{2\log(1/(P_{X|k}(B(h,k)))})^2}\\
&= e^{-\frac{1}{2\sigma_k^2}(\epsilon - \sigma_k\sqrt{2\log(1/\mathrm{err}(h|k))})^2} = e^{-\frac{1}{2\sigma_k^2}(\epsilon - \epsilon(h|k))^2}\\
&\leq e^{-\frac{1}{2\sigma_k^2}\epsilon(h|k)^2} = \mathrm{err}(h|k),\ \text{if } \epsilon \geq 2\epsilon(h|k).
\end{aligned}$$

On the other hand, it is clear that $\mathrm{acc}_\epsilon(h|k) \leq \mathrm{acc}(h|k)$ for any $\epsilon \geq 0$ since $B(h,k) \subseteq B(h,k)^\epsilon$ for any threat model. This concludes the proof of part *(A)*. For part *(B)*, define the random variable $Z := d(X, B(h,k))$ and note that

$$\begin{aligned}
&d(h|k) := \mathbb{E}_{X|k}[d(X, B(h,k))] = \int_0^\infty P_{X|k}(Z \geq \epsilon)d\epsilon\\
&= \int_0^{\epsilon(h|k)} P_{X|k}(Z \geq \epsilon)d\epsilon + \int_{\epsilon(h|k)}^\infty P_{X|k}(Z \geq \epsilon)d\epsilon\\
&\leq \epsilon(h|k) + \int_{\epsilon(h|k)}^\infty P_{X|k}(Z \geq \epsilon)d\epsilon,\ \text{ as } P_{X|k}(Z \geq \epsilon) \leq 1\\
&\leq \epsilon(h|k) + \int_{\epsilon(h|k)}^\infty e^{-\frac{1}{2\sigma_k^2}(\epsilon - \epsilon(h|k))^2}d\epsilon,\ \text{ by inequality (10)}\\
&= \epsilon(h|k) + \frac{\sigma_k\sqrt{2\pi}}{2}\left(\int_{-\infty}^\infty \frac{1}{\sigma_k\sqrt{2\pi}}e^{-\frac{1}{2\sigma_k^2}\epsilon^2}d\epsilon\right)\\
&= \epsilon(h|k) + \frac{\sigma_k\sqrt{2\pi}}{2} = \sigma_k\left(\sqrt{\log(1/\mathrm{err}(h|k))} + \sqrt{\frac{\pi}{2}}\right),
\end{aligned}$$

which is the desired inequality. □

Proof of Corollary 1. For flat geometry $\mathcal{X}_k = \mathbb{R}^p$; part *(A1)* of Corollary 1 then follows from Theorem 2 and the equivalence of ℓ_q norms, in particular

$$\|x\|_2 \leq p^{1/2-1/q}\|x\|_q, \tag{19}$$

for all $x \in \mathbb{R}^p$ and for all $q \in [1, \infty]$. Thus we have the blowup inclusion $B(h,k)^{\epsilon p^{1/2-1/q}}_{\ell_2} \subseteq B(h,k)^\epsilon_{\ell_q}$. Part *(B1)* is just the result restated for $q = \infty$. The proofs of parts *(A2)* and *(B2)* trivially follow from the inequality (19). □

[1]Criteo, Paris, France. [2]**AUTHORERR: Missing \icmlaffiliation.** . Correspondence to: Elvis Dohmatob <e.dohmatob@criteo.com>.

Proceedings of the 36^{th} International Conference on Machine Learning, Long Beach, California, PMLR 97, 2019.

Remark 2. Note that the particular structure of the error set $B(h,k)$ did not play any part in the proof of Theorem 2 or of Corollary 1, beyond the requirement that the set be Borel. This means that we can obtain and prove analogous bounds for much broader class of losses. For example, it is trivial to extend the theorem to targeted attacks, wherein the attacker can aim to change an images label from k to a particular k'.

Proof of Lemma 1. Let B be a Borel subset of $\mathcal{X} = (\mathcal{X}, d)$ with $\mu(B) > 0$, and let $\mu|_B$ be the restriction of μ onto B defined by $\mu|_B(A) := \mu(A \cap B)/\mu(B)$ for every Borel $A \subseteq \mathcal{X}$. Note that $\mu|_B \ll \mu$ with *Radon-Nikodym derivative* $\frac{d\mu|_B}{d\mu} = \frac{1}{\mu(B)} 1_B$. A direct computation then reveals that

$$\begin{aligned}\mathrm{kl}(\mu|_B \| \mu) &= \int \log\left(\frac{d\mu|_B}{d\mu}\right) d\mu|_B \\ &= \int_B \log\left(\frac{1}{\mu(B)}\right) d\mu|_B \\ &= \log(1/\mu(B))\mu|_B(B) = \log\left(\frac{1}{\mu(B)}\right).\end{aligned}$$

On the other hand, if X is a random variable with law $\mu|_B$ and X' is a random variable with law $\mu|_{\mathcal{X}\setminus B^\epsilon}$, then the definition of B^ϵ ensures that $d(X, X') \geq \epsilon$ μ-a.s, and so by definition (7), one has $W_2(\mu|_B, \mu|_{\mathcal{X}\setminus B^\epsilon}) \geq \epsilon$. Putting things together yields

$$\begin{aligned}\epsilon &\leq W_2(\mu|_B, \mu_{\mathcal{X}\setminus B^\epsilon}) \leq W_2(\mu|_B, \mu) + W_2(\mu|_{\mathcal{X}\setminus B^\epsilon}, \mu) \\ &\leq \sqrt{2c\,\mathrm{kl}(\mu|_B \| \mu)} + \sqrt{2c\,\mathrm{kl}(\mu|_{\mathcal{X}\setminus B^\epsilon} \| \mu)} \\ &\leq \sqrt{2c\log(1/\mu(B))} + \sqrt{2c\log(1/\mu(\mathcal{X}\setminus B^\epsilon))} \\ &= \sqrt{2c\log(1/\mu(B))} + \sqrt{2c\log(1/(1-\mu(B^\epsilon))},\end{aligned}$$

where the first inequality is the triangle inequality for W_2 and the second is the $\mathrm{T}_2(c)$ property assumed in the Lemma. Rearranging the above inequality gives

$$\sqrt{2c\log(1/(1-\mu(B^\epsilon)))} \geq \epsilon - \sqrt{2c\log(1/\mu(B))},$$

Thus, if $\epsilon \geq \sqrt{2c\log(1/\mu(B))}$, we can square both sides, multiply by $c/2$ and apply the increasing function $t \mapsto e^t$, to get the claimed inequality. □

B. Distributional No "Free Lunch" Theorem

As before, let $h : \mathcal{X} \to \mathcal{Y}$ be a classifier and $\epsilon \geq 0$ be a tolerance level. Let $\widetilde{\mathrm{acc}}_\epsilon(h)$ denote the *distributional robustness accuracy* of h at tolerance ϵ, that is the worst possible classification accuracy at test time, when the conditional distribution P is changed by at most ϵ in the Wasserstein-1 sense. More precisely,

$$\widetilde{\mathrm{acc}}_\epsilon(h) := \inf_{Q \in \mathcal{P}(\mathcal{X}\times\mathcal{Y}),\, W_1(Q,P)\leq\epsilon} Q(h(x) = y), \quad (20)$$

where the Wasserstein 1-distance $W_1(Q,P)$ (see equation (7) for definition) in the constraint is with respect to the pseudo-metric $\tilde{d}$ on $\mathcal{X}\times\mathcal{Y}$ defined by

$$\tilde{d}((x',y'),(x,y)) := \begin{cases} d(x',x), & \text{if } y' = y, \\ \infty, & \text{else.}\end{cases}$$

The choice of $\tilde{d}$ ensures that we only consider alternative distributions that conserve the marginals π_y; robustness is only considered w.r.t to changes in the class-conditional distributions $P_{X|k}$.

Note that we can rewrite $\widetilde{\mathrm{acc}}_\epsilon(h) = 1 - \widetilde{\mathrm{err}}_\epsilon(h)$,

$$\widetilde{\mathrm{err}}_\epsilon(h) := \sup_{Q \in \mathcal{P}(\mathcal{X}\times\mathcal{Y}),\, W_1(Q,P)\leq\epsilon} Q(X \in B(h,Y)), \quad (21)$$

where is the distributional robustness test error and $B(h,y) := \{x \in \mathcal{X} | h(x) \neq y\}$ as before. Of course, the goal of a machine learning algorithm is to select a classifier (perhaps from a restricted family) for which the average adversarial accuracy $\mathrm{acc}_\epsilon(h)$ is maximized. This can be seen as a two player game: the machine learner chooses a strategy h, to which an adversary replies by choosing a perturbed version $Q \in \mathcal{P}(\mathcal{X}\times\mathcal{Y})$ of the data distribution, used to measure the bad event "$h(X) \neq Y$".

It turns out that the lower bounds on adversarial accuracy obtained in Theorem 2 apply to distributional robustness as well.

Corollary 2 (No "Free Lunch" for distributional robustness)**.** Theorem 2 holds for distributional robustness, i.e with $\mathrm{acc}_\epsilon(h|k)$ replaced with $\widetilde{\mathrm{acc}}_\epsilon(h|k)$.

Proof. See Appendix A. □

Proof of Corollary 2. We will use a dual representation of $\widetilde{\mathrm{acc}}_\epsilon(h|k)$ to establish that $\widetilde{\mathrm{acc}}_\epsilon(h|k) \leq \mathrm{acc}_\epsilon(h|k)$. That is, distributional robustness is harder than adversarial robustness. In particular, this will allow us apply the lower bounds on adversarial accuracy obtained in Theorem 2 to distributional robustness as well!

So, for $\lambda \geq 0$, consider the convex-conjugate of $(x,y) \mapsto 1_{x\in B(h,y)}$ with respect to the pseudo-metric $\tilde{d}$, namely $1^{\lambda\tilde{d}}_{x\in B(h,y)} := \sup_{(x',y')\in\mathcal{X}\times\mathcal{Y}} 1_{x'\in B(h)} - \lambda\tilde{d}((x',y'),(x,y))$.

A straightforward computation gives

$$\begin{aligned}1^{\lambda\tilde{d}}_{x\in B(h,y)} &:= \sup_{(x',y')\in\mathcal{X}\times\mathcal{Y}} 1_{x'\in B(h,y')} - \lambda\tilde{d}((x',y'),(x,y)) \\ &= \max_{B\in\{B(h,y),\, \mathcal{X}\setminus B(h,y)\}} \sup_{x'\in B} 1_{x'\in B(h,y)} - \lambda d(x',x) \\ &= \max(1 - \lambda d(x, B(h,y)), -\lambda d(x, \mathcal{X}\setminus B(h,y))) \\ &= (1 - \lambda d(x, B(h,y)))_+.\end{aligned}$$

Now, since the transport cost function $\tilde{d}$ is nonnegative and lower-semicontinuous, strong-duality holds (Villani, 2008; Blanchet & Murthy, 2016) and one has

$$\begin{aligned}
&\sup_{W_1(Q,P)\leq\epsilon} Q(h(X)\neq Y)\\
&= \inf_{\lambda\geq 0}\sup_{Q}(Q(X\in B(h,Y))+\lambda(\epsilon - W_1(Q,P)))\\
&= \inf_{\lambda\geq 0}\left(\sup_{Q}(Q(X\in B(h,Y))-\lambda W_1(Q,P))+\lambda\epsilon\right)\\
&= \inf_{\lambda\geq 0}(\mathbb{E}_{(x,y)\sim P}[1^{\lambda\tilde{d}}_{x\in B(h,y)}]+\lambda\epsilon)\\
&= \inf_{\lambda\geq 0}(\mathbb{E}_{(x,y)\sim P}[(1-\lambda d(x,B(h,y)))_+]+\lambda\epsilon)\\
&= P(X\in B(h,Y)^{\lambda_*^{-1}}),
\end{aligned}$$

where $\lambda_* = \lambda_*(h)\geq 0$ is the (unique!) value of λ at which the infimum is attained and we have used the previous computations and the handy formula

$$\sup_{Q}(Q(X\in B(h,Y))-\lambda W_1(Q,P)) = \mathbb{E}_P[1^{\lambda\tilde{d}}_{X\in B(h,Y)}],$$

which is a direct consequence of Remark 1 of (Blanchet & Murthy, 2016). Furthermore, by Lemma 2 of (Blanchet & Murthy, 2016), one has

$$\begin{aligned}
\epsilon &\leq \sum_k \pi_k \int_{B(h,k)^{\lambda_*^{-1}}} d(x,B(h,k))dP_{X|k}(x)\\
&\leq \sum_k \pi_k \lambda_*^{-1} P_{X|k}(X\in B(h,k)^{\lambda_*^{-1}})\\
&= \lambda_*^{-1} P(X\in B(h,Y)^{\lambda_*^{-1}}) \leq \lambda_*^{-1}.
\end{aligned}$$

Thus $\lambda_*^{-1}\geq\epsilon$ and combining with the previous inequalities gives

$$\begin{aligned}
\sup_{Q\in\mathcal{P}(\mathcal{X}),\ W_1(Q,P)\leq\epsilon} Q(h(X)\neq Y) &\geq P(X\in B(h,Y)^{\lambda_*^{-1}})\\
&\geq P(X\in B(h,Y)^{\epsilon}).
\end{aligned}$$

Finally, noting that $\mathrm{acc}_\epsilon(h) = 1 - P(X\in B(h,Y)^\epsilon)$, one gets the claimed inequality $\widetilde{\mathrm{acc}}_\epsilon(h)\leq \mathrm{acc}_\epsilon(h)$. □