

Hierarchical Randomized Smoothing

Robustness Certificates for Images, Graphs, and More

Yan Scholten¹, Jan Schuchardt¹, Aleksandar Bojchevski², Stephan Gunnemann¹

¹Technical University of Munich ²University of Cologne

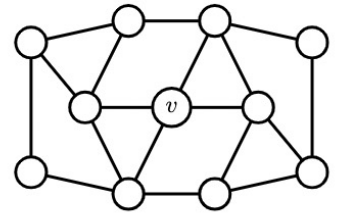
Motivation: Hierarchical Randomized Smoothing

Context

- Machine learning models are susceptible to adversarial perturbations
- [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

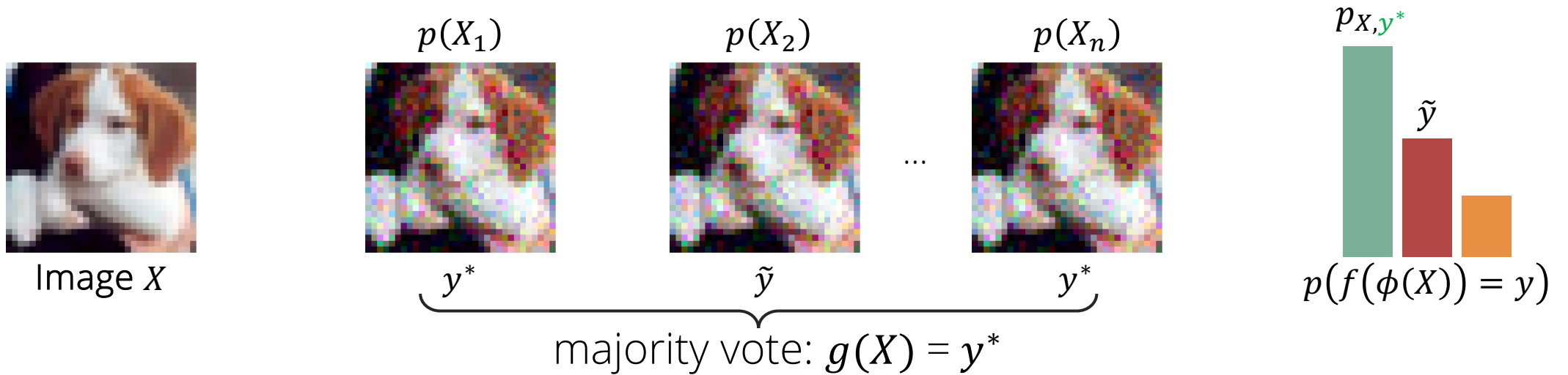


Solution: Hierarchical Randomized Smoothing

- Model-agnostic, efficient & highly flexible certification framework

Background: Randomized Smoothing

Majority vote under randomized smoothing of objects



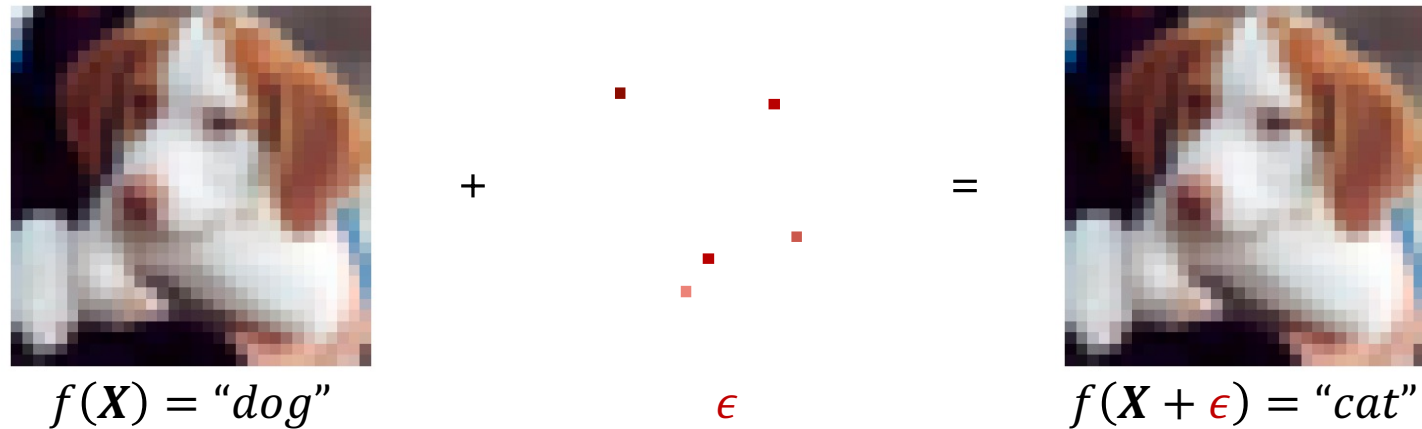
How to certify robustness under randomized smoothing?

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

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

Adversarial perturbations

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



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

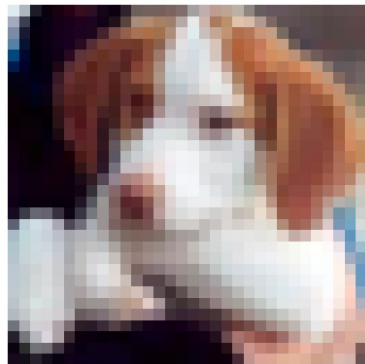
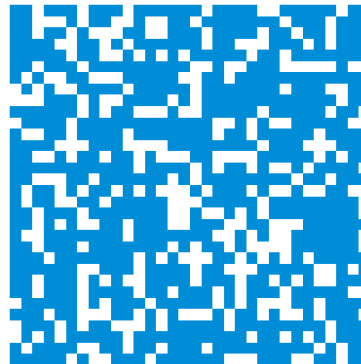


Image X



$$\tau_i \sim \text{Ber}(p)$$

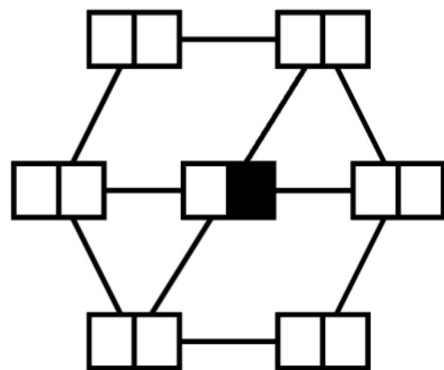


$$W \sim \mu_X(W|\tau)$$

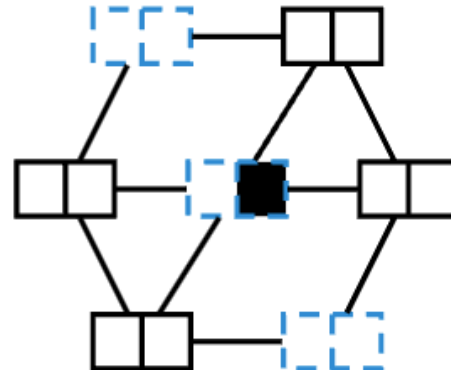
Hierarchical Smoothing Distribution

Hierarchical smoothing distribution: Partial smoothing of objects

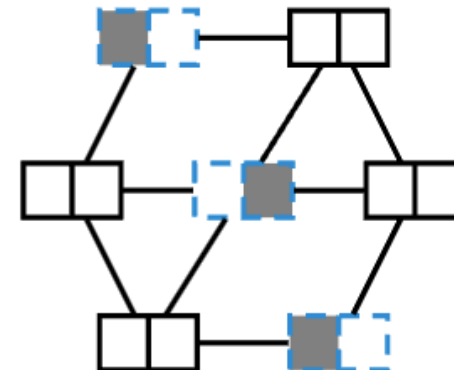
1. Upper-level smoothing: Sample an entity indicator
2. Lower-level smoothing: Sample additive noise for a subset of entities



Graph X



$$\tau_i \sim \text{Ber}(p)$$



$$W \sim \mu_X(W|\tau)$$

Robustness Certificates for Hierarchical Smoothing

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

- Append entity indicator τ to the object X

Image data



$$\tau_i \sim \text{Ber}(p)$$

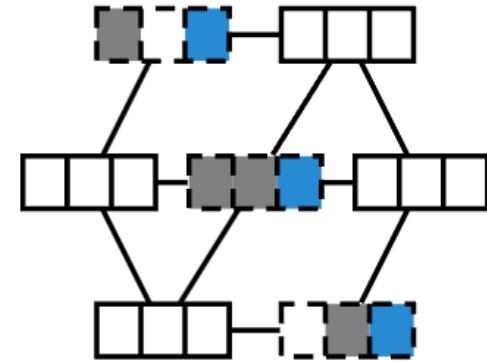


$$W|\tau \sim \mu_X(W|\tau)$$

Graph data



$$\tau_i \sim \text{Ber}(p)$$



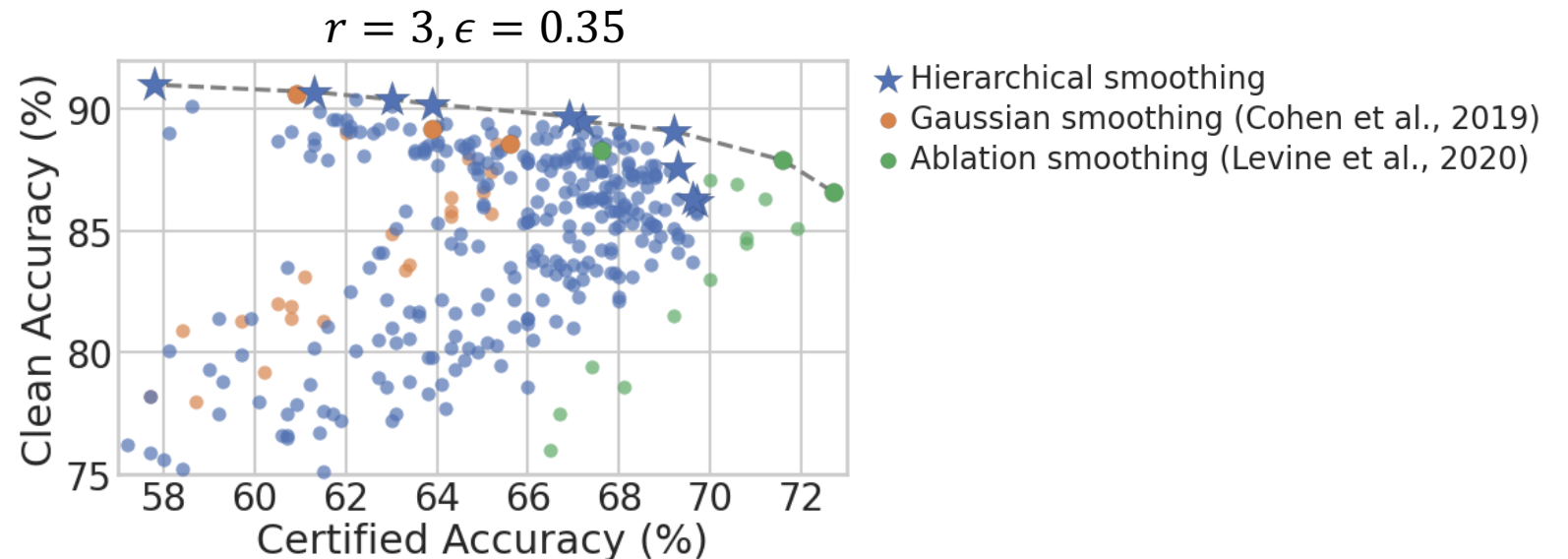
$$W|\tau \sim \mu_X(W|\tau)$$

- 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.

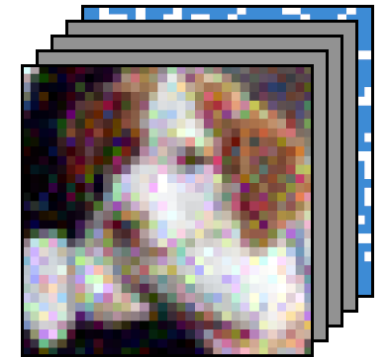
tl;dr Hierarchical Randomized Smoothing

Hierarchical Randomized Smoothing

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



$$\tau_i \sim \text{Ber}(p)$$



$$W|\tau \sim \mu_X(W|\tau)$$

