Nonlocal Neural Networks, Nonlocal Diffusion and Nonlocal Modeling

Yunzhe Tao[1], Qi Sun[2], Qiang Du[1], and Wei Liu[3]

[1]: Columbia University; [2]: CSRC; [3]: Tencent AI Lab

Introduction

We study the nature of diffusion and damping effect of nonlocal networks by doing spectrum analysis on the weight matrices of the well-trained networks, and then propose a new formulation of the nonlocal block from nonlocal modeling perspective, which is more robust.

Damping Effect of Nonlocal Net

Nonlocal block in [Wang et al., 2018]:

$$Z_i^{k+1} := Z_i^k + \frac{W_Z^k}{\mathcal{C}_i(Z^k)} \sum_{\forall j} \omega(Z_i^k, Z_j^k) g(Z_j^k)$$

weight matrix; affinity function; output; input; normalization factor; $g(Z_j^k) := W_g Z_j^k$

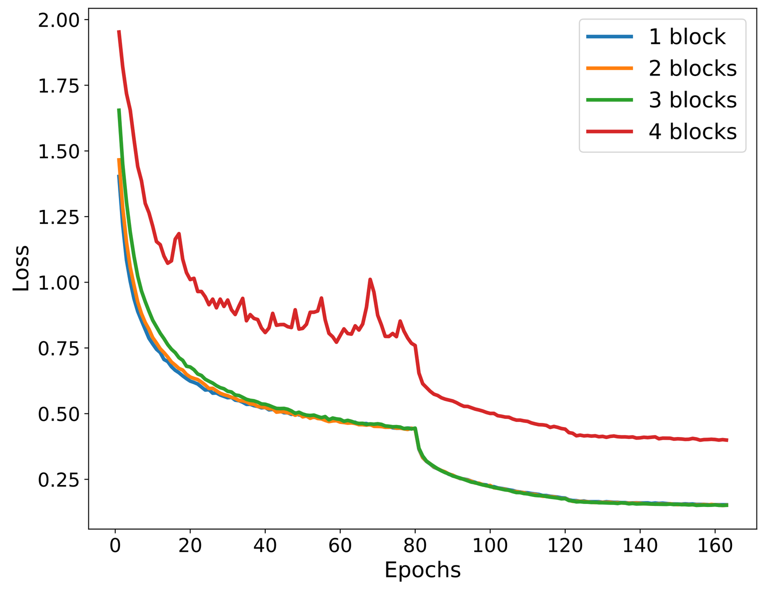
Incorporating nonlocal blocks into ResNet:

$$Z^{k+1} = Z^k + \mathcal{F}(Z^k; W^k)$$

nonlocal block or residual block

Training ResNet-20 with different number of nonlocal blocks on CIFAR-10:

- Works well when adding 1-3 nonlocal blocks
- Training becomes difficult when adding 4 blocks
- The original nonlocal network is not robust to multiple nonlocal blocks

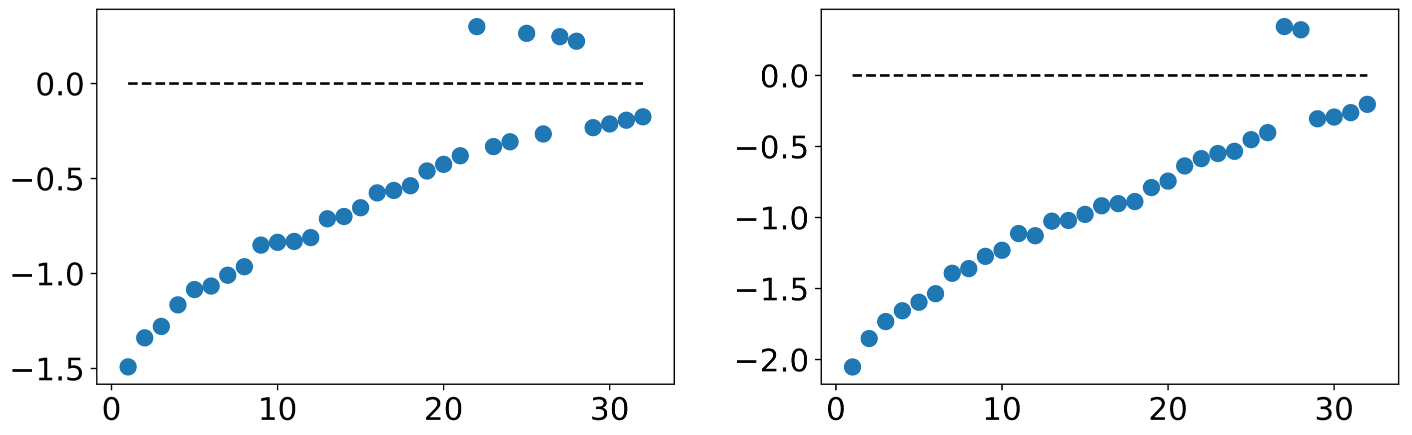


Doing **spectrum analysis** on *symmetrized* W^k when k-th block is a nonlocal block:

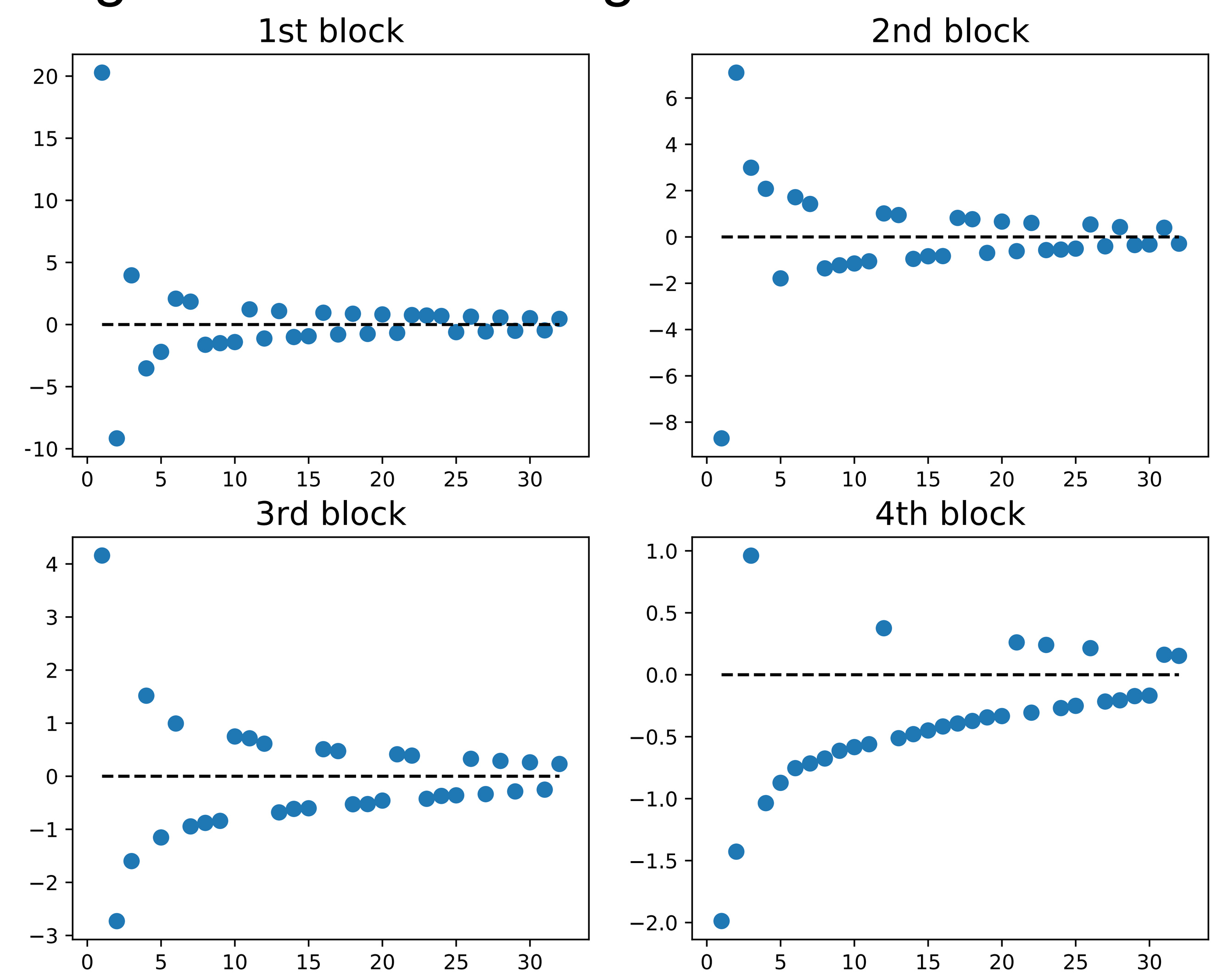
- Computing the eigenvalues of

$$\widetilde{W^k} := \frac{W_Z^k W_g + (W_Z^k W_g)'}{2}$$

- The eigenvalues are all real numbers
- The eigenvalues with greatest magnitudes describe the characteristics of the weights in nonlocal blocks
- Eigenvalues of adding 2 nonlocal blocks



- Eigenvalues of adding 4 nonlocal blocks



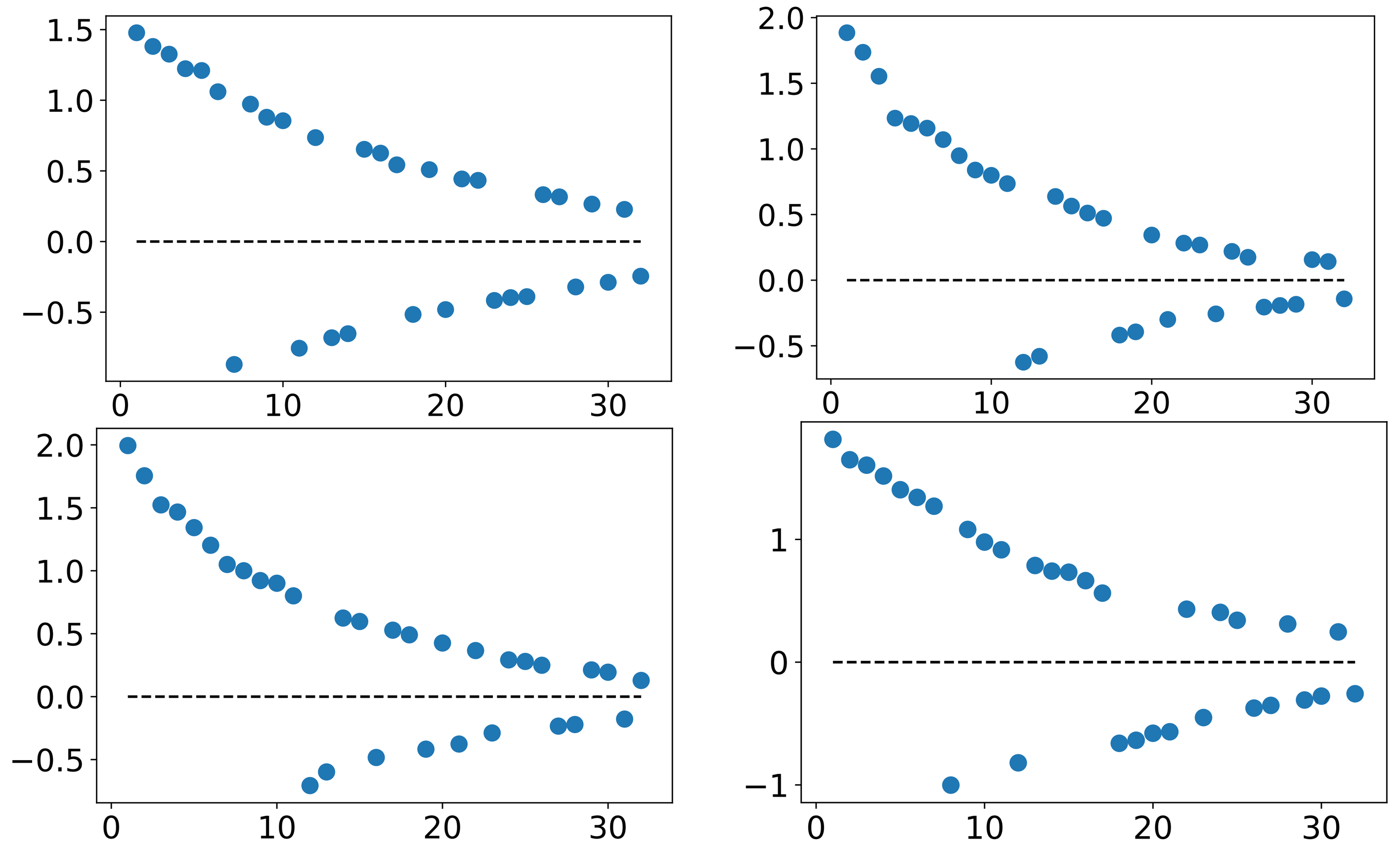
New Formulation of Nonlocal Net

New nonlocal ***stage*** (one or more blocks):

$$Z_i^{n+1} := Z_i^n + \frac{W^n}{\mathcal{C}_i(X)} \sum_{\forall j} \omega(X_i, X_j)(Z_j^n - Z_i^n)$$

stage input

- Eigenvalues of adding 4 nonlocal blocks



- Validation errors of different models (baseline: ResNet-20)

	Model	Error (%)
	baseline	8.19
The Same Place	2-block (original)	7.83
	3-block (original)	8.28
	4-block (original)	15.02
	2-block (proposed)	7.74
	3-block (proposed)	7.62
	4-block (proposed)	7.37
	5-block (proposed)	**7.29**
	6-block (proposed)	7.55
Different Places	3-block (original)	8.07
	3-block (proposed)	**7.33**

Connection to Nonlocal Modeling

- Nonlocal diffusion process

The discrete nonlocal diffusion:

$$Z^{n+1} = Z^n + \sum_{\forall j} \omega(X_i, X_j)(Z_j^n - Z_i^n)$$

Continuum form:

$$\begin{cases} z_t(x,t) - \mathcal{L}z(x) = 0 \\ z(x,0) = u(x) \end{cases}$$

where

$$\mathcal{L}z(x) := \int \rho(x,y)(z(y) - z(x))dy$$

is a Hilbert-Schmidt operator

⇒ Stable in finite time for both time directions (for both positive and negative weights)

- Markov jump process

Time-homogeneous Markov jump process:

$$z(x,t) = \int p(y,x) z(y,t) dy$$

Discrete-time Markov jump process:

$$Z_i^{n+1} - Z_i^n = \sum_{\forall j} \mathcal{K}_{ij} (Z_j^n - Z_i^n)$$

$\mathcal{K}$ is a Markov matrix, thus is Hilbert-Schmidt

⇒ Stable in finite time for both forward and backward jump process