On Certifying Non-uniform Bounds against Adversarial Attacks

Chen Liu†, Ryota Tomioka‡, Volkan Cevher†

†École Polytechnique Fédérale de Lausanne
‡Microsoft Research Cambridge

June 11th, 2019
Problem (Certification Problem)

Given the label set $\mathcal{C}$, a classification model $f : \mathbb{R}^n \rightarrow \mathcal{C}$ and an input data point $x \in \mathbb{R}^n$, we would like to find the largest neighborhood $S$ around $x$ such that $f(x) = f(x') \ \forall x' \in S$.

- Set $S$ is called adversarial budget and $x \in S$. 
Motivation

\[ S^{(p)}_\epsilon(x) = \{ x' = x + \epsilon v : \| v \|_p \leq 1 \} \]

\[ \epsilon \in \mathbb{R} \]

Advantages of non-uniform bounds:

- Larger overall volumes.
- Quantitative metric of feature robustness.

Liu et al. (EPFL)
Non-uniform Bounds
June 11th, 2019
Motivation

\[
\mathcal{S}_\varepsilon^{(p)}(x) = \{ x' = x + \varepsilon v \parallel v \parallel_p \leq 1 \}
\]
\(\varepsilon \in \mathbb{R}\)

Advantages of non-uniform bounds:

- Larger overall volumes.
- Quantitative metric of feature robustness.

Liu et al. (EPFL)
Motivation

\[ S^{(p)}_\epsilon(x) = \{ x' = x + \epsilon v \| v \|_p \leq 1 \} \]
\[ \epsilon \in \mathbb{R} \]

\[ S^{(p)}_\epsilon(x) = \{ x' = x + \epsilon \odot v \| v \|_p \leq 1 \} \]
\[ \epsilon \in \mathbb{R}^n \]

Advantages of non-uniform bounds:

- Larger overall volumes.
- Quantitative metric of feature robustness.
A $N$-layer fully connected neural network, parameterized by $\{W^{(i)}, b^{(i)}\}_{i=1}^{N-1}$

\[ z^{(i+1)} = W^{(i)} \hat{z}^{(i)} + b^{(i)} \quad i = 1, 2, \ldots, N-1 \]

\[ \hat{z}^{(i)} = \sigma(z^{(i)}) \quad i = 2, 3, \ldots, N-1 \] (1)
A $N$-layer fully connected neural network, parameterized by \( \{W^{(i)}, b^{(i)}\}_{i=1}^{N-1} \)

\[
\begin{align*}
\mathbf{z}^{(i+1)} &= W^{(i)}\hat{\mathbf{z}}^{(i)} + \mathbf{b}^{(i)} \quad i = 1, 2, ..., N - 1 \\
\hat{\mathbf{z}}^{(i)} &= \sigma(\mathbf{z}^{(i)}) \quad i = 2, 3, ..., N - 1
\end{align*}
\]

(1)

Given a model \( \{W^{(i)}, b^{(i)}\} \) and a data point \( \mathbf{x} \) labeled as \( c \in \mathcal{C} \), we want to

\[
\min_{\mathbf{e}} \left\{ -\sum_{j=0}^{n_1-1} \log e_j \right\}
\]

\[
\hat{\mathbf{z}}^{(1)} \in \mathcal{S}_{\mathbf{e}}(\mathbf{x})
\]

\[
\mathbf{z}^{(i+1)} = W^{(i)}\hat{\mathbf{z}}^{(i)} + \mathbf{b}^{(i)} \quad i = 1, 2, ..., N - 1 \\
\hat{\mathbf{z}}^{(i)} = \sigma(\mathbf{z}^{(i)}) \quad i = 2, 3, ..., N - 1
\]

\[
z_c^{(N)} - z_j^{(N)} \geq \delta \quad j = 0, 1, ..., n_N - 1; j \neq c
\]

(2)


**Formulation**

- A $N$-layer fully connected neural network, parameterized by $\{W^{(i)}, b^{(i)}\}_{i=1}^{N-1}$

\[
\begin{align*}
    z^{(i+1)} &= W^{(i)} \hat{z}^{(i)} + b^{(i)} & i &= 1, 2, ..., N - 1 \\
    \hat{z}^{(i)} &= \sigma(z^{(i)}) & i &= 2, 3, ..., N - 1
\end{align*}
\] (1)

- Given a model $\{W^{(i)}, b^{(i)}\}$ and a data point $x$ labeled as $c \in C$, we want to

\[
\min_{\epsilon} \left\{ - \sum_{j=0}^{n_1-1} \log \epsilon_j \right\}
\]

\[
\hat{z}^{(1)} \in S_\epsilon(x)
\]

\[
z^{(i+1)} = W^{(i)} \hat{z}^{(i)} + b^{(i)} & i = 1, 2, ..., N - 1 \\
\hat{z}^{(i)} = \sigma(z^{(i)}) & i = 2, 3, ..., N - 1
\]

\[
z_c^{(N)} - z_j^{(N)} \geq \delta & j = 0, 1, ..., n_N - 1; j \neq c
\]

- Generally intractable (at least NP-complete)! [Weng et al. 18]
A $N$-layer fully connected neural network, parameterized by $\{W^{(i)}, b^{(i)}\}_{i=1}^{N-1}$

$$z^{(i+1)} = W^{(i)}z^{(i)} + b^{(i)} \quad i = 1, 2, ..., N - 1$$

$$z^{(i)} = \sigma(z^{(i)}) \quad i = 2, 3, ..., N - 1$$

Given a model $\{W^{(i)}, b^{(i)}\}$ and a data point $\mathbf{x}$ labeled as $c \in C$, we want to

$$\min_{\epsilon} \left\{ - \sum_{j=0}^{n_1-1} \log \epsilon_j \right\}$$

$$\hat{z}^{(1)} \in S_{\epsilon}(\mathbf{x})$$

$$z^{(i+1)} = W^{(i)}\hat{z}^{(i)} + b^{(i)} \quad i = 1, 2, ..., N - 1$$

$$\hat{z}^{(i)} = \sigma(z^{(i)}) \quad i = 2, 3, ..., N - 1$$

$$I^{(N)}_c - u^{(N)}_j \geq \delta \quad j = 0, 1, ..., n_N - 1; j \neq c$$

Generally intractable (at least NP-complete)! [Weng et al. 18]

Relax the output logits!
Optimization

- \( l^{(N)} \) and \( u^{(N)} \) are differentiable w.r.t. \( \epsilon \).
**Optimization**

- $I^{(N)}$ and $u^{(N)}$ are differentiable w.r.t. $\epsilon$.
- The relaxation problem is tractable

\[
\begin{aligned}
\min_{\epsilon, y \geq 0} & \left\{ - \sum_{j=0}^{n_1-1} \log \epsilon_j \right\} \\
\text{s.t.} & \quad l^{(N)}_c - u^{(N)}_{j \neq c} - \delta = y
\end{aligned}
\] (3)
Optimization

- \( I^{(N)} \) and \( u^{(N)} \) are differentiable w.r.t. \( \epsilon \).
- The relaxation problem is tractable

\[
\min_{\epsilon, y \geq 0} \left\{ - \sum_{j=0}^{n_1-1} \log \epsilon_j \right\}
\]

s.t. \( I_c^{(N)} - u_{j \neq c}^{(N)} - \delta = y \) (3)

- The problem can be solved by Augmented Lagrangian Method

\[
\max_{\lambda} \min_{\epsilon, y \geq 0} \left( \sum_{j=0}^{n_1-1} \log \epsilon_j \right) + \langle \lambda, v - y \rangle + \frac{\rho}{2} \| v - y \|_2^2
\]

(4)

- \( v \) is defined as \( I_c^{(N)} - u_{j \neq c}^{(N)} - \delta \)
## Experiments

### General Result

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Architecture</th>
<th>Training Method</th>
<th>Uniform</th>
<th>Non-uniform</th>
<th>Ratio</th>
</tr>
</thead>
<tbody>
<tr>
<td>MNIST</td>
<td>100-100-100</td>
<td>-</td>
<td>0.0295</td>
<td>0.0349</td>
<td>1.183</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PGD, $\tau = 0.1$</td>
<td>0.0692</td>
<td>0.1678</td>
<td>2.425</td>
</tr>
<tr>
<td></td>
<td>300-300-300</td>
<td>-</td>
<td>0.0309</td>
<td>0.0350</td>
<td>1.133</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PGD, $\tau = 0.1$</td>
<td>0.0507</td>
<td>0.1404</td>
<td>2.769</td>
</tr>
<tr>
<td></td>
<td>500-500-500</td>
<td>-</td>
<td>0.0319</td>
<td>0.0360</td>
<td>1.129</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PGD, $\tau = 0.1$</td>
<td>0.0436</td>
<td>0.1167</td>
<td>2.677</td>
</tr>
<tr>
<td>Fashion-MNIST</td>
<td>1024-1024-1024</td>
<td>-</td>
<td>0.0397</td>
<td>0.0518</td>
<td>1.305</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PGD, $\tau = 0.1$</td>
<td>0.0446</td>
<td>0.1134</td>
<td>2.543</td>
</tr>
<tr>
<td>SVHN</td>
<td>1024-1024-1024</td>
<td>-</td>
<td>0.0022</td>
<td>0.0072</td>
<td>3.273</td>
</tr>
<tr>
<td></td>
<td></td>
<td>PGD, $\tau = 0.1$</td>
<td>0.0054</td>
<td>0.0281</td>
<td>5.204</td>
</tr>
</tbody>
</table>

**Table:** Average of uniform and non-uniform bounds in the test sets.

- Larger volumes covered by non-uniform bounds, especially for robust models.
Figure: Examples of distributions of bounds for normal and robust models among all pixels. (Left: MNIST, Right: SVHN)

- Features of very large bounds → Features dropped
We can visualize bounding map $\epsilon \in \mathbb{R}^n$ like an input data point.

The bounding maps demonstrate better interpretability of robust models.

**Figure:** Left: between digit 1 and 7. Right: between digit 3 and 8. Lighter pixels mean smaller bounds.
Welcome to Poster #63

Code on GitHub:
Certify_Nonuniform_Bounds
спасибо  GRACIAS  谢谢
ありがとう ございました  MERCI
DANKE  धन्यवाद
شُكرًا  OBRIGADO