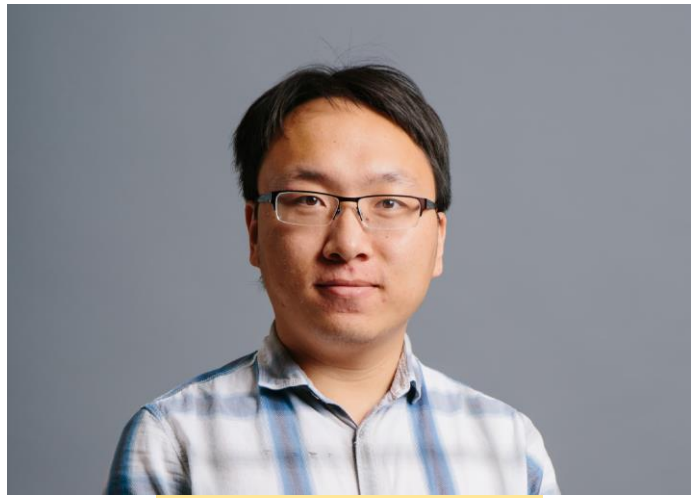
Provably Robust Deep Learning via Adversarially Trained Smoothed Classifiers

Hadi Salman, Greg Yang, Jerry Li,
Pengchuan Zhang, Huan Zhang, Ilya Razenshteyn, Sébastien Bubeck
MSR AI
NeurIPS 2019, Vancouver

Microsoft Research

Sébastien Bubeck

Greg Yang



Pengchuan Zhang

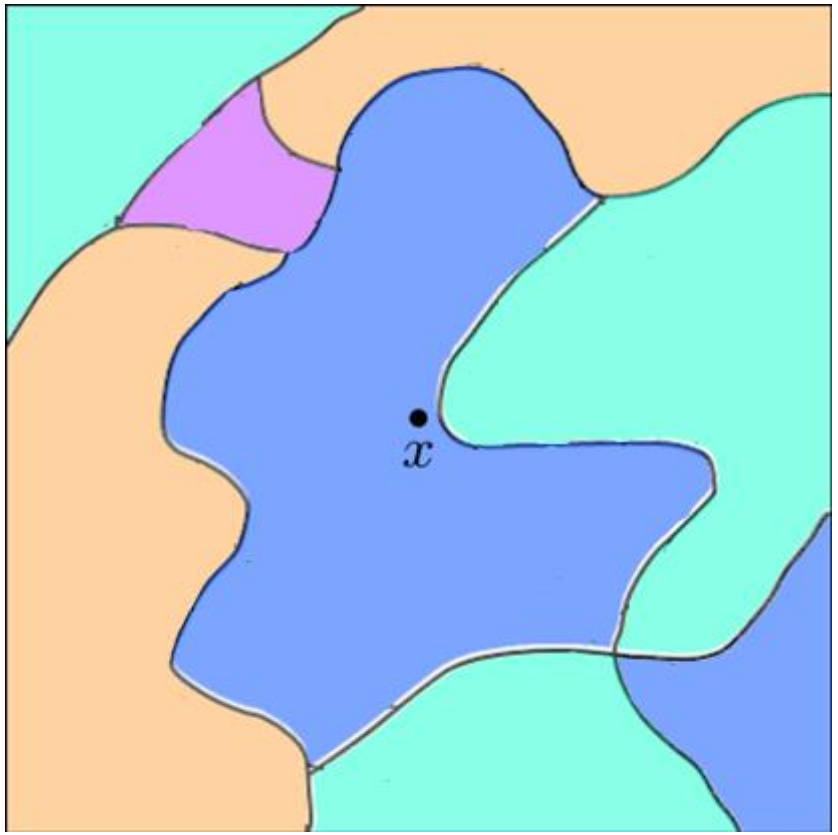
Ilya Razenshteyn

Jerry Li

Huan Zhang

Adversarial Examples

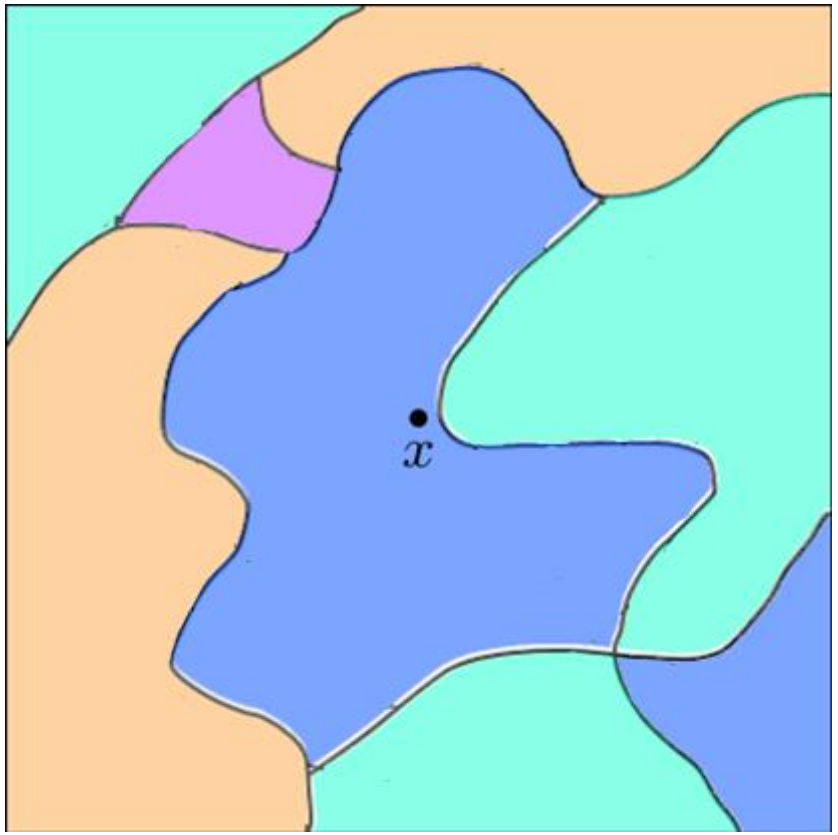
$f: \mathbb{R}^d \rightarrow \mathcal{Y}$



Adversarial Examples

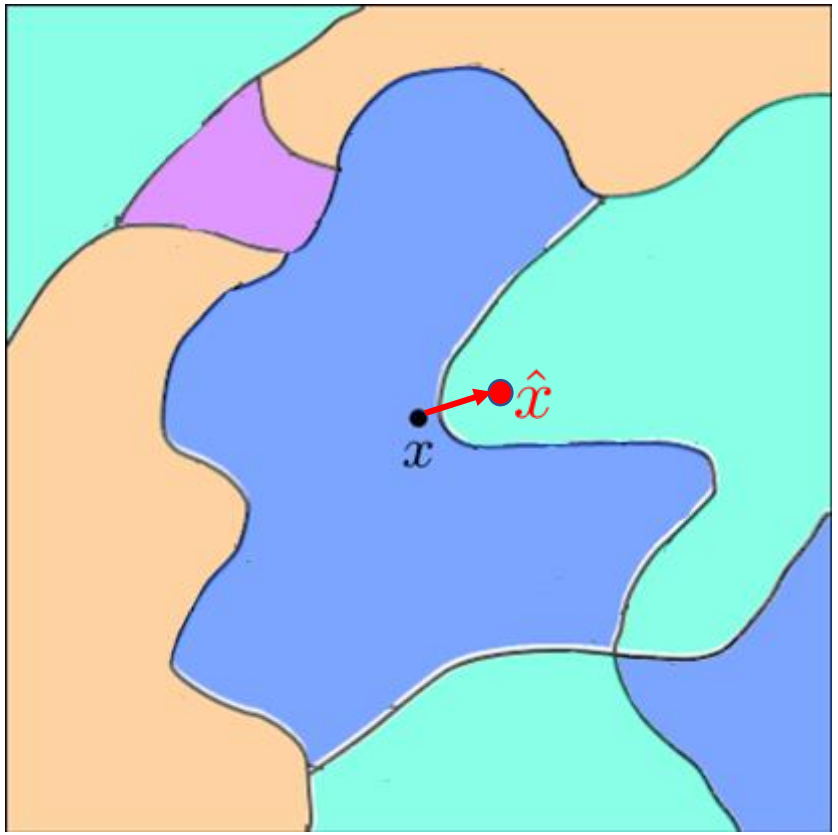
$f(x)$ 

$f: \mathbb{R}^d \rightarrow \mathcal{Y}$



Adversarial Examples

$f: \mathbb{R}^d \rightarrow \mathcal{Y}$



$f(x)$ →

$f(\hat{x})$ →

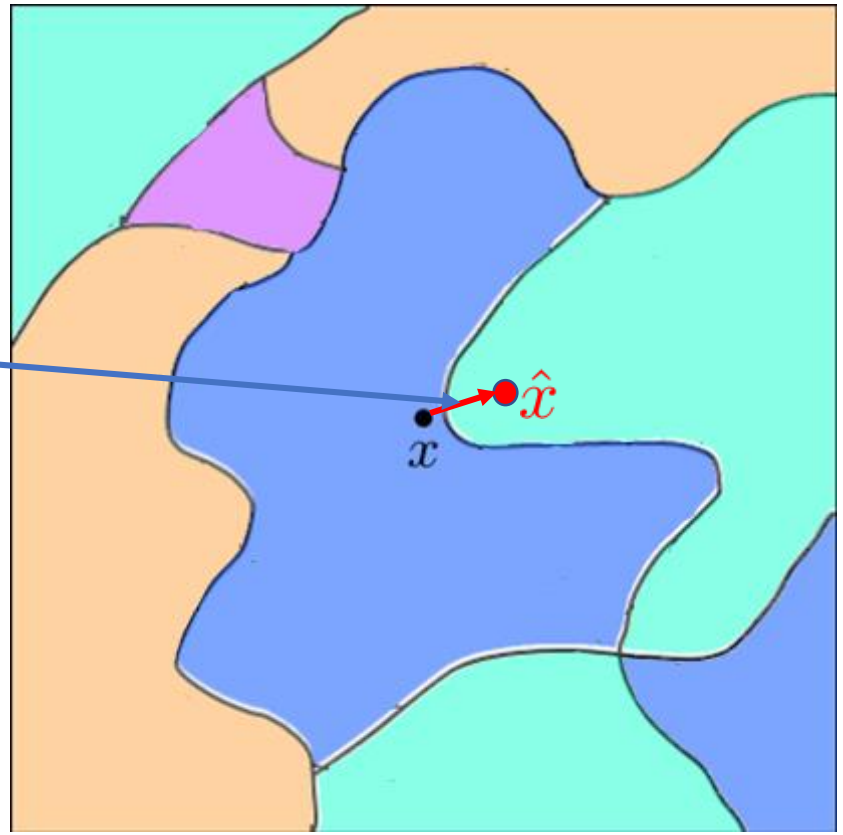
Adversarial Examples

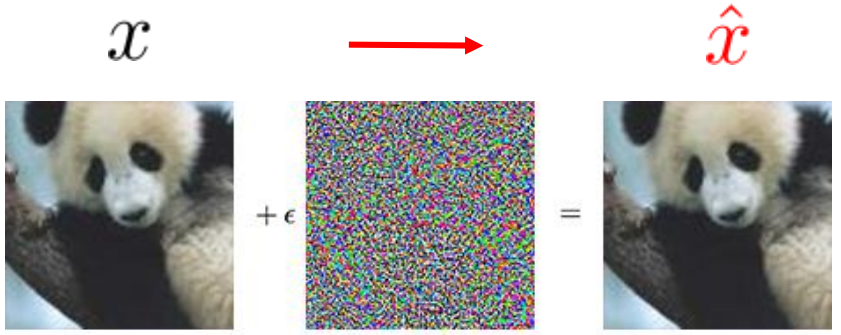
$f(x)$ → 

$f(\hat{x})$ → 

$f: \mathbb{R}^d \rightarrow \mathcal{Y}$

Imperceptible perturbation





Goodfellow et al. 2015

Adversarial Defenses

Empirical Defenses

- Upper bound on the **true robust accuracy**
- Nothing in principle prevents a stronger empirical attack from further lowering the empirical robust accuracy of a model

Certified Defenses

- Lower bound on the **true robust accuracy**
- Certified robustness!

Adversarial Defenses

Empirical Defenses

- Upper bound on the **true robust accuracy**
- Nothing in principle prevents a stronger empirical attack from further lowering the empirical robust accuracy of a model

Certified Defenses

- Lower bound on the **true robust accuracy**
- Certified robustness!

Randomized Smoothing

$f: \mathbb{R}^d \to \mathcal{Y}$

Randomized Smoothing

$$f: \mathbb{R}^d \to \mathcal{Y}$$

Randomized Smoothing

$$f \mapsto g$$

$$g(x) = \underset{c \in \mathcal{Y}}{\operatorname{argmax}}\, \mathbb{P}_\delta(f(x+\delta) = c)$$

$$\delta \sim N(0, \sigma^2 I)$$

Natural idea explored by several authors:
Lecuyer et al. 2018, Li et al. 2018,
Cohen et al. 2019

Randomized Smoothing

$$f: \mathbb{R}^d \to \mathcal{Y}$$

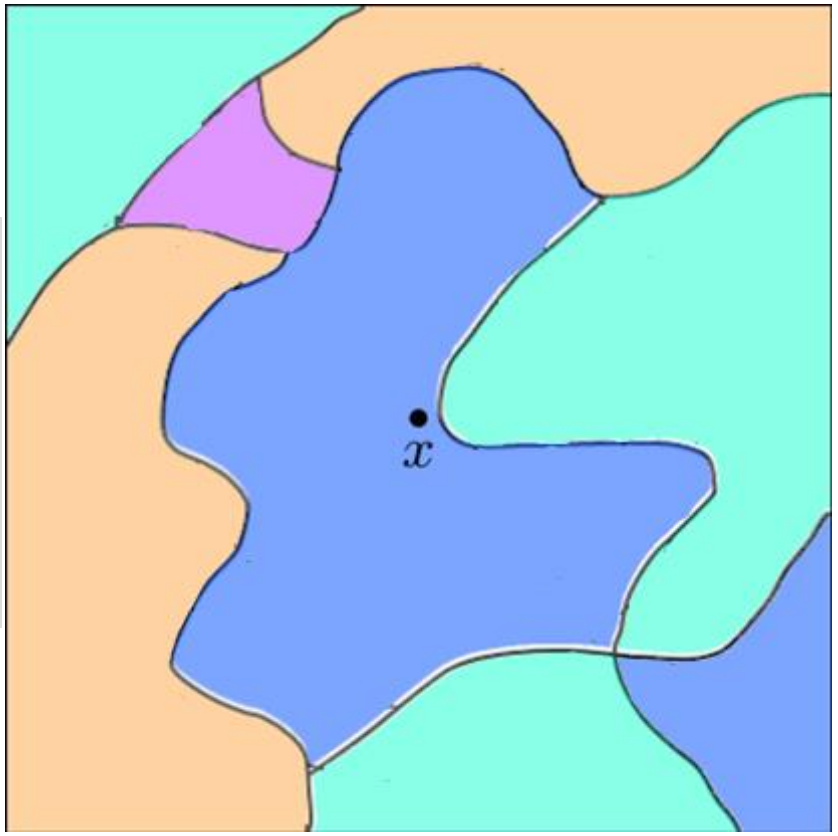
Randomized Smoothing

$$f \mapsto g$$

$$g(x) = \underset{c \in \mathcal{Y}}{\operatorname{argmax}}\, \mathbb{P}_\delta(f(x+\delta) = c)$$

$$\delta \sim N(0, \sigma^2 I)$$

Natural idea explored by several authors:
Lecuyer et al. 2018, Li et al. 2018,
Cohen et al. 2019



Randomized Smoothing

$$f: \mathbb{R}^d \to \mathcal{Y}$$

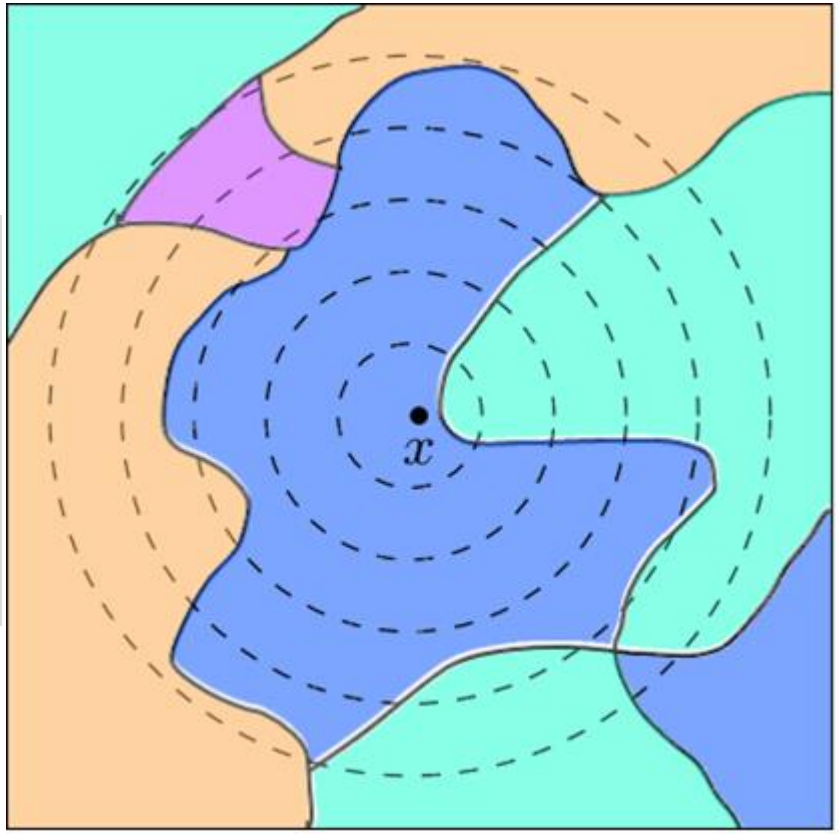
Randomized Smoothing

$$f \mapsto g$$

$$g(x) = \underset{c \in \mathcal{Y}}{\operatorname{argmax}}\, \mathbb{P}_\delta(f(x+\delta) = c)$$

$$\delta \sim N(0, \sigma^2 I)$$

Natural idea explored by several authors:
Lecuyer et al. 2018, Li et al. 2018,
Cohen et al. 2019



Randomized Smoothing

$$f: \mathbb{R}^d \to \mathcal{Y}$$

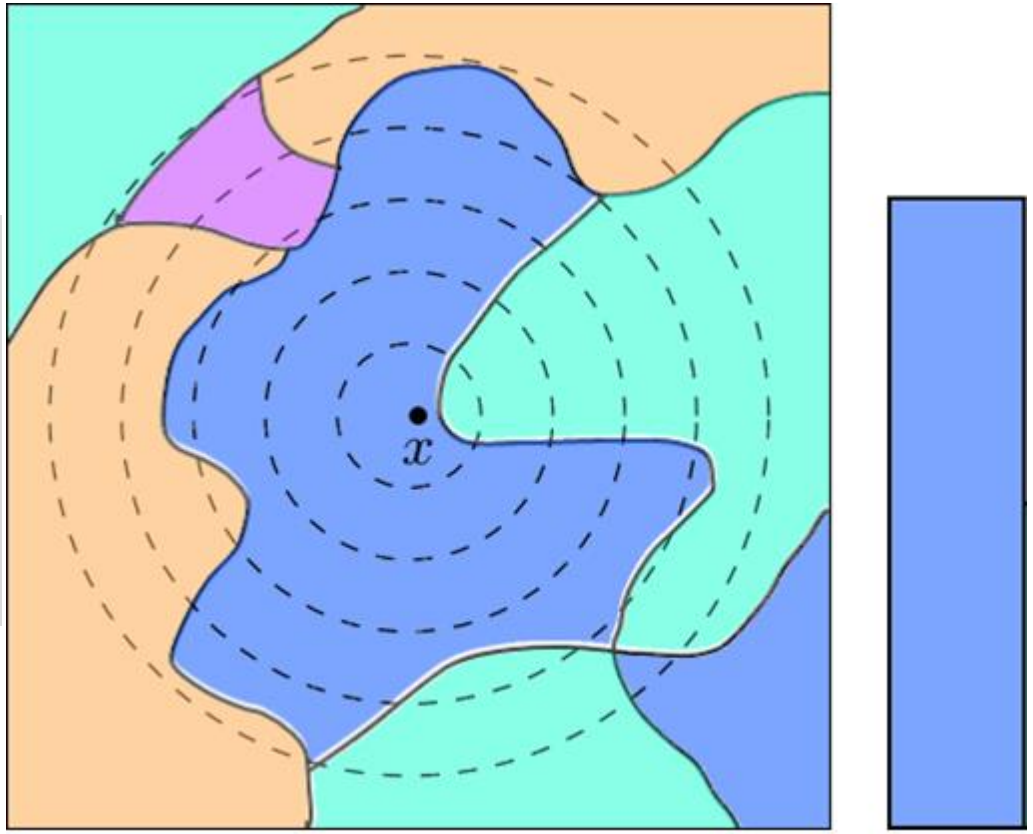
Randomized Smoothing

$$f \mapsto g$$

$$g(x) = \underset{c \in \mathcal{Y}}{\operatorname{argmax}}\, \mathbb{P}_\delta(f(x+\delta) = c)$$

$$\delta \sim N(0, \sigma^2 I)$$

Natural idea explored by several authors:
Lecuyer et al. 2018, Li et al. 2018,
Cohen et al. 2019



Randomized Smoothing

$$f: \mathbb{R}^d \to \mathcal{Y}$$

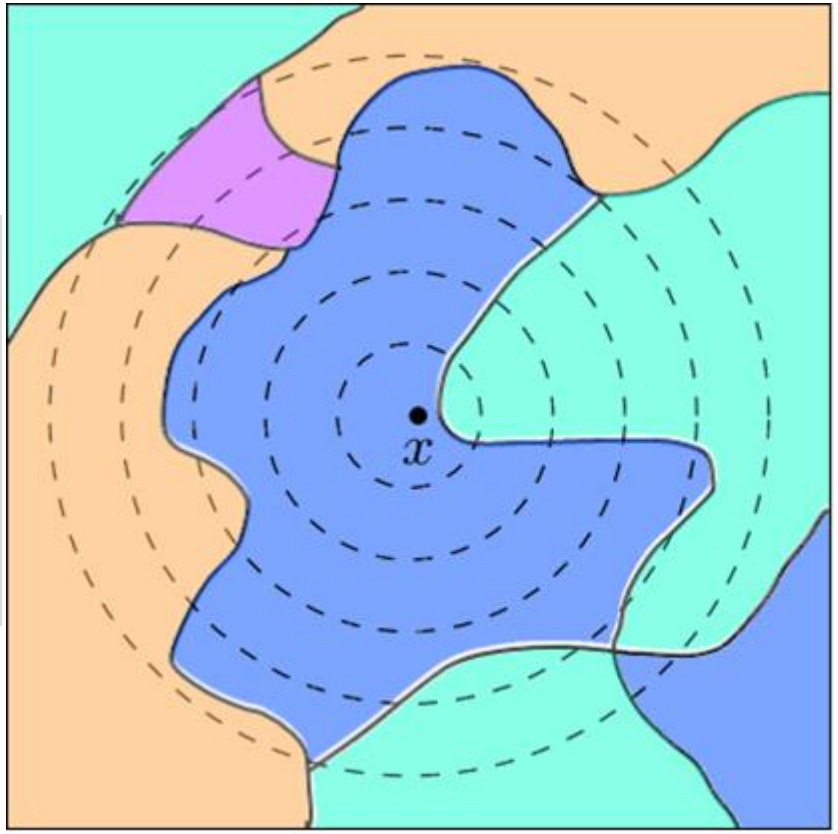
Randomized Smoothing

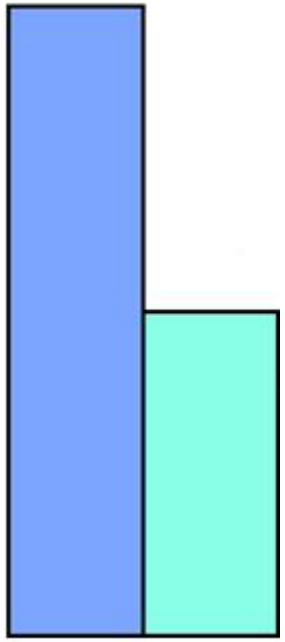
$$f \mapsto g$$

$$g(x) = \underset{c \in \mathcal{Y}}{\operatorname{argmax}}\, \mathbb{P}_\delta(f(x+\delta) = c)$$

$$\delta \sim N(0, \sigma^2 I)$$

Natural idea explored by several authors:
Lecuyer et al. 2018, Li et al. 2018,
Cohen et al. 2019





Randomized Smoothing

$$f: \mathbb{R}^d \to \mathcal{Y}$$

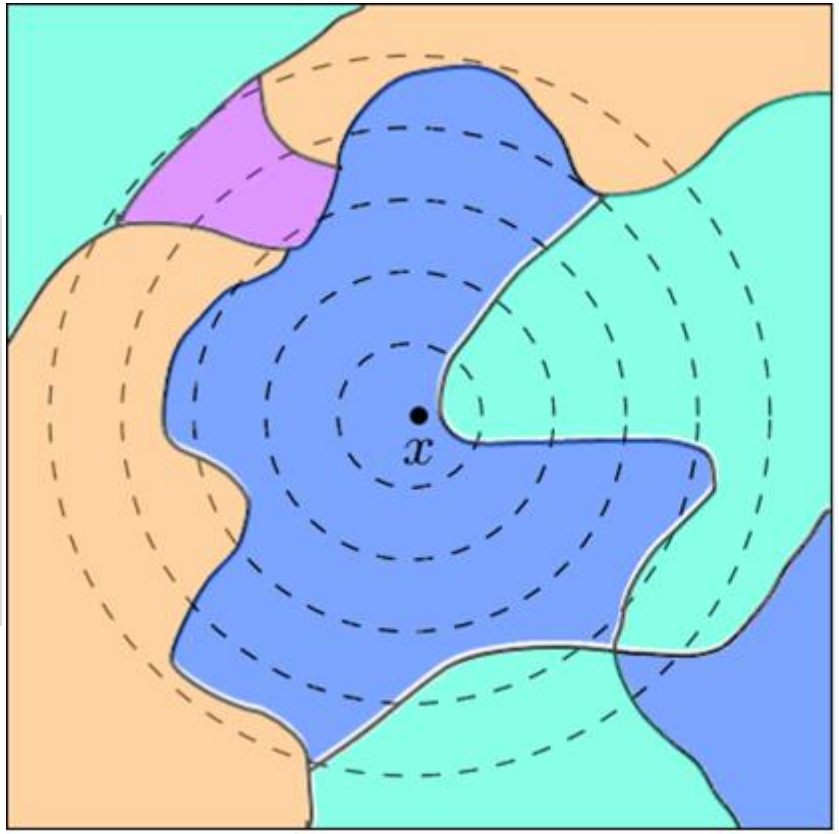
Randomized Smoothing

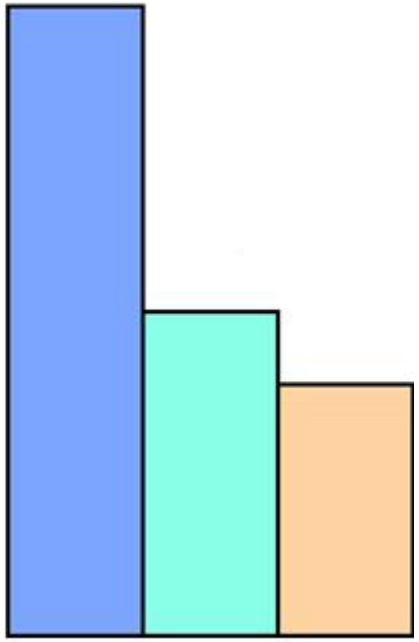
$$f \mapsto g$$

$$g(x) = \underset{c \in \mathcal{Y}}{\operatorname{argmax}}\, \mathbb{P}_\delta(f(x+\delta) = c)$$

$$\delta \sim N(0, \sigma^2 I)$$

Natural idea explored by several authors:
Lecuyer et al. 2018, Li et al. 2018,
Cohen et al. 2019





Randomized Smoothing

$$f: \mathbb{R}^d \to \mathcal{Y}$$

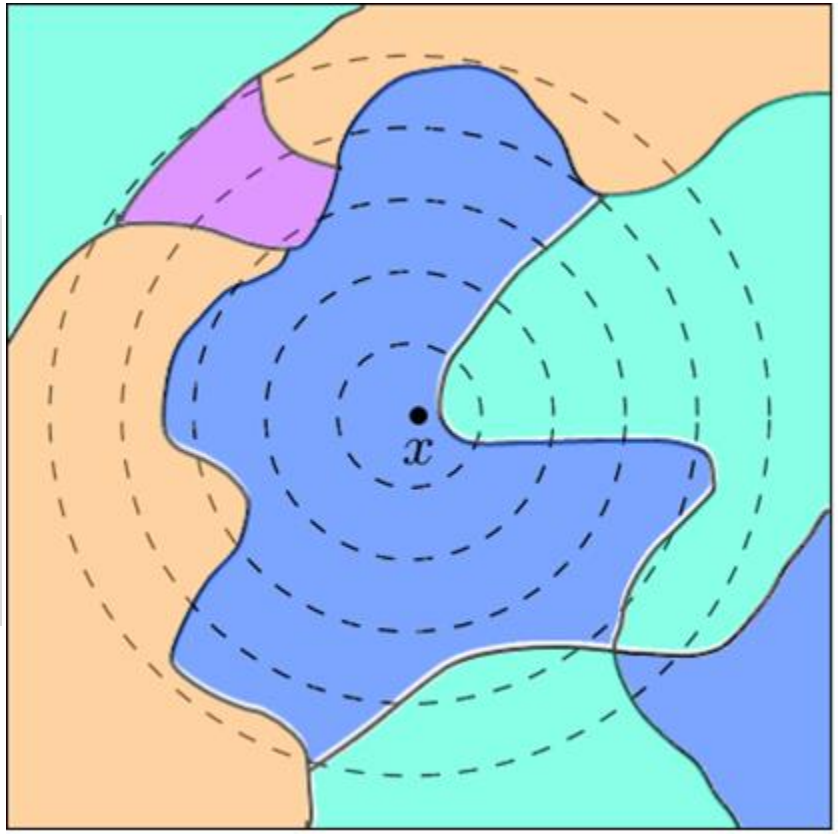
Randomized Smoothing

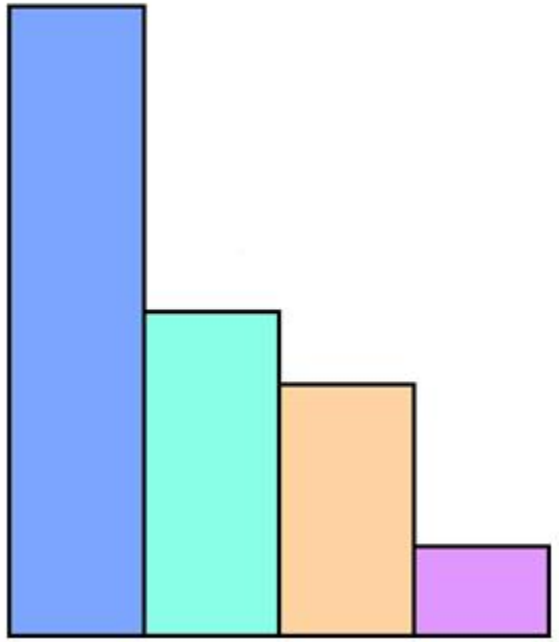
$$f \mapsto g$$

$$g(x) = \underset{c \in \mathcal{Y}}{\operatorname{argmax}}\, \mathbb{P}_\delta(f(x+\delta) = c)$$

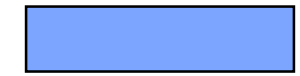
$$\delta \sim N(0, \sigma^2 I)$$

Natural idea explored by several authors:
Lecuyer et al. 2018, Li et al. 2018,
Cohen et al. 2019





Randomized Smoothing

$g(x)$ 

$$f: \mathbb{R}^d \to \mathcal{Y}$$

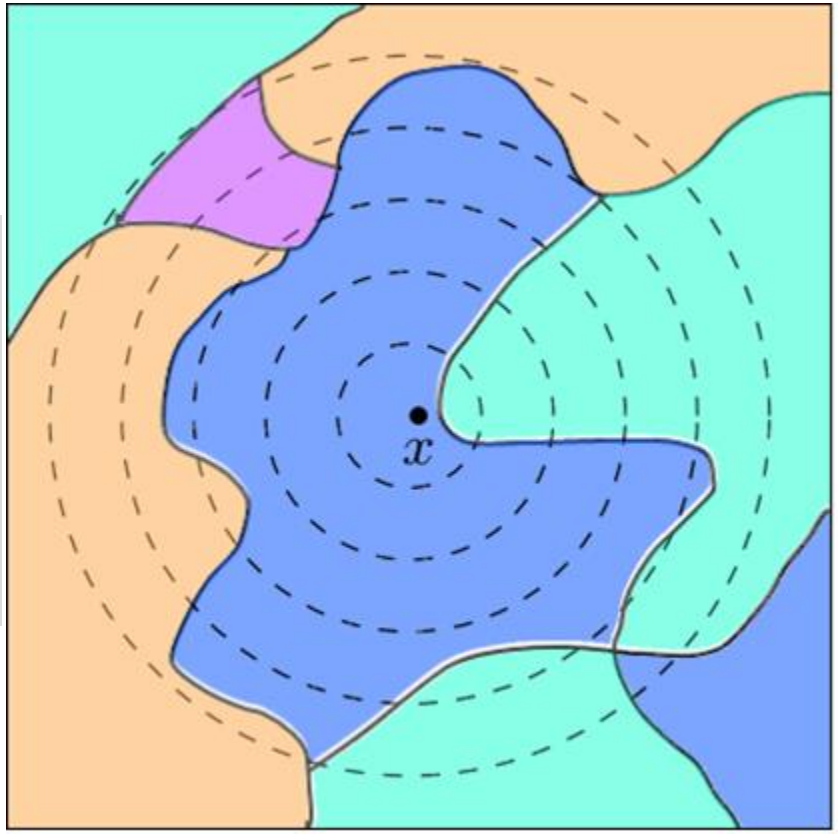
Randomized Smoothing

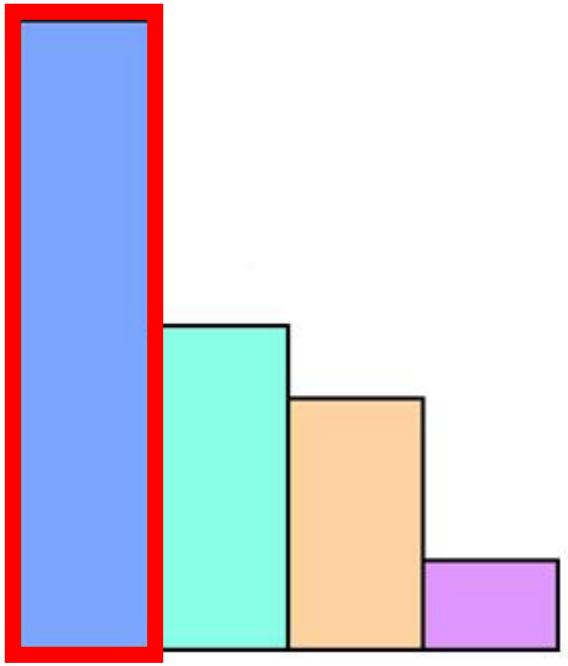
$$f \mapsto g$$

$$g(x) = \underset{c \in \mathcal{Y}}{\operatorname{argmax}}\, \mathbb{P}_\delta(f(x+\delta) = c)$$

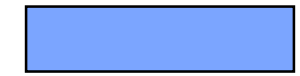
$$\delta \sim N(0, \sigma^2 I)$$

Natural idea explored by several authors:
Lecuyer et al. 2018, Li et al. 2018,
Cohen et al. 2019





Randomized Smoothing

$g(x)$ 

$$f: \mathbb{R}^d \to \mathcal{Y}$$

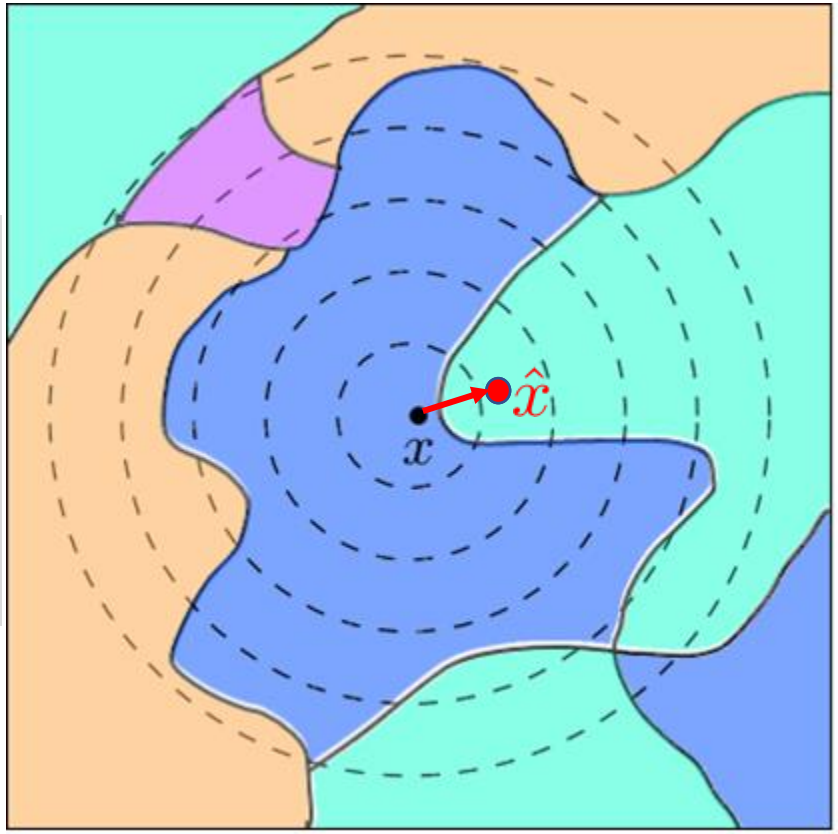
Randomized Smoothing

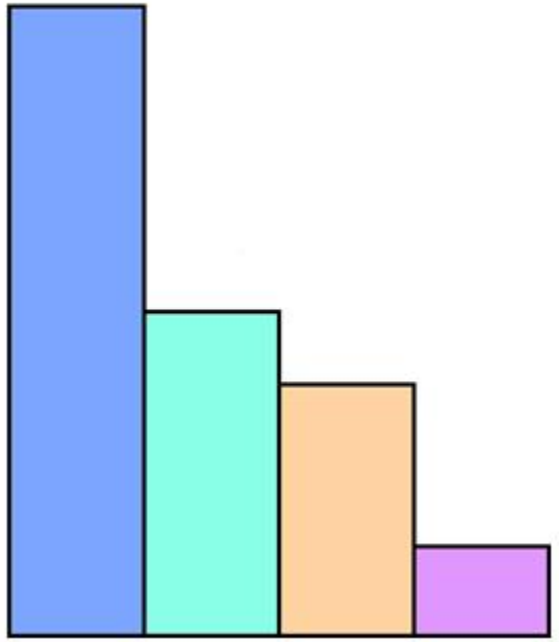
$$f \mapsto g$$

$$g(x) = \underset{c \in \mathcal{Y}}{\operatorname{argmax}}\, \mathbb{P}_\delta(f(x+\delta) = c)$$

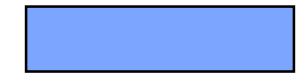
$$\delta \sim N(0, \sigma^2 I)$$

Natural idea explored by several authors:
Lecuyer et al. 2018, Li et al. 2018,
Cohen et al. 2019





Randomized Smoothing

$g(x)$ 

$$f: \mathbb{R}^d \to \mathcal{Y}$$

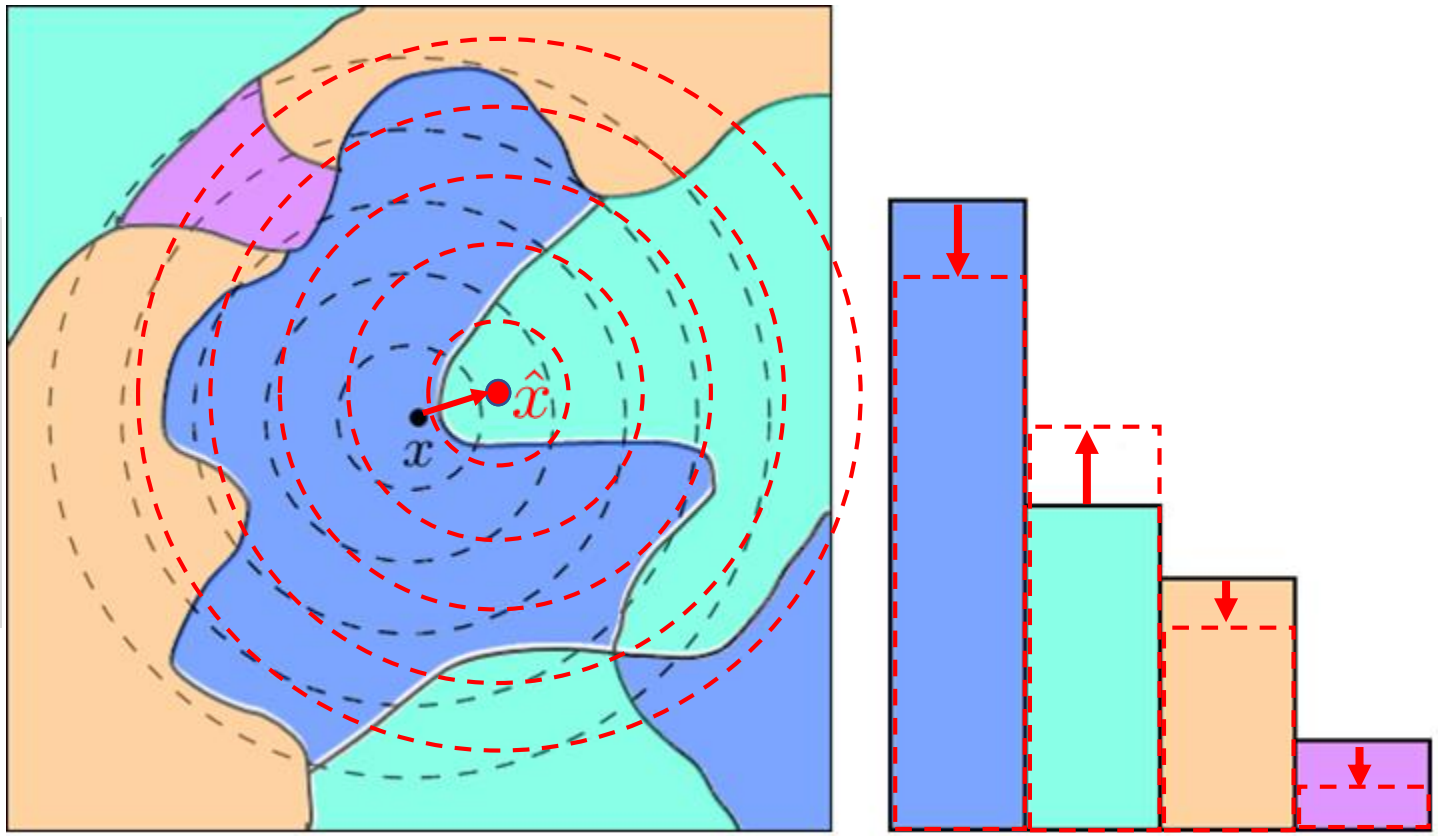
Randomized Smoothing

$$f \mapsto g$$

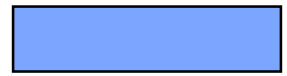
$$g(x) = \underset{c \in \mathcal{Y}}{\operatorname{argmax}}\, \mathbb{P}_\delta(f(x+\delta) = c)$$

$$\delta \sim N(0, \sigma^2 I)$$

Natural idea explored by several authors:
Lecuyer et al. 2018, Li et al. 2018,
Cohen et al. 2019



Randomized Smoothing

$g(x)$ → 

$$f: \mathbb{R}^d \to \mathcal{Y}$$

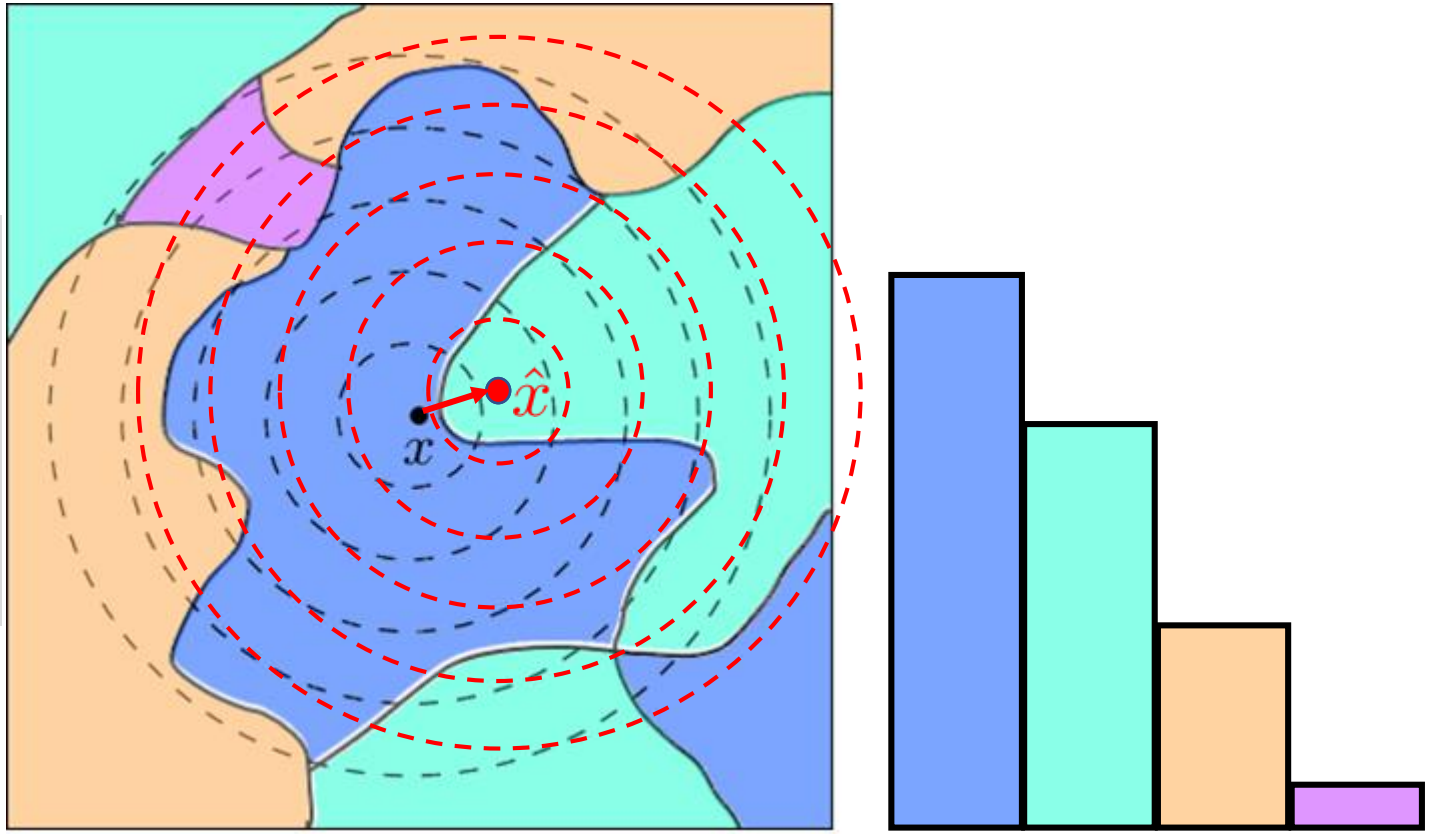
Randomized Smoothing

$$f \mapsto g$$

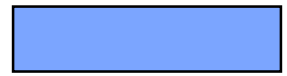
$$g(x) = \underset{c \in \mathcal{Y}}{\operatorname{argmax}}\, \mathbb{P}_\delta(f(x+\delta) = c)$$

$$\delta \sim N(0, \sigma^2 I)$$

Natural idea explored by several authors:
Lecuyer et al. 2018, Li et al. 2018,
Cohen et al. 2019



Randomized Smoothing

$g(x)$ → 

$$f: \mathbb{R}^d \to \mathcal{Y}$$

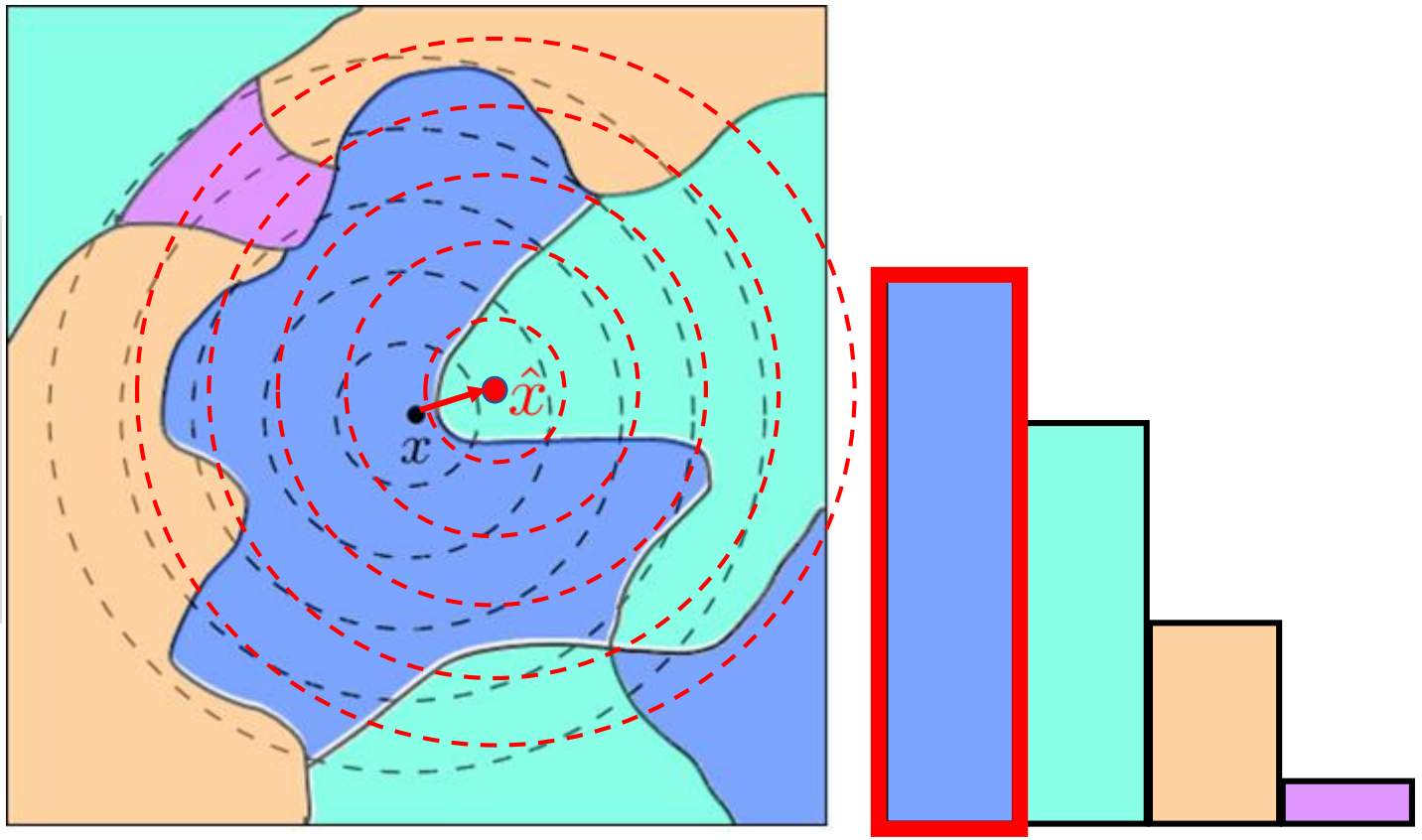
Randomized Smoothing

$$f \mapsto g$$

$$g(x) = \underset{c \in \mathcal{Y}}{\operatorname{argmax}}\, \mathbb{P}_\delta(f(x+\delta) = c)$$

$$\delta \sim N(0, \sigma^2 I)$$

Natural idea explored by several authors:
Lecuyer et al. 2018, Li et al. 2018,
Cohen et al. 2019



Randomized Smoothing

$$f: \mathbb{R}^d \to \mathcal{Y}$$

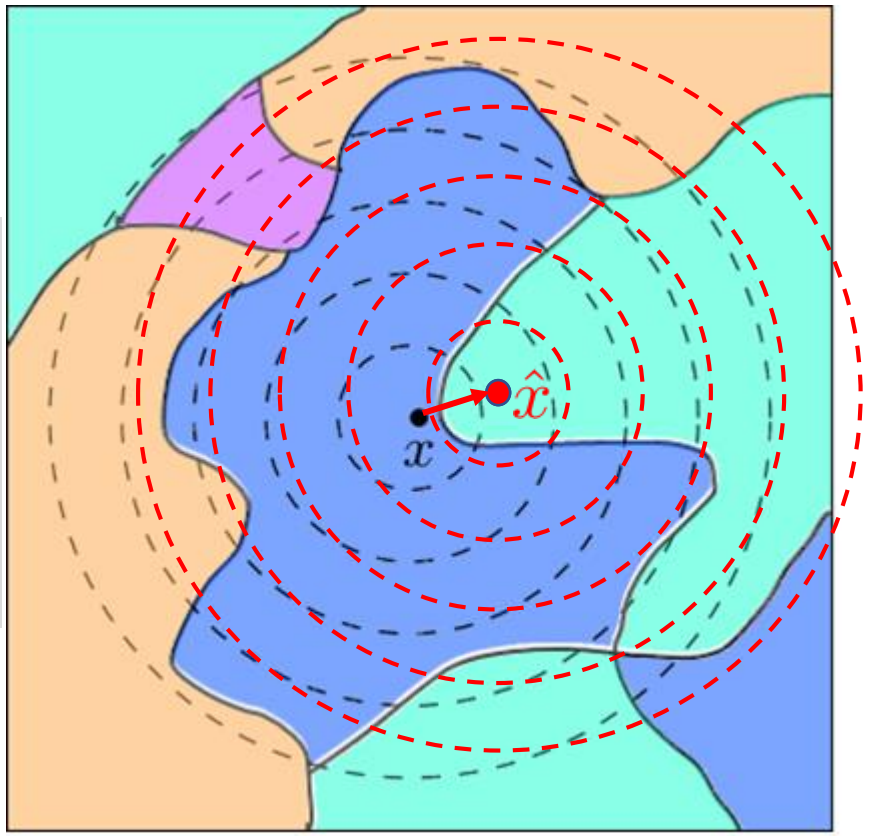
Randomized Smoothing

$$f \mapsto g$$

$$g(x) = \underset{c \in \mathcal{Y}}{\operatorname{argmax}}\, \mathbb{P}_\delta(f(x+\delta) = c)$$

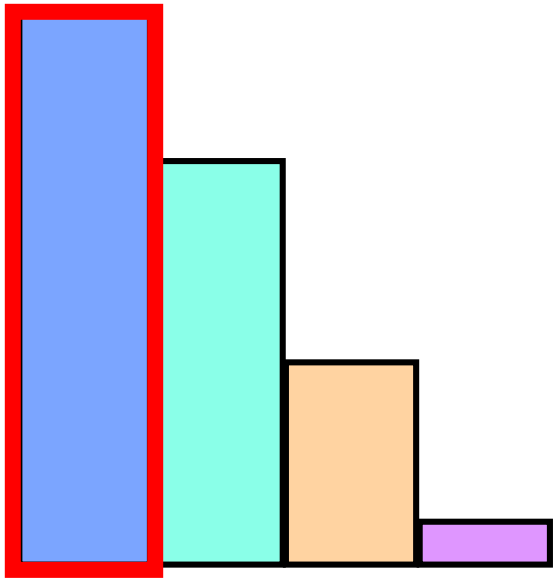
$$\delta \sim N(0, \sigma^2 I)$$

Natural idea explored by several authors:
Lecuyer et al. 2018, Li et al. 2018,
Cohen et al. 2019



$g(x)$ →

$g(\hat{x})$ →



Randomized Smoothing

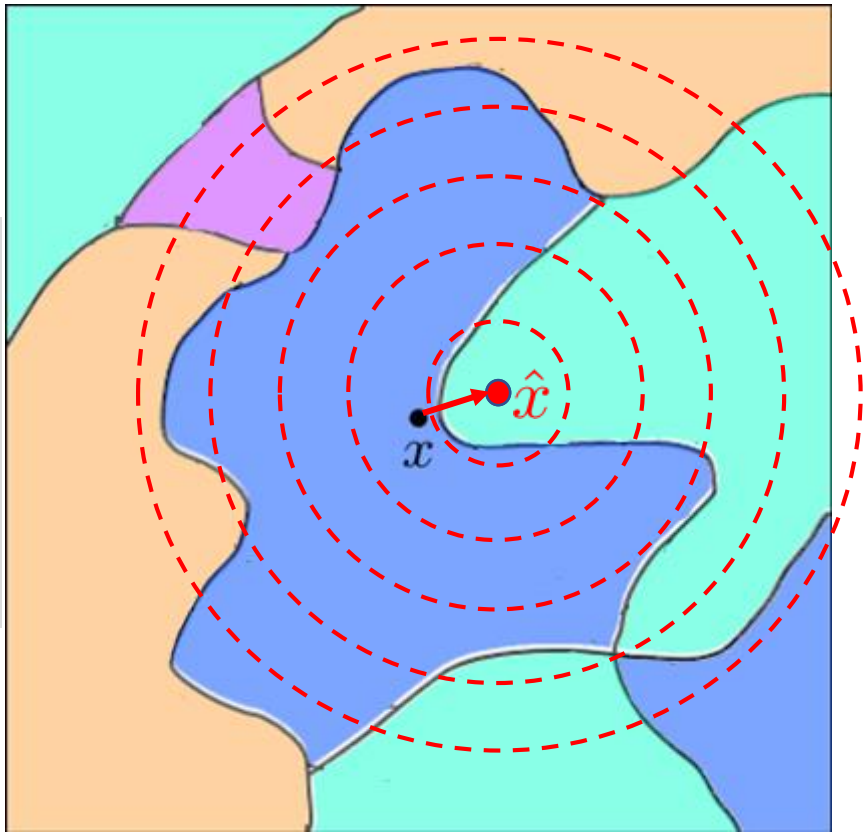
$$f: \mathbb{R}^d \to \mathcal{Y}$$

Randomized Smoothing

$$f \mapsto g$$

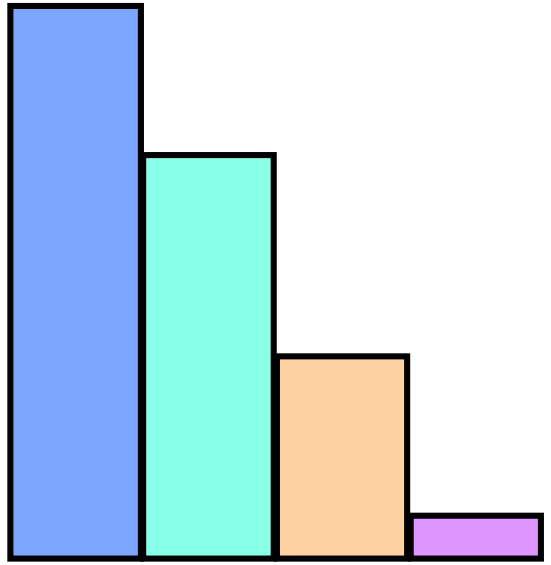
$$g(x) = \underset{c \in \mathcal{Y}}{\operatorname{argmax}}\, \mathbb{P}_\delta(f(x+\delta) = c)$$

$$\delta \sim N(0, \sigma^2 I)$$



$g(x)$ →

$g(\hat{x})$ →



Randomized Smoothing

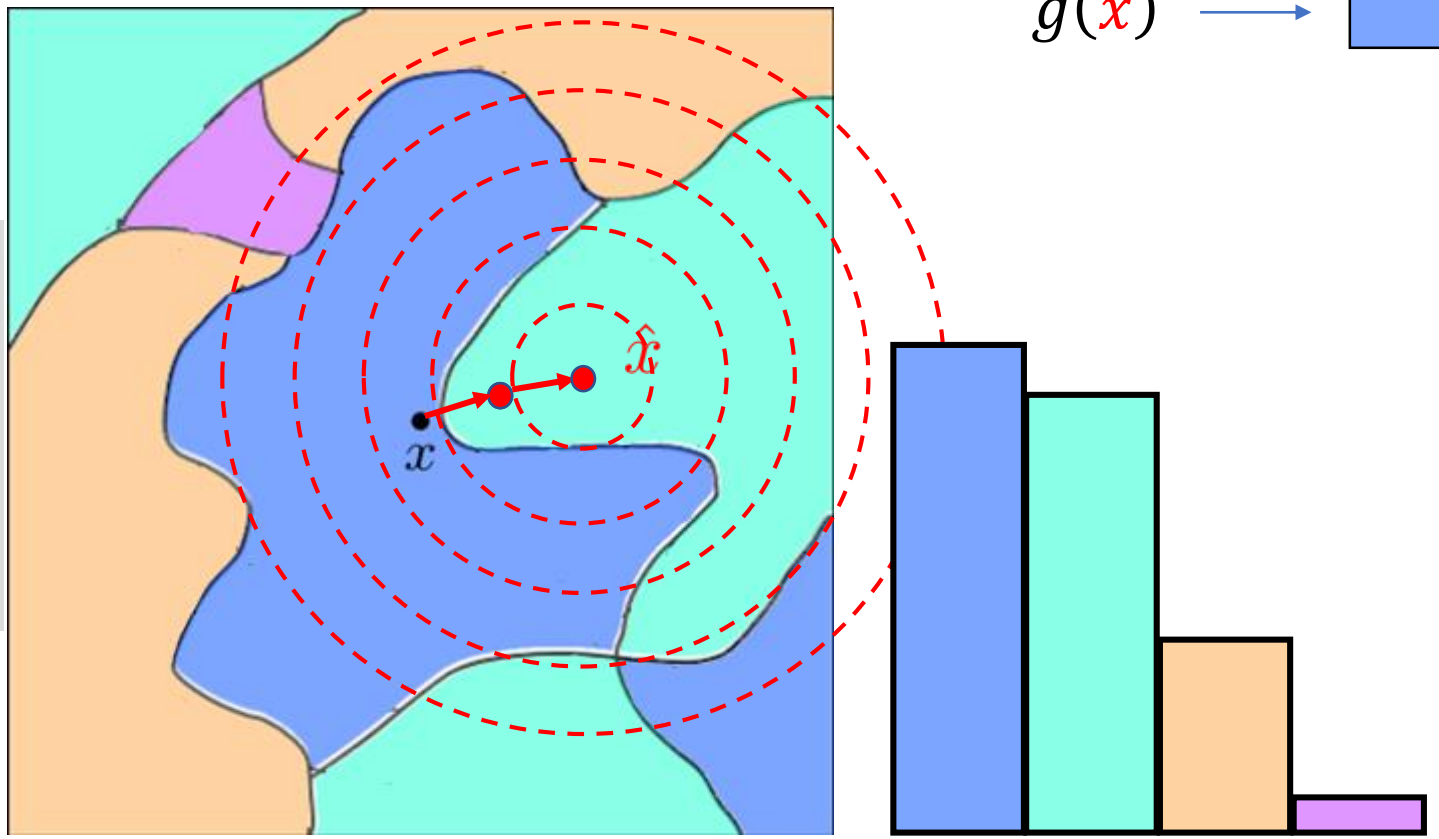
$f: \mathbb{R}^d \to \mathcal{Y}$

Randomized Smoothing

$f \mapsto g$

$g(x) = \underset{c \in \mathcal{Y}}{\operatorname{argmax}}\, \mathbb{P}_\delta(f(x+\delta) = c)$

$\delta \sim N(0, \sigma^2 I)$



$g(x)$ →

$g(\hat{x})$ →

Randomized Smoothing

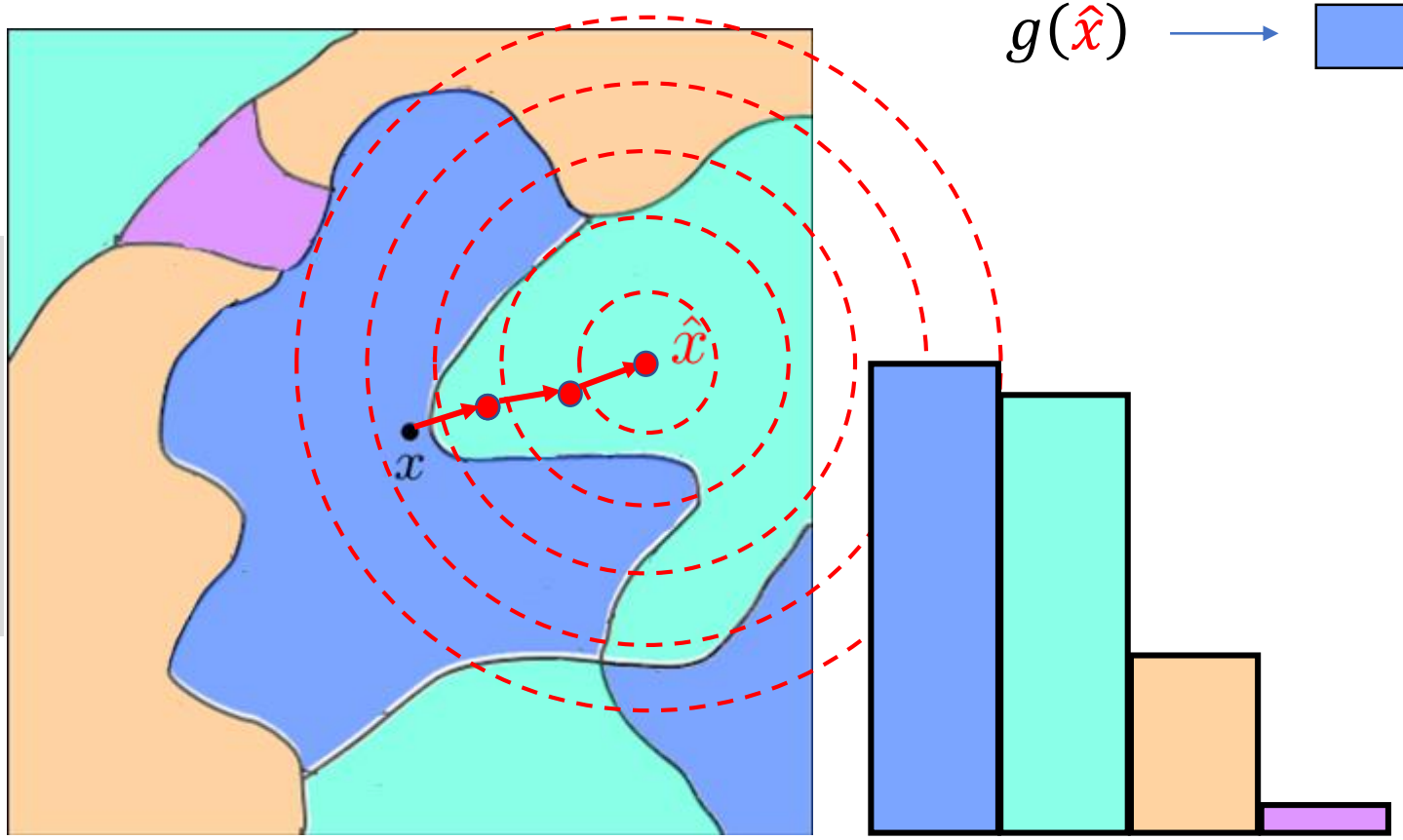
$f: \mathbb{R}^d \rightarrow \mathcal{Y}$

Randomized Smoothing

$f \mapsto g$

$g(x) = \underset{c \in \mathcal{Y}}{\operatorname{argmax}}\, \mathbb{P}_\delta(f(x+\delta) = c)$

$\delta \sim N(0, \sigma^2 I)$



$g(x)$

$g(\hat{x})$

Randomized Smoothing

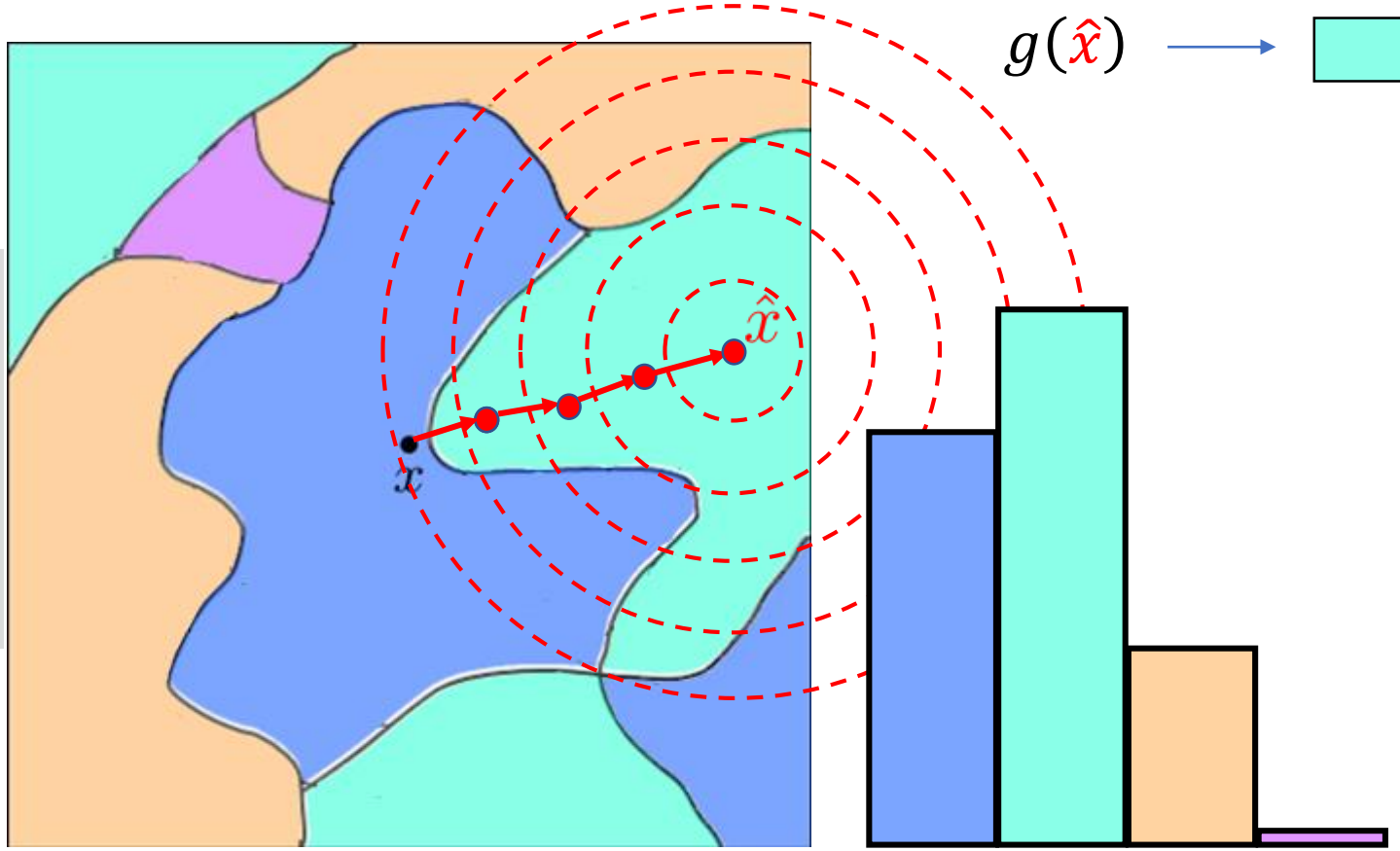
$$f: \mathbb{R}^d \to \mathcal{Y}$$

Randomized Smoothing

$$f \mapsto g$$

$$g(x) = \underset{c \in \mathcal{Y}}{\operatorname{argmax}}\, \mathbb{P}_\delta(f(x+\delta) = c)$$

$$\delta \sim N(0, \sigma^2 I)$$



Results

ImageNet

ℓ_2 radius (Imagenet)	0.5	1.0	1.5	2.0	2.5	3.0	3.5
Cohen et al. 2019	49	37	29	19	15	12	9
Ours	**56**	**45**	**38**	**28**	**26**	**20**	**17**

CIFAR-10

ℓ_2 radius (CIFAR10)	0.25	0.5	0.75	1.0	1.25	1.5	1.75	2.0	2.25
Cohen et al. 2019	61	43	32	22	17	14	10	7	4
Ours	**73**	**58**	**48**	**38**	**33**	**29**	**24**	**18**	**16**

Results

ImageNet

ℓ_2 radius (Imagenet)	0.5	1.0	1.5	2.0	2.5	3.0	3.5
Cohen et al. 2019	49	37	29	19	15	12	9
Ours	**56**	**45**	**38**	**28**	**26**	**20**	**17**

CIFAR-10

ℓ_2 radius (CIFAR10)	0.25	0.5	0.75	1.0	1.25	1.5	1.75	2.0	2.25
Cohen et al. 2019	61	43	32	22	17	14	10	7	4
Ours	**73**	**58**	**48**	**38**	**33**	**29**	**24**	**18**	**16**
Ours + pretraining	**80**	**62**	**52**	**38**	**34**	**30**	**25**	**19**	**16**

Thank you!

Today 10:45 AM -- 12:45 PM @ East Exhibition Hall B + C #24

Paper: https://arxiv.org/abs/1906.04584

Blog Post: https://decentdescent.org/smoothadv.html

Code: https://github.com/Hadisalman/smoothing-adversarial

Follow me on Twitter ☺

@hadisalmanX