Supplementary Materials for Implicit Euler Skip Connections: Enhancing Adversarial Robustness via Numerical Stability

Mingjie Li Lingshen He Zhouchen Lin

1 Preliminaries and Notations

We use $(\mathbf{x}_0, \mathbf{y}_0)$ to denote a pair of input and label for training or testing. $\mathcal{F}(\mathbf{x})$ represents the output of the network. For a function $f : \mathbb{R}^d \to \mathbb{R}^d$, we use $\nabla f(\mathbf{x})$ to denote its Jacobian at input $\mathbf{x}$. We let $\mathbb{B}^{(n)}(\mathbf{x}, r)$ denote n-dimensional ball centered at $\mathbf{x}$ with radius r. We call a N-stage network if the output $\mathbf{x}_N$ and input $\mathbf{x}_0$ of the network corresponds to the following equation:

$$\mathbf{x}_i = s_i(\mathbf{x}_{i-1}), \textit{ for } i = 1, ..., N,$$

meanwhile we call $s_i(\cdot)$ is the i-th stages of the network.

2 Stability of the ODE-based neural networks

As we will illustrate in the following, we find that the numerical stability on the initial value problem is similar to the network's robustness against adversarial attacks which add perturbations to the input, especially when training the network with least squared regression loss. First of all, we define the numerical stability for an N-stage neural network from the dynamic system perspective as follows:

Definition 1. *A network with N stages (s_i represents its i-th stage) is called C-stable for its initial value problem at input $\mathbf{x}_0 \in \mathbb{R}^n$, if for a small δ and all the perturbed inputs for each stage $\mathbf{x}'_{i-1} \in \mathbb{B}^{(n)}(\mathbf{x}_{i-1}, \delta)$, the following equations are satisfied for all the stages:*

$$\|s_i(\mathbf{x}'_{i-1}) - s_i(\mathbf{x}_{i-1})\|_2 \le C\delta,\ i = 1, ..., N.$$

where $C \le 1$ is a constant.

From the above definition, one can see that if the network is C-stable at certain input $\mathbf{x}_0$, then the impacts of the small adversarial perturbation will not enlarge, or even shrink, during the forward propagation. Furthermore, we can bound the increment of loss for any attacks $\boldsymbol{\eta} \in \mathbb{B}^{(n)}(0, \delta)$ for sample $\mathbf{x}_0$ if the network is C-stable at $\mathbf{x}_0$ using the least squared regression loss.

Proposition 1. *If a network with N stages is C-stable at $\mathbf{x}_0$, then the increment of the least squared regression loss under the adversarial attack $\eta \in \mathbb{B}^n(0, \delta)$ on input $\mathbf{x}_0$ is:*

$$\mathcal{L}(\mathcal{F}(\boldsymbol{\theta}; \mathbf{x}_0 + \boldsymbol{\eta}), \mathbf{y}_0) - \mathcal{L}(\mathcal{F}(\boldsymbol{\theta}; \mathbf{x}_0), \mathbf{y}_0) \le C^N\delta,$$

where $\mathcal{F}(\cdot)$ denotes the neural network and $\mathbf{y}_0$ is the label for clean data $\mathbf{x}_0$.

Proof. Since the ResNet ($\mathcal{F}$) is C-stable, then the following statement is satisfied:

$$\|\mathcal{F}(\mathbf{x}') - \mathcal{F}(\mathbf{x})\|_2 \le C^N\delta,$$

Since

$$\mathcal{L}(\mathcal{F}(\boldsymbol{\theta}; \mathbf{x}), \mathbf{y}) = \|\mathcal{F}(\boldsymbol{\theta}; \mathbf{x}) - \mathbf{y}\|_2$$

then

$$\begin{aligned}\max_{\boldsymbol{\eta}\in\mathcal{D}} \mathcal{L}(\mathcal{F}(\boldsymbol{\theta};\mathbf{x}+\boldsymbol{\eta}),\mathbf{y}) - \mathcal{L}(\mathcal{F}(\boldsymbol{\theta};\mathbf{x},\mathbf{y})) &= \max_{\boldsymbol{\eta}\in\mathcal{D}} \|\mathcal{F}(\boldsymbol{\theta};\mathbf{x}+\boldsymbol{\eta})-\mathbf{y}\|_2 - \|\mathcal{F}(\boldsymbol{\theta};\mathbf{x})-\mathbf{y}\|_2 \\ &= \max_{\boldsymbol{\eta}\in\mathcal{D}}\{\|\mathcal{F}(\boldsymbol{\theta};\mathbf{x}+\boldsymbol{\eta})-\mathbf{y}\|_2 - \|\mathcal{F}(\boldsymbol{\theta};\mathbf{x})-\mathbf{y}\|_2\} \\ &\le \max_{\boldsymbol{\eta}\in\mathcal{D}} \|\mathcal{F}(\boldsymbol{\theta};\mathbf{x}+\boldsymbol{\eta})-\mathcal{F}(\boldsymbol{\theta};\mathbf{x})\|_2 \\ &\le C^N\delta\end{aligned}$$

□

Therefore, if the network is C-stable at sample $\mathbf{x}_0$, then it can perform more stable or even finally defend the adversarial attacks on such sample. Since all the testing data and training data belong to a same distribution, then the more possible the network is to be C-stable under such data distribution, the more robust the network is under adversarial attacks. On this account, we call the network can defend the adversarial attacks on $\mathbf{x}_0$ if network is C-stable with $C \le 1$ in the following analysis.

Furthermore, we analyze the sufficient conditions that ResNet and our model can defend the adversarial attacks.

Proposition 2. *For a N-block Residual Neural Network with f_i representing its i-th residual Block and a small $\delta > 0$, if the following statements satisfied:*

$$\|\mathbf{I}+\nabla f_i(\mathbf{x}_{i-1})^\top\|_2 \le 1 \text{ for } i = 1,...,N,$$

where $\mathbf{x}_{i-1}$ denotes the input of the i-th block corresponding to the clean input $\mathbf{x}_0$ for the network, then the network with N blocks can defend the attack with perturbation $\boldsymbol{\eta} \in \mathbb{B}^{(n)}(0,\delta)$ on certain sample $\mathbf{x}_0$.

Proof. The original output for the first residual stage can be formulated as:

$$\mathbf{x}_1 = f_1(\mathbf{x}_0) + \mathbf{x}_0, \tag{1}$$

then, with the perturbed input $\widetilde{\mathbf{x}}_0 = \boldsymbol{\eta} + \mathbf{x}_0$, the perturbed output $\widetilde{\mathbf{x}}_1$ can be formulated as:

$$\widetilde{\mathbf{x}}_1 = f_1(\widetilde{\mathbf{x}}_0) + \mathbf{x}_0 + \boldsymbol{\eta}, \tag{2}$$

Then since the perturbation is small, we do Taylor expansion for $f_1(\cdot)$ at $\mathbf{x}_0$ for Eqn. 2. Then subtract Eqn. 1:

$$\boldsymbol{\Delta}_1 = \widetilde{\mathbf{x}}_1 - \mathbf{x}_1 = (\mathbf{I}+\nabla f_1(\mathbf{x}_0)^\top)\boldsymbol{\eta},$$

Then since

$$\|\mathbf{I}+\nabla f_i(\mathbf{x}_{i-1})^\top\|_2 \le 1\textit{for } i = 1,...,N,$$

we can obtain that:

$$\begin{aligned}\|\boldsymbol{\Delta}_1\|_2 = \|\widetilde{\mathbf{x}}_1 - \mathbf{x}_1\|_2 &= \|(\mathbf{I}+\nabla f_1(\mathbf{x}_0)^\top)\boldsymbol{\eta}\|_2 \\ &\le \|\mathbf{I}+\nabla f_1(\mathbf{x}_0)^\top\|_2\|\boldsymbol{\eta}\|_2 \\ &= C_1\|\boldsymbol{\eta}\|_2 \le \|\boldsymbol{\eta}\|_2\end{aligned}$$

with $C_1 = \|\mathbf{I}+\nabla f_1(\mathbf{x}_0)^\top\|_2 \le 1$. Then we figure out that the perturbations for output is smaller than the input perturbations which means $\boldsymbol{\Delta}_1 \in \mathbb{B}^{(n)}(0,\delta)$, then redo the above procedure on next residual stage, we can find that the perturbation also shrinks with $C_2 = \|\mathbf{I}+\nabla f_2(\mathbf{x}_1)^\top\|_2 \le 1$. Then for a ResNet with N blocks, the output perturbation $\boldsymbol{\Delta}_N$ corresponding to every initial perturbation $\boldsymbol{\eta}$ obeys:

$$\|\boldsymbol{\Delta}_N\|_2 \le \prod_{i=1}^{N} C_i\|\boldsymbol{\eta}\|_2 \le (\max_i C_i)^N\delta \le C^N\delta,$$

where $C = \max_i C_i \le 1$. So ResNet can defend the attack on $\mathbf{x}_0$. □

3 Roubst Analysis for our IR ResNet

Proposition 3. *For an N-block exact IE-ResNet with f_i representing its i-th residual block and a small $\delta > 0$, if the following statement is satisfied:*

$$\sigma_{min}(\mathbf{I} - \nabla f_i(\mathbf{x}_i)^\top) \geq 1 \text{ for } i = 1, ..., N, \tag{3}$$

where σ_{min} denotes the smallest singular value and $\mathbf{x}_i$ denotes the output of the i-th block corresponding to the clean input $\mathbf{x}_0$ for the network, then the network with N blocks can defend attacks with perturbation $\boldsymbol{\eta} \in \mathbb{B}^{(n)}(0, \delta)$ on sample $\mathbf{x}_0$.

Proof. The original output for the first residual stage can be formulated as:

$$\mathbf{x}_1 = f_1(\mathbf{x}_1) + \mathbf{x}_0, \tag{4}$$

then, with the perturbed input $\widetilde{\mathbf{x}}_0 = \boldsymbol{\eta} + \mathbf{x}_0$, the perturbed output $\widetilde{\mathbf{x}}_1$ can be formulated as:

$$\widetilde{\mathbf{x}}_1 = f_1(\widetilde{\mathbf{x}}_1) + \mathbf{x}_0 + \boldsymbol{\eta}, \tag{5}$$

Then since the perturbation is small, we do Taylor expansion for $f_1(\cdot)$ at $\mathbf{x}_1$ for Eqn. 5. Then subtract Eqn. 4:

Then since

$$\sigma_{min}(\mathbf{I} - \nabla f_i(\mathbf{x}_i)^\top) \geq 1 \, for\ i = 1, ..., N,$$

and use $\boldsymbol{\Delta}_1$ to denote $\widetilde{\mathbf{x}}_1 - \mathbf{x}_1$, we can obtain that:

$$\begin{aligned}
(\mathbf{I} - \nabla f_1(\mathbf{x}_1)^\top)\boldsymbol{\Delta}_1 &= \boldsymbol{\eta} \\
\sigma_{min}(\mathbf{I} - \nabla f_1(\mathbf{x}_1)^\top)\|\boldsymbol{\Delta}_1\|_2 &\leq \|\boldsymbol{\eta}\|_2 \\
\|\boldsymbol{\Delta}_1\|_2 &\leq \frac{1}{\sigma_{min}(\mathbf{I} - \nabla f_1^\top(\mathbf{x}_1))}\|\boldsymbol{\eta}\|_2 \\
&= C_1\|\boldsymbol{\eta}\|_2 \leq \|\boldsymbol{\eta}\|_2
\end{aligned}$$

with $C_1 = \frac{1}{\sigma_{min}(\mathbf{I} - \nabla f_1(\mathbf{x}_1)^\top)} \leq 1$. Then we figure out that the perturbations for output is smaller than the input perturbations which means $\boldsymbol{\Delta}_1 \in \mathbb{B}^{(n)}(0, \delta)$, then redo the above procedure on next residual stage, we can find that the perturbation also shrinks with $C_2 = \frac{1}{\sigma_{min}(\mathbf{I} - \nabla f_2(\mathbf{x}_2)^\top)} \leq 1$. hen for a ResNet with N blocks, the output perturbation $\boldsymbol{\Delta}_N$ corresponding to every initial perturbation $\boldsymbol{\eta}$ obeys:

$$\|\boldsymbol{\Delta}_N\|_2 \leq \prod_{i=1}^{N} C_i\|\boldsymbol{\eta}\|_2 \leq (\max_i C_i)^N \delta \leq C^N \delta,$$

where $C = \max_i C_i \leq 1$. So, IE-ResNet can defend the attack on $\mathbf{x}_0$. □

From the propositions above, one can see that our IE-ResNet is much easier to satisfy the Jacobian's condition for the stability. Furthermore, we have proved that our IE-ResNet has higher probability to defend the attack under our definitions above than its corresponding ResNet. For an N-block ResNet with g_i representing its i-th residual block and a N-block exact IE-ResNet with f_i denoting its i-th residual block. Furthermore, we use $\mathbf{x}_i$ to denote the input of i-th block for ResNet while we use $\mathbf{y}_i$ to represent the output of i-th block for IE-ResNet.

Theorem 1. *Suppose that for a input $\mathbf{x}$, which is sampled from a data distribution, its corresponding $\nabla g_i(\mathbf{x}_i)$ and $\nabla f_i(\mathbf{y}_i)$ obey the same distribution since they enjoy the same strategies and Jacobians $\{\nabla g_i(\mathbf{x}_i), \nabla f_i(\mathbf{y}_i)\}$ are independent. Then, we can obtain the following relations:*

$$\begin{aligned}
&\mathbb{P}[\cap_{i=1,..N}\{\|\mathbf{I} + \nabla g_i(\mathbf{x}_i)^\top\|_2 \leq 1\}] \leq \\
&\qquad \mathbb{P}[\cap_{i=1,..N}\{\sigma_{min}(\mathbf{I} - \nabla f_i(\mathbf{y}_i)^\top) \geq 1\}].
\end{aligned}$$

From the above theorem, one can see that the possibility for our IE-ResNet to maintain stable on a sample is higher than the vanilla ResNet. On this account, the robustness of our IE-ResNet is superior to the vanilla ResNet.

Proof. First of all, we define the following sets:

$$\begin{aligned}\Omega_{1i} =& \{\nabla g_i(\mathbf{x}_i) | \|\nabla g_i(\mathbf{x}_i)^\top + \mathbf{I}\|_2 \leq 1\},\\ \Omega_{2i} =& \{\nabla f_i(\mathbf{y}_i) | \|\nabla f_i(\mathbf{y}_i)^\top + \mathbf{I}\|_2 \leq 1\},\\ \Omega_{3i} =& \{\nabla f_i(\mathbf{y}_i) | \sigma_{min}(\mathbf{I} - \nabla f_i(\mathbf{y}_i)^\top) \geq 1\}.\end{aligned}$$

Since $\nabla f_i(\mathbf{x}_i)$ and $\nabla g_i(\mathbf{y}_i)$ corresponding to a same input $\mathbf{x}_0$ obeys the same distribution, we can obtain the following equation:

$$\mathbb{P}[\cap_{i=1,..N}\{\|\mathbf{I} + \nabla f_i(\mathbf{x}_i)^\top\|_2 \leq 1\}] = \mathbb{P}[\cap_{i=1,..N}\{\|\nabla g_i(\mathbf{y}_i)^\top + \mathbf{I}\|_2 \leq 1\}],$$

Secondly, we prove that $\Omega_{2i} \subset \Omega_{3i}$. Before that, we first prove that $(\mathbf{I} - \nabla f_i^\top)$ is invertible if $\nabla f_i \in \Omega_{2i}$. If $\mathbf{I} - \nabla f_i$ is not invertible, which means:

$$\exists \alpha \neq \mathbf{0},\ s.t.\ (\mathbf{I} - \nabla f_i^\top)\alpha = 0$$

Then,

$$\begin{aligned}(\mathbf{I} + \nabla f_i^\top)\alpha &= 2\alpha\\ \|(\mathbf{I} + \nabla f_i^\top)\alpha\|_2 &= 2\alpha \leq \|(\mathbf{I} + \nabla f_i^\top)\|_2 \|\alpha\|_2\end{aligned}$$

so that:

$$\|(\mathbf{I} + \nabla f_i^\top)\|_2 \geq 2.$$

which is contradicts to the above facts that $\nabla f_i \in \Omega_{2i}$. So $(\mathbf{I} - \nabla f_i^\top|_{\mathbf{x}})$ is invertible if $\nabla f_i \in \Omega_{2i}$.

Next, we can get that if $\nabla f_i \in \Omega_{2i}$, we can obtain that:

$$\begin{aligned}-(2\mathbf{I})^{-1}(\mathbf{I} + \nabla f_i^\top)(\mathbf{I} - \nabla f_i^\top)^{-1} &= -\frac{1}{2}(\mathbf{I} + \nabla f_i^\top)(\mathbf{I} - \nabla f_i^\top)\\ &= [-\mathbf{I} + \frac{1}{2}(\mathbf{I} - \nabla f_i^\top)](\mathbf{I} - \nabla f_i^\top)^{-1}\\ &= \frac{1}{2}\mathbf{I} - (\mathbf{I} - \nabla f_i^\top)^{-1}\\ (\mathbf{I} - \nabla f_i^\top)^{-1} &= \frac{1}{2}\mathbf{I} + (2\mathbf{I})^{-1}(\mathbf{I} + \nabla f_i^\top)(\mathbf{I} - \nabla f_i^\top)^{-1}\end{aligned}$$

So that

$$\begin{aligned}\|(\mathbf{I} - \nabla f_i^\top)^{-1}\|_2 &= \|\frac{1}{2}\mathbf{I} + (2\mathbf{I})^{-1}(\mathbf{I} + \nabla f_i^\top)(\mathbf{I} - \nabla f_i^\top)^{-1}\|_2\\ &\leq \frac{1}{2} + \frac{1}{2}\|(\mathbf{I} + \nabla f_i^\top)\|_2 \|(\mathbf{I} - \nabla f_i^\top)^{-1}\|_2\\ (1 - \frac{1}{2}\|\mathbf{I} + \nabla f_i^\top\|_2)\|(\mathbf{I} - \nabla f_i^\top)^{-1}\|_2 &\leq \frac{1}{2}\end{aligned}$$

Since $\|\mathbf{I} + \nabla f_i^\top\|_2 \leq 1$, we can acquire that:

$$\|(\mathbf{I} - \nabla f_i^\top)^{-1}\|_2 \leq \frac{1}{2 - \|\mathbf{I} + \nabla f_i^\top\|_2} \leq 1.$$

Since

$$\|(\mathbf{I} - \nabla f_i^\top)^{-1}\|_2 = \frac{1}{\sigma_{min}(\mathbf{I} - \nabla f_i^\top)}.$$

Then the following equation is hold:

$$\sigma_{min}(\mathbf{I} - \nabla f_i^\top) \geq 1.$$

Therefore, we proved that if $\nabla f_i \in \Omega_{2i}$, then it also belongs to Ω_{3i}. Furthermore, one can see that $-3\mathbf{I}$ only belongs to Ω_{3i}. So we can conduct that:

$$\Omega_{2i} \subset \Omega_{3i}$$

So we can conduct,

$$\begin{aligned}\mathbb{P}[\cap_{i=1,..N}\{\|\mathbf{I} + \nabla f_i(\mathbf{x}_i)^\top\|_2 \leq 1\}] &\leq \mathbb{P}[\cap_{i=1,..N}\{\sigma_{min}(\mathbf{I} - \nabla f_i(\mathbf{y}_i)^\top) \geq 1\}]\\ \mathbb{P}[\cap_{i=1,..N}\{\|\mathbf{I} + \nabla g_i(\mathbf{x}_i)^\top\|_2 \leq 1\}] &\leq \mathbb{P}[\cap_{i=1,..N}\{\sigma_{min}(\mathbf{I} - \nabla f_i(\mathbf{y}_i)^\top) \geq 1\}]\end{aligned}$$

□

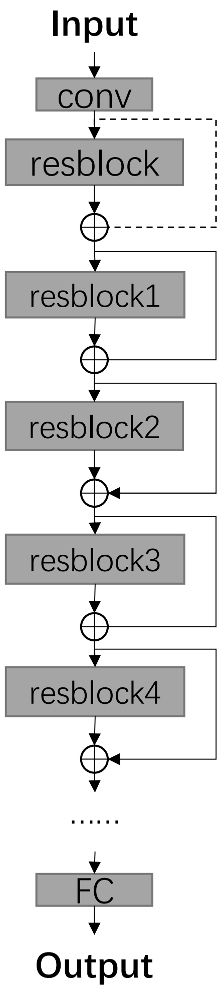

Figure 1: The structure sketch for the swResNet. The wights for resblock1 and resblock2 are the same. Meanwhile, the weights for resblock3 and resblock4 are the same.

4 swResNet Skectch

In Section 5.2, we design a share weight ResNet (swResNet) to compare with our model. The sketch of the swResNet can be depict as follow: As one can see from the above figure, we share weights for two adjacent residual block with the same dimension. We let the blocks unchanged as the original ResNet if there are dimension changes in such blocks. Therefore, the parameter size for swResNet-58 is the same as ResNet-34 and IE-ResNet-34.

5 Hyper-Parameters for TRADES training

We set perturbation $\delta = 0.031$, perturbation step size $\alpha = 0.007$, number of iterations K = 10 and $1/\lambda = 6$ for the TRADES adversarial training, which is consistent to the settings in Zhang *et al.* [2019].

References

Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric P Xing, Laurent El Ghaoui, and Michael I Jordan. Theoretically principled trade-off between robustness and accuracy. *arXiv preprint arXiv:1901.08573*, 2019.