Preventing Gradient Attenuation in Lipschitz Constrained Convolutional Networks

Qiyang Li*, Saminul Haque*, Cem Anil, James Lucas, Roger Grosse, Jörn-Henrik Jacobsen

*Equal Contribution University of Toronto, Vector Institute

Objective

Training an expressive convolutional neural network with a **known, tight upper-bound** on its **Lipschitz constant** by enforcing **gradient norm preservation (GNP)**.

Motivation

Why Lipschitz-constrained Networks?

1. Provable adversarial robustness via large-margin training.
2. 1-Wasserstein distance estimation via Kantorovich and Rubinstein duality [8].

Why Gradient Norm Preservation (GNP)?

1-Lipschitz-constrained networks suffer from two common problems solved by GNP:

1. Loose upper-bound obtained by $\mathrm{Lip}(f_1 \circ f_2) \leq \mathrm{Lip}(f_1)\,\mathrm{Lip}(f_2)$.
2. Gradient attenuation during backpropagation since $\|\nabla_{\mathbf{x}}\mathcal{L}\|_2 \leq \mathrm{Lip}(f)\,\|\nabla_{\mathbf{y}}\mathcal{L}\|_2$, where $\mathbf{y} = f(\mathbf{x})$.

Challenges of Enforcing GNP for Convolutional Networks

1. Optimization over the space of GNP convolutions does not have an established method.
2. Topology is unknown for GNP convolutions.

Background

GNP Functions: f is GNP if $\|\nabla f(\mathbf{x})^T\mathbf{g}\|_2 = \|\mathbf{g}\|_2, \forall \mathbf{g}$.

- GNP functions have a Lipschitz constant of 1; Composition of GNP functions are GNP.
- GNP linear functions are **orthogonal**; GNP convolutions are **orthogonal convolutions**.

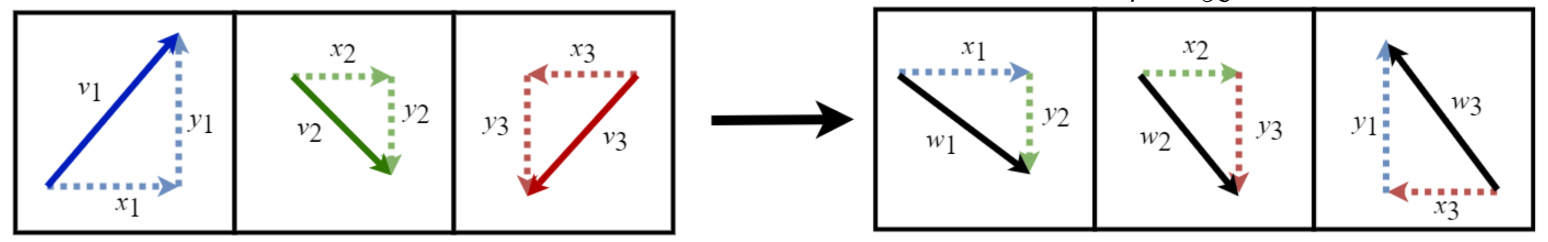
Symmetric Projectors: $\mathbb{P}(n,k) = \{P | P = P^T = P^2, \mathrm{rank}(P) = k, P \in \mathbb{R}^{n\times n}\}$.
$\mathbb{P}(n) = \bigcup_k \mathbb{P}(n,k)$ has $n+1$ connected components: $\{\mathbb{P}(n,0), \cdots, \mathbb{P}(n,k), \cdots, \mathbb{P}(n,n)\}$.

Orthogonal Convolutions Are Disconnected

Block Convolution Parameterization in 1-D [6]

$$\mathcal{W}(H, P_{1:K-1}) = H \Box \left[P_1\ (I-P_1)\right] \Box \cdots \Box \left[P_{K-1}\ (I-P_{K-1})\right],$$

$$\text{where } P_i \in \mathbb{P}(n), H \in O(n), [X \Box Y]_i = \sum_{i'=-\infty}^{\infty} X_{i'} Y_{i-i'}.$$



v $[P \;\; I\text{-}P] \,\Box\, v$

Theorem 1: 1-D orthogonal convolution space has $2(K-1)n+2$ connected components.

Extension to 2-D: Analogous parameterization and disconnectedness results as 1-D [10].
Implication: Gradient-based optimization would be trapped in the initial connected component.

Overcoming Disconnectedness

Theorem 2: For any convolution $C = \mathcal{W}(H, P_{1:K-1}, Q_{1:K-1})$ with input and output channel sizes of n ($P_i, Q_i \in \mathbb{P}(n)$), there exists a convolution $C' = \mathcal{W}(H', P'_{1:K-1}, Q'_{1:K-1})$ with input and output channels sizes of $2n$ constructed from only n-rank projectors ($P'_i, Q'_i \in \mathbb{P}(2n,n)$) such that $C'(\mathbf{x})_{1:n} = C(\mathbf{x}_{1:n})$. That is, the first n channels of the output is the same with respect to the first n channels of the input under both convolutions.

Implication: Using this, one can double the number of channels of a BCOP constructed network to represent all the connected components of the original network in a *single* connected component.

Block Convolution Orthogonal Parameterization (BCOP)

A BCOP orthogonal convolution of $2n$ channel size is

$$\mathcal{W}(H, P_{1:K-1}, Q_{1:K-1}),\ P_i, Q_i \in \mathbb{P}(2n, n)$$

We can use any unconstrained matrix $\tilde{R} \in \mathbb{R}^{2n\times n}$ to parameterize $T \in \mathbb{P}(2n,n)$,

$$T = RR^T, R = \psi(\tilde{R})$$

where ψ can be any differentiable orthogonalization procedure that results in a matrix of the same size, $R \in \mathbb{R}^{2n\times n}$, with orthonormal columns: $R^TR = I$ (e.g., Björck orthogonalization [2]).

Design Rationale: $\mathbb{P}(2n,n)$ is the largest connected component of $\mathbb{P}(2n)$ by dimensionality and using $\mathbb{P}(2n,n)$ to construct BCOP layers represents all networks with channel size of n.

Building GNP Convolutional Networks

Network Components	Problems under GNP	Solutions
Residual connection	Degenerates into identity	Removed
Batch normalization	Not GNP	Removed
Zero-padding	Degenerates into 1×1 convolutions	Cyclic padding instead
Strided convolution	Orthogonality properties unknown	Invertible downsampling [5]
Linear layer	Not GNP in general	Orthogonalize the matrix [1]
Nonlinear activation	Not GNP in general	GroupSort [1]

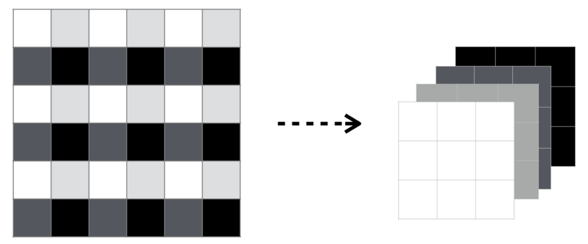


Figure: Invertible Downsampling [5]

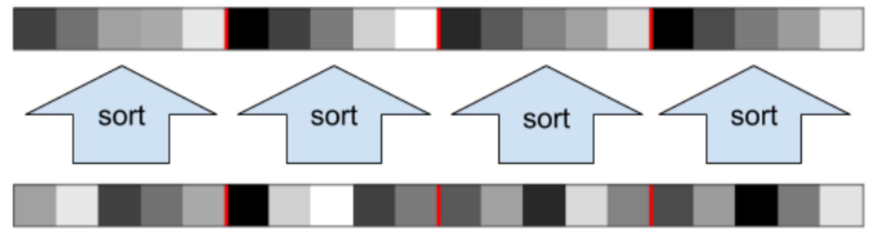


Figure: GroupSort [1]

Empirical Results: Provable Adversarial Robustness Under L_2 Norm

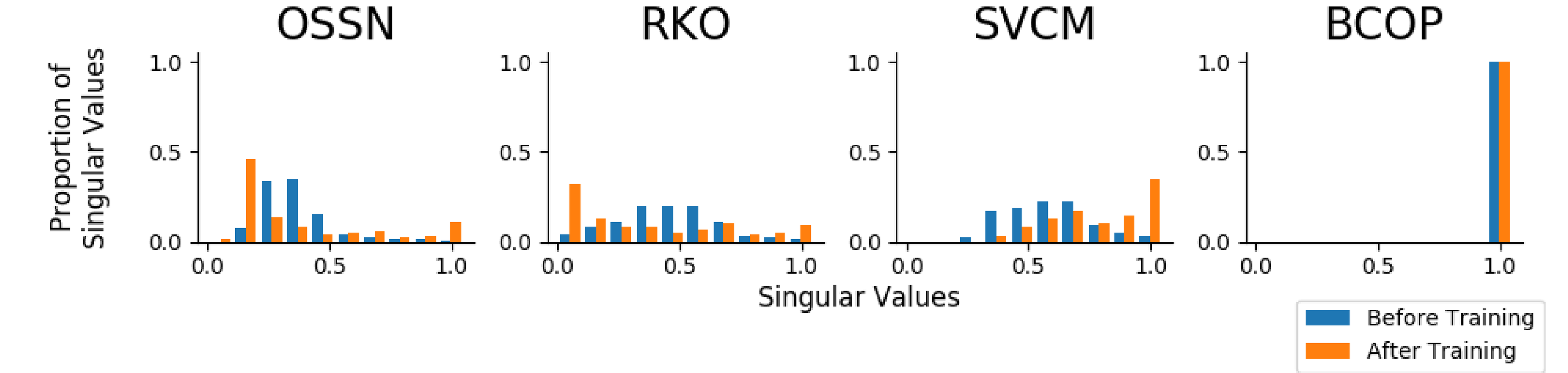
Ablation Study (Provable Adversarial Robustness with L_2 Metric)

Dataset			OSSN [4]	RKO [3]	SVCM [7]	BCOP
MNIST ($\varepsilon = 1.58$)	Small	Clean	96.86	97.28	97.24	**97.54**
		Robust	42.95	43.58	28.94	**45.84**
	Large	Clean	98.31	98.44	97.93	**98.69**
		Robust	53.77	55.18	38.00	**56.37**
CIFAR10 ($\varepsilon = 36/255$)	Small	Clean	62.18	61.77	62.39	**64.53**
		Robust	48.03	47.46	47.59	**50.01**
	Large	Clean	67.51	70.01	69.65	**72.16**
		Robust	53.64	55.76	53.61	**58.26**

State-of-the-art Comparison (L_2)

Dataset		BCOP-Large	FC-3	KW-Large [9]	KW-Resnet [9]
MNIST ($\varepsilon = 1.58$)	Clean	98.69	**98.71**	88.12	–
	Robust	**56.37**	54.46	44.53	–
CIFAR10 ($\varepsilon = 36/255$)	Clean	**72.16**	62.60	59.76	61.20
	Robust	**58.26**	49.97	50.60	51.96

Singular Value Distribution of a Conv Layer Jacobian Before and After Training



Empirical Results: 1-Wasserstein Distance Estimation

	BCOP	RKO	OSSN
MaxMin	**9.91**	8.95	7.39
ReLU	**8.28**	7.82	7.06

Note: All the methods give a lower bound on the Wasserstein distance (higher is better).

References

[1] C. Anil, J. Lucas, and R. Grosse. Sorting out Lipschitz function approximation. In K. Chaudhuri and R. Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 291–301, Long Beach, California, USA, 09–15 Jun 2019. PMLR.
[2] Å. Björck and C. Bowie. An iterative algorithm for computing the best estimate of an orthogonal matrix. *SIAM Journal on Numerical Analysis*, 8(2):358–364, 1971.
[3] M. Cisse, P. Bojanowski, E. Grave, Y. Dauphin, and N. Usunier. Parseval networks: Improving robustness to adversarial examples. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 854–863. JMLR. org, 2017.
[4] H. Gouk, E. Frank, B. Pfahringer, and M. Cree. Regularisation of neural networks by enforcing Lipschitz continuity. *arXiv preprint arXiv:1804.04368*, 2018.
[5] J.-H. Jacobsen, A. W. Smeulders, and E. Oyallon. i-RevNet: Deep invertible networks. In *International Conference on Learning Representations*, 2018.
[6] J. Kautsky and R. Turcajová. A matrix approach to discrete wavelets. In *Wavelet Analysis and Its Applications*, volume 5, pages 117–135. Elsevier, 1994.
[7] H. Sedghi, V. Gupta, and P. M. Long. The singular values of convolutional layers. In *International Conference on Learning Representations*, 2019.
[8] C. Villani. *Optimal transport: old and new*, volume 338. Springer Science & Business Media, 2008.
[9] E. Wong, F. Schmidt, J. H. Metzen, and J. Z. Kolter. Scaling provable adversarial defenses. In *Advances in Neural Information Processing Systems*, pages 8400–8409, 2018.
[10] L. Xiao, Y. Bahri, J. Sohl-Dickstein, S. Schoenholz, and J. Pennington. Dynamical isometry and a mean field theory of CNNs: How to train 10,000-layer vanilla convolutional neural networks. In *Proceedings of the 35th International Conference on Machine Learning*, pages 5393–5402, 2018.