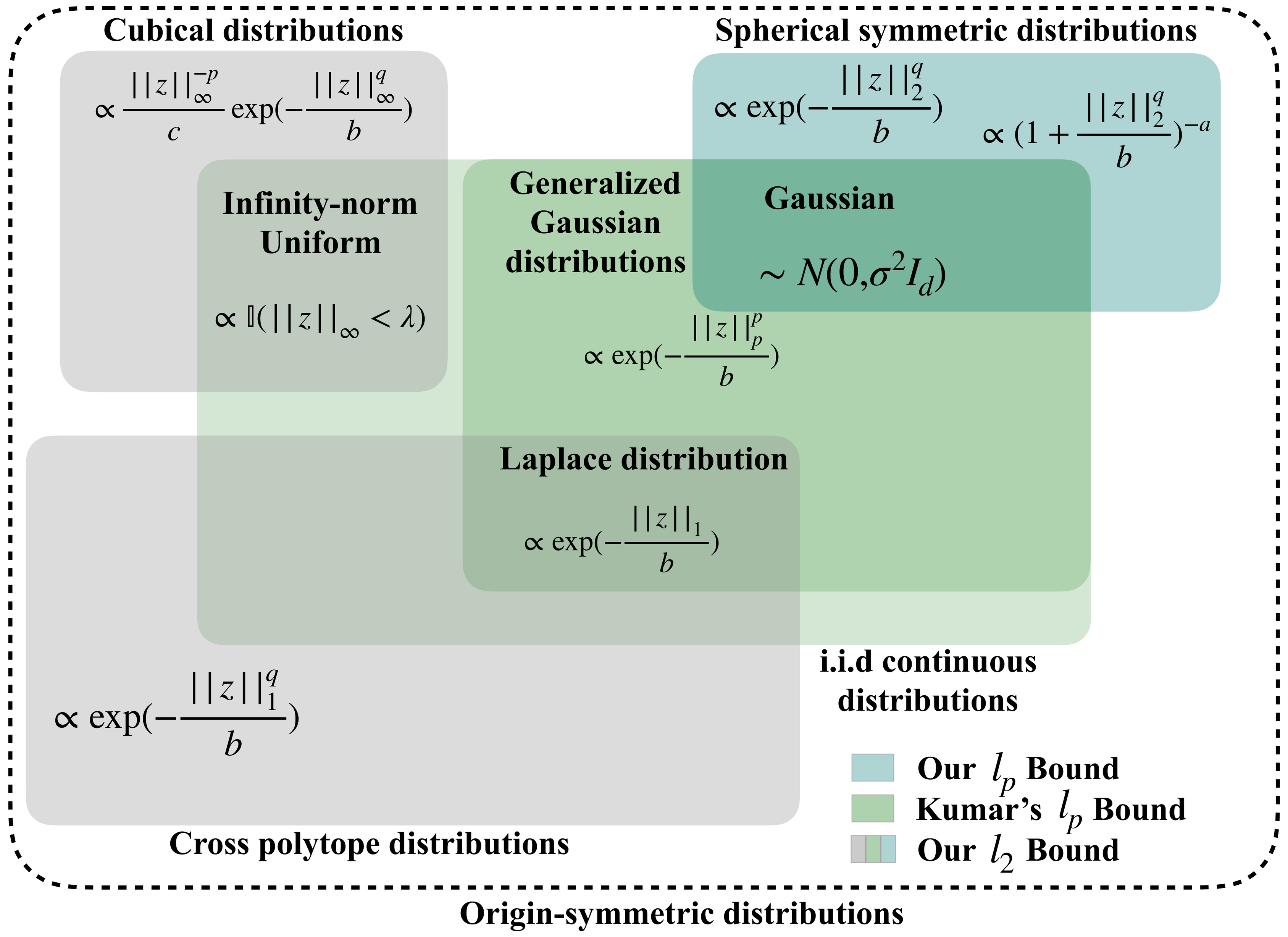
Data Analytics and Machine Learning Group
Department of Informatics
Technical University of Munich

Completing the Picture: Randomized Smoothing Suffers from the Curse of Dimensionality for a Large Family of Distributions

Yihan Wu, Aleksandar Bojchevski, Aleksei Kuvshinov, Stephan Günnemann

www.daml.in.tum.de/curse/

TL;DR: We derived two upper bounds on the certified radius with randomized smoothing method for origin-symmetric distributions. The radius scales as $O(1/\sqrt{d})$.



Main Research Question

What is the best result we can hope for with randomized smoothing?

Randomized smoothing is important

Randomized smoothing is currently the only SOTA certificate that scales to large networks and different settings.

Smoothed classifier: given a base classifier f and a sample x with label c, the smoothed classifier g with distribution q is

$$g_f(x) := E_{z\sim q}[1_{f(x+z)=c}]$$

Certified radius: the largest r such that for any perturbation δ with norm $\|\delta\|_2 < r$, the prediction of perturbed samples does not change.

Certified radius in randomized smoothing: We calculate a **lower bound** of certified radius of the smoothed classifier by finding a lower bound of $g_f(x+\delta)$.

Methods to find the upper bound

Functional optimization method:

Provide a tight lower bound of $g_f(x+\delta)$.

Optimization problem:

$$\underline{g(x+\delta)} := \min_h g_h(x+\delta)$$
$$s.t. g_h(x) = g_f(x)$$

where h is in the feasible set of all classifiers.

The certified radius is the largest norm of δ such that

$$\underline{g(x+\delta)} > 0.5$$

Intuition: we can find an upper bound of certified radius by selecting a classifier h and perturbation δ such that $g_h(x) = g_f(x)$ and $g_h(x+\delta) < 0.5$, then

$$\underline{g(x+\delta)} < g_h(x+\delta) < 0.5$$

and the certified radius is upper bounded by the norm of δ.

We need to find a classifier and a perturbation that simplify the calculation.

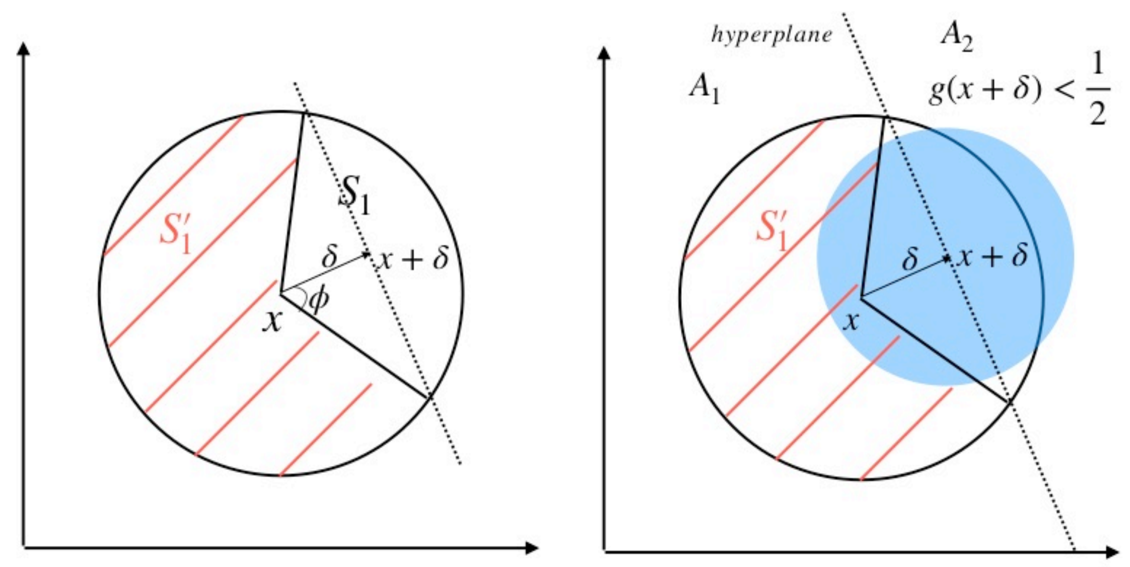
Bound the ℓ_2 radius

S_1 : a hyperspherical sector of a d-dimensional ball.

The classifier we defined is $1_{\{x\in S_1'\}}$ and δ is orthogonal to the hyperplane.

1st upper bound:

$$r < \|\delta\|_2 < \frac{5}{\sqrt{d}}\Psi^{-1}(\frac{g(x)}{1-5\times10^{-7}};q), \quad \Psi(R;q) = \int_{B_R} q(z)dz$$



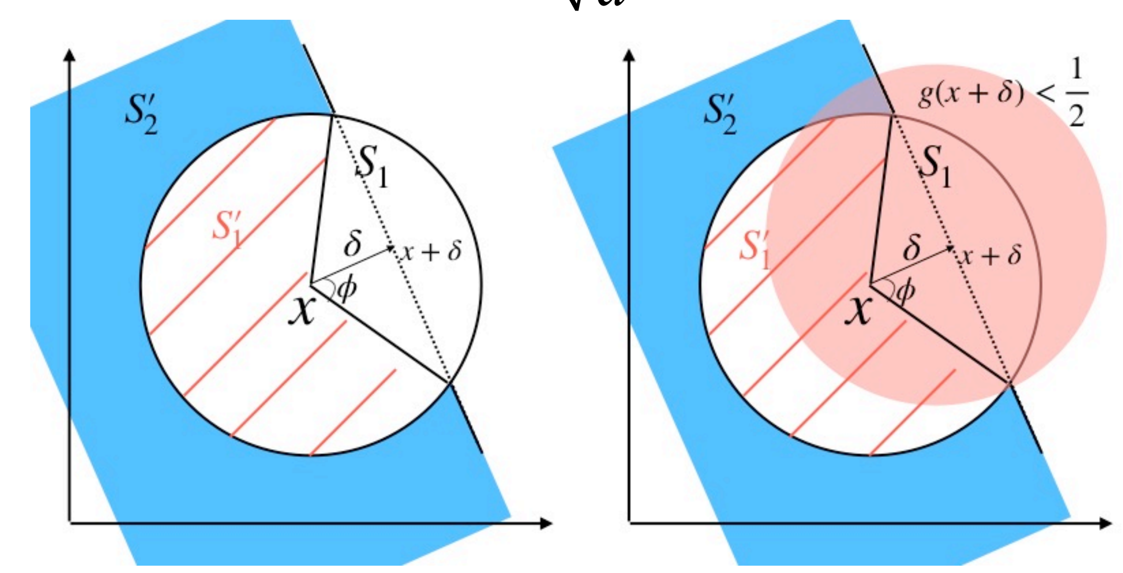
Alternative bound for ℓ_2

S_1: a hyperspherical sector of a d-dimensional ball.

The classifier we defined is $1_{\{x\in S_1'\}\cup\{x\in S_2'\}}$ and δ is orthogonal to the hyperplane. R_x is a sample and distribution based radius.

2nd upper bound:

$$r < \|\delta\|_2 < \frac{5}{\sqrt{d}}R_x$$



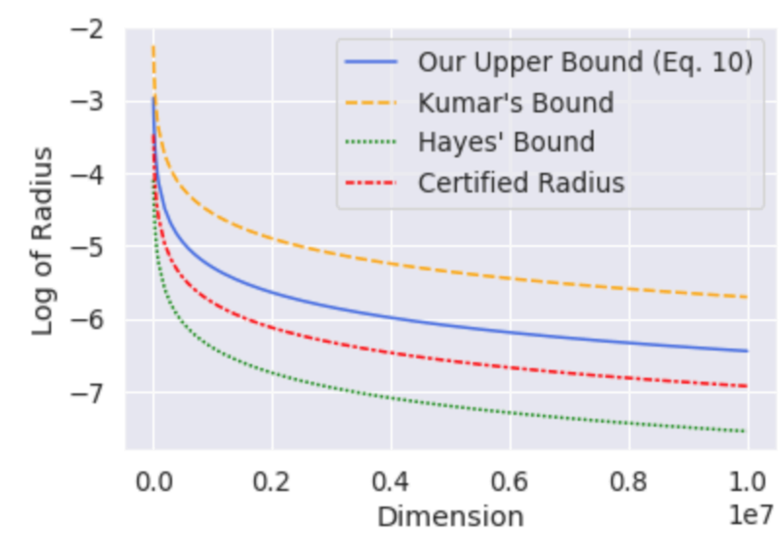
Extension to ℓ_p bounds

We extend our ℓ_2 bounds to ℓ_p bounds with spherical symmetric distributions.

$$r < \frac{5}{d^{1-1/p}}\Psi^{-1}(\frac{g(x)}{1-5\times10^{-7}};q) \qquad r < \frac{5}{d^{1-1/p}}R_x$$

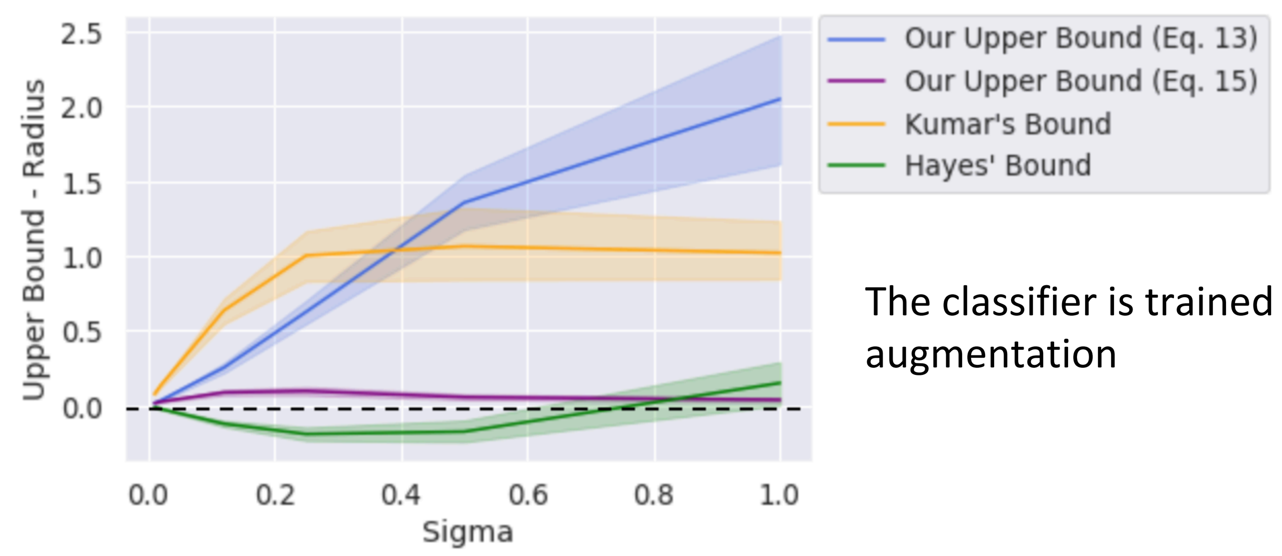
Experiment with Gaussian smoothing

Evaluating our first upper bound for different parameters, we compare our result with SOTA bounds.



$\sigma = 1/\sqrt{d}, g(x) = 0.999$

Evaluating our second upper bound on CIFAR10 dataset with 100 random samples.



The classifier is trained with noise augmentation