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

Hierarchical Randomized Smoothing

Yan Scholten[1] Jan Schuchardt[1] Aleksandar Bojchevski[2] Stephan Günnemann[1]
[1]Technical University of Munich [2]University of Cologne

Zuse Schools
Konrad Zuse Schools of Excellence in Artificial Intelligence

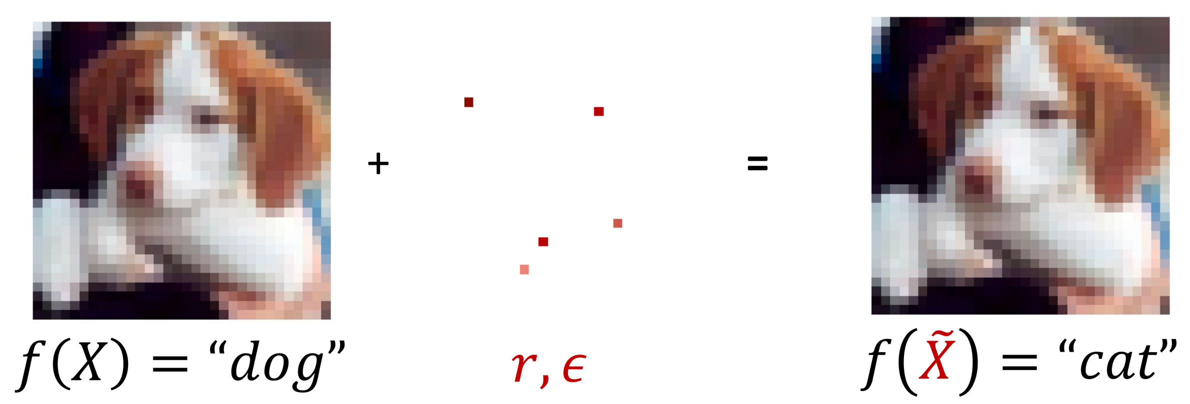
tl;dr: Novel robustness certificates for decomposable data
- Adversaries can perturb a subset of all entities of an object (e.g. pixels of an image, nodes of a graph)
- We propose a highly flexible certification framework for continuous and discrete domains
- Superior robustness-accuracy trade-offs under our threat model

Context

- **Machine learning models are susceptible to adversarial perturbations**
- **Robustness certificates provide provable robustness guarantees**

Problem

Certifying robustness on decomposable data (e.g. images, graphs, …) is challenging when adversarial perturbations are bounded by both: (1) the number of perturbed entities r, and (2) perturbation strength ϵ

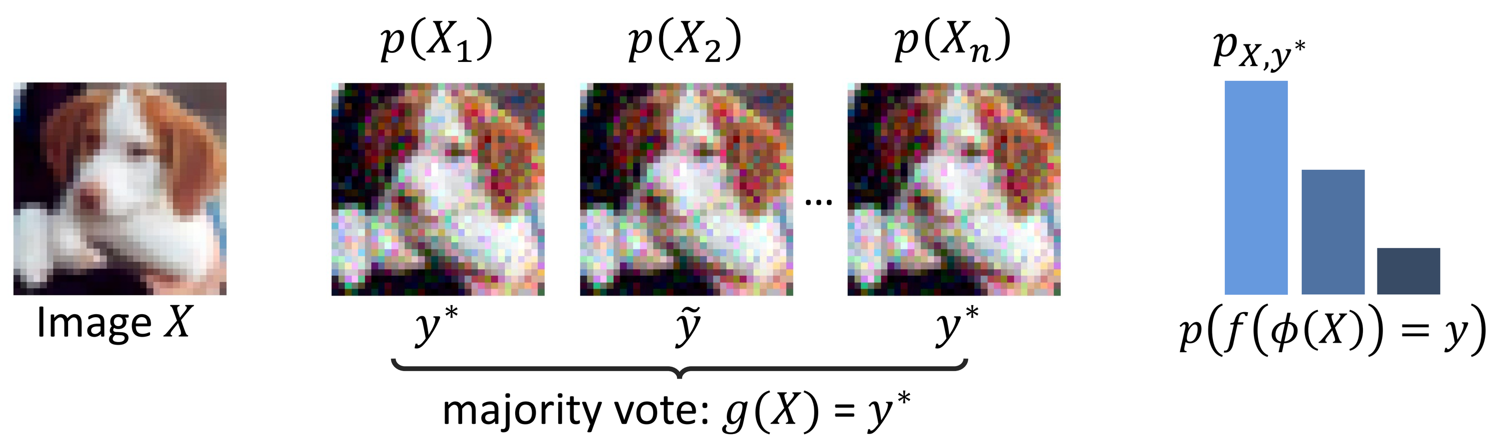


How can we guarantee robustness under such adversarial perturbations?

Existing approaches sacrifice robustness over accuracy or vice versa

Background: Randomized smoothing

- Sample smoothed images $X_i \sim \phi(X)$ from smoothing distribution ϕ
- Classify them with base classifier f and certify the majority vote

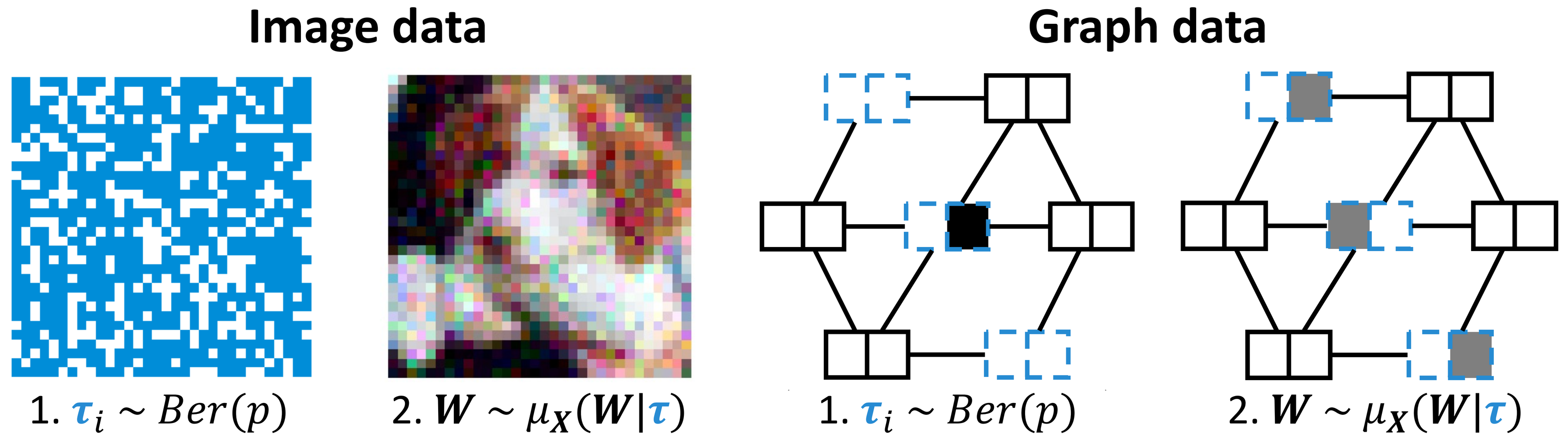


How to certify robustness under randomized smoothing?
- Derive lower bound $\underline{p_{\tilde{X},y^*}}(p_{X,y^*})$ on probability $p_{\tilde{X},y^*}$ to classify $\tilde{X}$ as y^*
- Smoothed classifier g is certifiably robust if

$$\underline{p_{\tilde{X},y^*}}(p_{X,y^*}) > 0.5 \text{ for any perturbed } \tilde{X} \in \mathcal{B}(X)$$

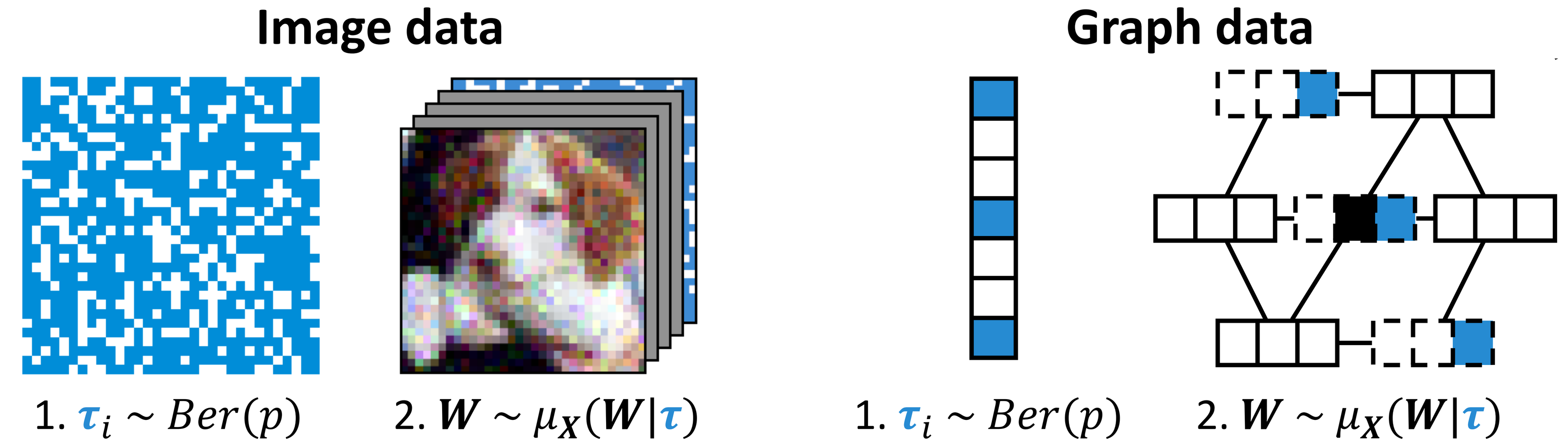
Hierarchical smoothing distribution

1. **Upper-level smoothing**: Sample indicator $\tau_i \sim Ber(p)$ with probability p
2. **Lower-level smoothing** μ: Sample additive noise for indicated entities only



How to certify robustness?

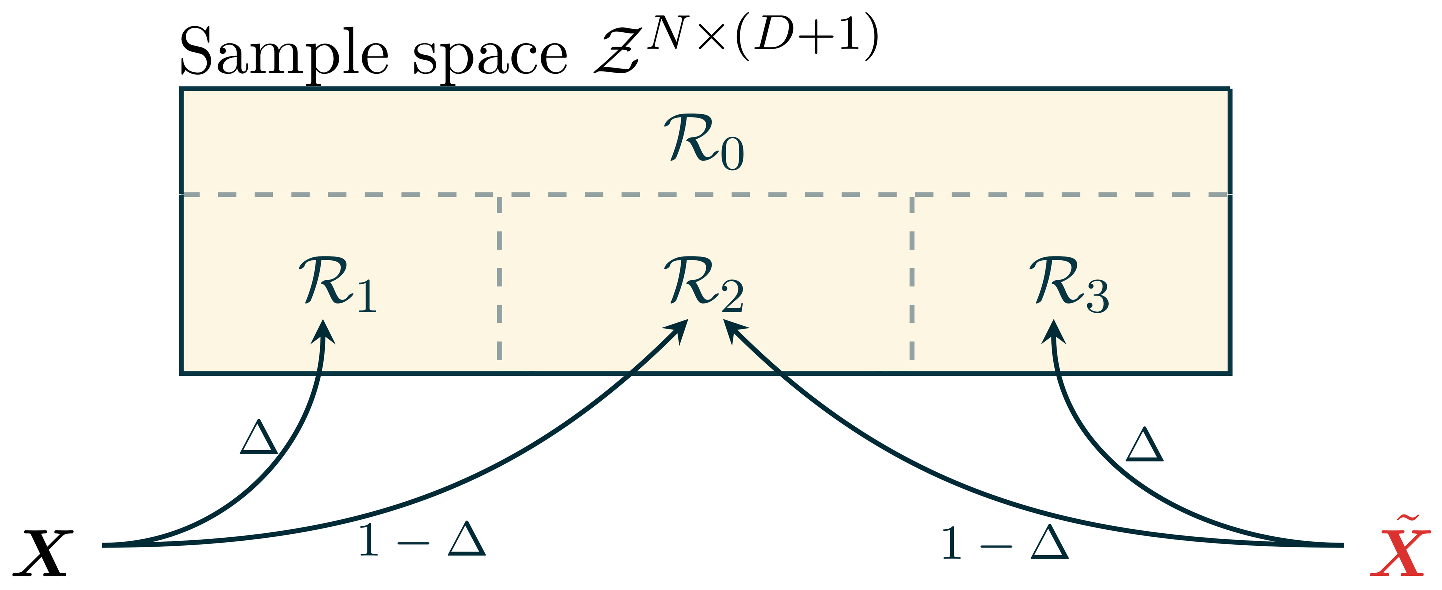
- Append indicator τ to the object X
- Construct a new base classifier f operating on this higher-dimensional space



Certify robustness by reusing existing bounds
1. Compute constant $\Delta = 1 - p^r$ for selection probability p and radius r
2. Plug Δ into existing lower bound $\underline{p_{\tilde{X},y^*}}(p_{X,y^*})$ for the lower-level distribution

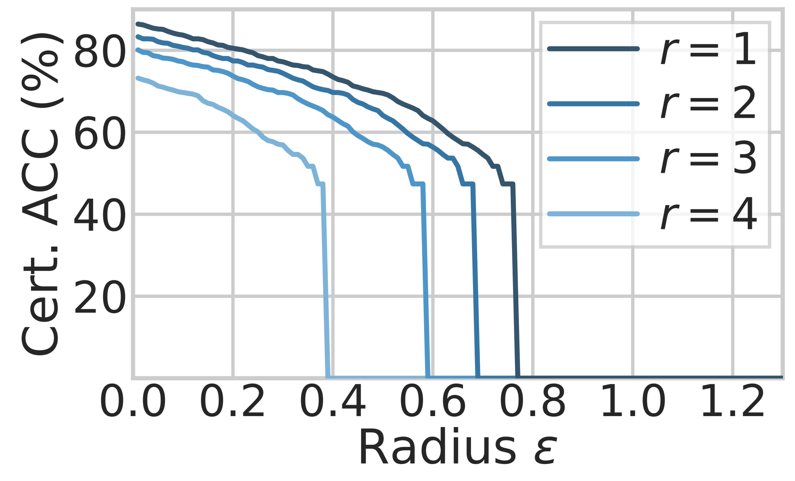
Why can we integrate robustness guarantees for the lower-level smoothing distribution?

- Partition the sample space into disjoint regions R_0, R_1, R_2, R_3
- Supports S_X and $S_{\tilde{X}}$ for hierarchical smoothing around X and $\tilde{X}$ intersect only for samples where all perturbed entities are selected by τ (Region R_2)
- This allows the certificate to separate clean from perturbed entities



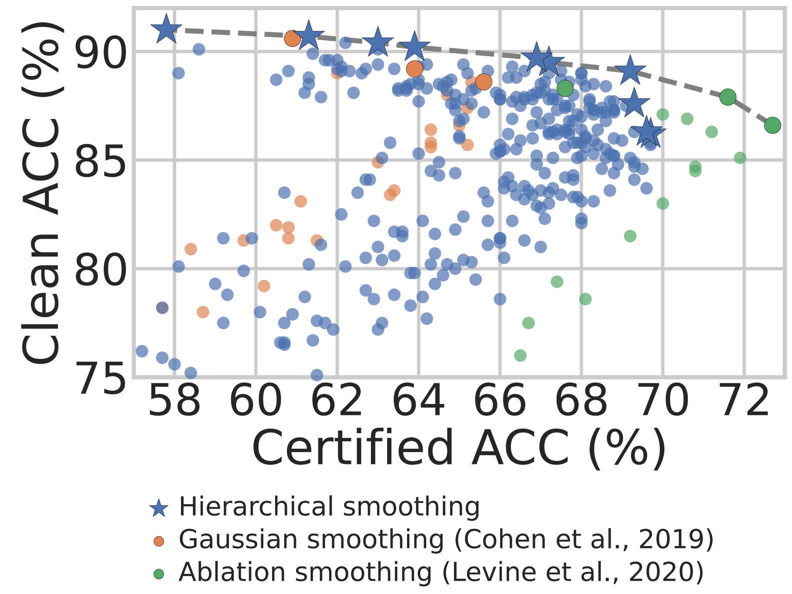
Experimental evaluation

Certificate captures robustness w.r.t. both radii r and ϵ

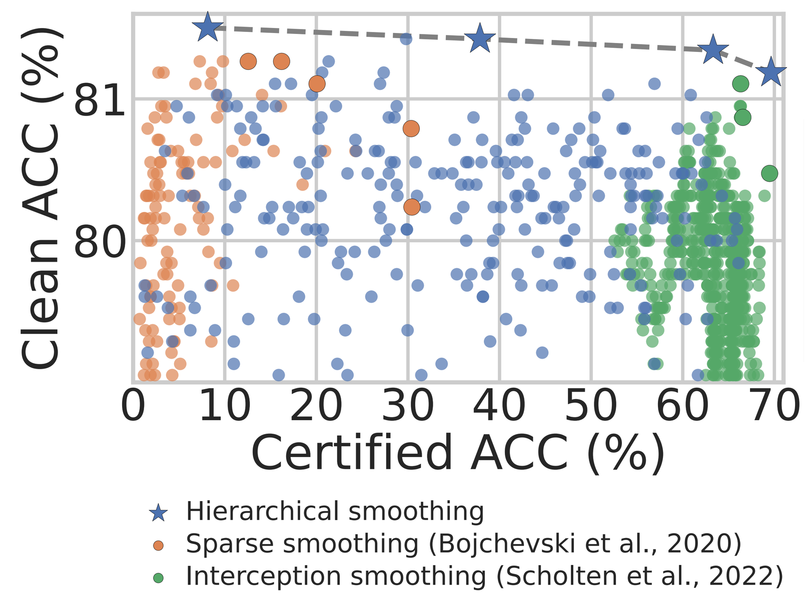


We expand the Pareto-front w.r.t. robustness and accuracy

Image classification
- Perturbation strength bounded under ℓ_2-norm ($r=3, \epsilon=0.35$)
- Hierarchical smoothing with Gaussian smoothing
- Smoothed ResNet50 on CIFAR10



Node classification
- Sparse attribute perturbations ($r=1, r_a=0, r_d=40$)
- Hierarchical smoothing with sparse smoothing
- Smoothed GAT on CoraML



Paper, code, and more

SPONSORED BY THE
Federal Ministry of Education and Research