Existence, Stability and Scalability of Orthogoonal Convolutional Neural Networks

El Mehdi Achour[1] François Malgouyres[1] Franck Mamalet[2]

[1]Institut de Mathématiques de Toulouse ; UMR 5219
Université de Toulouse ; CNRS
UPS IMT F-31062 Toulouse Cedex 9, France

[2]Institut de Recherche Technologique Saint Exupéry, Toulouse, France

GdR ISIS: Théorie du deep learning, June 2021

Motivation

Cons of deep CNNs :

- hard to train : exploding/vanishing gradient
- subject to : adversarial attacks
- suffers from : feature correlation
- can overfit

Orthogonality helps with all of these problems.
Applied to

- fully-connected networks [1]
- RNNs [2]

---

[1] Anil, Lucas, and Grosse. “Sorting out lipschitz function approximation”. ICML 2019

[2] Arjovsky, Shah, and Bengio. “Unitary evolution recurrent neural networks”. ICML 2016

Notation

2D Convolutional layer (M, C, k, S), where

- M : number of **output** channels
- C : number of **input** channels
- $k \times k$: size of the filters
- S : stride parameter

Circular boundary conditions:

- $SN \times SN$: size of input channels
- $N \times N$: size of output channels
- $\mathbf{K} \in \mathbb{R}^{M \times C \times k \times k}$: kernel tensor
- $\mathcal{K} \in \mathbb{R}^{MN^2 \times CS^2N^2}$: convolutional layer matrix

$$Y = \text{conv}(\mathbf{K}, X) \iff \text{Vect}(Y) = \mathcal{K}\,\text{Vect}(X)$$

Kernel Orthogonality

Input $X \in \mathbb{R}^{C \times SN \times SN}$, Output $Y = \mathbf{conv}(\mathbf{K}, X) \in \mathbb{R}^{M \times N \times N}$.

- Kernel orthogonality: $\overline{\mathbf{K}} = \text{Reshape}(\mathbf{K}) \in \mathbb{R}^{M \times Ck^2}$ s.t. $Y = \text{Reshape}(\overline{\mathbf{K}} U(X))$ where $U(X) \in \mathbb{R}^{Ck^2 \times N^2}$. Constrains the orthogonality of $\overline{\mathbf{K}}$.
  - Hard [5]: Parameterization, Optim on Stiefel manifold
  - Soft[3] [4]: $L_{task} + \lambda \| \overline{\mathbf{K}}\, \overline{\mathbf{K}}^T - Id \|^2$

---

[5] Li, Li, and Todorovic. “Efficient Riemannian Optimization on the Stiefel Manifold via the Cayley Transform”. ICLR 2019

[3] Bansal, Chen, and Wang. “Can We Gain More from Orthogonality Regularizations in Training Deep Networks?” NeurIPS 2018

[4] Cisse et al. “Parseval networks: improving robustness to adversarial examples”. ICML 2017

Layer Orthogonality

Input $X \in \mathbb{R}^{C \times SN \times SN}$, Output $Y = \mathbf{conv}(\mathbf{K}, X) \in \mathbb{R}^{M \times N \times N}$.

- Layer orthogonality: $\mathcal{K}$ s.t. $\mathrm{Vect}(Y) = \mathcal{K}\,\mathrm{Vect}(X)$. Constrains the orthogonality of $\mathcal{K}$.
  - Hard [7]: Parameterization
  - Soft[8] [6]: $L_{task} + \lambda \|\mathcal{K}\mathcal{K}^T - Id\|^2$
    $L_{task} + \lambda L_{orth}$

,

[7] Trockman and Kolter. “Orthogonalizing Convolutional Layers with the Cayley Transform”. ICLR 2021

[8] Wang et al. “Orthogonal convolutional neural networks”. CVPR 2020

[6] Qi et al. “Deep isometric learning for visual recognition”. ICLR 2020

Layer Orthogonality

Input $X \in \mathbb{R}^{C \times SN \times SN}$, Output $Y = \mathbf{conv}(\mathbf{K}, X) \in \mathbb{R}^{M \times N \times N}$.

- Layer orthogonality: $\mathcal{K}$ s.t. $\mathrm{Vect}(Y) = \mathcal{K}\,\mathrm{Vect}(X)$. Constrains the orthogonality of $\mathcal{K}$.
  - Hard [7]: Parameterization
  - Soft [8] [6]: $L_{task} + \lambda \|\mathcal{K}\mathcal{K}^T - Id\|^2$
    $L_{task} + \lambda L_{orth}$

,

---

[7] Trockman and Kolter. “Orthogonalizing Convolutional Layers with the Cayley Transform”. ICLR 2021

[8] Wang et al. “Orthogonal convolutional neural networks”. CVPR 2020

[6] Qi et al. “Deep isometric learning for visual recognition”. ICLR 2020

$$\mathcal{K} \in \mathbb{R}^{MN^2 \times CS^2 N^2}$$

- **Row Orthogonality (RO):** $M \leq CS^2$, $\mathcal{K} = \begin{pmatrix} \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix}$

$$\mathcal{K} \text{ is orthogonal } \iff \mathcal{K}\mathcal{K}^T = \mathrm{Id}_{MN^2} .$$

$$\mathcal{K} \in \mathbb{R}^{MN^2 \times CS^2N^2}$$

- **Row Orthogonality (RO):** $M \leq CS^2$, $\mathcal{K} = \begin{pmatrix} \cdots & \cdots & \cdots & \cdots \\ \cdots & \cdots & \cdots & \cdots \end{pmatrix}$

$$\mathcal{K} \text{ is orthogonal} \iff \mathcal{K}\mathcal{K}^T = \mathrm{Id}_{MN^2} .$$

- **Column Orthogonality (CO):** $M \geq CS^2$, $\mathcal{K} = \begin{pmatrix} \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \\ \vdots & \vdots \end{pmatrix}$

$$\mathcal{K} \text{ is orthogonal} \iff \mathcal{K}^T\mathcal{K} = \mathrm{Id}_{CS^2N^2} .$$

Regularization penalty

- $P = \lfloor \frac{k-1}{S} \rfloor S$
- $\mathbf{conv}(\mathbf{K}, \mathbf{K}, \text{padding zero} = P, \text{stride} = S) \in \mathbb{R}^{M \times M \times (2P/S+1) \times (2P/S+1)}$
  - easy to implement
  - independent of N
- $I_{r0} \in \mathbb{R}^{M \times M \times (2P/S+1) \times (2P/S+1)}$ s.t. $[I_{r0}]_{a,b,c,d} = 0$, except for $[I_{r0}]_{:,:,P/S,P/S} = Id_M$
- $L_{orth} : \mathbb{R}^{M \times C \times k \times k} \to \mathbb{R}^+$
  - RO case : $\| \mathbf{conv}(\mathbf{K}, \mathbf{K}, \text{padding zero} = P, \text{stride} = S) - I_{r0} \|_F^2$
  - CO case : $\| \mathbf{conv}(\mathbf{K}, \mathbf{K}, \text{padding zero} = P, \text{stride} = S) - I_{r0} \|_F^2 - (M - CS^2)$

Regularization penalty

- L_{orth} is independent of N.
- $L_{orth}(\mathbf{K}) = 0 \quad \Longleftrightarrow \quad \mathcal{K}$ orthogonal
- Impressive experimental results on various tasks using $L_{task} + \lambda L_{orth}$ [8] [6]

Questions:

- **Existence:** Condition on (M, C, k, S) and N s.t. $\exists \mathcal{K}$ orthogonal
- **Stability:** Behaviour of $\mathcal{K}$ when $L_{orth}(\mathbf{K}) = \varepsilon > 0$
- **Scalability:** Behaviour of $\mathcal{K}$ when N grows
- **Landscape:** Can we always find L_{orth}'s global minimizer ?

[8] Wang et al. "Orthogonal convolutional neural networks". CVPR 2020
[6] Qi et al. "Deep isometric learning for visual recognition". ICLR 2020

Existence of orthogonal convolutional layers

Theorem

Let N be such that $SN \geq k$.

- RO case, i.e. $M \leq CS^2$: $\exists\, \mathcal{K}$ orthogonal if and only if $M \leq Ck^2$.
- CO case, i.e. $M \geq CS^2$: $\exists\, \mathcal{K}$ orthogonal if and only if $S \leq k$.

Rk: In the RO case: $S \leq k \implies M \leq CS^2 \leq Ck^2 \implies \exists \mathcal{K}$ orthogonal

Frobenius norm scalability

$$\mathrm{err}_N^F(\mathbf{K}) = \begin{cases} \|\mathcal{K}\mathcal{K}^T - \mathrm{Id}_{MN^2}\|_F & \text{, in the RO case,} \\ \|\mathcal{K}^T\mathcal{K} - \mathrm{Id}_{CS^2N^2}\|_F & \text{, in the CO case,} \end{cases}$$

Theorem

Let N be such that $SN \geq 2k-1$. We have, both in the RO and CO case,

$$(\mathrm{err}_N^F(\mathbf{K}))^2 = N^2 L_{orth}(\mathbf{K})$$

$$L_{task} + \lambda \sum_l L_{orth}(\mathbf{K}_l)$$

becomes

$$L_{task} + \sum_l \frac{\lambda}{N_l^2} (\mathrm{err}_{N_l}^F(\mathbf{K}_l))^2.$$

Spectral norm scalability

$$\text{err}_N^s(\mathbf{K}) = \begin{cases} \|\mathcal{K}\mathcal{K}^T - \text{Id}_{MN^2}\|_2 & \text{, in the RO case,} \\ \|\mathcal{K}^T\mathcal{K} - \text{Id}_{CS^2N^2}\|_2 & \text{, in the CO case.} \end{cases}$$

Theorem

Let N be such that $SN \geq 2k-1$. We have,

$$(\text{err}_N^s(\mathbf{K}))^2 \leq \alpha \, L_{orth}(\mathbf{K})$$

with:

$$\alpha = \begin{cases} \left(2\left\lfloor \frac{k-1}{S} \right\rfloor + 1\right)^2 M & \text{in the RO case } (M \leq CS^2), \\ (2k-1)^2 C & \text{in the CO case } (M \geq CS^2). \end{cases}$$

α is independent of N.

Approximate Isometry Property

Definition

A layer transform matrix $\mathcal{K} \in \mathbb{R}^{MN^2 \times CS^2N^2}$ satisfies the ε-Approximate Isometry Property if and only if

- RO case, $M \leq CS^2$:

$$\begin{cases} \forall x \in \mathbb{R}^{CS^2N^2} & \|\mathcal{K}x\|^2 \leq (1+\varepsilon)\|x\|^2 \\ \forall y \in \mathbb{R}^{MN^2} & (1-\varepsilon)\|y\|^2 \leq \|\mathcal{K}^T y\|^2 \leq (1+\varepsilon)\|y\|^2 \end{cases}$$

- CO case, $M \geq CS^2$:

$$\begin{cases} \forall x \in \mathbb{R}^{CS^2N^2} & (1-\varepsilon)\|x\|^2 \leq \|\mathcal{K}x\|^2 \leq (1+\varepsilon)\|x\|^2 \\ \forall y \in \mathbb{R}^{MN^2} & \|\mathcal{K}^T y\|^2 \leq (1+\varepsilon)\|y\|^2 \end{cases}$$

Proposition

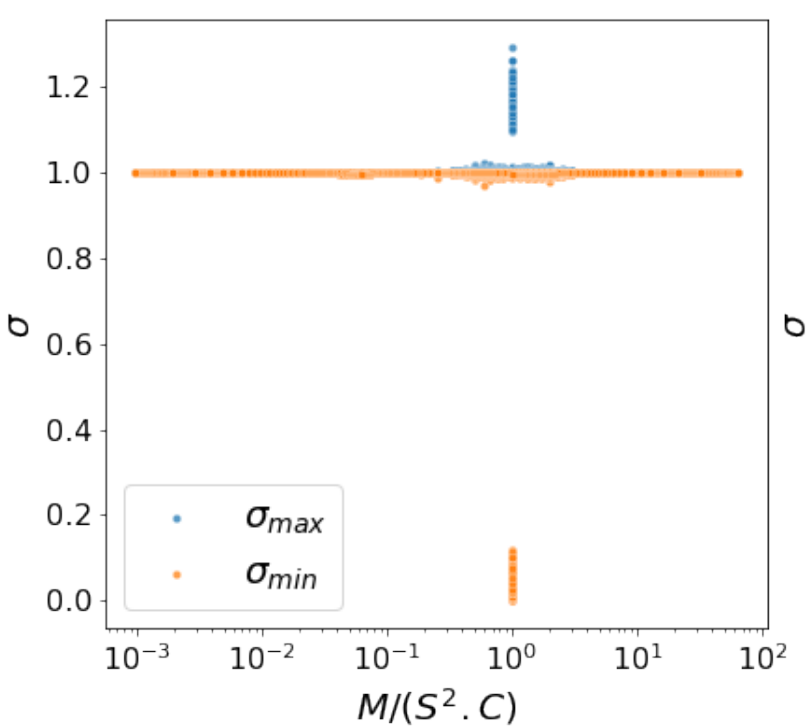
Let N be such that $SN \geq k$. We have $\mathcal{K}$ is $\mathrm{err}^s_N(\mathbf{K})$-AIP.

Benefits of AIP

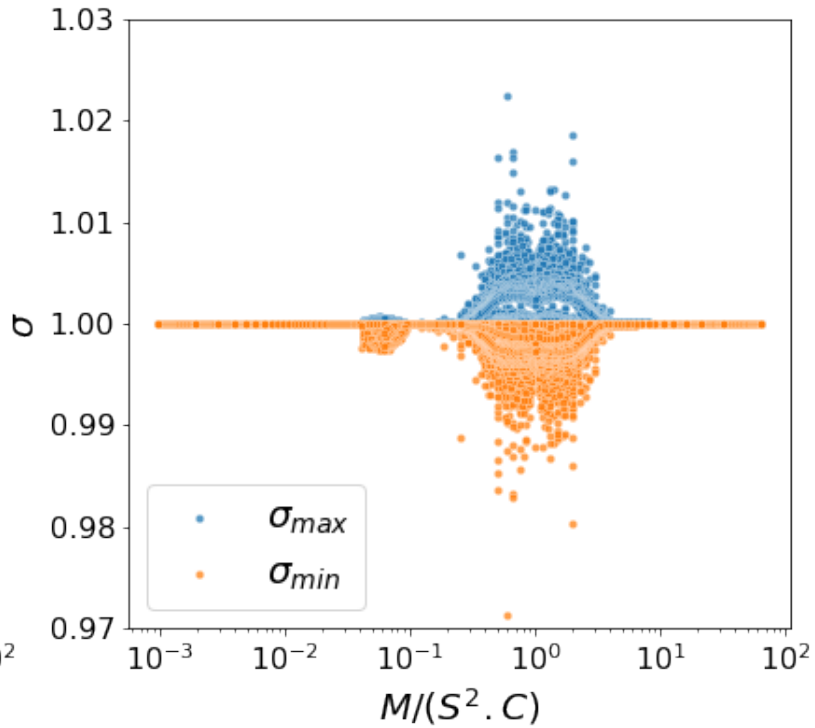
		Forward pass		Backward pass	
		Lipschitz Forward pass	Keep examples separated	Prevent gradient exploding	Prevent gradient vanishing
Conv layer	$M < CS^2$	✓	✗	✓	✓
	$M > CS^2$	✓	✓	✓	✗
Deconv layer	$M < CS^2$	✓	✓	✓	✗
	$M > CS^2$	✓	✗	✓	✓
Both	$M = CS^2$	✓	✓	✓	✓

Landscape of L_{orth}

Optimization of L_{orth}



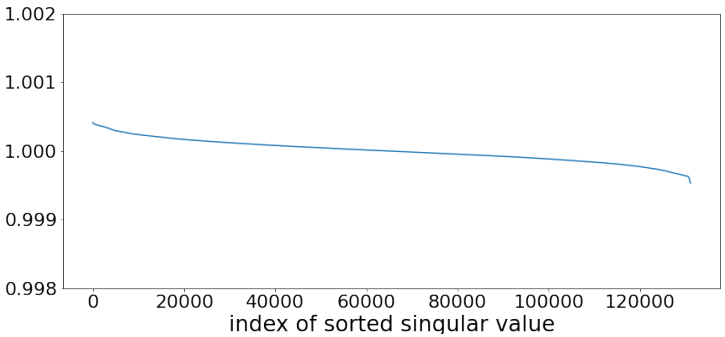
Scenarios where $\exists \mathcal{K}$ orthogonal



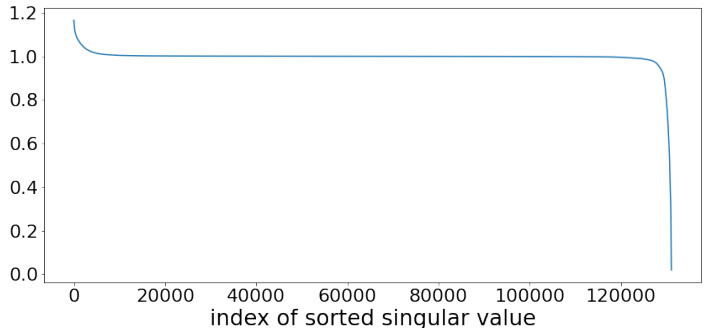
Zoom on $M \neq CS^2$

Analysis of the $M = CS^2$ cases

Singular values of $\mathcal{K}$

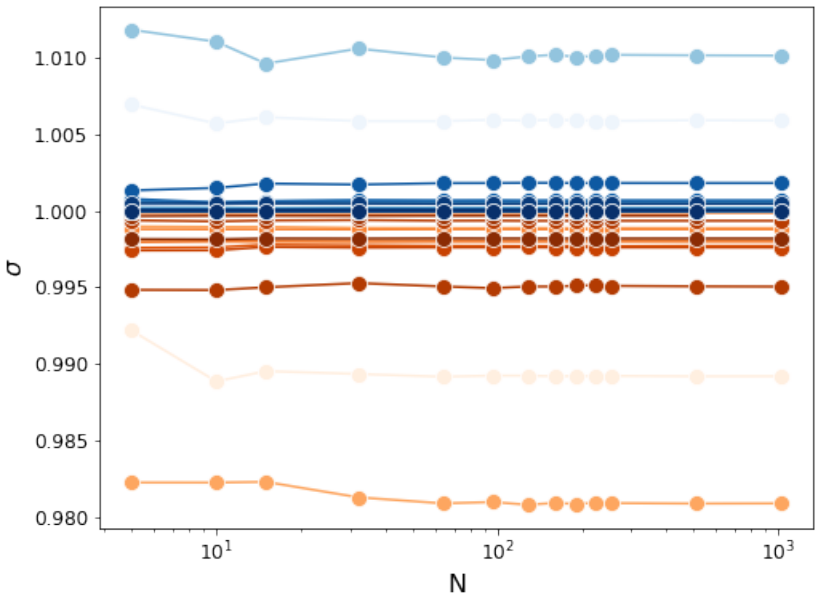


L_{orth} small

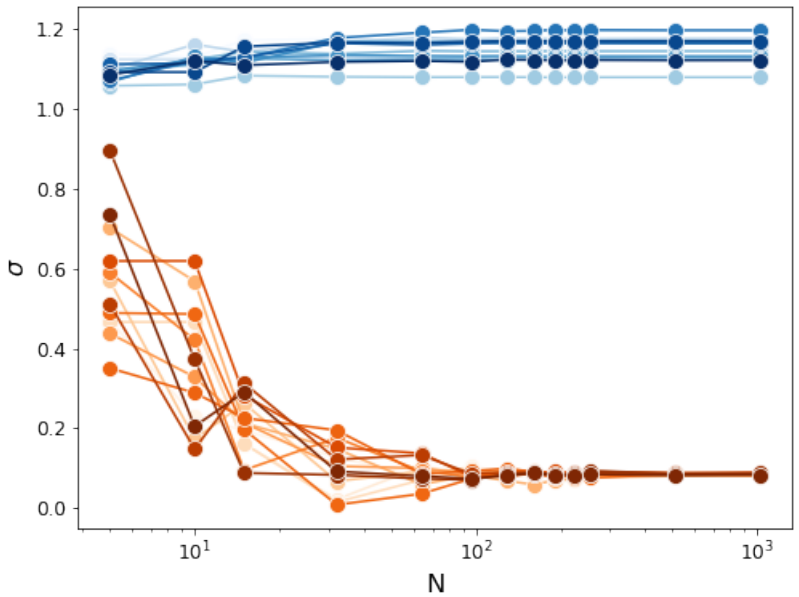


L_{orth} large

Stability of singular values when N varies



L_{orth} small



L_{orth} large

Conclusion

Positive results:

- Orthogonal Convolutional layers exist for most architectures.
- L_{orth} has a favorable landscape in most cases.
- Singular values of $\mathcal{K}$ are stable around 1, even when N grows.

Questions

- What causes the difference for $M = CS^2$, and what is its impact?

Thank you
Soon on Arxiv