Wasserstein Distributional Robustness of Neural Networks

Xingjian Bai, Guangyi He, Yifan Jiang, Jan Obłój

xingjian.bai@sjc.ox.ac.uk, guangyihe2002@outlook.com, yifan.jiang@maths.ox.ac.uk, jan.obloj@maths.ox.ac.uk

UNIVERSITY OF OXFORD

Main Contribution

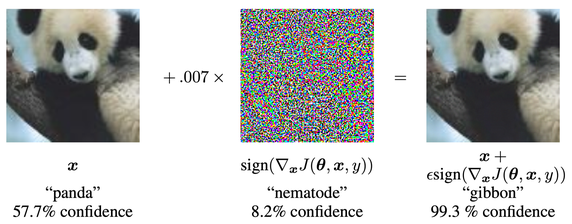
1. A unified approach to adversarial attacks and training based on sensitivity analysis for Wasserstein DRO from Bartl et al. (2021).
2. A fast-to-compute first order adversarial attack method for *distributional* threat models. As a special case, this recovers the classical FGSM attack, lending it further theoretical underpinning.
3. Asymptotically certified bounds and out-of-sample performance of adversarial accuracy, applicable to a general threat, including classical pointwise perturbations.

Adversarial Attack of NNs

Deep neural networks have achieved a great success in image recognition tasks. However, they are notoriously known to be vulnerable under adversarial attacks, which are small perturbations of the input data that lead to misclassification.
To generate adversarial images, we reverse the training process by maximizing the loss over input data:

$$E_P\Big[\sup_{\|x-x'\|_\infty\le\delta} J_\theta(x',y)\Big] \rightsquigarrow x^* \approx x+\delta\mathrm{sgn}(\nabla_x J_\theta(x,y)),$$

where $J_\theta(x,y) = L(f_\theta(x),y)$, and P is the input distribution.



A demonstration of Fast Gradient Sign Method (FGSM).

Distributional Threat Models

Let $d((x,y),(x',y')) = \|x-x'\|_s + \infty\mathbb{1}\{y\neq y'\}$ and $\mathcal{W}_p$ be the Wasserstein distance given by

$$\mathcal{W}_p(P,Q) = \inf\Big\{E\big[d(X,Y)^p\big]^{1/p} : X\sim P,\, Y\sim Q\Big\}.$$

We propose a novel *distributional* threat model

$$\sup_{Q:\mathcal{W}_p(P,Q)\le\delta} E_Q[J_\theta(x,y)].$$

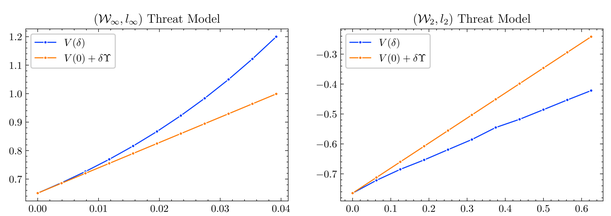
In *distributional* threat models, the attacker has a greater flexibility and can perturb images close to the decision boundary only slightly while spending more of the attack budget on images farther away from the boundary.
The classical *pointwise* threat model can be covered as an extreme case $p=s=\infty$

$$E_P\Big[\sup_{\|x-x'\|_\infty\le\delta} J_\theta(x',y)\Big] = \sup_{Q:\mathcal{W}_\infty(P,Q)\le\delta} E_Q[J_\theta(x,y)].$$

Wasserstein DRO Sensitivity

We write $V(\delta) = \sup_{Q:\mathcal{W}_p(P,Q)\le\delta} E_Q\big[J_\theta(x,y)\big]$ and assume J_θ is Lipschitz under d.
The following result follows readily from Bartl et al. and its proof.

- $V(\delta) = V(0) + \delta\Upsilon + o(\delta)$, where
$$\Upsilon = \Big(E_P\|\nabla_x J_\theta(x,y)\|_*^q\Big)^{1/q}.$$
- $V(\delta) = E_{Q_\delta}[J_\theta(x,y)] + o(\delta)$, where Q_δ is the pushforward of P under the map
$$x\mapsto x+\delta h(\nabla_x J_\theta(x,y))\|\Upsilon^{-1}\nabla_x J_\theta(x,y)\|_*^{q-1},$$
and h is uniquely determined by $\langle h(x),x\rangle = \|x\|_*$.



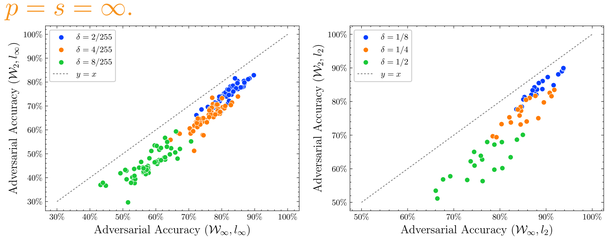
Performance of the first order approximation for the W-DRO value on CIFAR-10.

Wasserstein Distributional Adversarial Attacks

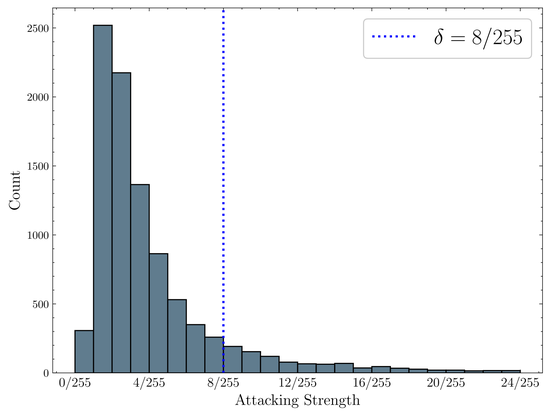
Our proposed PGD-type attack is:

$$x^{n+1} = \mathrm{pj}\big(x^n + \alpha h(\nabla_x J_\theta(x^n,y))\|\Upsilon^{-1}\nabla_x J_\theta(x^n,y)\|_*^{q-1}\big),$$

where α is the step size and pj is a projection.
The classical FGSM attack is retrieved by taking $p=s=\infty$.



Shortfall of WD-adversarial accuracy on CIFAR-10.



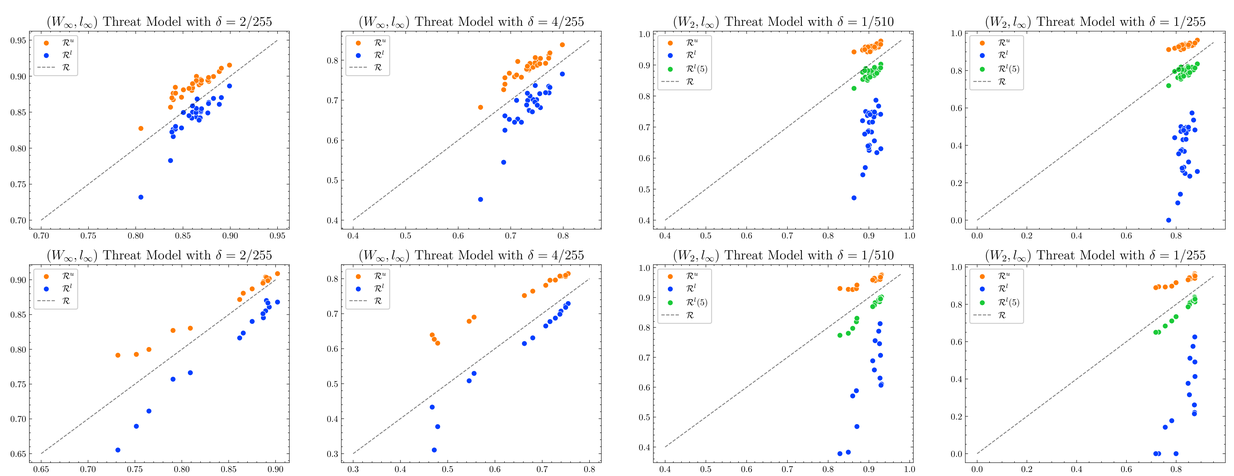
Histogram of *distributional* attacking strength on CIFAR-10.

Certified Adversarial Accuracy Bounds

- S the set of images equipped with their labels generated by f_θ, i.e.,
$$S = \Big\{(x,y) : \arg\max_{1\le i\le m} f_\theta(x)_i = \{y\}\Big\}.$$
- A the clean accuracy given by $A = E_P[\mathbb{1}_S]$.
- A_δ the adversarial accuracy given by $A_\delta = \inf_{Q:\mathcal{W}_p(P,Q)\le\delta} E_{Q_\delta}[\mathbb{1}_S]$.
- W_δ the loss condition on the misclassified image, given by
$$W(\delta) = \sup_{Q\in B_\delta(P)} E_Q[J_\theta(x,y)|S^c].$$

We propose a metric of robustness: $\mathcal{R}_\delta = A_\delta/A$.

- Upper bound: $\mathcal{R}_\delta \le \mathcal{R}_\delta^u := Q_\delta(S)/A$.
- An asymptotic lower bound:
$$\mathcal{R}_\delta \ge \frac{W(0)-V(\delta)}{W(0)-V(0)} + o(\delta) = \widetilde{\mathcal{R}}_\delta^l + o(\delta) = \overline{\mathcal{R}}_\delta^l + o(\delta),$$
where we utilize the first order approximations of $V(\delta)$ and write $\widetilde{\mathcal{R}}_\delta^l = \frac{W(0)-E_{Q_\delta}[J_\theta(x,y)]}{W(0)-V(0)}$ and $\overline{\mathcal{R}}_\delta^l = \frac{W(0)-V(0)-\delta\Upsilon}{W(0)-V(0)}$.



Certified bounds $\mathcal{R}^u$ & $\mathcal{R}^l$ versus $\mathcal{R}$ CIFAR-100 (top) and ImageNet (bottom).