Technical University of Munich
Department of Computer Science
Data Analytics and Machine Learning Group

DAAD Zuse Schools
Konrad Zuse Schools of Excellence in Artificial Intelligence

Hierarchical Randomized Smoothing

Robustness Certificates for Images, Graphs, and More

Yan Scholten[1], Jan Schuchardt[1], Aleksandar Bojchevski[2], Stephan Günnemann[1]

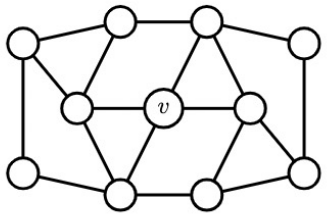
[1]Technical University of Munich [2]University of Cologne

SPONSORED BY THE
Federal Ministry of Education and Research

Motivation: Hierarchical Randomized Smoothing

Context
- Machine learning models are susceptible to adversarial examples
- Robustness certificates provide provable robustness guarantees

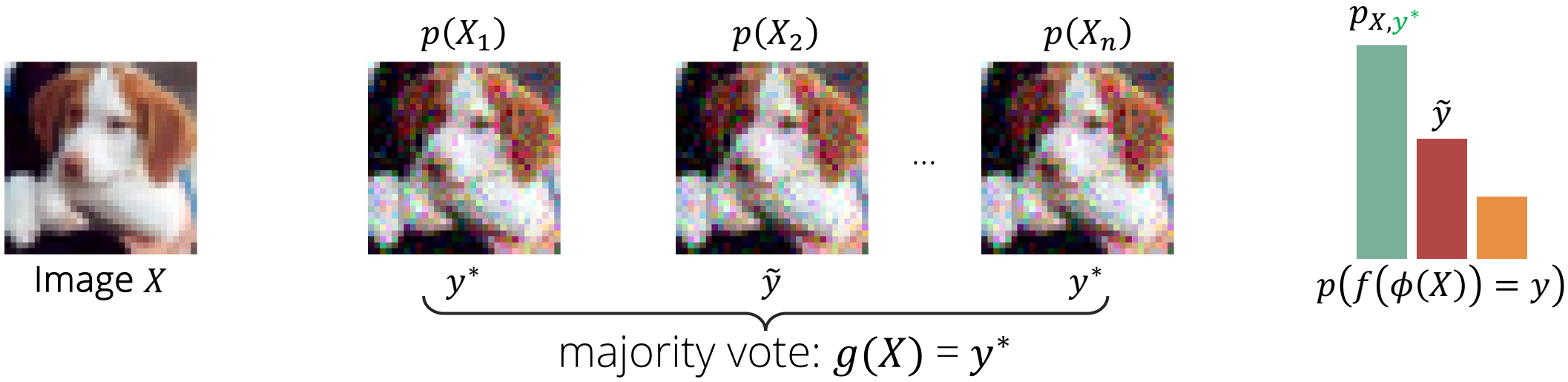
Problem
- Challenging to certify robustness on decomposable data (e.g. images, graphs, ...)
- Existing approaches sacrifice robustness over accuracy or vice versa



Solution: Hierarchical Randomized Smoothing
- Model-agnostic, efficient & highly flexible certification framework

Background: Randomized Smoothing

Majority vote under randomized smoothing of objects



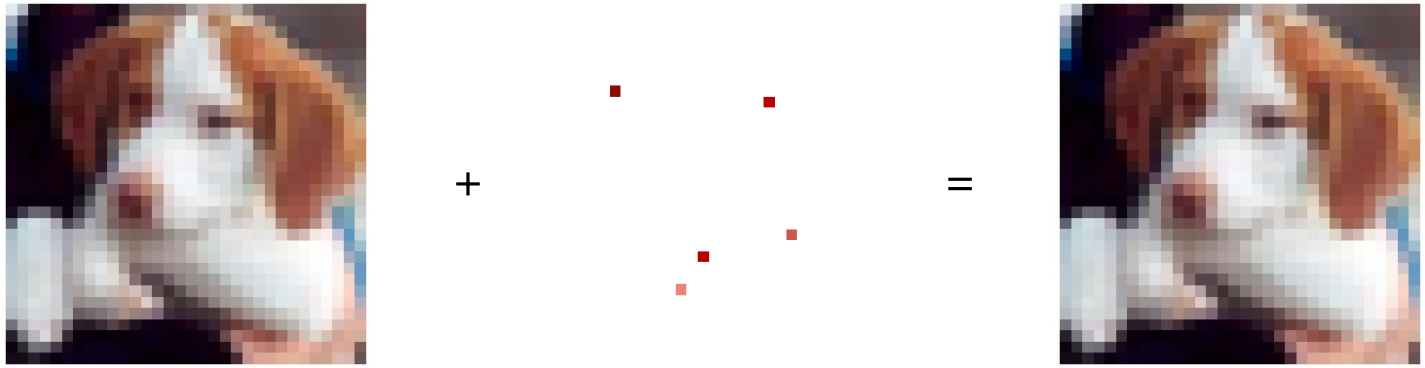
How to certify robustness under randomized smoothing?

- Derive a lower bound on $p_{\tilde{X},y^*}$
- If $\underline{p_{\tilde{X},y^*}} > 0.5$ for any $\tilde{X} \in \mathcal{B}(X)$ then the smoothed classifier g is certifiably robust

Threat Model $\mathcal{B}^{r}_{p,\epsilon}(\boldsymbol{X})$

Adversarial perturbations

- Adversaries can perturb at most entities r
- Perturbation strength bounded by ϵ under a ℓ_p-norm



$f(\boldsymbol{X}) = \text{“}dog\text{”}$ ϵ $f(\boldsymbol{X} + \epsilon) = \text{“}cat\text{”}$

How can we guarantee robustness under such adversarial perturbations?

Hierarchical Smoothing Distribution

Hierarchical smoothing distribution: Partial smoothing of objects

1. **Upper-level smoothing**: Sample an entity indicator
2. **Lower-level smoothing**: Sample additive noise for a subset of entities

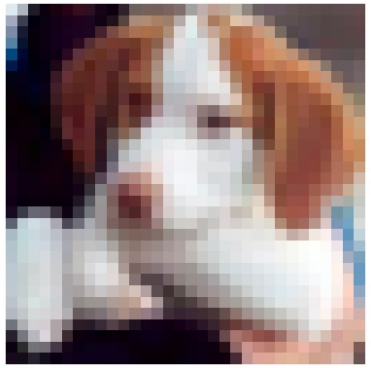
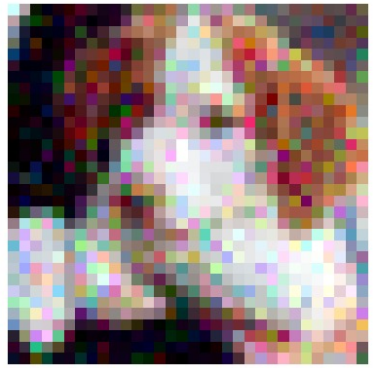


Image $\boldsymbol{X}$

$\boldsymbol{\tau}_i \sim Ber(p)$

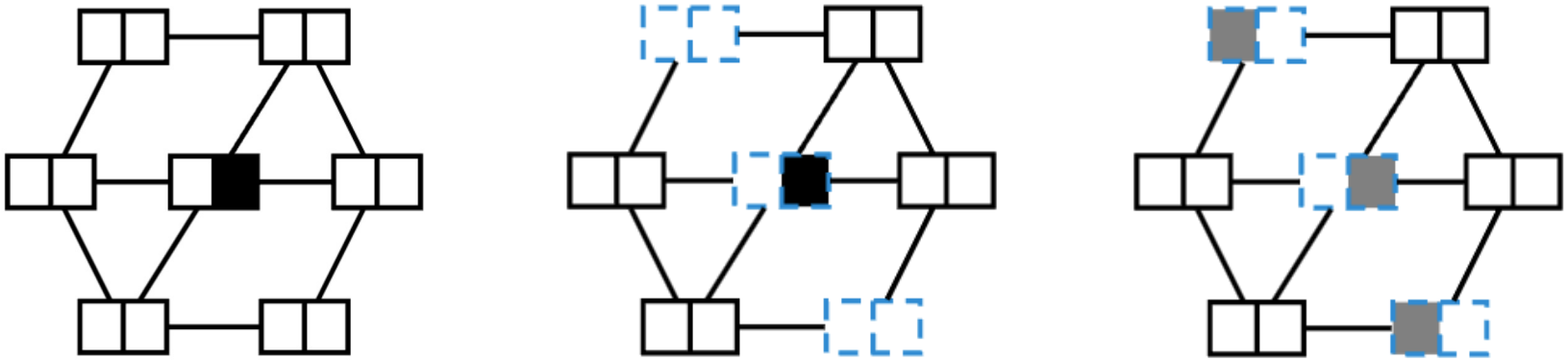


$\boldsymbol{W} \sim \mu_{\boldsymbol{X}}(\boldsymbol{W}|\boldsymbol{\tau})$

Hierarchical Smoothing Distribution

Hierarchical smoothing distribution: Partial smoothing of objects

1. **Upper-level smoothing**: Sample an entity indicator
2. **Lower-level smoothing**: Sample additive noise for a subset of entities

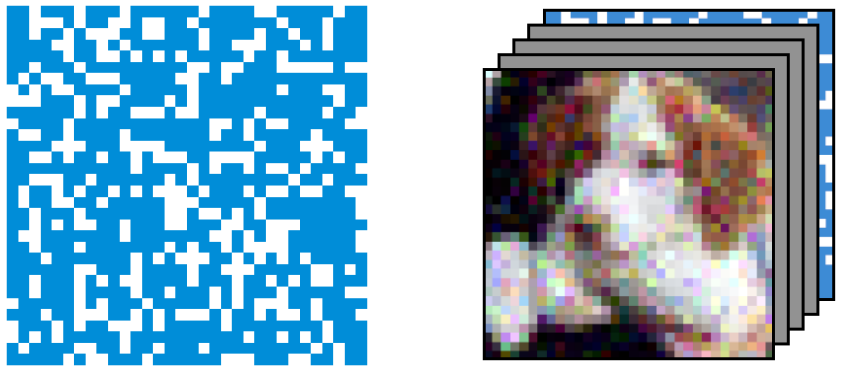


Robustness Certificates for Hierarchical Smoothing

How to compute a lower bound on $p_{\tilde{X}, y^*}$ under hierarchical smoothing?

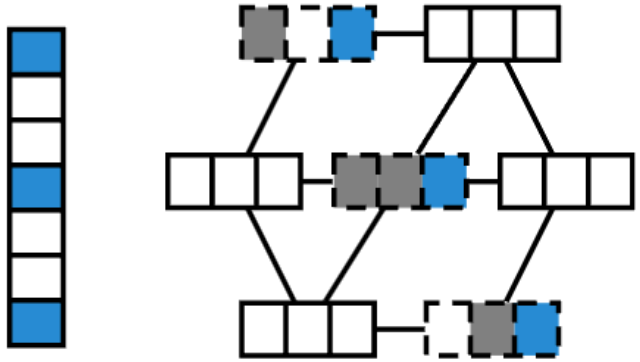
- Append entity indicator $\boldsymbol{\tau}$ to the object $\boldsymbol{X}$

Image data



$\boldsymbol{\tau}_i \sim Ber(p)$ $\quad$ $\boldsymbol{W}|\boldsymbol{\tau} \sim \mu_{\boldsymbol{X}}(\boldsymbol{W}|\boldsymbol{\tau})$

Graph data



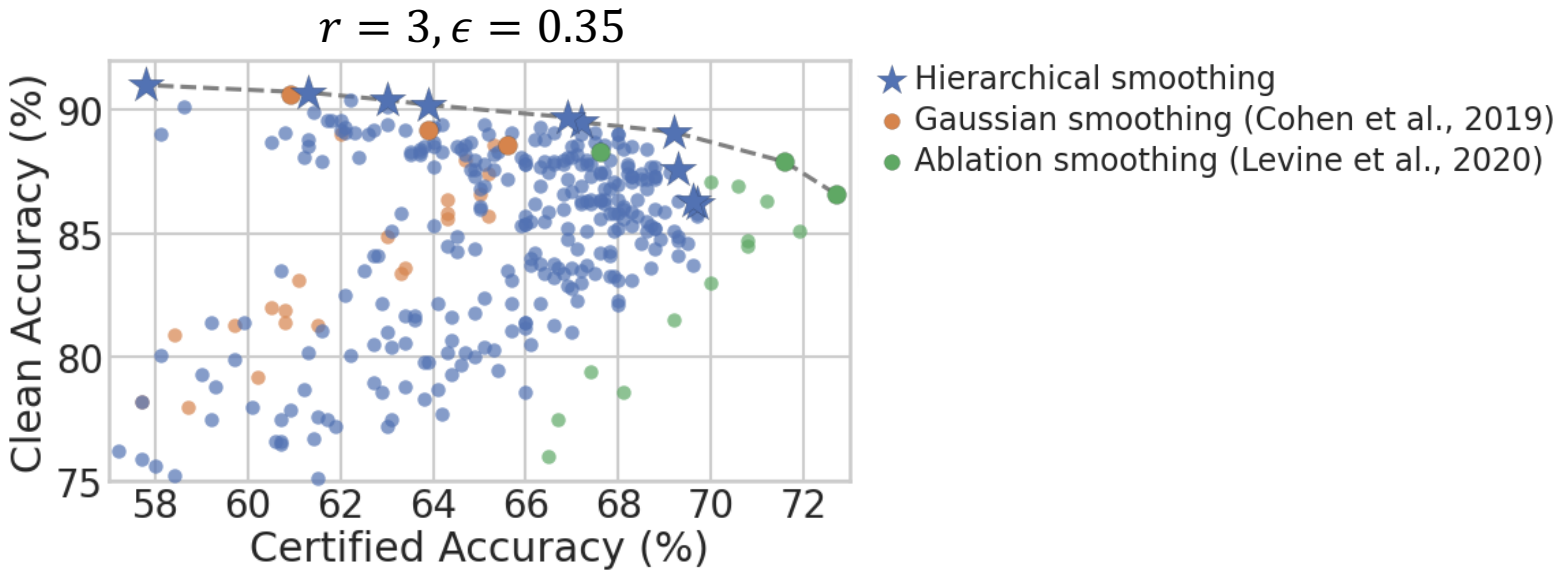
$\boldsymbol{\tau}_i \sim Ber(p)$ $\quad$ $\boldsymbol{W}|\boldsymbol{\tau} \sim \mu_{\boldsymbol{X}}(\boldsymbol{W}|\boldsymbol{\tau})$

- Allows to reuse existing bound for the lower-level smoothing distribution

Hierarchical Smoothing Certificates for Images

Initializing hierarchical smoothing with Gaussian smoothing

- Perturbation strength bounded under the ℓ_2-norm
- ResNet50 on CIFAR10

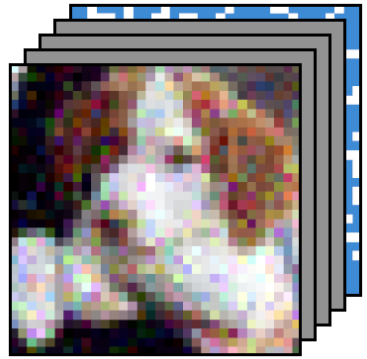


Jeremy Cohen, Elan Rosenfeld, J. Zico Kolter. Certified Adversarial Robustness via Randomized Smoothing. ICML 2019.
Alexander Levine, Soheil Feizi. Robustness Certificates for Sparse Adversarial Attacks by Randomized Ablation. AAAI 2020.

tl;dr Hierarchical Randomized Smoothing

Hierarchical Randomized Smoothing
- First certificate for hierarchical (mixture) smoothing
- Superior robustness-accuracy trade-offs
- Model-agnostic, efficient and highly flexible

$\tau_i \sim Ber(p)$



$\boldsymbol{W}|\boldsymbol{\tau} \sim \mu_{\boldsymbol{X}}(\boldsymbol{W}|\boldsymbol{\tau})$