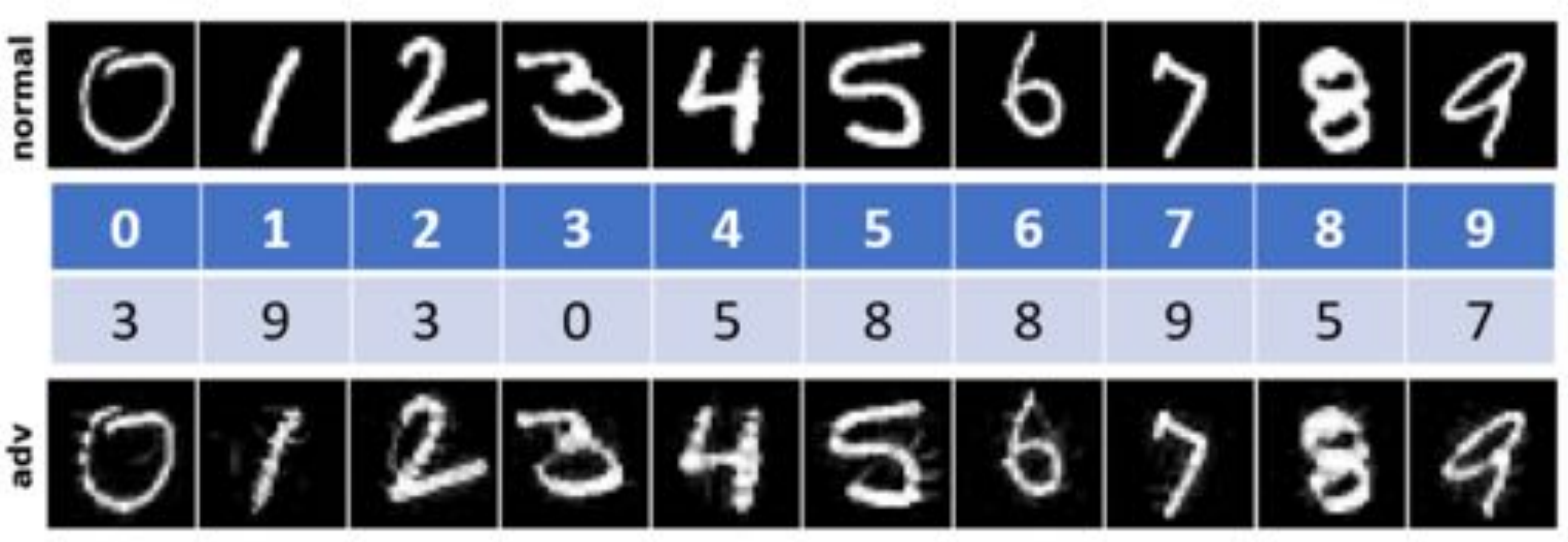
On the Convergence and Robustness of Adversarial Training

Yisen Wang*, Xingjun Ma*
James Bailey, Jinfeng Yi, Bowen Zhou, Quanquan Gu

JD.com University of Melbourne UCLA

Adversarial Examples:

Handwritten Digits: MNIST



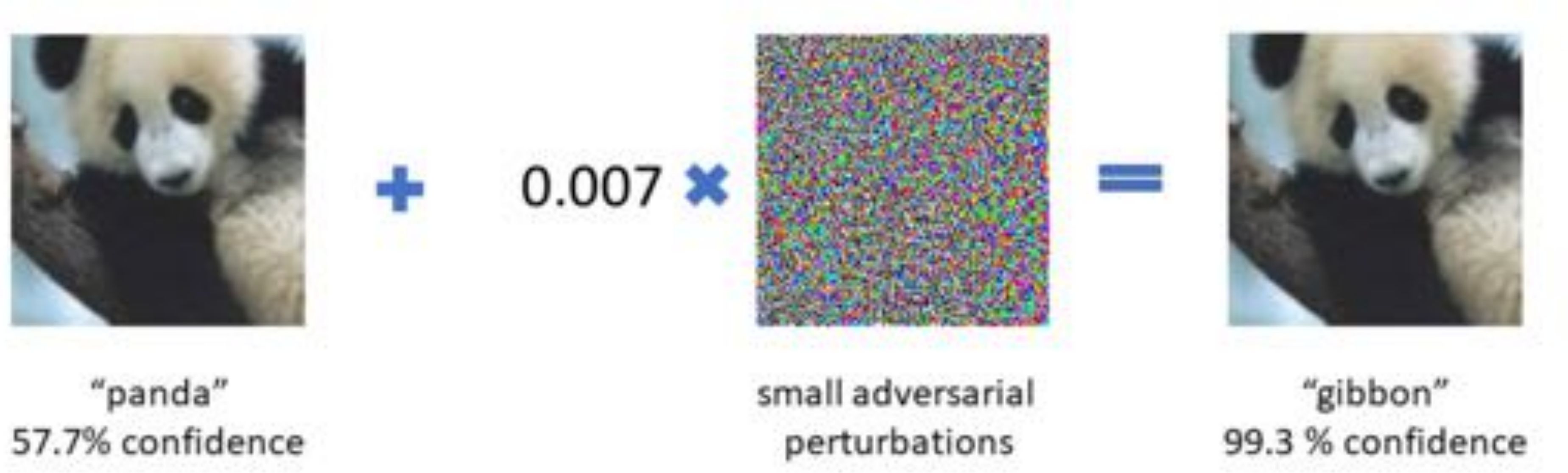
← original class

← adversarial class

✓ Small perturbations added to normal inputs can easily fool a DNN.

Adversarial Examples:

Natural Images



✓ Perturbations are small, imperceptible to human eyes.

Making DNN robust to adversarial examples is crucial !

Szegedy et al. 2013, Goodfellow et al. 2014

Adversarial Defense -- Adversarial Training:

Core idea: training robust DNNs on adversarial examples.

- Min-max formulation:

$$\min_{\boldsymbol{\theta}} \frac{1}{n}\sum_{i=1}^{n} \max_{\|x_i - x_i^0\| \leq \epsilon} \ell(h_\theta(x_i, y_i))$$

where, x_i^0 is a natural (clean) training sample, y_i is the label of x_i^0.

Inner Maximization:

- Inner maximization is to generate adversarial examples, by maximizing classification loss (e.g. $\ell(\cdot)$).
- It is a **constrained** optimization problem: $\|x_i - x_i^0\| \leq \epsilon$.
- First order method Projected Gradient Descent (**PGD)** usually gives good solution.

Outer Minimization:

- Outer minimization is to train a robust model on adversarial examples generated in the inner maximization.
- **It is hugely influenced by how well the maximization is solved.**

Convergence Quality of Adversarial Training Examples:

Question: How to measure the convergence quality of the inner maximization?

Definition (First-Order Stationary Condition (FOSC))

Given a data sample $x^0 \in X$, let x^k be an intermediate example found at the k^{th} step of the inner maximization. The First-Order Stationary Condition of x^k is

$$c(x^k) = \max_{x \in \chi} \langle x - x^k, \nabla_x f(\boldsymbol{\theta}, x^k) \rangle,$$

where $\chi = \{x \,|\, \|x - x^0\|_\infty \leq \epsilon\}$ is the input domain of the ϵ-ball around normal example x^0, $f(\boldsymbol{\theta}, x^k) = \ell(h_{\boldsymbol{\theta}}(x^k, y))$, and $\langle \cdot \rangle$ is the inner product.

FOSC:

- A smaller value of $c(x^k)$ indicates a better solution of the inner maximization, or equivalently, better convergence quality of the adversarial example x^k.
- It has a closed-form solution.

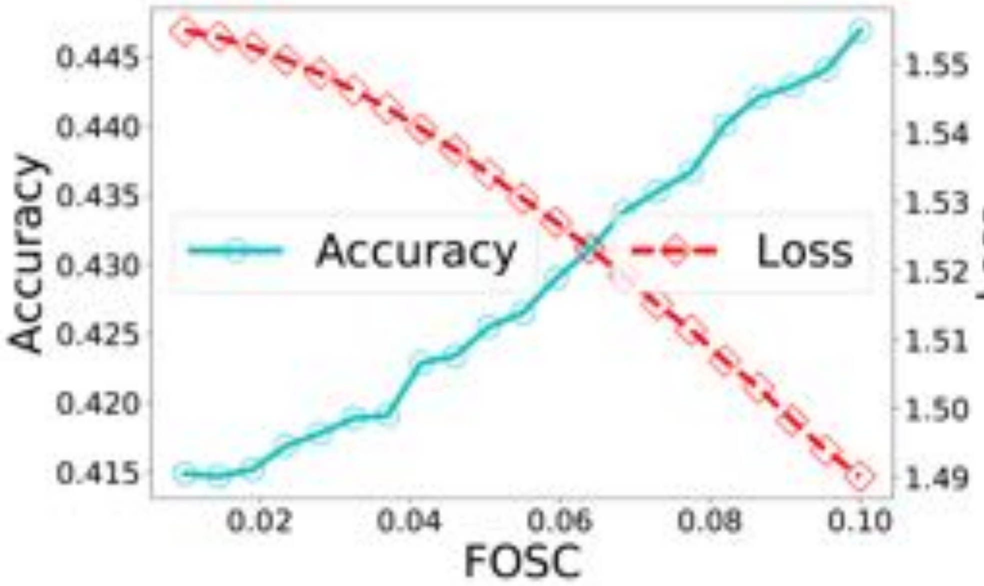
Closed-form Solution of FOSC:

FOSC have the following closed-form solution:

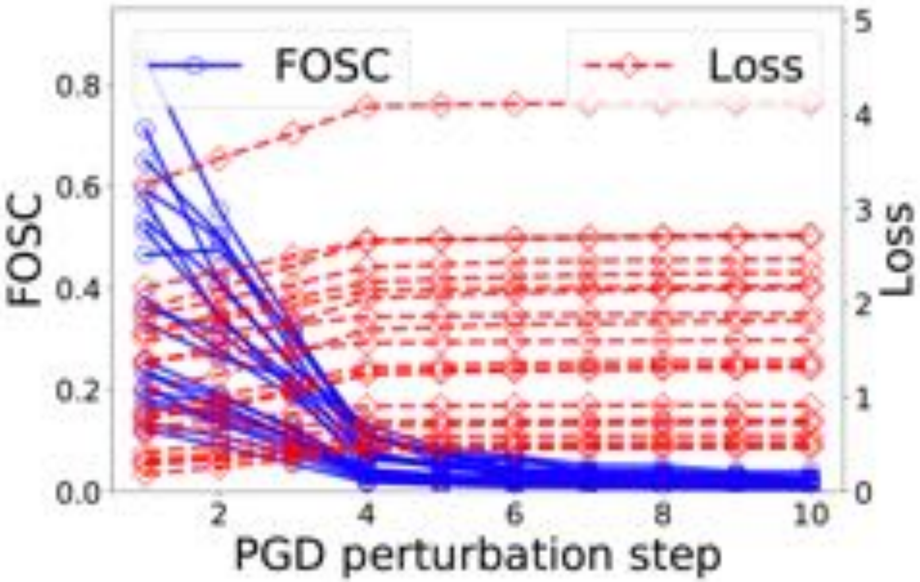
$$
\begin{aligned}
c(x^k) &= \max_{x \in \chi} \langle x - x^k,\ \nabla_x f(\boldsymbol{\theta}, x^k) \rangle \\
&= \max_{x \in \chi} \langle x - x^0 + x^0 - x^k,\ \nabla_x f(\boldsymbol{\theta}, x^k) \rangle \\
&= \max_{x \in \chi} \langle x - x^0, \nabla_x f(\boldsymbol{\theta}, x^k) \rangle + \langle x^k - x^0, -\nabla_x f(\boldsymbol{\theta}, x^k) \rangle \\
&= \epsilon \cdot \left\| \nabla_x f(\boldsymbol{\theta}, x^k) \right\|_1 - \langle x^k - x^0, \nabla_x f(\boldsymbol{\theta}, x^k) \rangle
\end{aligned}
$$

- The last equality is because the dual norm of $\max\langle\cdot\rangle$ is the L_1-norm under ∞ case.
- $c(x^k) = 0$ indicates x^k is the optimal solution, and can be achieved when:
  1. $\nabla_{\mathrm{x}} f(\boldsymbol{\theta}, \mathrm{x}^{\mathrm{k}}) = 0$: x^{k} is a stationary point in the interior of χ.
  2. $\mathrm{x}^{\mathrm{k}} - \mathrm{x}^0 = \epsilon \cdot \mathrm{sign}(\nabla_{\mathrm{x}} \mathrm{f}(\boldsymbol{\theta}, \mathrm{x}^{\mathrm{k}}))$: local maximum point of $\mathrm{f}(\boldsymbol{\theta}, \mathrm{x}^{\mathrm{k}})$ is reached on the boundary of χ.

FOSC View of Adversarial Strength:



(a) Accuracy, Loss vs. FOSC



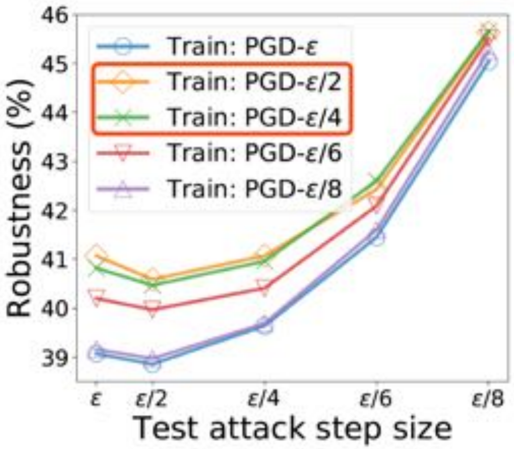
(b) FOSC, Loss vs. Step

- The lower the FOSC, the lower the accuracy, and the higher the loss. Meaning the stronger attack

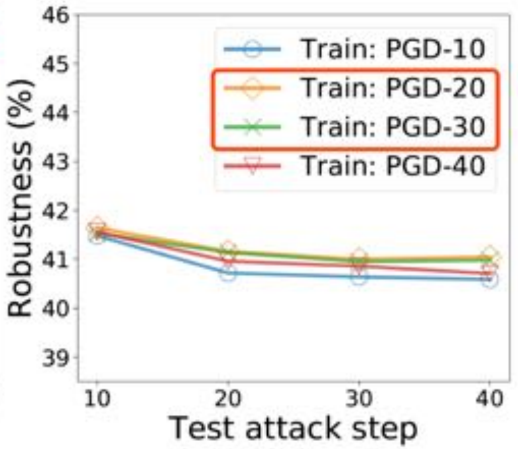
- The closer FOSC to 0, the stronger the attack. While the loss varies a large range.

FOSC provides a comparable and consistent measurement of adversarial strength.

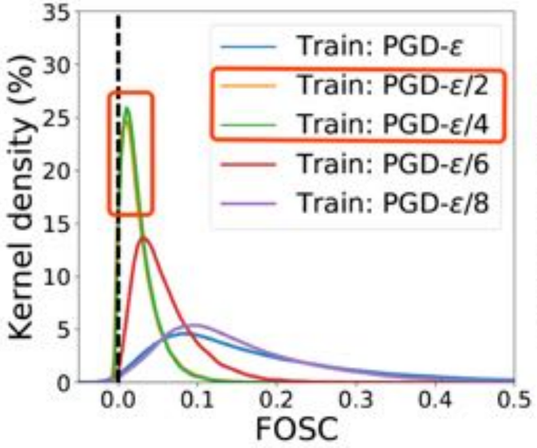
FOSC View of Adversarial Robustness:



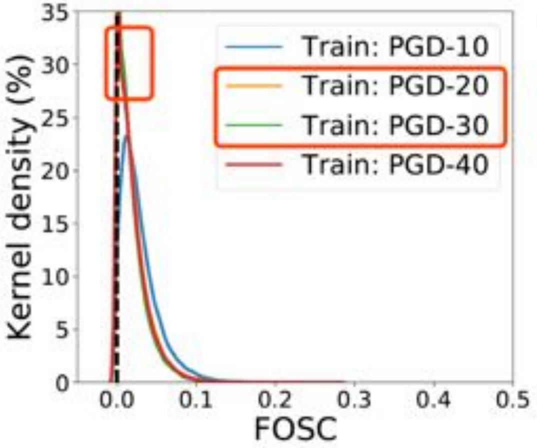
(a) Robustness vs. Step size



(b) Robustness vs. Step number

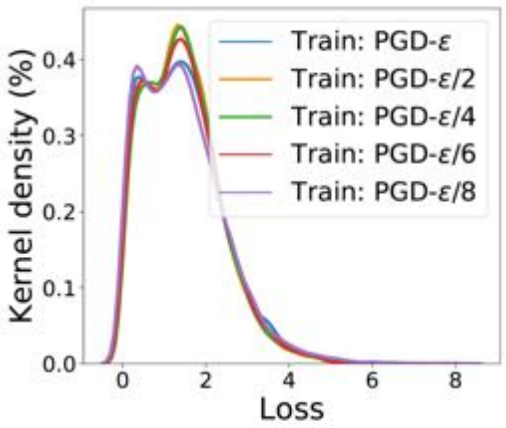


(c) FOSC vs. Step size

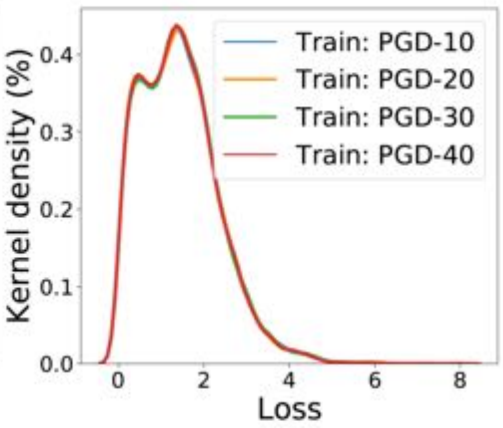


(d) FOSC vs. Step number

FOSC is a good and reliable indicator of the final robustness



(e) Loss vs. Step size

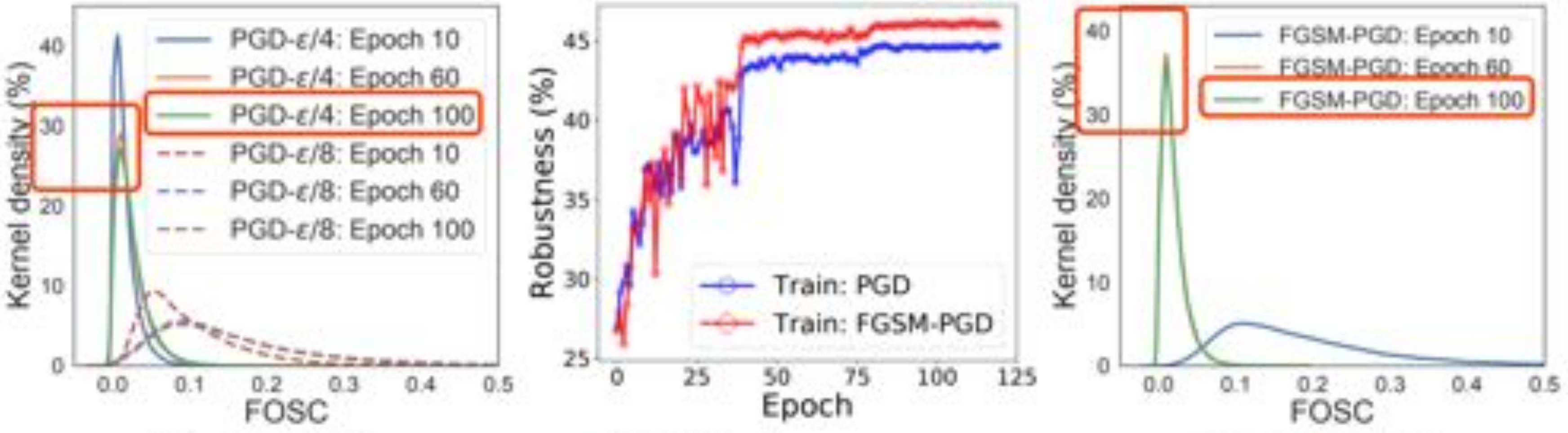


(f) Loss vs. Step number

Adversarial Training with different settings for PGD-based inner maximization.

- **PGD step size**: PGD-$\frac{\epsilon}{2}$ / PGD-$\frac{\epsilon}{4}$ produces the best robustness, their FOSC values are also concentrated around 0.
- **PGD step number**: similar robustness, with PGD-20/30 are slightly better, reflected by the distribution of FOSC.
- **Loss distributions** are similar for different robustness.

FOSC View of Adversarial Training Process:



(a) FOSC (b) Robustness (c) FOSC

- Standard adversarial training **overfits** to strong PGD adversarial examples at the **early stage**.

- Simply use **weak attack FGSM** at the **early stage** can improve robustness.

- Improvement in robustness is also reflected in FOSC distribution.

Proposed Dynamic Adversarial Training (Dynamic):

Adversarial training with **dynamic convergence control** of the inner maximization:
gradually increasing convergence quality, i.e., gradually decreasing FOSC.

Algorithm 1 Dynamic Adversarial Training

Input: Network $h_{\boldsymbol{\theta}}$, training data S, initial model parameters $\boldsymbol{\theta}^0$, step size η_t, mini-batch $\mathcal{B}$, maximum FOSC value c_{max}, training epochs T, FOSC control epoch T', PGD step K, PGD step size α, maximum perturbation ϵ.

for $t = 0$ **to** $T-1$ **do**

  $c_t = \max(c_{\max} - t \cdot c_{\max}/T', 0)$

  for each batch $\mathbf{x}^0_{\mathcal{B}}$ **do**

    $V = \mathbb{1}_{\mathcal{B}}$ *# control vector of all elements is 1*

    while $\sum V > 0$ & $k < K$ **do**

      $\mathbf{x}^{k+1}_{\mathcal{B}} = \mathbf{x}^k_{\mathcal{B}} + \boxed{V \cdot} \alpha \cdot \mathrm{sign}(\nabla_{\mathbf{x}} \ell(h_{\boldsymbol{\theta}}(\mathbf{x}^k_{\mathcal{B}}), y))$

      $\mathbf{x}^k_{\mathcal{B}} = clip(\mathbf{x}^k_{\mathcal{B}}, \mathbf{x}^0_{\mathcal{B}} - \epsilon, \mathbf{x}^0_{\mathcal{B}} + \epsilon)$

      $V = \mathbb{1}_{\mathcal{B}}(c(\mathbf{x}^k_{1 \dots \mathcal{B}}) \le c_t)$ *# The element of V becomes 0 at which FOSC is smaller than* c_t

    end while

    $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \eta_t \mathbf{g}(\boldsymbol{\theta}^t)$ # $\mathbf{g}(\boldsymbol{\theta}^t)$ *: stochastic gradient*

  end for

end for

Comparing to Standard Adv Training:

✓ At each perturbation step

✓ Monitoring the FOSC value

✓ Stopping the perturbation process once FOSC $\le c_t$ (enabled by control vector V)

Convergence Analysis:

Assumption 1. $f(\boldsymbol{\theta}; x)$ *satisfies the gradient Lipschitz conditions as follows*

$$\sup_x \|\nabla_\theta f(\theta, x) - \nabla_\theta f(\theta', x)\|_2 \le L_{\theta\theta}\|\theta - \theta'\|_2$$

$$\sup_\theta \|\nabla_\theta f(\theta, x) - \nabla_\theta f(\theta, x')\|_2 \le L_{\theta x}\|x - x'\|_2$$

$$\sup_x \|\nabla_x f(\theta, x) - \nabla_{\boldsymbol{x}} f(\theta', x)\|_2 \le L_{\theta\theta}\|\theta - \theta'\|_2$$

Assumption 2. $f(\boldsymbol{\theta}; x)$ *is locally μ-strongly concave in the gradient Lipschitz conditions as follows* $\chi_i = \{x: \|x_i - x_i^0\|_\infty \le \epsilon\}$ *for all* $i \in [n]$, *i.e., for any* x_1, $x_2 \in \chi_i$, *it holds that*

$$f(\theta, x_1) \le f(\theta, x_2) + \langle \nabla_x f(\boldsymbol{\theta}, x_2), x_1 - x_2 \rangle - \frac{\mu}{2}\|x_1 - x_2\|_2^2$$

Assumption 3. *The variance of the stochastic gradient* $g(\theta)$ *is bounded by a constant* $\sigma^2 > 0$,

$$\mathbb{E}[\|g(\theta) - \nabla L_S(\theta)\|_2^2] \le \sigma^2$$

Convergence Theorem:

Theorem 1. *Under certain assumptions, let* $\Delta = L_S(\theta^0) - \min_\theta L_S(\theta)$*. If the step size of the outer minimization is set to* $\eta_t = min\left(\frac{1}{L}, \sqrt{\frac{\Delta}{L\sigma^2 T}}\right)$*. Then the output of* ***Dynamic Adversarial Training*** *satisfies*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}[\|\nabla L_S(\theta^t)\|_2^2] \leq 4\sigma\sqrt{\frac{L\Delta}{T}} + \frac{5L_{\theta x}^2\delta}{\mu},$$

where $L = \left(\frac{L_{\theta x}L_{\theta x}}{\mu} + L_{\theta\theta}\right)$.

- If the inner maximization is solved up to a precision that FOSC is less than δ, **Dynamic** can converge to a first-order stationary point at a sublinear rate up to a precision of $\frac{5L_{\theta x}^2\delta}{\mu}$.

- If δ is sufficiently small such that $\frac{5L_{\theta x}^2\delta}{\mu}$ small enough, **Dynamic** can find a robust model θ^T.

Robustness Evaluation of Dynamic:

Table 1. White-box robustness (accuracy (%) on white-box test attacks) of different defense models on MNIST and CIFAR-10 datasets.

Defense	MNIST					CIFAR-10				
	Clean	FGSM	PGD-10	PGD-20	C&W$_\infty$	Clean	FGSM	PGD-10	PGD-20	C&W$_\infty$
Unsecured	**99.20**	14.04	0.0	0.0	0.0	**89.39**	2.2	0.0	0.0	0.0
Standard	97.61	94.71	91.21	90.62	91.03	66.31	48.65	44.39	40.02	36.33
Curriculum	98.62	**95.51**	91.24	90.65	91.12	72.40	50.47	45.54	40.12	35.77
Dynamic	97.96	95.34	**91.63**	**91.27**	**91.47**	72.17	**52.81**	**48.06**	**42.40**	**37.26**

Table 2. Black-box robustness (accuracy (%) on black-box test attacks) of different defense models on MNIST and CIFAR-10 datasets.

Defense	MNIST				CIFAR-10			
	FGSM	PGD-10	PGD-20	C&W$_\infty$	FGSM	PGD-10	PGD-20	C&W$_\infty$
Standard	96.12	95.73	95.73	97.20	65.65	65.80	65.60	66.12
Curriculum	96.59	95.87	96.09	97.52	71.25	71.44	71.13	71.94
Dynamic	**97.60**	**97.01**	**96.97**	**98.36**	**71.95**	**72.15**	**72.02**	**72.85**

- Network: 4-layer CNN on MNIST and 8-layer CNN on CIFAR-10
- $\epsilon = 0.3$ for MNIST and $\epsilon = 8/255$ for CIFAR-10 (Standard defense settings)
- Better robustness than the state-of-the-art against 4 white-box and black-box attacks

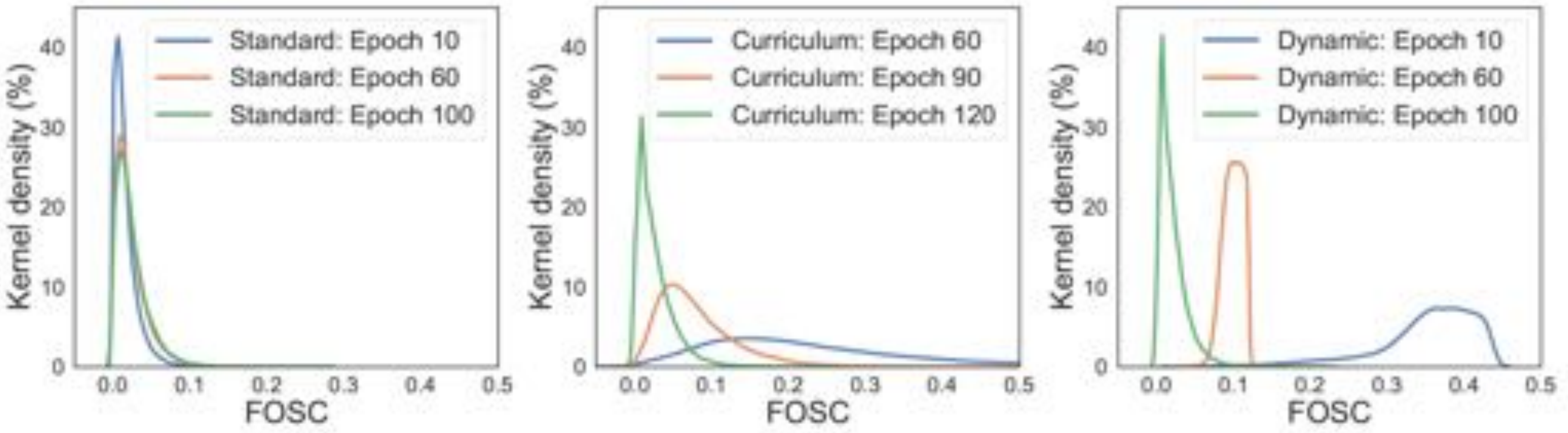
Benchmarking the State-of-the-art on WideResNet:

Table 3. White-box robustness (%) of different defense models on CIFAR-10 dataset using WideResNet setting in Madry's baselines.

Defense	Clean	FGSM	PGD-20	C&W$_\infty$
Madry's	**87.3**	56.1	45.8	46.8
Curriculum	77.43	57.17	46.06	42.28
Dynamic	85.03	**63.53**	**48.70**	**47.27**

- **Network: WideResNet (10 times wider than ResNet)**
- **$\epsilon = 8/255$ for CIFAR-10 (Standard defense settings)**
- **Achieving the state-of-the-art robustness against various attacks on CIFAR-10**

FOSC View of Dynamic Adversarial Training:



(a) *Standard* (b) *Curriculum* (c) *Dynamic*

✓ **Dynamic** has more precise control over the convergence quality with FOSC criterion.

- **More concentrated FOSC distributions at each stages of training.**
- **More separated FOSC distributions at different stages of training.**

Thank you!

Poster @ Pacific Ballroom #151 Wed 6:30 pm

eewangyisen@gmail.com
qgu@cs.ucla.edu