[ICLR 2022]
How to Robustify Black-Box ML Models? A Zeroth-Order Optimization Perspective

Yimeng Zhang
Michigan State University

Yuguang Yao
Michigan State University

Jinghan Jia
Michigan State University

Jinfeng Yi
JD AI Research

Mingyi Hong
University of Minnesota

Shiyu Chang
UC Santa Barbara

Sijia Liu
Michigan State University

MICHIGAN STATE
UNIVERSITY

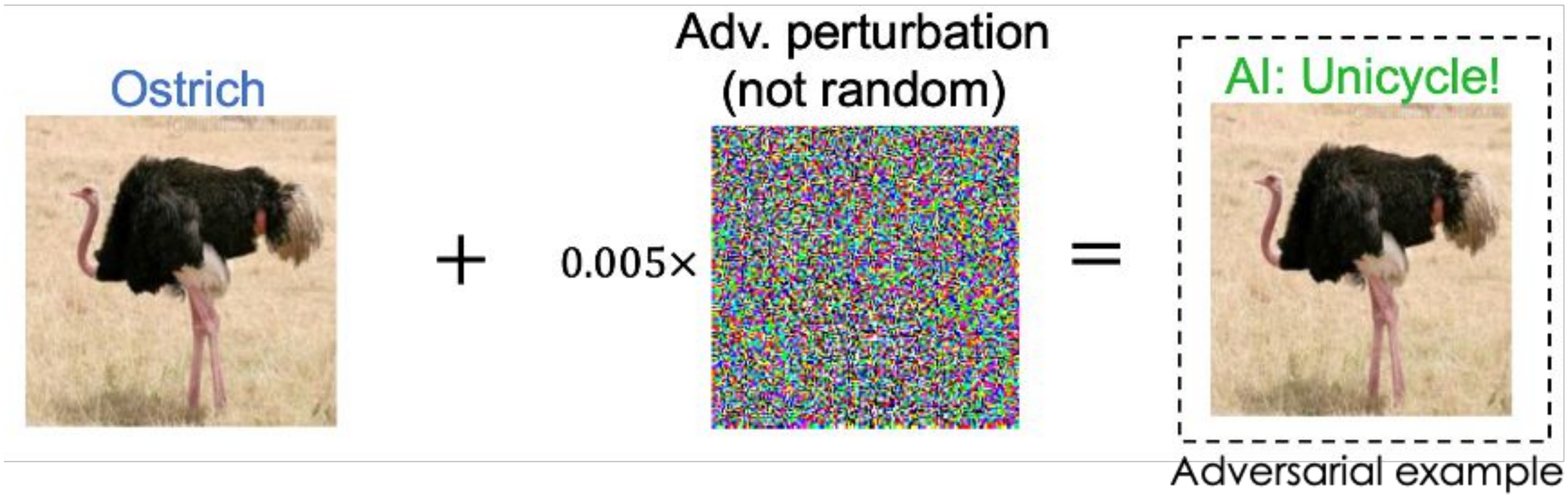
ICLR

OPTML

How to Robustify Black-Box ML Models?
A Zeroth-Order Optimization Perspective

Background

- DNN is **not robust** to the adversarial perturbations.



How to Robustify Black-Box ML Models?
A Zeroth-Order Optimization Perspective

Background

- DNN is **not robust** to the adversarial perturbations.
- Nearly all existing works ask a defender to perform **over white-box ML models**.

MICHIGAN STATE UNIVERSITY

OPTML

How to Robustify Black-Box ML Models?
A Zeroth-Order Optimization Perspective

Background

- DNN is **not robust** to the adversarial perturbations.
- Nearly all existing works ask a defender to perform **over white-box ML models**.

What if a model owner may refuse to share the model details ?

(e.g., APIs)

MICHIGAN STATE
UNIVERSITY

OPTML

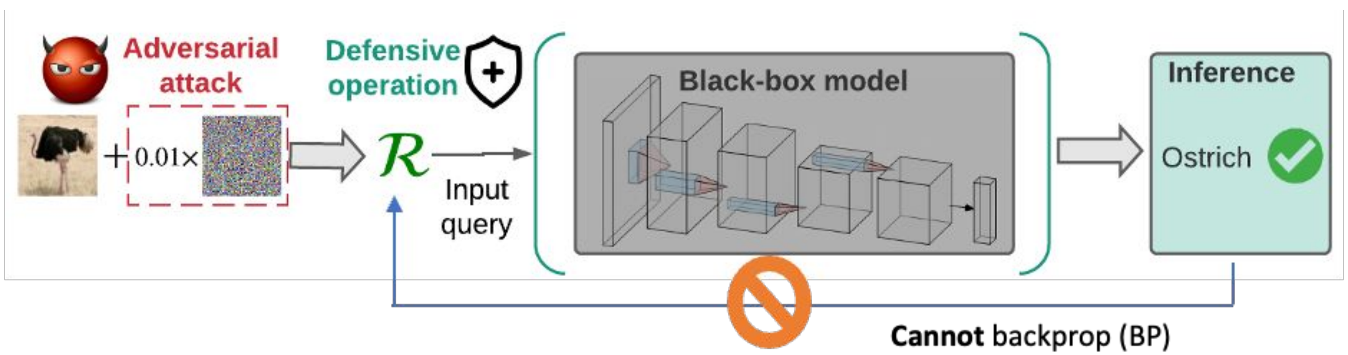
How to Robustify Black-Box ML Models?
A Zeroth-Order Optimization Perspective

Background

- DNN is **not robust** to the adversarial perturbations.
- Nearly all existing works ask a defender to perform **over white-box ML models**.

What if a model owner may refuse to share the model details ?

(e.g., APIs)



Black-Box Defense

MICHIGAN STATE UNIVERSITY

OPTML

How to Robustify Black-Box ML Models?
A Zeroth-Order Optimization Perspective

Background

- DNN is **not robust** to the adversarial perturbations.
- Nearly all existing works ask a defender to perform **over white-box ML models**.
- Zeroth-Order (ZO) Optimization can be utilized for black-box defense but suffers **high variance** for high-dimension variables.

MICHIGAN STATE
UNIVERSITY

OPTML

How to Robustify Black-Box ML Models?
A Zeroth-Order Optimization Perspective

Background

- DNN is **not robust** to the adversarial perturbations.
- Nearly all existing works ask a defender to perform **over white-box ML models**.
- Zeroth-Order (ZO) Optimization can be utilized for black-box defense but suffers **high variance** for high-dimension variables.

What Is ZO Optimization?

MICHIGAN STATE UNIVERSITY

OPTML

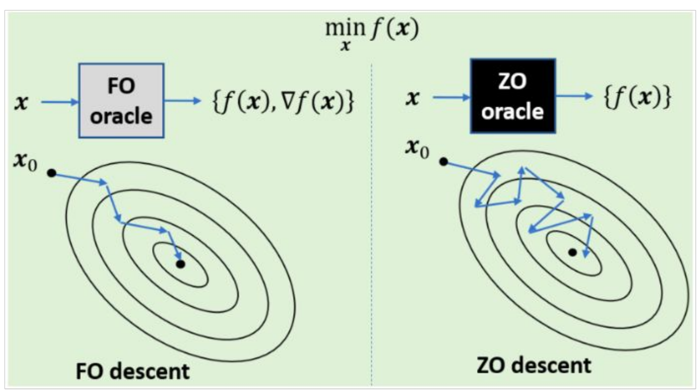
How to Robustify Black-Box ML Models?
A Zeroth-Order Optimization Perspective

Background

- DNN is **not robust** to the adversarial perturbations.
- Nearly all existing works ask a defender to perform **over white-box ML models**.
- Zeroth-Order (ZO) Optimization can be utilized for black-box defense but suffers **high variance** for high-dimension variables.

What Is ZO Optimization?

- **ZO Optimization:** Gradient-free optimization that leverages **finite differences of function values to estimate gradients**, rather than requesting explicit gradient information



Advantages:
- Simple, easy to implement
- Provable convergence as first-order optimization

Challenges:
- Slow convergence
- Lack of scalability in high dimensions

Liu, et al. "A primer on zeroth-order optimization in signal processing and machine learning", IEEE Signal Processing Magazine, 2020

MICHIGAN STATE UNIVERSITY

OPTML

How to Robustify Black-Box ML Models? A Zeroth-Order Optimization Perspective

Background

- DNN is **not robust** to the adversarial perturbations.
- Nearly all existing works ask a defender to perform **over white-box ML models**.
- Zeroth-Order (ZO) Optimization can be utilized for black-box defense but suffers **high variance** for high-dimension variables.

Randomized Gradient Estimate (RGE)

$$\hat{\nabla}_{\mathbf{w}}\ell(\mathbf{w}) = \frac{1}{q}\sum_{i=1}^{q}\left[\frac{d}{\mu}\left(\ell(\mathbf{w}+\mu\mathbf{u}_i) - \ell(\mathbf{w})\right)\mathbf{u}_i\right]$$

Coordinate-wise Gradient Estimate (CGE)

$$\hat{\nabla}_{\mathbf{w}}\ell(\mathbf{w}) = \sum_{i=1}^{d}\left[\frac{\ell(\mathbf{w}+\mu\mathbf{e}_i) - \ell(\mathbf{w})}{\mu}\mathbf{e}_i\right],$$

$\ell(w)$	:	black-box function
w	:	the ***d*-dimension** parameter
$\{\boldsymbol{u}_i\}_{i=1}^{q}$	:	q random vectors
μ	:	step size, known as smoothing parameter
$e_i \in R^d$	:	ith elementary basis vector (1 at the ith coordinate and 0s elsewhere)

MICHIGAN STATE UNIVERSITY

OPTML

How to Robustify Black-Box ML Models? A Zeroth-Order Optimization Perspective

Background

- DNN is **not robust** to the adversarial perturbations.
- Nearly all existing works ask a defender to perform **over white-box ML models**.
- Zeroth-Order (ZO) Optimization can be utilized for black-box defense but suffers **high variance** for high-dimension variables.

Randomized Gradient Estimate (RGE)

$$\hat{\nabla}_{\mathbf{w}}\ell(\mathbf{w}) = \frac{1}{q}\sum_{i=1}^{q}\left[\frac{d}{\mu}\left(\ell(\mathbf{w}+\mu\mathbf{u}_i)-\ell(\mathbf{w})\right)\mathbf{u}_i\right]$$

⇨ **High variances**

Coordinate-wise Gradient Estimate (CGE)

$$\hat{\nabla}_{\mathbf{w}}\ell(\mathbf{w}) = \sum_{i=1}^{d}\left[\frac{\ell(\mathbf{w}+\mu\mathbf{e}_i)-\ell(\mathbf{w})}{\mu}\mathbf{e}_i\right],$$

$\ell(w)$	:	black-box function
w	:	the ***d*-dimension** parameter
$\{\boldsymbol{u}_i\}_{i=1}^{q}$	:	q random vectors
μ	:	step size, known as smoothing parameter
$e_i \in R^d$	:	ith elementary basis vector (1 at the ith coordinate and 0s elsewhere)

MICHIGAN STATE UNIVERSITY

OPTML

How to Robustify Black-Box ML Models? A Zeroth-Order Optimization Perspective

Background

- DNN is **not robust** to the adversarial perturbations.
- Nearly all existing works ask a defender to perform **over white-box ML models**.
- Zeroth-Order (ZO) Optimization can be utilized for black-box defense but suffers **high variance** for high-dimension variables.

Randomized Gradient Estimate (RGE)

$$\hat{\nabla}_{\mathbf{w}}\ell(\mathbf{w}) = \frac{1}{q}\sum_{i=1}^{q}\left[\frac{d}{\mu}\left(\ell(\mathbf{w}+\mu\mathbf{u}_i)-\ell(\mathbf{w})\right)\mathbf{u}_i\right]$$

⇨ **High variances**

Coordinate-wise Gradient Estimate (CGE)

$$\hat{\nabla}_{\mathbf{w}}\ell(\mathbf{w}) = \sum_{i=1}^{d}\left[\frac{\ell(\mathbf{w}+\mu\mathbf{e}_i)-\ell(\mathbf{w})}{\mu}\mathbf{e}_i\right],$$

⇨ **High Computation Cost**

$\ell(w)$	:	black-box function
w	:	the ***d*-dimension** parameter
$\{u_i\}_{i=1}^{q}$	:	q random vectors
μ	:	step size, known as smoothing parameter
$e_i \in R^d$	:	ith elementary basis vector (1 at the ith coordinate and 0s elsewhere)

MICHIGAN STATE UNIVERSITY

OPTML

How to Robustify Black-Box ML Models? A Zeroth-Order Optimization Perspective

Background

- DNN is **not robust** to the adversarial perturbations.
- Nearly all existing works ask a defender to perform **over white-box ML models**.
- Zeroth-Order (ZO) Optimization can be utilized for black-box defense but suffers **high variance** for high-dimension variables.

Randomized Gradient Estimate (RGE)

$$\hat{\nabla}_{\mathbf{w}}\ell(\mathbf{w}) = \frac{1}{q}\sum_{i=1}^{q}\left[\frac{d}{\mu}\left(\ell(\mathbf{w}+\mu\mathbf{u}_i)-\ell(\mathbf{w})\right)\mathbf{u}_i\right]$$

⇨ **High variances**

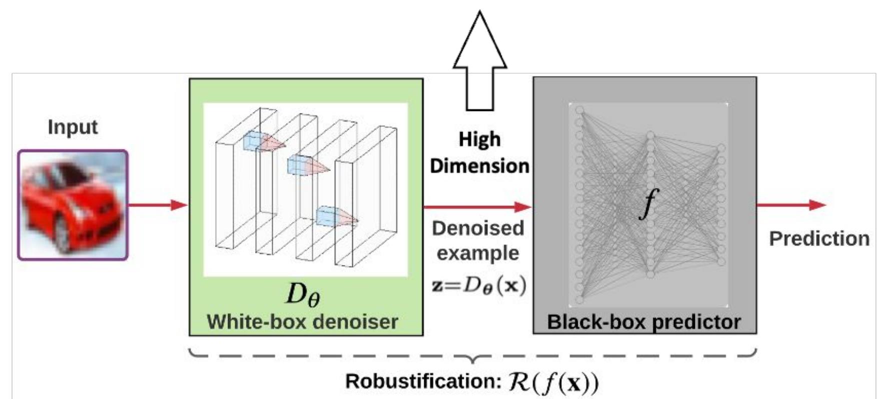
Coordinate-wise Gradient Estimate (CGE)

$$\hat{\nabla}_{\mathbf{w}}\ell(\mathbf{w}) = \sum_{i=1}^{d}\left[\frac{\ell(\mathbf{w}+\mu\mathbf{e}_i)-\ell(\mathbf{w})}{\mu}\mathbf{e}_i\right],$$

⇨ **High Computation Cost**

$\ell(w)$	:	black-box function
w	:	the ***d*-dimension** parameter
$\{\boldsymbol{u}_i\}_{i=1}^{q}$	:	q random vectors
μ	:	step size, known as smoothing parameter
$e_i \in R^d$	:	ith elementary basis vector (1 at the ith coordinate and 0s elsewhere)

Zeroth-Order Optimization for high-dimension variables suffers high variance ⚠️ ⚠️ ⚠️



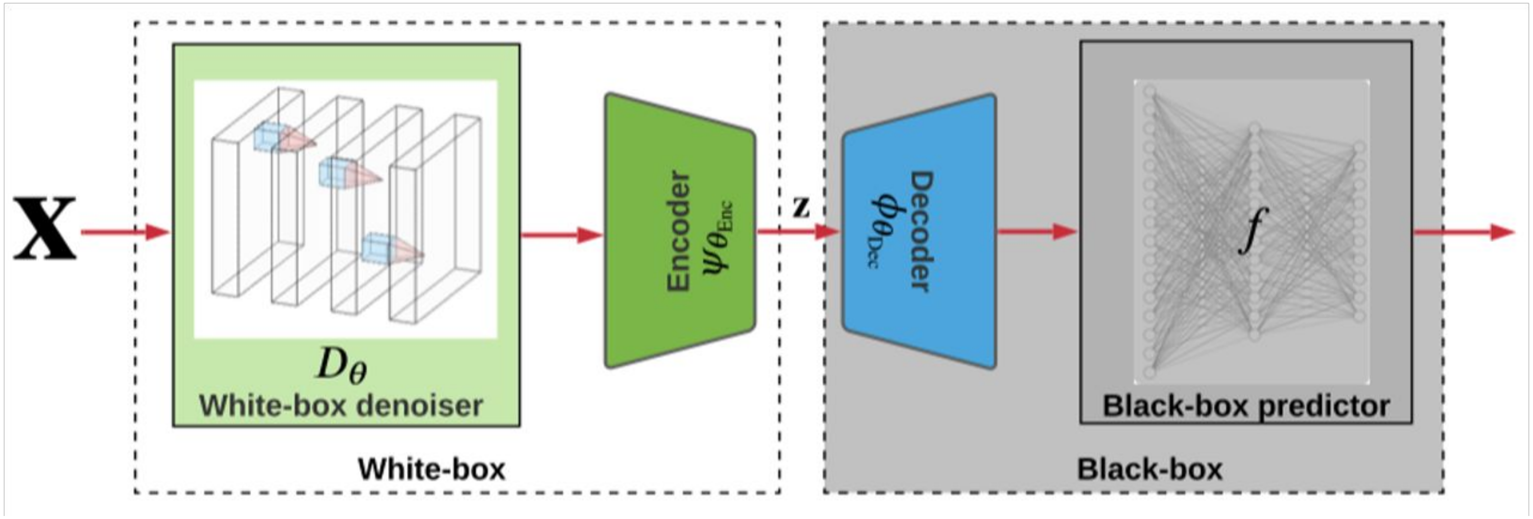
D_θ	:	white-box denoiser with parameter θ
f	:	black-box predictor
x	:	input

MICHIGAN STATE UNIVERSITY

OPTML

How to Robustify Black-Box ML Models?
A Zeroth-Order Optimization Perspective

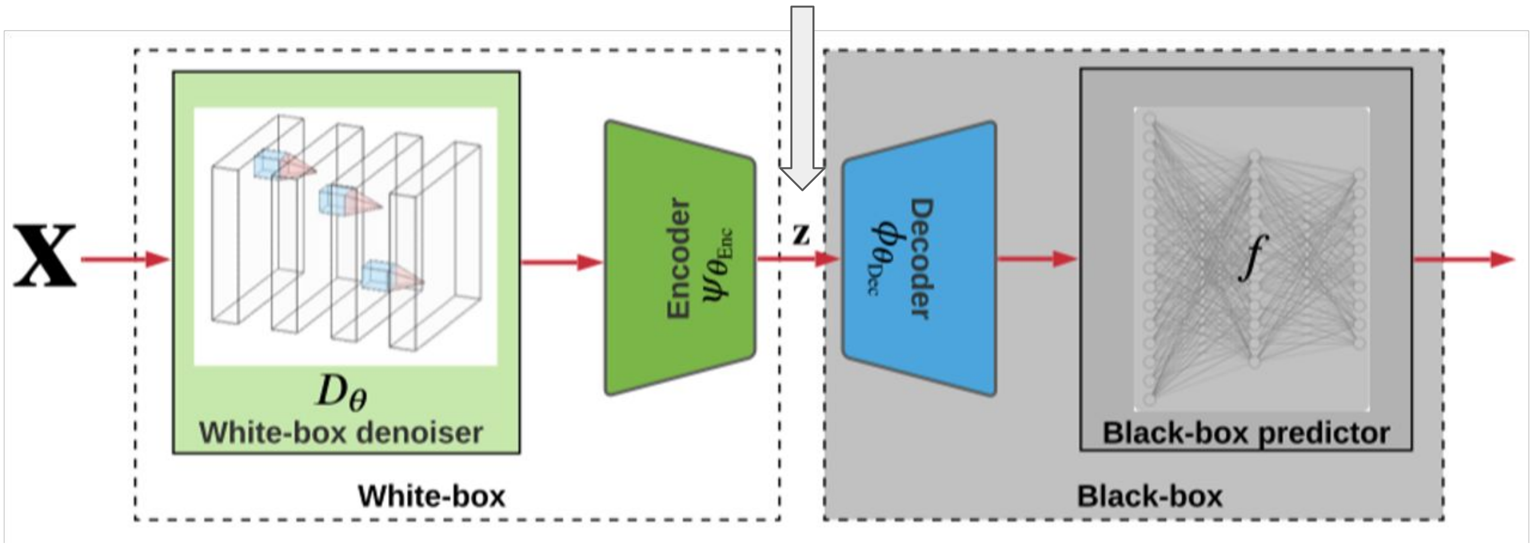
Method



How to Robustify Black-Box ML Models?
A Zeroth-Order Optimization Perspective

Method

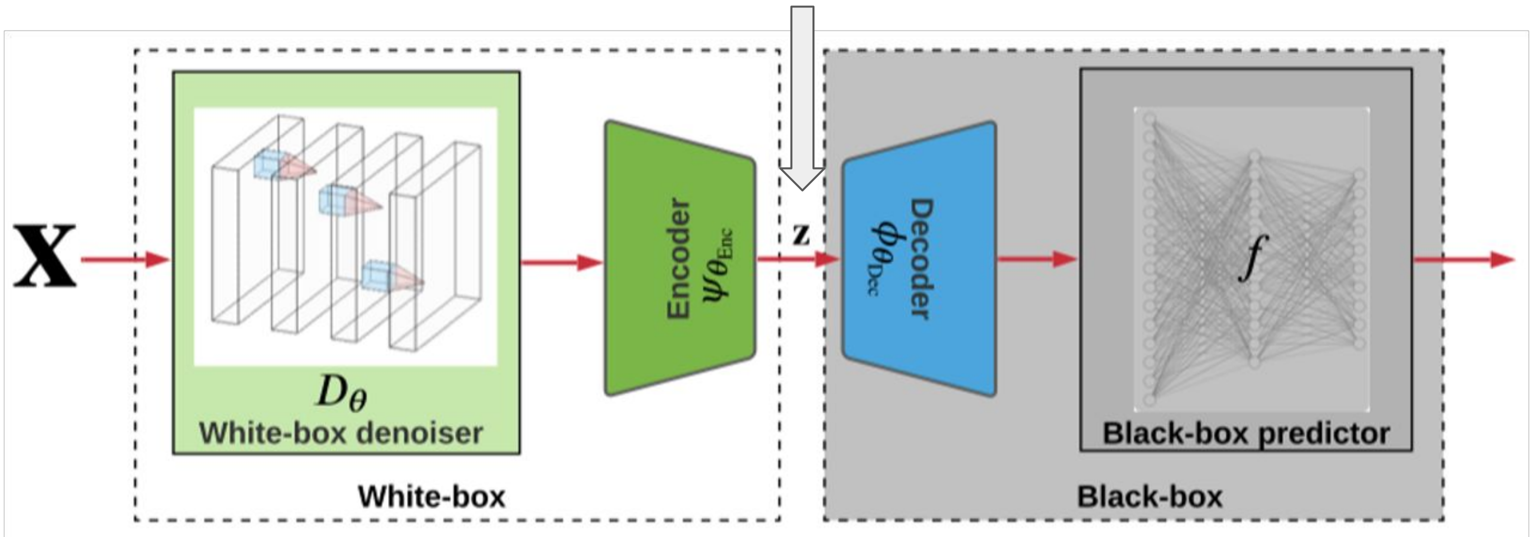
The dimension of z is reduced!!



How to Robustify Black-Box ML Models? A Zeroth-Order Optimization Perspective

Method

The dimension of z is reduced!!



$$\nabla_{\boldsymbol{\theta}} \mathcal{R}(f(\mathbf{x})) = \frac{dD_{\boldsymbol{\theta}}(\mathbf{x})}{d\boldsymbol{\theta}} \frac{df(\mathbf{z})}{d\mathbf{z}} \Big|_{\mathbf{z}=D_{\boldsymbol{\theta}}(\mathbf{x})} \approx \overbrace{\boxed{\frac{dD_{\boldsymbol{\theta}}(\mathbf{x})}{d\boldsymbol{\theta}}}}^{\text{FO Gradient}} \overbrace{\boxed{\hat{\nabla}_{\mathbf{z}} f(\mathbf{z})}}^{\text{ZO Gradient Estimation}} \Big|_{\mathbf{z}=D_{\boldsymbol{\theta}}(\mathbf{x})}$$

MICHIGAN STATE UNIVERSITY

OPTML

How to Robustify Black-Box ML Models? A Zeroth-Order Optimization Perspective

Performance

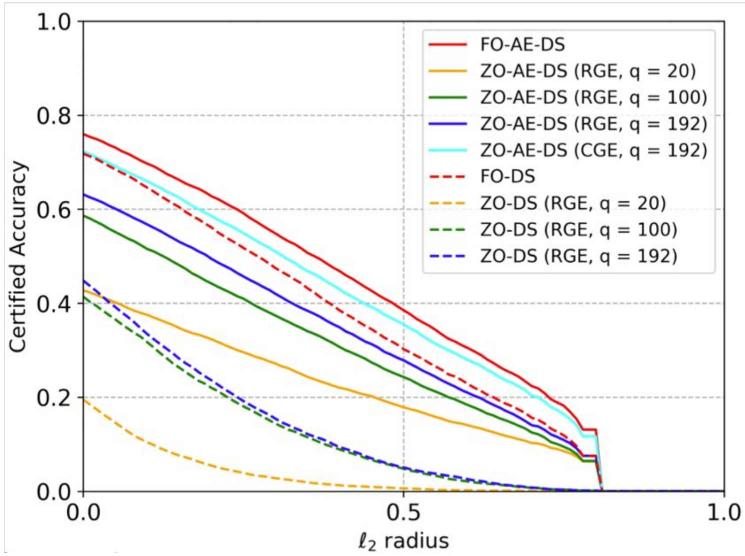
(White-box baseline) **(Black-box baseline)**

ℓ_2-radius r	FO: RS	FO: FO-DS	FO: FO-AE-DS	ZO-DS: $q=20$ (RGE)	ZO-DS: $q=100$ (RGE)	ZO-DS: $q=192$ (RGE)	ZO-AE-DS (Ours): $q=20$ (RGE)	ZO-AE-DS (Ours): $q=100$ (RGE)	ZO-AE-DS (Ours): $q=192$ (RGE)	ZO-AE-DS (Ours): $q=192$ (CGE)
0.00 (SA)	**76.44**	71.80	75.97	19.50	41.38	44.81	42.72	58.61	63.13	**72.23**
0.25	**60.64**	51.74	59.12	3.89	18.05	19.16	29.57	40.96	45.69	**54.87**
0.50	**41.19**	30.22	38.50	0.60	4.78	5.06	17.85	24.28	27.84	**35.50**
0.75	**21.11**	11.87	18.18	0.03	0.32	0.30	8.52	9.45	10.89	**16.37**

Dataset: CIFAR-10
Black-box classifier: ResNet-110
White-box denoiser: DnCNN

- *FO* : First-Order optimization
- *ZO* : Zeroth-Order optimization
- *RGE* : Randomized Gradient Estimate
- CGE : Coordinate-wise Gradient Estimate
- *q* : the number of queries

- *RS* : Randomized Smoothing
- *DS* : Denoised Smoothing
- *AE- DS* : AutoEncoder-based Denoised Smoothing (Ours)



MICHIGAN STATE UNIVERSITY

OPTML