Intriguing Properties of Input-Dependent Randomized Smoothing (IPIDRS)

Peter Súkeník [1,2], Aleksei Kuvshinov [2], Stephan Günnemann [2]

[1]Institute of Science and Technology Austria (ISTA)
[2]Technical University of Munich (TUM)

July 13, 2022

Input-Dependent Randomized Smoothing - Motivation

- Problems of standard Randomized Smoothing (RS):

Input-Dependent Randomized Smoothing - Motivation

- Problems of standard Randomized Smoothing (RS):

➢ Certified accuracy "waterfalls".

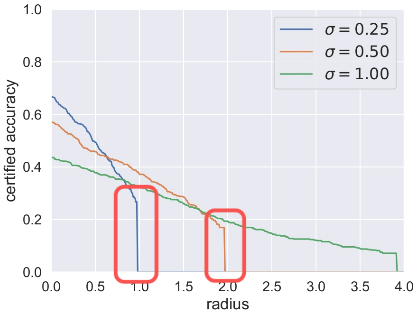


Figure: Source: [Cohen et al., 2019], modified.

Input-Dependent Randomized Smoothing - Motivation

- Problems of standard Randomized Smoothing (RS):
- ➢ Certified accuracy “waterfalls”.
- ➢ Robustness vs. accuracy tradeoff [Gao et al., 2020]

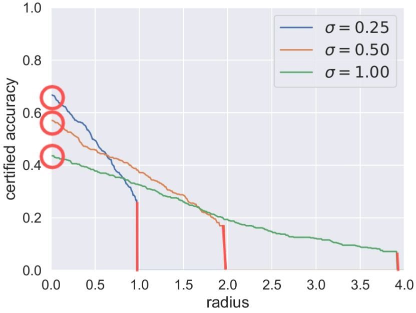
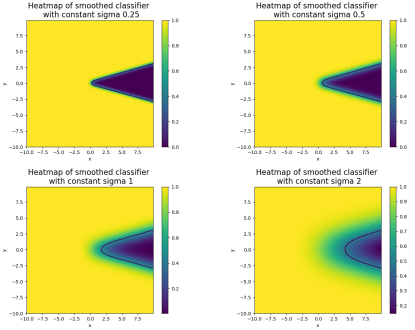


Figure: Source: [Cohen et al., 2019], modified.

Input-Dependent Randomized Smoothing - Motivation

- Problems of standard Randomized Smoothing (RS):
- ➢ Certified accuracy "waterfalls".
- ➢ Robustness vs. accuracy tradeoff [Gao et al., 2020]
- ➢ Shrinking phenomenon (and subsequent class-wise unfairness) [Mohapatra et al., 2020]



ICML International Conference On Machine Learning

Input-Dependent Randomized Smoothing - Motivation

- Problems of standard Randomized Smoothing (RS):
  - ➢ Certified accuracy "waterfalls".
  - ➢ Robustness vs. accuracy tradeoff [Gao et al., 2020]
  - ➢ Shrinking phenomenon (and subsequent class-wise unfairness) [Mohapatra et al., 2020]
- Use input-dependent $\sigma(x)$ instead of σ!

Curse of dimensionality

Curse of dimensionality

Theorem 2.4

Let x_0 be certified point, x_1 potential adversary, p_B probability of runner-up class at point x_0, σ_i^2 the smoothing variance at x_i and N the dimension. The following two implications hold:

- If $\sigma_0 > \sigma_1$ and

$$\log\left(\frac{\sigma_1^2}{\sigma_0^2}\right) + 1 - \frac{\sigma_1^2}{\sigma_0^2} < \frac{2\log(p_B)}{N},$$

  then x_1 **cannot be certified** w.r.t. x_0.

- If $\sigma_0 < \sigma_1$ and

$$\log\left(\frac{\sigma_1^2}{\sigma_0^2}\frac{N-1}{N}\right) + 1 - \frac{\sigma_1^2}{\sigma_0^2}\frac{N-1}{N} < \frac{2\log(p_B)}{N},$$

  then x_1 **cannot be certified** w.r.t. x_0.

Curse of dimensionality

Table: Theoretical lower-thresholds for σ_1/σ_0 for different data dimensions and runner-up class probabilities p_B.

p_A	0.1	0.01	0.001	0.00007	N
MNIST	0.946	0.924	0.908	0.892	784
CIFAR10	0.973	0.961	0.953	0.945	3072
ImageNet	0.997	0.995	0.994	0.993	196608

Curse of dimensionality

Table: Theoretical lower-thresholds for σ_1/σ_0 for different data dimensions and runner-up class probabilities p_B.

p_A	0.1	$\mathbf{0.01}$	0.001	0.00007	N
MNIST	0.946	0.924	0.908	0.892	784
CIFAR10	0.973	0.961	0.953	0.945	**3072**
ImageNet	0.997	0.995	0.994	0.993	196608

Curse of dimensionality

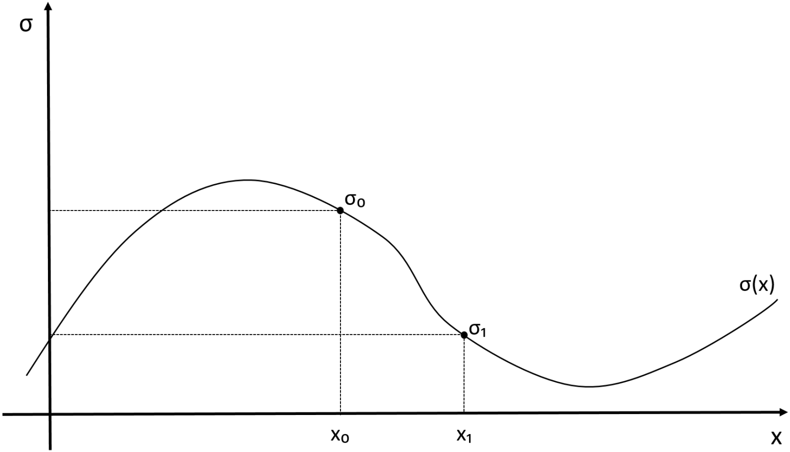


Figure: Problems with the curse of dimensionality.

Curse of dimensionality

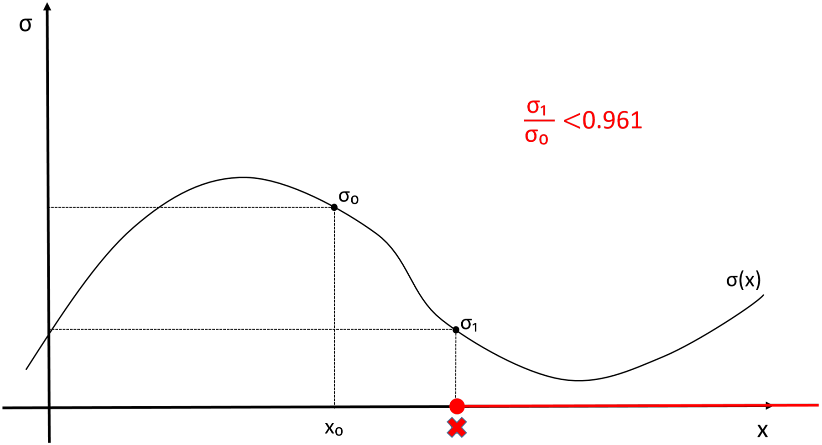


Figure: Problems with the curse of dimensionality.

Curse of dimensionality

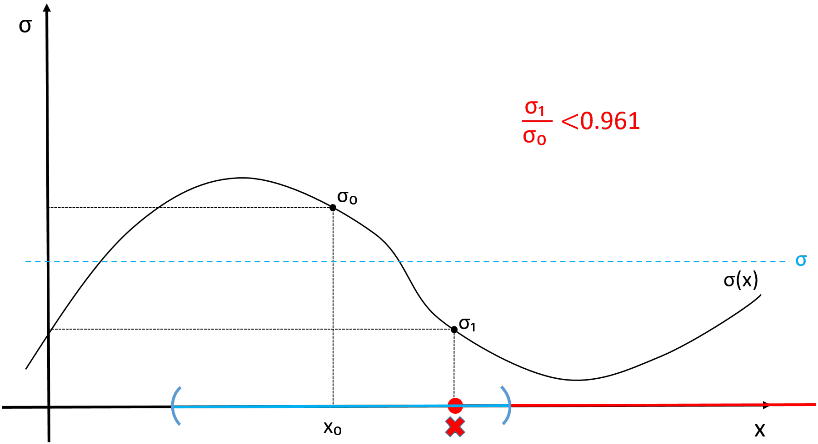


Figure: Problems with the curse of dimensionality.

IDRS can still work!

- If we are careful, IDRS can still be useful!

IDRS can still work!

- If we are careful, IDRS can still be useful!
- We just need that $\sigma(x)$ is r-semi-elastic.

IDRS can still work!

- If we are careful, IDRS can still be useful!
- We just need that $\sigma(x)$ is r-semi-elastic.

Certified radius

Let $\sigma(x)$ be an r-semi-elastic function and x_0, p_B, N, σ_0 as usual. Then, the certified radius at x_0 guaranteed by our method is

$$\text{CR}(x_0) = \sup\{R \geq 0 : \xi(R) < 0.5\}$$

IDRS can still work!

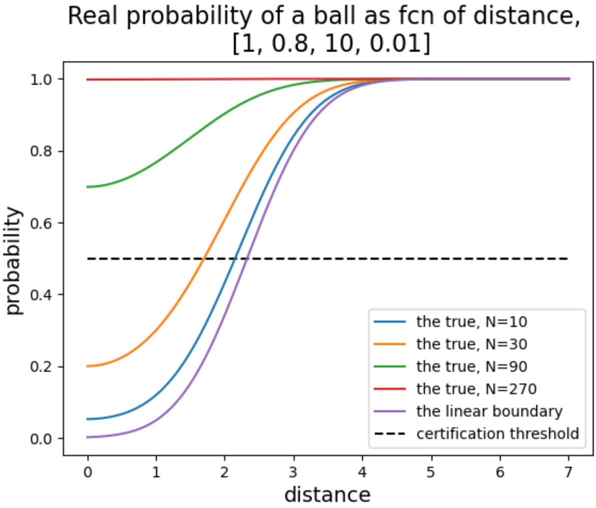
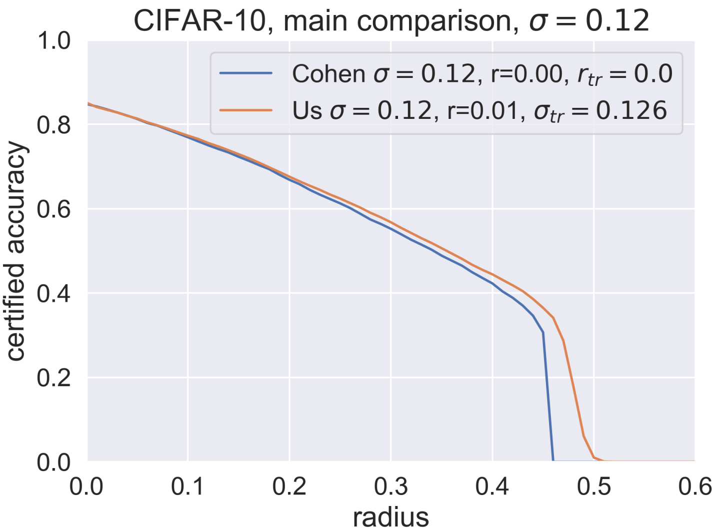


Figure: Numerical evaluation of the certified radii. The function $\xi_>$ and the threshold for different values of N.

Experiments



IPIDRS - Contributions

- Generalize framework of [Cohen et al., 2019].
- Point out the curse of dimensionality for IDRS.
- Build abstract framework which enables justified use of IDRS.
- Demonstrate correctly used IDRS for newly proposed $\sigma(x)$ and compare it to RS.
- Provide additional insights in many aspects of RS and IDRS.

ICML
International Conference
On Machine Learning

References I

Cohen, J., Rosenfeld, E., and Kolter, Z. (2019).
Certified adversarial robustness via randomized smoothing.
In *International Conference on Machine Learning*, pages 1310–1320. PMLR.

Gao, Y., Rosenberg, H., Fawaz, K., Jha, S., and Hsu, J. (2020).
Analyzing accuracy loss in randomized smoothing defenses.
arXiv preprint arXiv:2003.01595.

Mohapatra, J., Ko, C.-Y., Liu, S., Chen, P.-Y., Daniel, L., et al. (2020).
Rethinking randomized smoothing for adversarial robustness.
arXiv preprint arXiv:2003.01249.

ICML
International Conference On Machine Learning

APPENDIX: The $\sigma(x)$ design

Let σ_b be a *base standard deviation*, r the required semi-elasticity, $\{x_i\}_{i=1}^{d}$ the training set, $\mathcal{N}_k(x)$ the k nearest neighbors of x and m the normalization constant. Then:

$$\sigma(x) = \sigma_b \exp\left(r \left(\frac{1}{k} \sum_{x_i \in \mathcal{N}_k(x)} \|x - x_i\| - m \right) \right).$$

APPENDIX: Randomized Smoothing (RS)

- Classifier f susceptible against adversarial attacks $\implies$ robust smoothed classifier g
- $g(x) = \underset{C \in \text{CLASSES}}{\arg\max}\ \mathbb{P}(f(\tilde{x}) = C)$, $\tilde{x} \sim \mathcal{N}(x, \sigma^2 I)$.
- σ does *not* depend on x.
- g has provably large certified l_2 robustness.

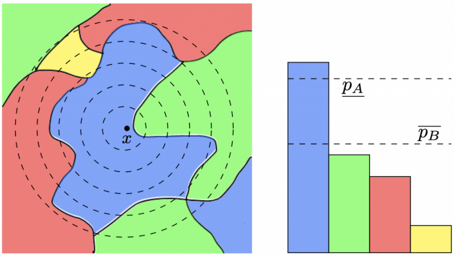


Figure: [Cohen et al., 2019]