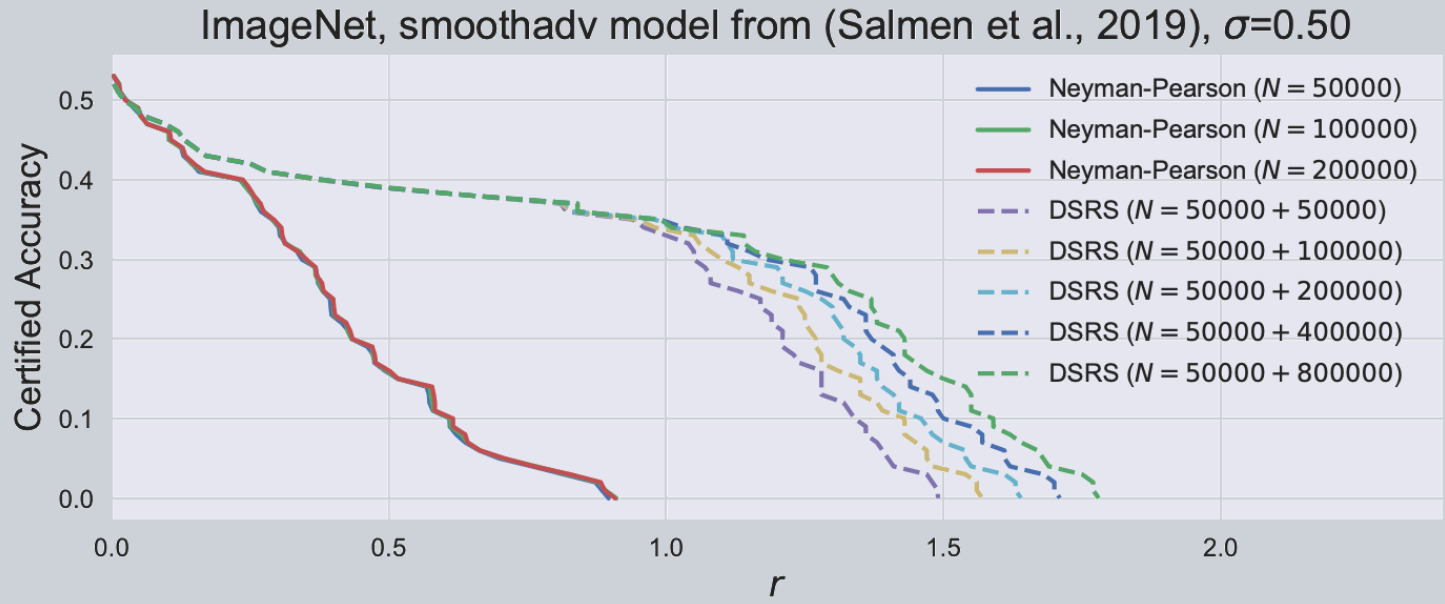


Double Sampling Randomized Smoothing

Linyi Li (UIUC), Jiawei Zhang (UIUC),
Tao Xie (Peking University), Bo Li (UIUC)

Background: Randomized Smoothing

- Train an NN f (the "base classifier") under Gaussian data corruption

Clean Image

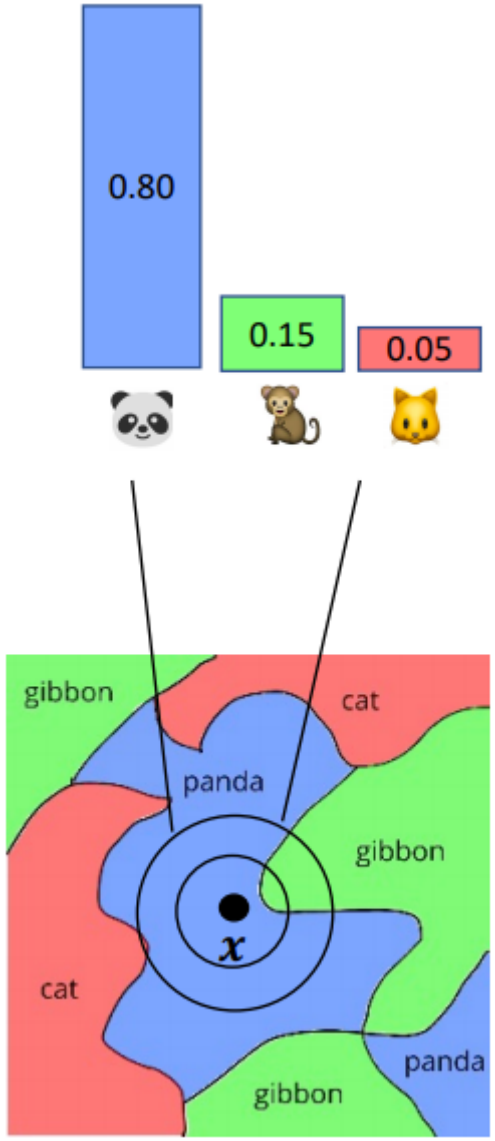
Corrupted by Gaussian Noise

- Then, smooth f into a new classifier g (the "smoothed classifier"), defined as follows:

Adapted from Cohen et al's slides

Classical Randomized Smoothing

- $g(x)$ = the most probable prediction by f under random Gaussian corruptions of x
- Example:
  - Consider the input x with label panda.
  - Suppose that when f classifies $N(x, \sigma^2 I)$:
    - Panda is returned with probability 0.80
    - Gibbon is returned with probability 0.15
    - Cat is returned with probability 0.05
  - Then $g(x)$ = panda



Adapted from Cohen et al's slides

Robustness Guarantee of Classical Randomized Smoothing

- Let P_A be the probability of the top class (panda)
- Then g probably returns the top-class panda within an ℓ_2 ball around x of radius

$$R = \sigma \Phi^{-1}(P_A)$$

- Where Φ^{-1} is the inverse standard Gaussian CDF

Adapted from Cohen et al's slides

Can we achieve better certified robustness?

Improving classical RS is a hot research topic these years

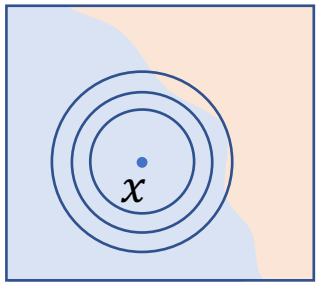
- Improving smoothing distribution
  - Uniform smoothing > Gaussian smoothing for ℓ_1 [ICML 2020, Yang et al]
  - Generalized Gaussian smoothing > Gaussian smoothing for ℓ_2 [NeurIPS 2020, Zhang et al]
  - Dimention-dependent discrete smoothing achieves SOTA for ℓ_1 [ICML 2021, Levine et al]
  - …
- Improving training
  - SmoothAdv [NeurIPS 2019, Salman et al]
  - MACER [ICLR 2020, Zhai et al]
  - DRT [ICLR 2022, Yang et al]
  - …

Can we improve the certification itself?

Negative results:

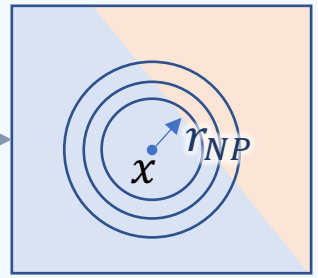
- RS may be unable to certify high ℓ_∞ robustness
  - [Yang et al, ICML 2020] [Blum et al, JMLR 2020] [Kumar et al, ICML 2020] [Wu et al, AISTATS 2021] …
  - Since certified radius under ℓ_2 is almost constant w.r.t. input dim. d
    - ➢ Shrinking radius $r_{\ell_2}/\sqrt{d}$ under ℓ_∞
- However, this barrier **only** holds for RS certification using top-class prob.
- Use more information in RS certification may be able to circumvent the barrier!

DSRS Pipeline



Given model and smoothing distribution $\mathcal{P}$

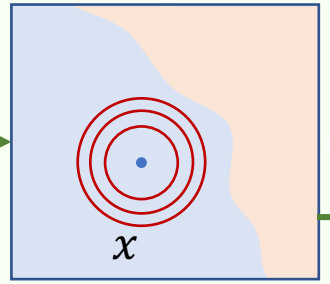
Get P_A from sampling



Standard Certification

- Construct worst-case classifier with P_A
- Derive certified radius r_{NP}

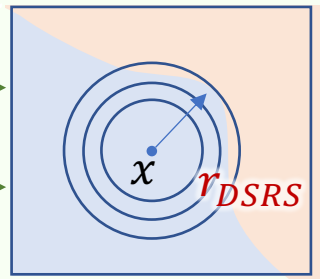
$$\begin{aligned} \underset{f}{\text{minimize}} \quad & \mathrm{E}_{\epsilon\sim\mathcal{P}}[f(\epsilon+\delta)] \\ \text{s.t.} \quad & \mathrm{E}_{\epsilon\sim\mathcal{P}}[f(\epsilon)] = P_A, \\ & 0 \le f(\epsilon) \le 1 \quad \forall \epsilon \in \mathbb{R}^d. \end{aligned}$$



Smooth the model with another distribution $\mathcal{Q}$

Get P_A from sampling

Get Q_A from sampling



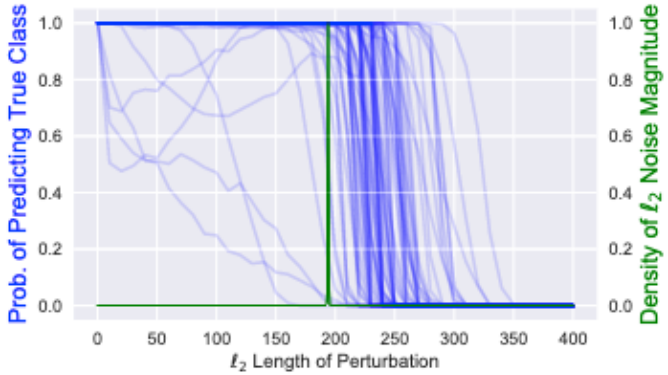
Our DSRS Certification

- Construct worst-case classifier with both P_A and Q_A
- Derive tighter certified radius $r_{DSRS} > r_{NP}$

$$\begin{aligned} \underset{f}{\text{minimize}} \quad & \mathrm{E}_{\epsilon\sim\mathcal{P}}[f(\epsilon+\delta)] \\ \text{s.t.} \quad & \mathrm{E}_{\epsilon\sim\mathcal{P}}[f(\epsilon)] = P_A, \\ & \mathrm{E}_{\epsilon\sim\mathcal{Q}}[f(\epsilon)] = Q_A, \\ & 0 \le f(\epsilon) \le 1 \quad \forall \epsilon \in \mathbb{R}^d. \end{aligned}$$

Theoretical Benefits of DSRS

- Common classifiers satisfies concentration property
  - Correct prediction prob. is high when adding small magnitude (in terms of ℓ_2) noises



(On ImageNet)

- ***(Informal, Theorem 2) Suppose ideal concentration property holds, if we use generalized Gaussian for smoothign, the ℓ_2 certified radius of DSRS:**

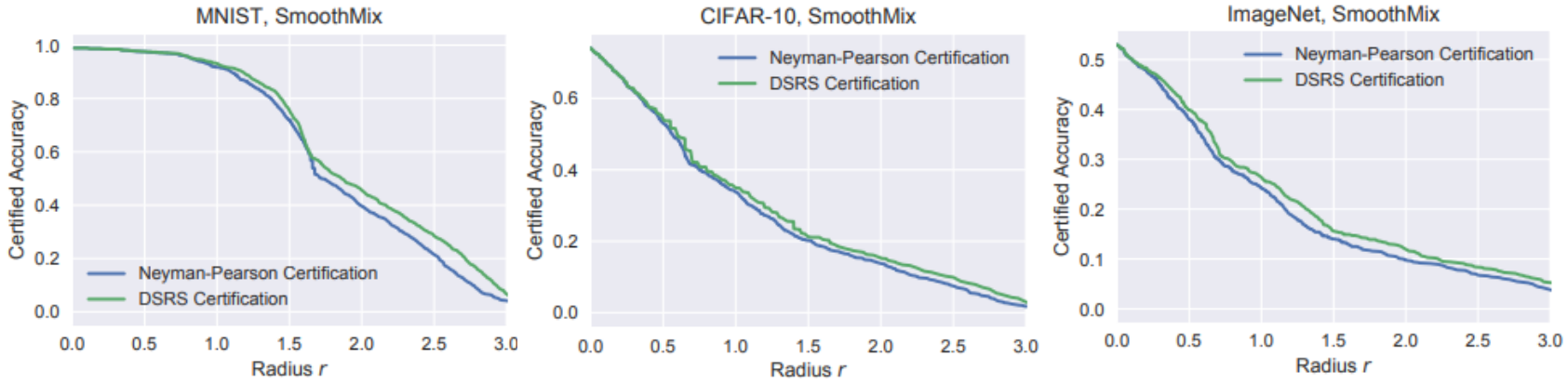
$$r_{DSRS} \geq 0.02\sigma\sqrt{d}$$

➢Translates to constant ℓ_∞ radius and circumvents the well-known ℓ_∞-barrier!

Empirical Benefits of DSRS

- Consistent certified radius improvement on MNIST, CIFAR-10, and ImageNet

Training Method	Certification	MNIST	CIFAR-10	ImageNet
Gaussian Augmentation	Neyman-Pearson	1.550	0.447	0.677
	DSRS	1.629	0.469	0.750
	Relative Improvement	**+5.10%**	**+4.92%**	**+10.78%**
Consistency	Neyman-Pearson	1.645	0.636	0.796
	DSRS	1.730	0.659	0.862
	Relative Improvement	**+5.17%**	**+3.62%**	**+8.29%**
SmoothMix	Neyman-Pearson	1.716	0.676	0.490
	DSRS	1.806	0.712	0.525
	Relative Improvement	**+5.24%**	**+5.33%**	**+7.14%**



- Large gain on average ℓ_2 certified radius:

5% - 6% on MNIST, 3% - 6% on CIFAR-10, 7% - 10% on ImageNet

Open Problems

- Better auxiliary smoothing distribution $\mathcal{Q}$?
- Suitable training methods for DSRS?
- Efficient sampling with both $\mathcal{P}$ and $\mathcal{Q}$?

Thank you!

- Paper (latest version): arxiv.org/abs/2206.07912
- Code & Model & Data: github.com/llylly/DSRS