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

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

original class  adversarial class
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:
  \[
  \min_{\theta} \frac{1}{n} \sum_{i=1}^{n} \max_{\|x_i - x_i^0\| \leq \epsilon} \ell(h_\theta(x_i, y_i))
  \]
  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

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

where $\chi = \{x \mid \|x - x^0\|_\infty \leq \epsilon\}$ is the input domain of the $\epsilon$-ball around normal example $x^0$, $f(\theta, x^k) = \ell(h_\theta(x^k, y))$, 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:

\[
c(x^k) = \max_{x \in \chi} \langle x - x^k, \nabla_x f(\theta, x^k) \rangle
= \max_{x \in \chi} \langle x - x^0 + x^0 - x^k, \nabla_x f(\theta, x^k) \rangle
= \max_{x \in \chi} \langle x - x^0, \nabla_x f(\theta, x^k) \rangle + \langle x^k - x^0, -\nabla_x f(\theta, x^k) \rangle
= \epsilon \cdot \|\nabla_x f(\theta, x^k)\|_1 - \langle x^k - x^0, \nabla_x f(\theta, x^k) \rangle
\]

- The last equality is because the dual norm of \(\max(\cdot)\) is the \(L_1\)-norm under \(\infty\) case.
- \(c(x^k) = 0\) indicates \(x^k\) is the optimal solution, and can be achieved when:
  1. \(\nabla_x f(\theta, x^k) = 0\): \(x^k\) is a stationary point in the interior of \(\chi\).
  2. \(x^k - x^0 = \epsilon \cdot \text{sign}(\nabla_x f(\theta, x^k))\): local maximum point of \(f(\theta, x^k)\) is reached on the boundary of \(\chi\).
FOSC View of Adversarial Strength:

- 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:

- Adversarial Training with different settings for PGD-based inner maximization.
  - **PGD step size**: $\text{PGD-}\frac{\epsilon}{2}$ / $\text{PGD-}\frac{\epsilon}{4}$ 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 is a good and reliable indicator of the final robustness.
FOSC View of Adversarial Training Process:

- 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 $h_\theta$, training data $S$, initial model parameters $\theta^0$, step size $\eta_t$, mini-batch $B$, maximum FOSC value $c_{max}$, training epochs $T$, FOSC control epoch $T'$, PGD step $K$, PGD step size $\alpha$, maximum perturbation $\epsilon$.

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

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

**for** each batch $x_B^k$ **do**

$V = 1_B$ # control vector of all elements is 1

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

$x_B^{k+1} = x_B^k + V \cdot \alpha \cdot \text{sign}(\nabla_x \ell(h_\theta(x_B^k), y))$

$x_B^k = \text{clip}(x_B^k, x_B^0 - \epsilon, x_B^0 + \epsilon)$

$V = 1_B(c(x_1^k...B) \leq c_t)$ # The element of $V$ becomes 0 at which FOSC is smaller than $c_t$

end while

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

end for

Comparing to Standard Adv Training:

✓ At each perturbation step
✓ Monitoring the FOSC value
✓ Stopping the perturbation process once FOSC $\leq c_t$ (enabled by control vector $V$)
Convergence Analysis:

Assumption 1. \( f(\theta; x) \) satisfies the gradient Lipschitz conditions as follows

\[
\sup_x \| \nabla_\theta f(\theta, x) - \nabla_\theta f(\theta', x) \|_2 \leq L_{\theta\theta} \| \theta - \theta' \|_2 \\
\sup_\theta \| \nabla_\theta f(\theta, x) - \nabla_\theta f(\theta, x') \|_2 \leq L_{\theta x} \| x - x' \|_2 \\
\sup_x \| \nabla_x f(\theta, x) - \nabla_x f(\theta', x) \|_2 \leq L_{\theta\theta} \| \theta - \theta' \|_2
\]

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

\[
f(\theta, x_1) \leq f(\theta, x_2) + \langle \nabla_x f(\theta, x_2), x_1 - x_2 \rangle - \frac{\mu}{2} \| x_1 - x_2 \|_2^2
\]

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

\[
\mathbb{E}[\| g(\theta) - \nabla L_S(\theta) \|_2^2] \leq \sigma^2
\]
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 $\eta_t = \min \left( \frac{1}{L}, \sqrt{\frac{\Delta}{L\sigma^2T}} \right)$. Then the output of Dynamic Adversarial Training satisfies

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

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

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

- If $\delta$ is sufficiently small such that $\frac{5L^2\theta_x}{\mu}$ small enough, Dynamic can find a robust model $\theta^T$. 
Robustness Evaluation of Dynamic:

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

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

<table>
<thead>
<tr>
<th>Defense</th>
<th>Clean</th>
<th>FGSM</th>
<th>PGD-10</th>
<th>PGD-20</th>
<th>C&amp;W$_\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsecured</td>
<td>99.20</td>
<td>14.04</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Standard</td>
<td>97.61</td>
<td>94.71</td>
<td>91.21</td>
<td>90.62</td>
<td>91.03</td>
</tr>
<tr>
<td>Curriculum</td>
<td>98.62</td>
<td>95.51</td>
<td>91.24</td>
<td>90.65</td>
<td>91.12</td>
</tr>
<tr>
<td>Dynamic</td>
<td>97.96</td>
<td>95.34</td>
<td>91.63</td>
<td>91.27</td>
<td>91.47</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Defense</th>
<th>Clean</th>
<th>FGSM</th>
<th>PGD-10</th>
<th>PGD-20</th>
<th>C&amp;W$_\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unsecured</td>
<td>89.39</td>
<td>2.2</td>
<td>0.0</td>
<td>0.0</td>
<td>0.0</td>
</tr>
<tr>
<td>Standard</td>
<td>66.31</td>
<td>48.65</td>
<td>44.39</td>
<td>40.02</td>
<td>36.33</td>
</tr>
<tr>
<td>Curriculum</td>
<td>72.40</td>
<td>50.47</td>
<td>45.54</td>
<td>40.12</td>
<td>35.77</td>
</tr>
<tr>
<td>Dynamic</td>
<td>72.17</td>
<td>52.81</td>
<td>48.06</td>
<td>42.40</td>
<td>37.26</td>
</tr>
</tbody>
</table>

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

<table>
<thead>
<tr>
<th>Defense</th>
<th>FGSM</th>
<th>PGD-10</th>
<th>PGD-20</th>
<th>C&amp;W$_\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>96.12</td>
<td>95.73</td>
<td>95.73</td>
<td>97.20</td>
</tr>
<tr>
<td>Curriculum</td>
<td>96.59</td>
<td>95.87</td>
<td>96.09</td>
<td>97.52</td>
</tr>
<tr>
<td>Dynamic</td>
<td>97.60</td>
<td>97.01</td>
<td>96.97</td>
<td>98.36</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Defense</th>
<th>FGSM</th>
<th>PGD-10</th>
<th>PGD-20</th>
<th>C&amp;W$_\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Standard</td>
<td>65.65</td>
<td>65.80</td>
<td>65.60</td>
<td>66.12</td>
</tr>
<tr>
<td>Curriculum</td>
<td>71.25</td>
<td>71.44</td>
<td>71.13</td>
<td>71.94</td>
</tr>
<tr>
<td>Dynamic</td>
<td>71.95</td>
<td>72.15</td>
<td>72.02</td>
<td>72.85</td>
</tr>
</tbody>
</table>
Benchmarking the State-of-the-art on WideResNet:

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

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

<table>
<thead>
<tr>
<th>Defense</th>
<th>Clean</th>
<th>FGSM</th>
<th>PGD-20</th>
<th>C&amp;W$_\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Madry’s</td>
<td>87.3</td>
<td>56.1</td>
<td>45.8</td>
<td>46.8</td>
</tr>
<tr>
<td>Curriculum</td>
<td>77.43</td>
<td>57.17</td>
<td>46.06</td>
<td>42.28</td>
</tr>
<tr>
<td>Dynamic</td>
<td>85.03</td>
<td>63.53</td>
<td>48.70</td>
<td>47.27</td>
</tr>
</tbody>
</table>
FOSC View of Dynamic Adversarial Training:

- **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