ETH zürich
SRILAB

Dual Randomized Smoothing: Beyond Global Noise Variance

Chenhao Sun | Yuhao Mao | Martin Vechev

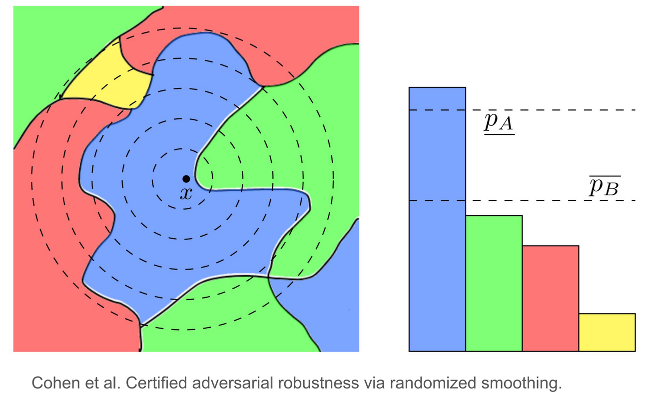
Paper
Code

Traditional Randomized Smoothing is limited by the **constant** noise variance.

We extend it to **model-predicted adaptive noise variance**, significantly improving the accuracy-robustness tradeoff

Background

Randomized smoothing



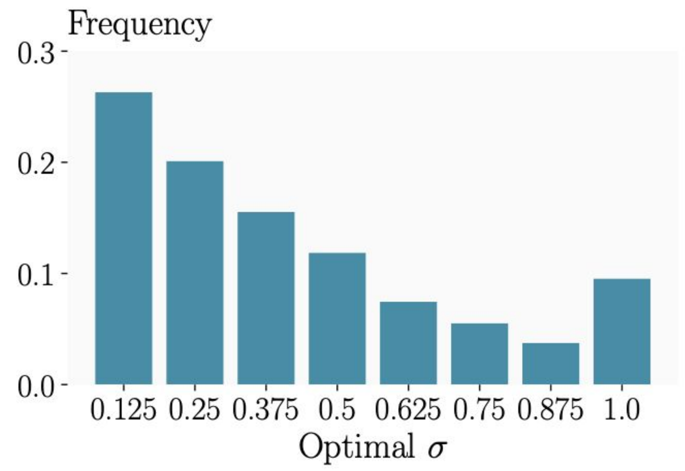
$$g(x) = \arg\max_{c \in \mathcal{Y}} \mathbb{P}(f(x+\epsilon) = c)$$

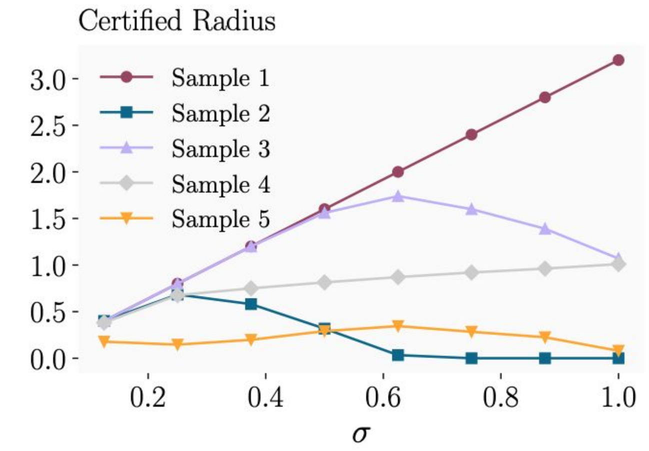
Smooth model | Base model | Gaussian noise

Certified radius

$g(x)$ is adversarially robust in an ℓ_2 norm neighborhood of x with the radius $R(x)$

Constant Noise Variance is not Optimal

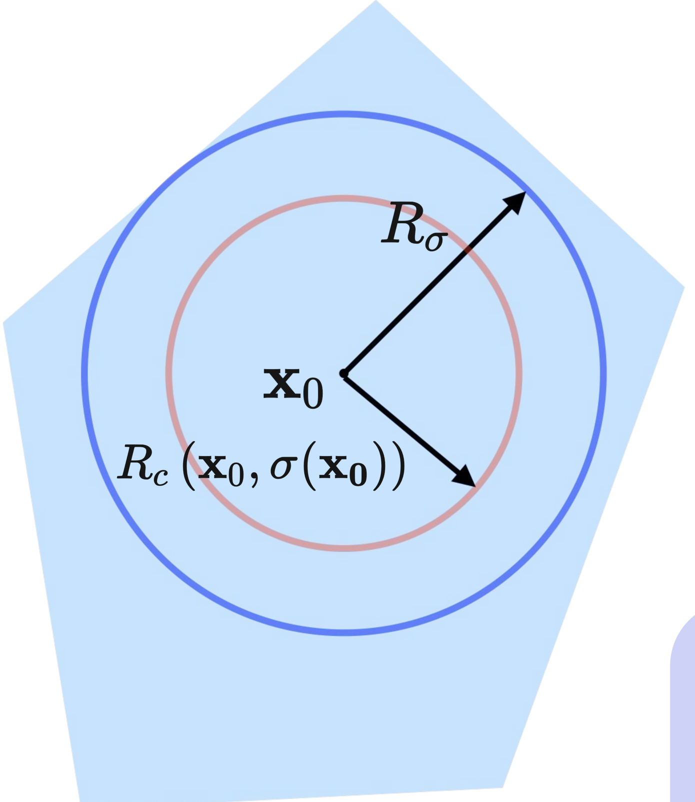




The optimal σ varies across different samples

→ Accuracy-robustness trade off

Certification with Locally Constant Noise Variance



$\sigma(\mathbf{x})$ constant area

$R_\sigma : \sigma(\mathbf{x}) = \sigma(\mathbf{x}_0), \|\mathbf{x} - \mathbf{x}_0\|_2 \le R_\sigma$

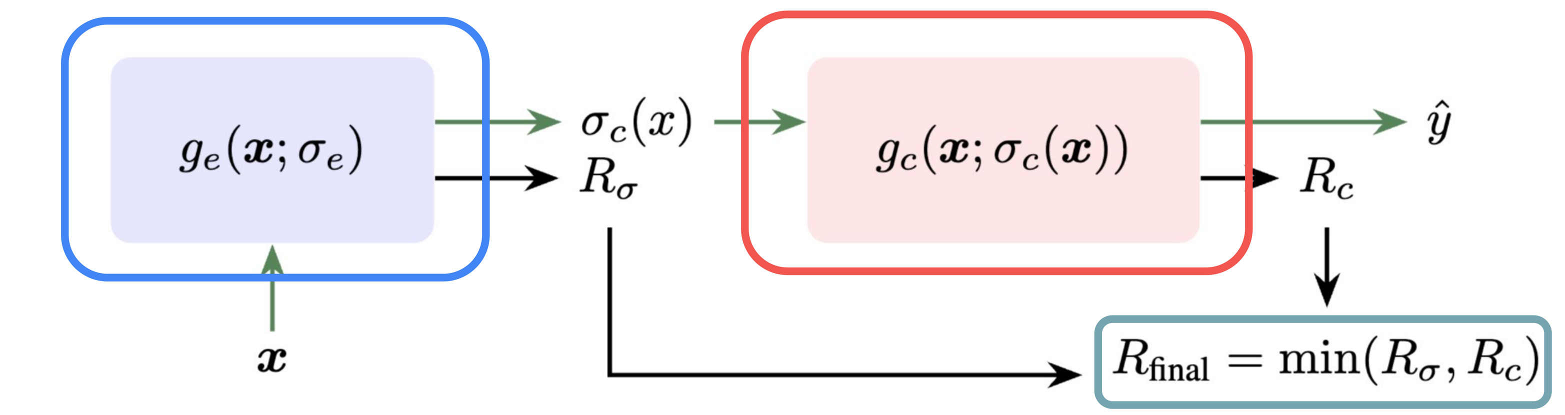
R_c : Classification certified radius of $\mathbf{x}_0$ with noise variance $\sigma(\mathbf{x}_0)^2$

Final certified radius:

$$R_{\text{final}} = \min\left(R_\sigma, R_c(\mathbf{x}_0, \sigma(\mathbf{x}_0))\right)$$

Take away: RS certification remains **valid** if $\sigma(\mathbf{x})$ **is constant inside the neighborhood** around $\mathbf{x}_0$ — **It does not need to be globally constant**

Dual Randomized Smoothing



Final certified radius

Training Method

Construct dataset
Certified radii with all σ
Maximum radius
Optimal σ

Train the σ-Estimator
Train a ResNet to predict optimal σ on denoised images

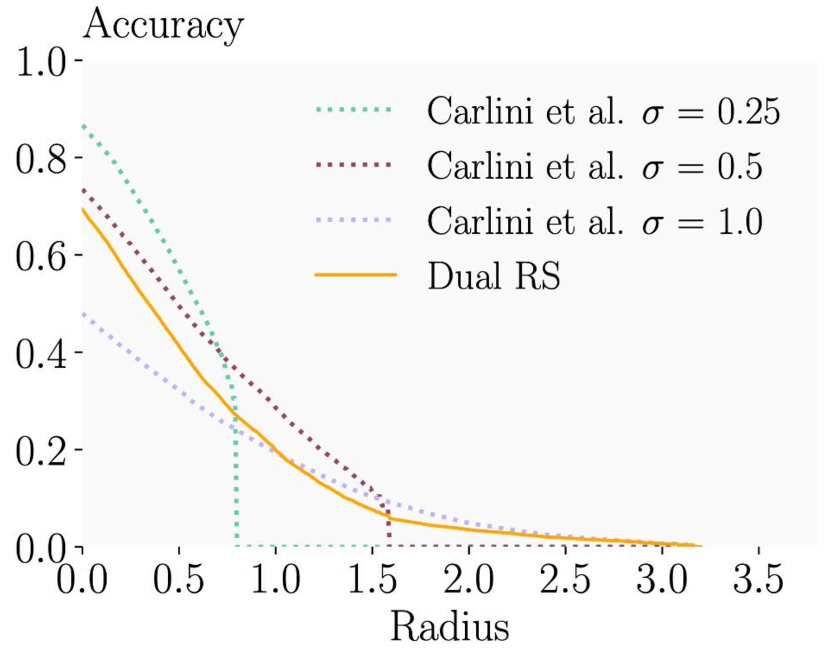
Predict optimal σ
Predict optimal σ on the whole dataset

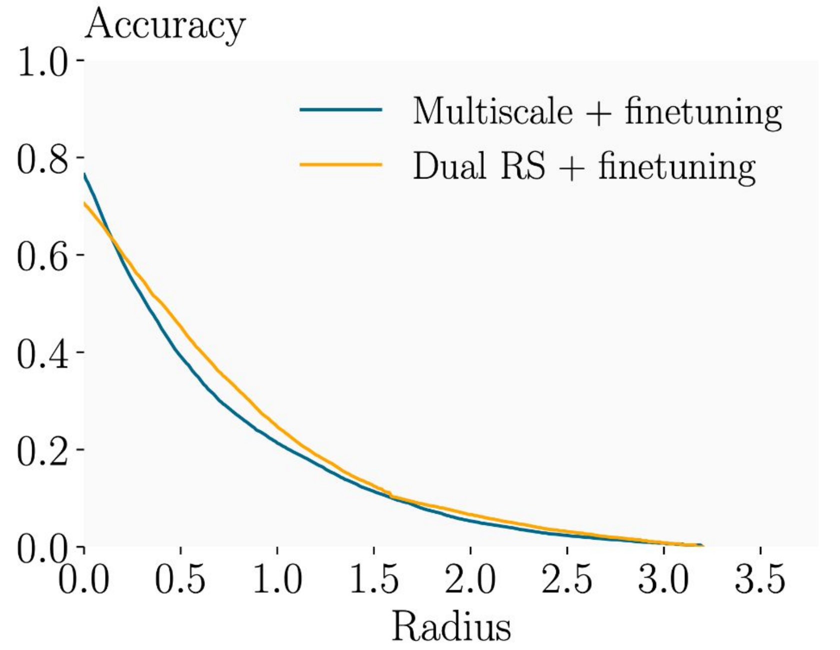
Fine-tune the classifier
Fine-tune the classifier on images augmented by Gaussian noise with the predicted σ

Optional iteration

Better than SOTA input-dependent RS

CIFAR-10 relative gains: **15.6%** at r=0.5, **20.0%** at r=0.75, **15.7%** at r=1.0



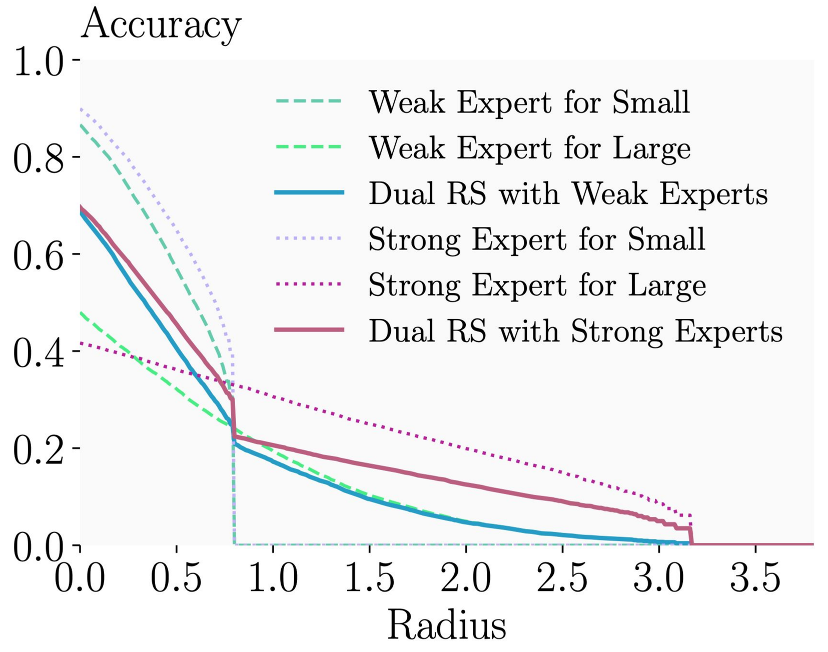


IMAGENET relative gains: **8.6%** at r=0.5, **17.1%** at r=1.0, **9.1%** at r=1.5

Method		σ	0.0	0.5	1.0	1.5	2.0
Off-the-shelf	Carlini at al.	0.5	74.4	64.4	52.4	34.8	0.0
		1.0	56.0	47.8	37.4	29.4	24.0
	Ours	{0.5, 1.0}	67.8	54.6	40.4	28.0	15.6
Adaptive finetuning	Multiscale	{0.5, 1.0}	75.0	55.8	41.0	30.8	25.2
	Ours		74.0	**60.6**	**48.0**	**33.6**	17.0

Delve into Dual RS

Different σ Routing to Different Experts:



Trade-off between different σ candidates: