

Certified Adversarial Robustness Under the Bounded Support Set

Authors: Yiwen Kou¹, Qinyuan Zheng², Yisen Wang^{2,3}

¹Yuanpei College, Peking University

²Key Lab. of Machine Perception (MoE), School of Artificial Intelligence, Peking University

³Institute for Artificial Intelligence, Peking University

Background: Certified Adversarial Robustness

- DNNs' vulnerability to adversarial samples



- One solution: adversarial training
 - Drawback: stronger or adaptive attacks decrease its effectiveness
- Certified robustness: algorithms that are provably robust to the worst-case attacks
 - Method: randomized smoothing

Background: Randomized Smoothing

- Base classifier f : maps inputs R^d to classes Y
- Randomized smoothing transforms any arbitrary base classifier f into a new “smoothed classifier” g
- For any input x , the μ -smoothed classifier’s prediction $g(x)$ is defined to be the class which f is most likely to classify the random variable $x + \mu$ as.
 - $g(x) = \operatorname{argmax}_{c \in Y} P(f(x + \epsilon) = c)$ where $\epsilon \sim \mu$

Background: f -divergence based framework

- Dvijotham et al. (2020a)¹ introduce a robustness framework by utilizing convex relaxation technique for f -divergence.
 - Binary Classifier: $f: x \in R^d \mapsto Y = \{\pm 1\}$
 - Adversarial perturbations: size ϵ and norm $\|\cdot\|$
 - Solve the optimization problem $\text{OPT} = \min_{\|x-x'\| \leq \epsilon} P(f(x' + \mu) = +1)$
- Rewrite: $\min_{v \in \{x' + \mu: \|x-x'\| \leq \epsilon\} \subseteq P(X)} P(f(v) = +1)$
 - Relaxation: $\{x' + \mu: \|x-x'\| \leq \epsilon\} \subseteq \{v: D_f(v||x + \mu) \leq \epsilon\}$, D_f : f -divergence
 - $\min_{v \in P(X): D_f(v||x+\mu) \leq \epsilon} P(f(v) = +1)$, easy to dualize

¹A framework for robustness certification of smoothed classifiers using f -divergences

Motivation of our work

- Definition of f -divergence: $D_f(p||q) = \int q(x)f\left(\frac{p(x)}{q(x)}\right)$,
 - When $\text{supp}(p(x)) := K$ is bounded, $q(x) = p(x - v)$ then $\text{supp}(p(x)) := v + K$ where v is a displacement vector
 - If $x \in K \setminus (v + K)$, $\frac{p(x)}{q(x)} = \frac{p(x)}{0}$ which is undefined
- Definition of Wasserstein distance: $W_1(\mu, \nu) = \sup_{\|f\|_L \leq 1} E_{X \sim \mu - \nu}[f(X)]$
- Definition of total variation distance: $\text{TV}(\mu, \nu) = \sup_A |\mu(A) - \nu(A)|$
 - not involve likelihood ratio $\frac{p(x)}{q(x)}$ and are only related to difference density function $p(x) - q(x)$
 - Able to analyze smoothing distribution with bounded support set

Problem Setting: Randomized smoothing using Wasserstein distance and total variation distance

- Binary base classifier: $f: x \in R^d \mapsto Y = \{\pm 1\}$
- Smoothing probability measure $\mu \in P(X)$
- Adversarial perturbation: size at most ϵ , given norm $\|\cdot\|_q$
- Solve optimization problem: $OPT(f, D) := \min_{\nu \in D} E_{X \in \nu} [f(X)]$
 - $D_{x, \epsilon, q} := \{x' + \mu: \|x - x'\|_q \leq \epsilon\}$
 - Check $OPT(f, D_{x, \epsilon, q}) \geq 0$ for $f(x) = +1$

Our Certification Procedure

- Relaxation Using Wasserstein Distance
 - find a $\delta_q(\epsilon)$ such that
 - $D_{x,\epsilon,q} := \{x' + \mu: \|x - x'\|_q \leq \epsilon\} \subseteq \{v: W_p(x + \mu, v) \leq \delta_q(\epsilon)\} := D_{x,\delta_q(\epsilon),p}$
- Relaxation Using Total Variation Distance
 - find a $\xi_q(\epsilon)$ such that
 - $D_{x,\epsilon,q} := \{x' + \mu: \|x - x'\|_q \leq \epsilon\} \subseteq \{v: TV(x + \mu, v) \leq \xi_q(\epsilon)\} = D_{x,\xi_q(\epsilon)}$
- Since $D_{x,\epsilon,q} \subseteq D_{x,\delta_q(\epsilon),p} \cap D_{x,\xi_q(\epsilon)}$,
 - $OPT(f, D_{x,\epsilon,q}) \geq OPT\left(f, D_{x,\delta_q(\epsilon),p} \cap D_{x,\xi_q(\epsilon)}\right)$
 - Our procedure verifies whether $OPT\left(f, D_{x,\delta_q(\epsilon),p} \cap D_{x,\xi_q(\epsilon)}\right) \geq 0$

Dualize the optimization problem

- The Lagrange function of $OPT \left(f, D_{x, \delta_q(\epsilon), p} \cap D_{x, \xi_q(\epsilon)} \right)$ is
 - $L(\lambda) = E_{X \sim x + \mu} [f(X)] - 2\xi_q(\epsilon) - \lambda C, \lambda \geq 0$
- Using the duality result, the optimal value can be obtained by computing
 - $\max_{\lambda \geq 0} L(\lambda) = E_{X \sim x + \mu} [f(X)] - 2\xi_q(\epsilon)$
 - Which is only related to total variation distance based radius
- Applying above formula, we can obtain following table

Certification formulas

Table 1. Certification objectives and prerequisites.

| Smoothing Measure | Perturbation | Certification Objective | Prerequisite |
|-------------------------------|------------------|--|--|
| $\mathcal{U}(B_2(O, r))$ | $l_q (q \leq 2)$ | $\mathbb{E}_{X \sim x+\mu}[\phi(X)] - 2 \left(1 - \frac{\int_0^{\arccos(\frac{\epsilon}{2r})} \sin^n(t) dt}{\int_0^{\frac{\pi}{2}} \sin^n(t) dt} \right)$ | $\epsilon \leq 2r$ |
| | $l_q (q > 2)$ | $\mathbb{E}_{X \sim x+\mu}[\phi(X)] - 2 \left(1 - \frac{\int_0^{\arccos(\frac{\epsilon d^{1/2-1/q}}{2r})} \sin^n(t) dt}{\int_0^{\frac{\pi}{2}} \sin^n(t) dt} \right)$ | $\epsilon \leq 2r d^{\frac{1}{q} - \frac{1}{2}}$ |
| $\mathcal{U}(B_\infty(O, r))$ | l_1 | $\mathbb{E}_{X \sim x+\mu}[\phi(X)] - \frac{\epsilon}{r}$ | $\epsilon \leq 2r$ |
| | l_2 | $\mathbb{E}_{X \sim x+\mu}[\phi(X)] - 2 \left(1 - \left(1 - \frac{\epsilon}{2d^{\frac{1}{2}} r} \right)^d \right)$ | $\epsilon \leq 2t_n r$ |
| | l_∞ | $\mathbb{E}_{X \sim x+\mu}[\phi(X)] - 2 \left(1 - \left(1 - \frac{\epsilon}{2r} \right)^d \right)$ | $\epsilon \leq 2r$ |
| $\mathcal{N}(0, \sigma^2 I)$ | $l_q (q \leq 2)$ | $\mathbb{E}_{X \sim x+\mu}[\phi(X)] - 2 \left(2G\left(\frac{\epsilon}{2\sigma}\right) - 1 \right)$ | - |
| | $l_q (q > 2)$ | $\mathbb{E}_{X \sim x+\mu}[\phi(X)] - 2 \left(2G\left(\frac{\epsilon d^{\frac{1}{2} - \frac{1}{q}}}{2\sigma}\right) - 1 \right)$ | - |

Relationship with Previous Work

- Curse of dimensionality:
 - Hardness results of previous papers^{1,2} work for all measurable base classifier, which might be overtight
 - We provide an efficient certification procedure related to specific base classifier
- Performance of different smoothing measure:
 - Our framework is able to obtain certification formula for smoothing measures with bounded support sets
- When applying to smoothing measure $N(\mathbf{0}, \sigma^2 I)$, our framework and f -divergence based framework³ have the same certified formula

¹Random smoothing might be unable to certify l_∞ robustness for high-dimensional images

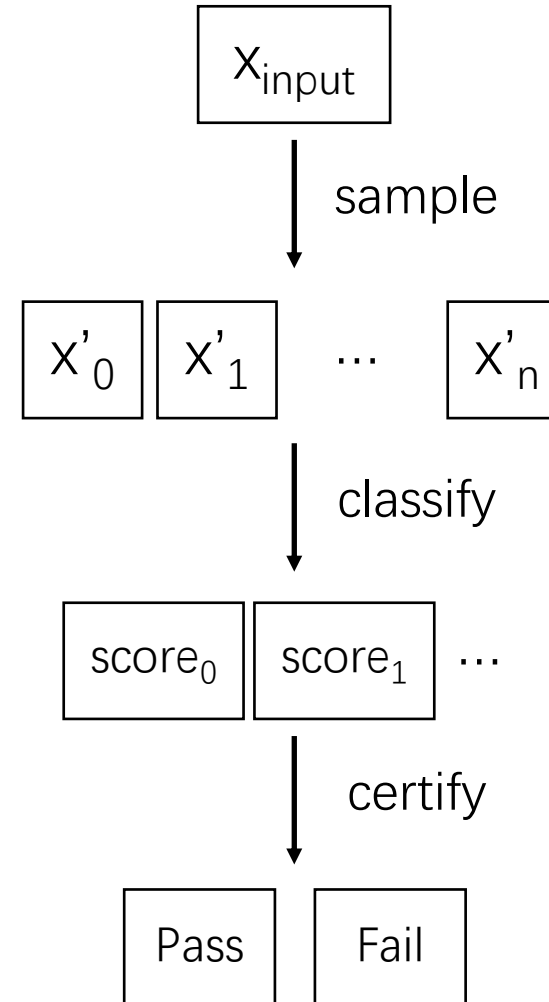
²Curse of dimensionality on randomized smoothing for certifiable robustness

³A framework for robustness certification of smoothed classifiers using f -divergences

Algorithm Demonstration

Algorithm 1 Certification Process

```
1: Input:  $T$ : test set,  $\text{target}(x)$ : true class of image  $x$ ,  $f(x)$ : base classifier,  $D(x)$ : smoothing distribution,  $n$ : sample amount,  $\epsilon$ : perturbation radius,  $\text{cert}(\text{score}_a, \text{score}_b, \epsilon)$ : certification object
2: Output:  $\text{acc}$ : test set certified accuracy
3:  $\text{certifiedCount} \leftarrow 0, \text{allCount} \leftarrow 0$ 
4: for all  $x \in T$  do
5:    $S \leftarrow \{n \text{ samples from } D(x)\}$ 
6:    $\text{count}_c \leftarrow 0$  for every class  $c$ 
7:   for all  $x' \in S$  do
8:      $\text{count}_{f(x')} \leftarrow \text{count}_{f(x')} + 1$ 
9:   end for
10:   $\text{score}_c \leftarrow \text{count}_c / \text{card}(S)$  for every class  $c$ 
11:   $\text{predict} \leftarrow \arg \max_c \{\text{score}_c\}$ 
12:  if  $\text{predict} = \text{target}(x) \wedge \text{cert}(\text{score}_c, 1 - \text{score}_c, \epsilon)$  then
13:     $\text{certifiedCount} \leftarrow \text{certifiedCount} + 1$ 
14:  end if
15:   $\text{allCount} \leftarrow \text{allCount} + 1$ 
16: end for
17: return  $\text{acc} \leftarrow \text{certifiedCount} / \text{allCount}$ 
```



Experiment results with different smoothing distributions

- Similar characteristics:
 - Increase of smoothing distribution variance
 1. drop of initial certification accuracy
 2. stronger robustness that can endure more significant perturbation
- Major difference:
 - Bounded support set => cutoff perturbation 50 times earlier
 - May be inherent and unavoidable when using uniform distribution

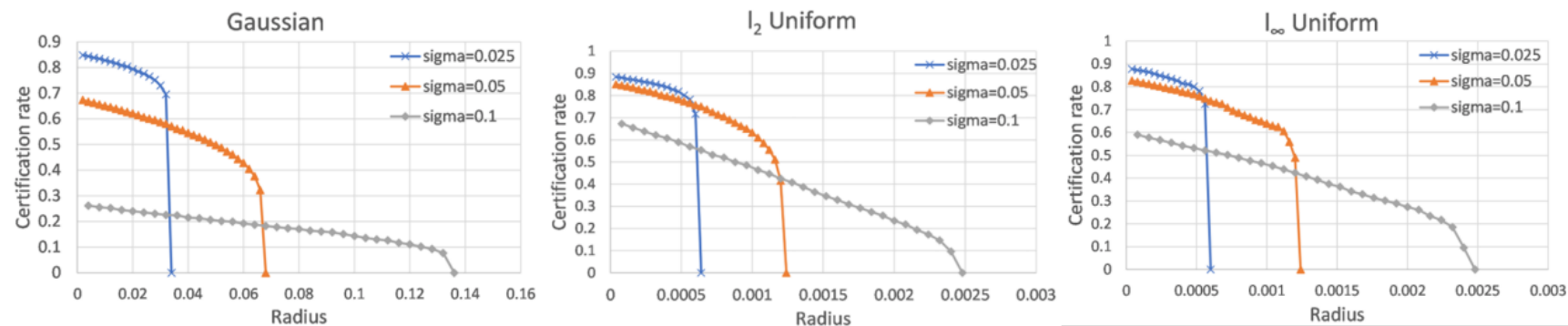


Figure 3. Results of different smoothing distributions using our W-distance and TV-distance based framework. 'Sigma' refers to parameter σ for Gaussian distribution and parameter r for uniform distribution.

Compare with previous method

- Method of comparison: Dvijotham et al. (2020a)
- Result:
 1. Possibility to utilize more types of smoothing distributions
 2. Minor improvement of certification rate (larger and closer to real value)
 3. Significant improvement of time cost (ten time less)

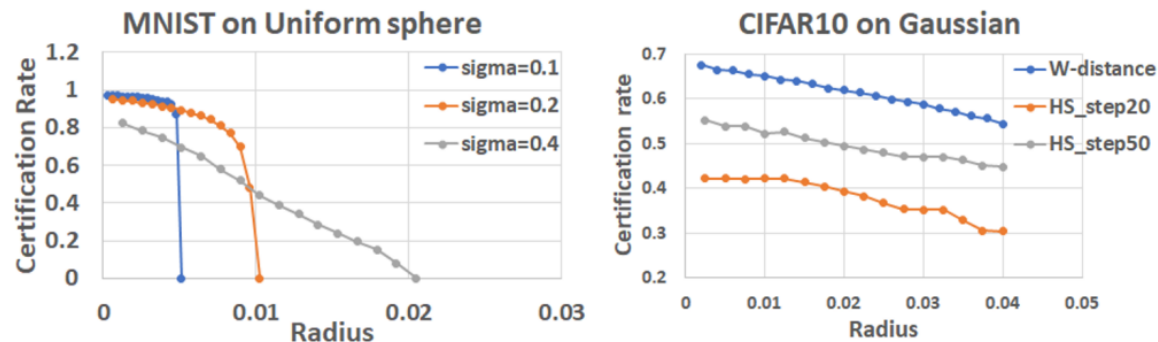


Figure 2. Left: our method's performance on MNIST dataset utilizing l_2 distribution. Right: result of Dvijotham et al. (2020a)'s method with different optimizing steps on CIFAR10 dataset and ours using W-distance.

Conclusion

- We introduce a certified robustness framework using
 - Wasserstein distance
 - Total variation distance
 - Lagrange duality
- We apply the framework to Gaussian $N(\mathbf{0}, \sigma^2 I)$ and uniform smoothing measure $U(B_p(\mathbf{0}, r))$ and provide several theoretical analyses
- We experimentally verify the badness of $U(B_p(\mathbf{0}, r))$ -smoothed classifier and the time efficiency of our method