Effects of Exponential Gaussian Distribution on (Double Sampling) Randomized Smoothing

Youwei Shu[1], Xi Xiao[1], Derui Wang[2], Yuxin Cao[1], Siji Chen[1], Minhui Xue[2], Linyi Li[3,4], Bo Li[3,5]

[1]Tsinghua University [2]CSIRO's Data61 [3]UIUC [4]Simon Fraser University [5]University of Chicago

BACKGROUND

- **Randomized smoothing**: a provable certified robustness method against vulnerability issues of neural networks.

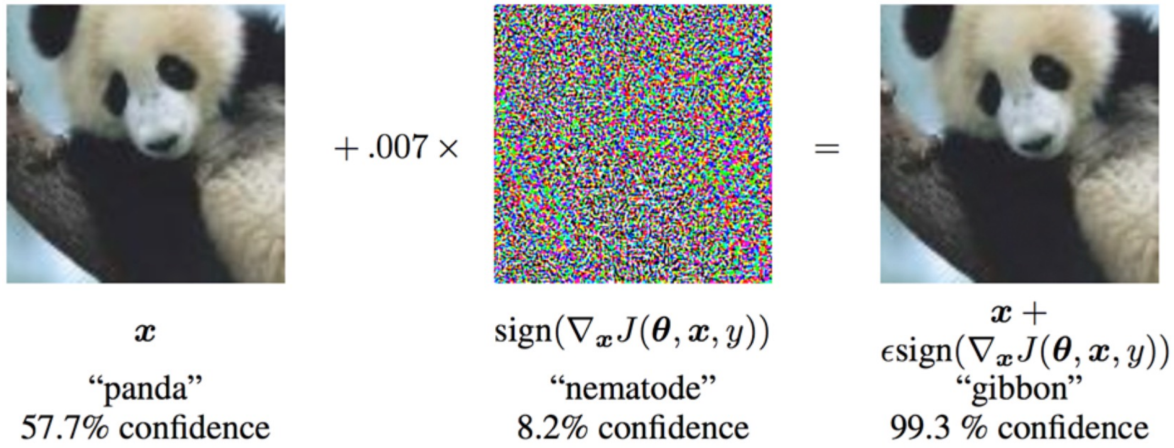


Figure from Goodfellow et al. (2015)

- Definition of smoothed classifier: $\bar{f}(x) \triangleq argmax_c\, \mathbb{P}_{\epsilon \sim D}\{f(x+\epsilon) = c\}$
- When $\epsilon \sim N(0, \sigma^2 I)$, Cohen et al. (2019) derive a concise formula for the certified radius:

$$R = \sigma \Phi^{-1}(\underline{p})$$

Refs: [1] Goodfellow et al. Explaining and Harnessing Adversarial Examples. In ICLR 2015.
[2] Cohen et al. Certified Adversarial Robustness via Randomized Smoothing. In ICML 2019.

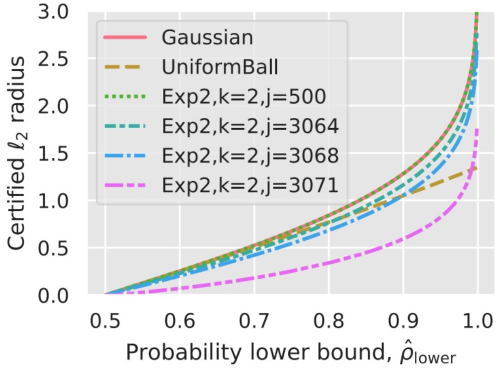
BACKGROUND

- **Essential problems** in Randomized Smoothing:
- 1. the working mechanism of smoothing distribution on base classifiers
- E.g. The dispute between Zhang et al. (2020) and Yang et al. (2020):
- Zhang et al. (2020): General Gaussian enhances the ℓ_2 certified robustness provided by Gaussian.
- Yang et al. (2020): Gaussian is the best distribution; General Gaussian does not defeat Gaussian.

ℓ_2 RADIUS (CIFAR-10)	0.25	0.5	0.75	1.0	1.25	1.5	1.75	2.0	2.25
BASELINE (%)	60	43	34	23	17	14	12	10	8
OURS (%)	**61**	**46**	**37**	**25**	**19**	**16**	**14**	**11**	**9**

ℓ_2 RADIUS (IMAGENET)	0.5	1.0	1.5	2.0	2.5	3.0	3.5
BASELINE (%)	49	37	29	19	15	12	9
OURS (%)	**50**	**39**	**31**	**21**	**17**	**13**	**10**

Zhang et al. (2020)'s results



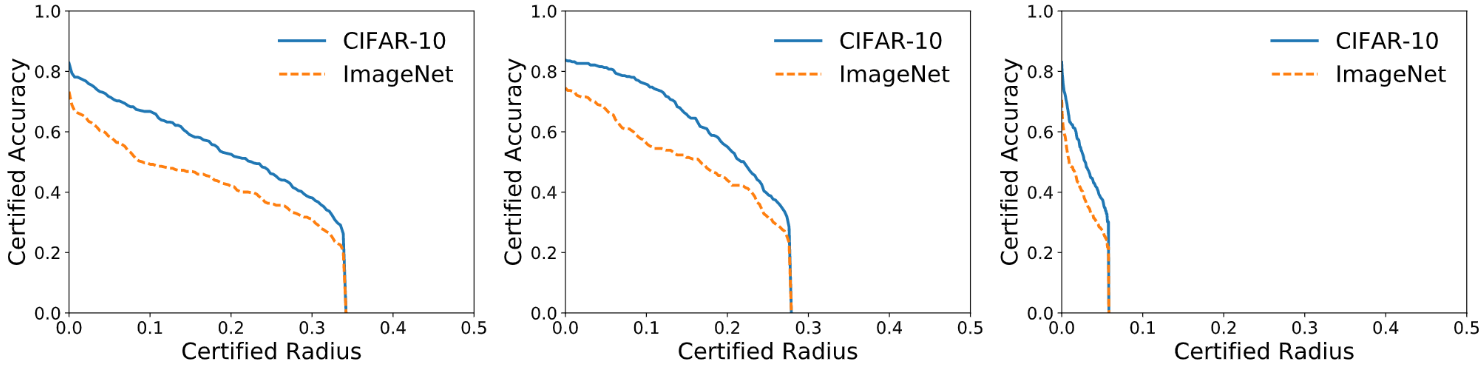
Yang et al. (2020)'s results

How General Gaussian works?

Refs: [1] Zhang et al. Black-Box Certification with Randomized Smoothing: A Functional Optimization Based Framework. In NeurIPS 2020.
[2] Yang et al. Randomized Smoothing of All Shapes and Sizes. In ICML 2020.

BACKGROUND

- **Essential problems** in Randomized Smoothing:
- 2. the curse of dimensionality in Randomized Smoothing



(a) ℓ_3 (b) ℓ_4 (c) ℓ_5

Figure from Hayes (2020)

- Li et al. (2022) first proposed a theoretical solution to the curse of dimensionality in randomized smoothing by introducing General Gaussian into Double Sampling Randomized Smoothing (DSRS).

Can Li et al.'s theoretical solution be improved by improving General Gaussian?

Refs: [1] Hayes. Extensions and Limitations of Randomized Smoothing for Robustness Guarantees. In CVPR 2020.
[2] Li et al. Double Sampling Randomized Smoothing. In ICML 2022.

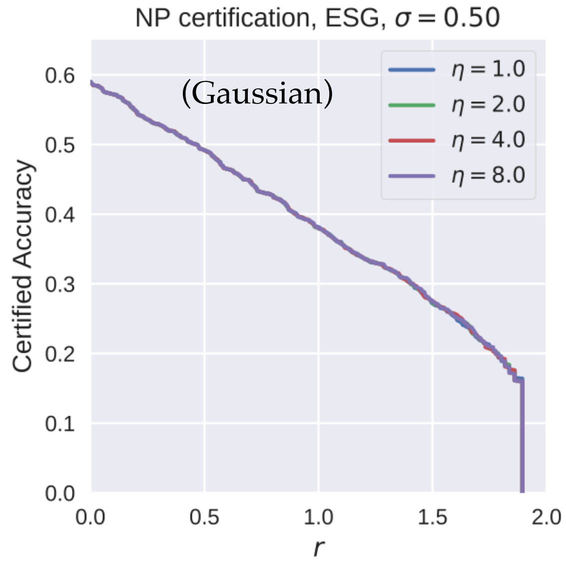
DISTRIBUTION

Distribution	PDF	Notation
Gaussian	$\propto \exp(-\frac{r^2}{2\sigma^2})$	$\mathcal{N}(\sigma)$
Exponential Standard Gaussian (ESG)	$\propto \exp(-\frac{r^\eta}{2\sigma_s^\eta})$	$\mathcal{S}(\sigma, \eta)$
Exponential General Gaussian (EGG)	$\propto \mathrm{r}^{-2\mathrm{k}}\exp(-\frac{r^\eta}{2\sigma_g^\eta})$	$\mathcal{G}(\sigma, \eta, k)$
Truncated Exponential Standard Gaussian (TESG)	$\propto \exp\left(-\frac{r^\eta}{2\sigma_s^\eta}\right)\mathbf{1}_{r\leq T}$	$\mathcal{S}_t(\sigma, \eta, T)$
Truncated Exponential General Gaussian (TEGG)	$\propto \mathrm{r}^{-2\mathrm{k}}\exp(-\frac{r^\eta}{2\sigma_g^\eta})\ \mathbf{1}_{r\leq T}$	$\mathcal{G}_t(\sigma, \eta, k, T)$

ESSENTIAL PROBLEM 1

Which is the optimal distribution for providing ℓ_2 certified robustness in randomized smoothing?

- **Mainstream view in academia:** Gaussian is the best.
- **Ours:** Many Exponential Standard Gaussian (ESG) distributions perform as well as Gaussian. i.e., **many ESG tie the best**.

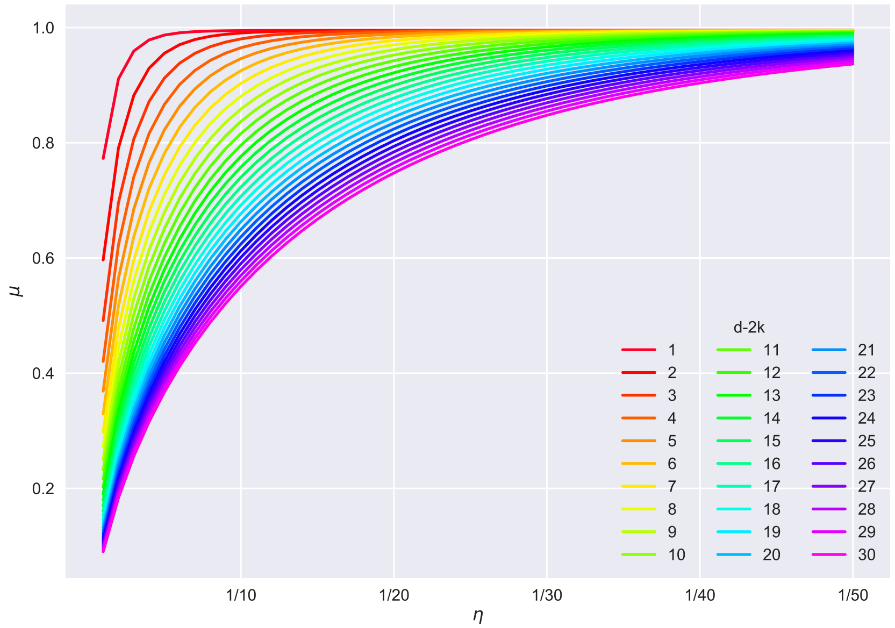


ESSENTIAL PROBLEM 2

How to solve the curse of dimensionality in randomized smoothing?

- **The SOTA theoretical solution:** Introducing General Gaussian in DSRS can provide an $\Omega(\sqrt{d})$ lower bound for the certified radius, which breaks the curse of dimensionality.

- **Ours:** Exponential General Gaussian (EGG) distribution can improve the lower bound of General Gaussian distribution. i.e., **EGG improves the SOTA theoretical solution**.



CONTRIBUTION

On ESG:

- We **first derive the integral solution** to the certified radius of ESG distributions in randomized smoothing. (Kumar et al. (2020) have noticed the sampling probability stays almost constant with the exponent of ESG-like distributions, but no computation of certified radius is derived due to the lack of math tools at that time.)
- We propose 2 asymptotically mild assumptions and **derive a highly approximate analytic formula** for the certified radius-sampling probability relation of ESG.
- We prove the analytic formula derived above **converges to** Cohen et al. (2019)'s origin formula for randomized smoothing with an $O\left(\frac{1}{\sqrt{d}}\right)$ error bound.

Refs: [1] Kumar et al. Curse of Dimensionality on Randomized Smoothing for Certifiable Robustness. In ICML 2020.
[2] Cohen et al. Certified Adversarial Robustness via Randomized Smoothing. In ICML 2019.

CONTRIBUTION

On EGG:

- Under appropriate concentration assumptions for the base classifier, we prove Exponential General Gaussian can provide tighter lower bounds for the certified radius in DSRS, **improving the SOTA theoretical solution to the curse of dimensionality** in randomized smoothing.
- Our experiments demonstrate that EGG can **significantly improve the robustness certification** provided by General Gaussian on real-world datasets. On ImageNet, the increment in certified accuracy reaches up to 6.4%.

EXPERIMENTAL RESULTS

Table 2. Certified radius at r for standardly augmented models, certified by ESG under DSRS

Dataset	Method	Certified accuracy at r													
		0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50
CIFAR10	ESG, $\eta = 1.0$	57.6%	42.6%	31.3%	21.5%	15.8%	12.8%	8.6%	6.8%	4.3%	2.3%	1.3%	0.8%	0.3%	0.1%
	ESG, $\eta = 2.0$ (Gaussian)	57.6%	42.6%	31.6%	21.5%	15.8%	12.7%	8.8%	6.8%	4.5%	2.4%	1.3%	0.7%	0.2%	0.2%
	ESG, $\eta = 4.0$	57.6%	42.6%	31.3%	21.5%	15.9%	12.9%	8.6%	6.9%	4.3%	2.4%	1.3%	0.8%	0.2%	0.1%
	ESG, $\eta = 8.0$	57.8%	42.6%	31.6%	21.6%	15.9%	12.9%	8.9%	6.7%	4.2%	2.4%	1.3%	0.9%	0.2%	0.1%
ImageNet	ESG, $\eta = 1.0$	59.6%	51.5%	43.2%	37.9%	33.0%	26.8%	23.1%	21.5%	19.9%	17.4%	13.8%	11.5%	10.3%	7.7%
	ESG, $\eta = 2.0$, (Gaussian)	59.6%	51.6%	43.1%	38.0%	32.9%	26.9%	23.1%	21.5%	19.7%	17.4%	13.6%	11.4%	10.1%	8.3%
	ESG, $\eta = 4.0$	59.6%	51.5%	43.2%	38.0%	32.9%	27.2%	23.1%	21.6%	19.9%	17.2%	13.6%	11.4%	10.2%	8.0%
	ESG, $\eta = 8.0$	59.6%	51.5%	43.2%	38.0%	33.0%	26.8%	23.1%	21.6%	19.7%	17.3%	13.6%	11.5%	10.1%	8.4%

Table 3. Certified radius at r for standardly augmented models, certified by EGG under DSRS

Dataset	Method	Certified accuracy at r													
		0.25	0.50	0.75	1.00	1.25	1.50	1.75	2.00	2.25	2.50	2.75	3.00	3.25	3.50
CIFAR10	EGG, $\eta = 0.25$	54.2%	37.6%	23.5%	16.5%	9.4%	4.5%	0.5%	0.1%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	EGG, $\eta = 0.5$	55.5%	40.4%	25.2%	19.1%	13.4%	8.5%	5.5%	2.0%	0.4%	0.1%	0.0%	0.0%	0.0%	0.0%
	EGG, $\eta = 1.0$	56.3%	41.7%	28.2%	20.0%	15.1%	10.5%	7.1%	4.2%	1.9%	0.9%	0.1%	0.0%	0.0%	0.0%
	DSRS (Li et al., 2022) (EGG, $\eta = 2.0$)	56.7%	42.4%	29.3%	20.2%	15.7%	11.5%	8.0%	5.5%	2.6%	1.5%	0.6%	0.1%	0.0%	0.0%
	EGG, $\eta = 4.0$	57.5%	42.5%	30.0%	20.2%	15.9%	12.2%	8.5%	6.5%	3.4%	1.8%	0.9%	0.4%	0.0%	0.0%
	Ours (EGG, $\eta = 8.0$)	**57.6%**	**42.5%**	**30.9%**	**20.6%**	**15.8%**	**12.3%**	**8.6%**	**6.6%**	**3.7%**	**2.1%**	**1.1%**	**0.5%**	**0.2%**	0.0%
ImageNet	EGG, $\eta = 0.25$	53.8%	41.4%	28.4%	20.1%	7.1%	0.8%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
	EGG, $\eta = 0.5$	54.9%	46.3%	36.4%	26.3%	22.1%	15.2%	8.7%	3.1%	0.8%	0.0%	0.0%	0.0%	0.0%	0.0%
	EGG, $\eta = 1.0$	57.0%	47.8%	39.9%	32.8%	24.9%	22.0%	18.5%	13.1%	9.2%	5.0%	2.1%	0.5%	0.0%	0.0%
	DSRS (Li et al., 2022) (EGG, $\eta = 2.0$)	58.4%	48.5%	41.5%	35.2%	28.9%	23.3%	21.3%	18.8%	14.1%	11.1%	8.9%	6.1%	2.2%	1.4%
	EGG, $\eta = 4.0$	58.7%	49.9%	42.6%	36.4%	31.0%	23.9%	22.3%	20.2%	17.3%	13.2%	10.7%	9.2%	6.8%	4.0%
	Ours (EGG, $\eta = 8.0$)	**59.1%**	**50.8%**	**42.9%**	**36.8%**	**31.8%**	**24.6%**	**22.6%**	**20.7%**	**18.9%**	**14.5%**	**11.7%**	**10.1%**	**8.6%**	**5.2%**