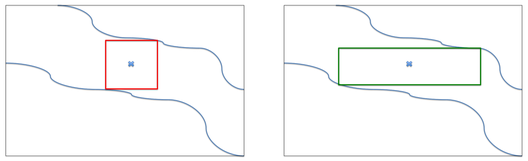
On Certifying Non-uniform Bounds against Adversarial Attacks

Chen Liu†, Ryota Tomioka‡, Volkan Cevher† † EPFL, ‡ Microsoft Research

Motivation

Compared with uniform bounds, the advantages of certified non-uniform bounds:

- Larger volumes.
- Quantitation of feature robustness.
- Tools to study the decision boundary.



Formulation

Given a fully-connected network parameterized by $\{\mathbf{W}^{(i)}, \mathbf{b}^{(i)}\}_{i=1}^{(N-1)}$, a data point $\mathbf{x}$ and adversarial budget $\mathcal{S}_{\boldsymbol{\epsilon}}(\mathbf{x}) = \{\mathbf{x}' = \mathbf{x} + \boldsymbol{\epsilon} \odot \mathbf{v} | \|\mathbf{v}\|_\infty \leq 1\}$, we want to maximize the volume of adversarial budget while guaranteeing the model to give consistent predictions.

$$
\begin{aligned}
\min_{\boldsymbol{\epsilon}} \quad & -\sum_{j=0}^{n_1-1} \log \epsilon_j \\
\hat{\mathbf{z}}^{(1)} & \in \mathcal{S}_{\boldsymbol{\epsilon}}(\mathbf{x}) \\
\mathbf{z}^{(i+1)} &= \mathbf{W}^{(i)} \hat{\mathbf{z}}^{(i)} + \mathbf{b}^{(i)} \qquad i = 1, 2, ..., N-1 \\
\hat{\mathbf{z}}^{(i)} &= \sigma(\mathbf{z}^{(i)}) \qquad\qquad i = 2, 3, ..., N-1 \\
z_c^{(N)} - z_j^{(N)} &\geq \delta \qquad\quad j = 0, 1, ..., n_N - 1; j \neq c
\end{aligned}
\tag{1}
$$

Solve (1) is at least NP-Complete. We make relaxations based on the bounds $\mathbf{l}^{(N)}$, $\mathbf{u}^{(N)}$ of $\mathbf{z}^{(N)}$.

Code on GitHub:

github.com/liuchen11/Certify_Nonuniform_Bounds

Method

Linearizing Activation Functions

For bounded input $\mathbf{l} \leq \mathbf{x} \leq \mathbf{u}$, we can use diagonal matrix $\mathbf{D}$ and vector $\mathbf{m}_1$, $\mathbf{m}_2$ to bound function σ: $\mathbf{D}\mathbf{x} + \mathbf{m}_1 \leq \sigma(\mathbf{x}) \leq \mathbf{D}\mathbf{x} + \mathbf{m}_2$.
Equivalently:

$$
\begin{aligned}
&\exists \mathbf{D}, \mathbf{m}_1, \mathbf{m}_2 : \forall \mathbf{x} \in [\mathbf{l}, \mathbf{u}], \\
&\qquad \exists \mathbf{m} \in [\mathbf{m}_1, \mathbf{m}_2] \; s.t. \; \sigma(\mathbf{x}) = \mathbf{D}\mathbf{x} + \mathbf{m}
\end{aligned}
\tag{2}
$$

So, any intermediate activation can be linearized as:

$$
\begin{aligned}
\mathbf{z}^{(i)} &= \mathbf{W}^{(i-1)}(\sigma(...(\mathbf{W}^{(1)}(\mathbf{x} + \mathbf{m}^{(1)}) + \mathbf{b}^{(1)})...)) + \mathbf{b}^{(i-1)} \\
&= \left(\Pi_{j=2}^{i-1} \mathbf{W}^{(j)} \mathbf{D}^{(j)}\right) \mathbf{W}^{(1)} \mathbf{x} + \sum_{h=1}^{i-1} \left(\Pi_{j=h+1}^{i-1} \mathbf{W}^{(j)} \mathbf{D}^{(j)}\right) \mathbf{b}^{(h)} \\
&\quad + \sum_{h=1}^{i-1} \left(\Pi_{j=h+1}^{i-1} \mathbf{W}^{(j)} \mathbf{D}^{(j)}\right) \mathbf{W}^{(h)} \mathbf{m}^{(h)}
\end{aligned}
\tag{3}
$$

The RHS is a linear function of $\mathbf{m}$, the only variable. The bound of $\mathbf{m}^{(i)}$, thus $\mathbf{z}^{(i)}$, can be calculated iteratively.

Augmented Lagrangian Method

The relaxed problem by bound estimation:

$$
\begin{aligned}
&\min_{\boldsymbol{\epsilon}, \mathbf{y} \geq 0} -\sum_{j=0}^{n_1-1} \log \epsilon_j \\
&s.t. \; l_c^{(N)} - \mathbf{u}_{j \neq c}^{(N)} - \delta = \mathbf{y}
\end{aligned}
\tag{4}
$$

By replacement $\mathbf{v} = l_c^{(N)} - \mathbf{u}_{j \neq c}^{(N)} - \delta$, the objective function to solve in Augmented Lagrangian method:

$$
\max_{\boldsymbol{\lambda}} \min_{\boldsymbol{\epsilon}, \mathbf{y} \geq 0} -\sum_{j=0}^{n_1-1} \log \epsilon_j + \langle \boldsymbol{\lambda}, \mathbf{v} - \mathbf{y} \rangle + \frac{\rho}{2} \|\mathbf{v} - \mathbf{y}\|_2^2 \tag{5}
$$

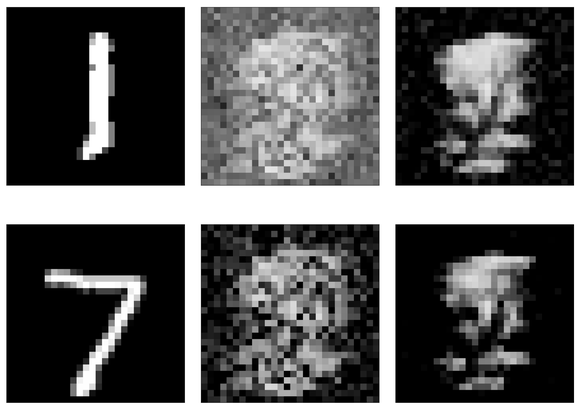
$\boldsymbol{\lambda}$ is an estimate of the Lagrange multiplier, so the penalty coefficient ρ does not need to go to ∞. For inner minimization, the optimal $\mathbf{y} = \max(0, \mathbf{v} + \frac{1}{\rho}\boldsymbol{\lambda})$ and the near-optimal $\boldsymbol{\epsilon}$ can be found by gradient method.

In practice, $\boldsymbol{\lambda}$ and $\boldsymbol{\epsilon}$, $\mathbf{y}$ are alternately updated. ρ is gradually increased and $\boldsymbol{\epsilon}$ is slightly shrinked to satisfy the hard constraint $\mathbf{v} \geq 0$.

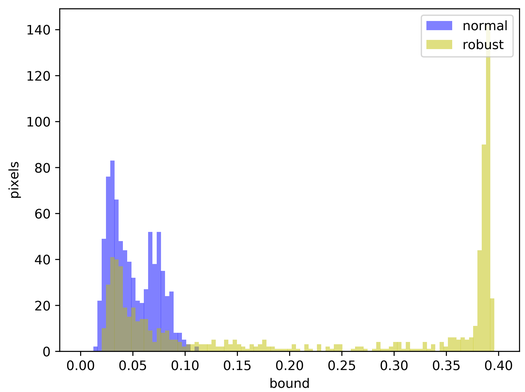
Experiments

Model[1]	Type[2]	Uniform	Non-uniform[3]	Ratio	Mean Cosine[4]	Min Cosine
MNIST-100	normal	0.0295	0.0349	1.183	0.9548	0.2304
	robust	0.0692	0.1678	2.425	0.9957	0.9155
MNIST-300	normal	0.0309	0.0350	1.133	0.9774	0.5038
	robust	0.0507	0.1404	2.769	0.9964	0.9104
MNIST-500	normal	0.0319	0.0360	1.129	0.9874	0.6367
	robust	0.0436	0.1167	2.677	0.9941	0.8920
FMNIST-1024	normal	0.0397	0.0518	1.305	0.9804	0.5257
	robust	0.0446	0.1134	2.543	0.9931	0.8891
SVHN-1024	normal	0.0022	0.0072	3.273	0.9836	0.7129
	robust	0.0054	0.0281	5.204	0.9952	0.9339

Notes: 1) All models have three hidden layers. They are named by the dataset and number of each hidden layer's neurons. 2) Robust models are ones from adversarial training by PGD and uniform budget of 0.1. 3) For non-uniform bounds, we use geometric average $\left(\prod_j^n \boldsymbol{\epsilon}_j\right)^{\frac{1}{n}}$. 4) The cosine value of $\boldsymbol{\epsilon}$ for all data pairs in the dataset.



Orignal image (left), bounding maps of normal model (middle) and robust model (right).



Histogram of bound per feature.

Robustness and Volume of Bounds

Non-uniform bounds indeed can certify a much larger region than uniform bounds. The difference is more significant when we certify robust models. When comparing histogram of bound per feature for different models, we can find robust models tend to drop irrelevant features and rely on fewer features when making predictions.

Robustness and Interpretability

When we plot vector $\boldsymbol{\epsilon}$ like an image, we can obtain a bounding map. Compare bounding maps between normal and robust models, we find non-uniform bounds of robust models are more interpretable.

Robustness and Geometric Similarity

We fine values of $\boldsymbol{\epsilon}$ for different data points but the same model are almost collinear, which indicates we can draw a quantitative and data-agnostic metric measuring the robustness of different features. The cosine similarity of $\boldsymbol{\epsilon}$ for robust models are consistently higher. This indicates higher geometric similarity of robust models' decision boundary around data manifold.