Generalizability and Sample Complexity of Quadratic Shallow Neural Networks Under Low-Rank Learning

Xiaolin WANG[1], Olivier RIOUL[1], Anissa MOKRAOUI[2], Pierre DUHAMEL[3], Jacob BENESTY[4]

[1] Télécom Paris, Institut Polytechnique de Paris [2] Université Sorbonne Paris Nord [3] CentraleSupélec, Université Paris-Saclay

[4] Université du Québec

25 August 2025

GRETSI'25

0. Summary

This paper studies the impact of data correlation on the performance of low rank approximation during training of Quadratic Shallow Neural Networks (QSNN)

1. Background

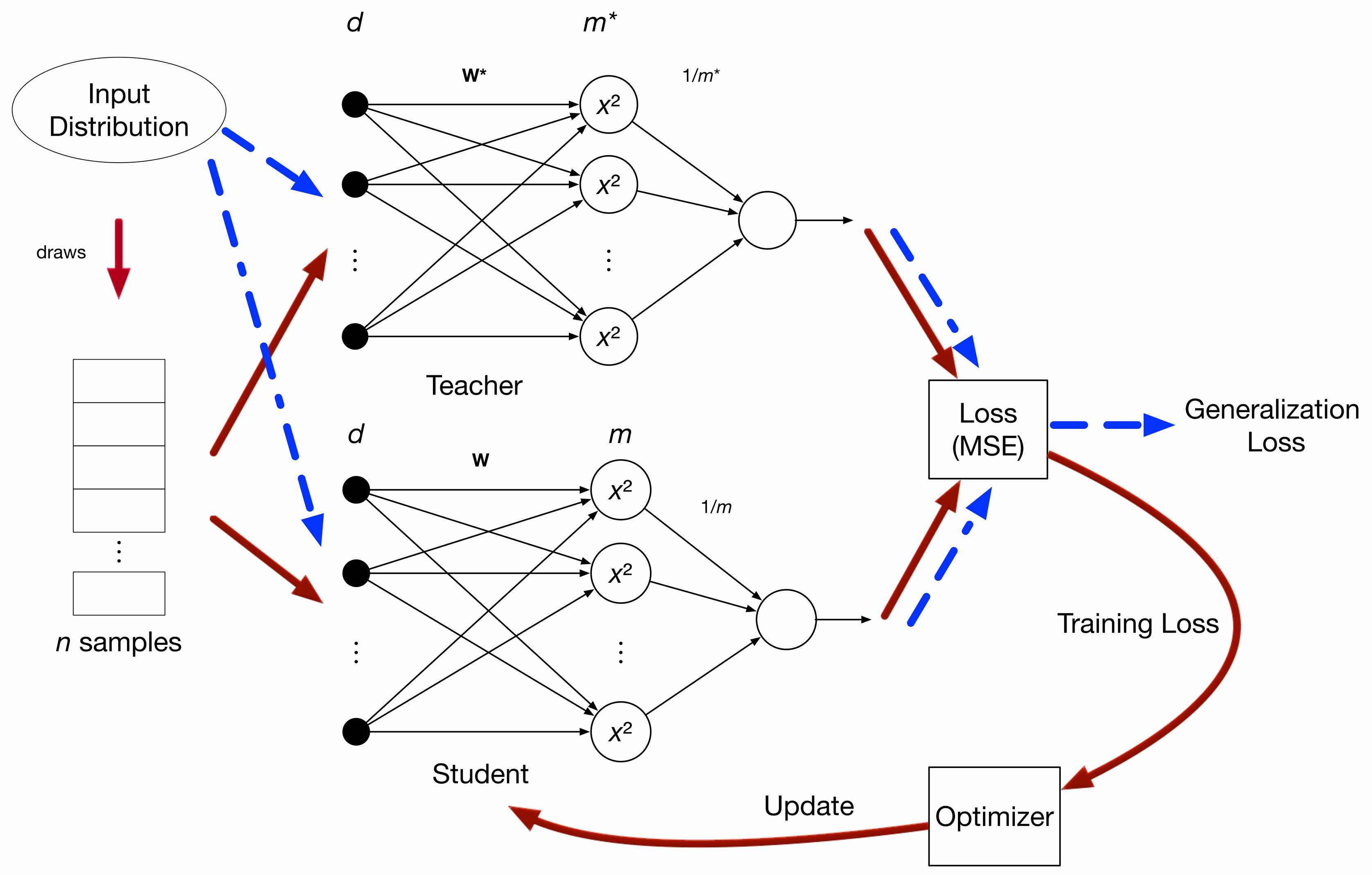
- Neural networks are over-parameterized, leading to high computational costs [1, 8]
- Post-training model compression obtained by low-rank approximation effectively reduces over-parameterization [3, 7, 4]

2. Key Questions

- Can low rank approximation be applied during training? With which consequences on the performance?
- What is the sample size required for target generalization loss?
- What data properties facilitate model compression?

3. Problem Formulation (Toy Example)

Teacher-Student Learning of QSNN



Current State-of-the-Art (Hard Sample Complexity) [2, 6]

Full-rank student needs at least

$$n_c = \begin{cases} d(m^*+1) - \frac{m^*(m^*+1)}{2} & \text{if } m^* < d \\ \frac{d(d+1)}{2} & \text{if } m^* \geq d \end{cases} \qquad (1)$$

samples to enable **perfect generalization** (stronger than simple convergence)

4. Low-Rank Learning

- Factorize: $W = W_1 W_2^T$ where $W_1 \in \mathbb{R}^{m\times r}$, $W_2 \in \mathbb{R}^{d\times r}$
- Degree-of-freedom (DOF) reduction when $r < \frac{dm}{d+m}$

Theorem: For Gaussian data distribution, the loss-optimal rank-r student (W^o) is determined by the teacher's singular values (s_i^*) [2]

$$W^{oT} W^o = U \operatorname{diag}\left(s_1^* + \tfrac{S}{r+2}, \ldots, s_r^* + \tfrac{S}{r+2}, 0, \ldots, 0\right) U^T \qquad (2)$$

where $S = \sum_{i=r+1}^{d} s_i^*$.

5. Doubly-Correlated Teacher-Student Framework

- Teacher weights generated by stable 2D AR(1) process with correlation parameter ϕ ($0 < \phi < 1$)
- Captures correlation behaviors in real-world data
- Higher ϕ concentrates SV near largest components (see Fig.)

Key Insight: Parameter correlations in teacher significantly enhance generalizability of rank-reduced students [5]

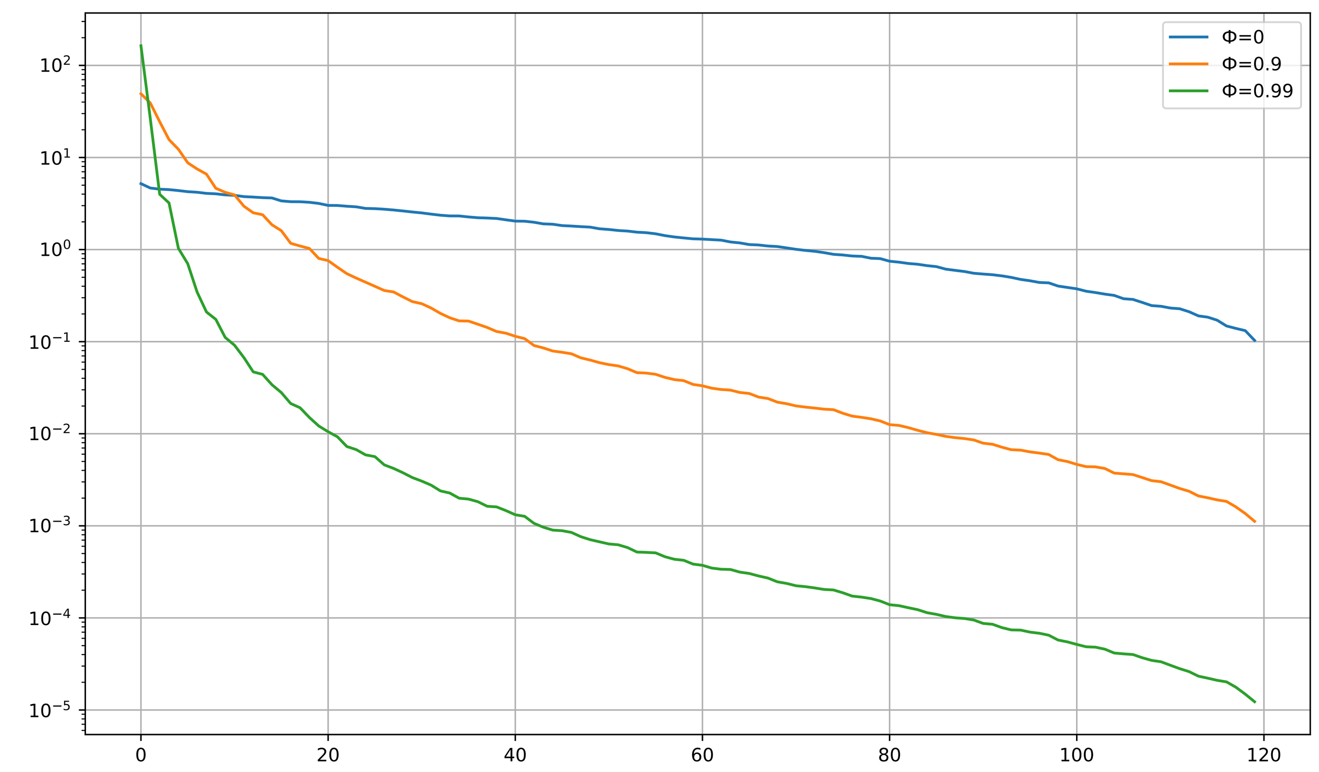


Figure: SV of $A^* = \frac{1}{m^*} W^{*T} W^*$ to indices, where $d = 200$, $m^* = 120$. W^* generated by stable 2D AR(1) process with different ϕ

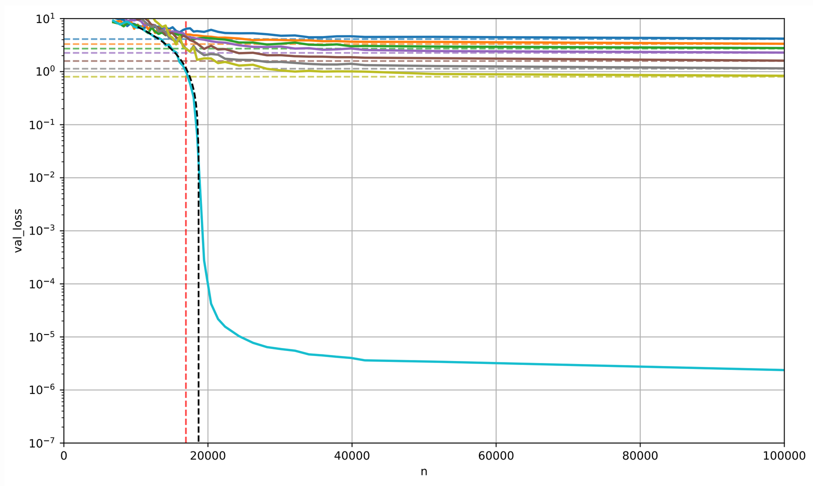
6. Generalized Sample Complexity

New: Generalization loss in terms of the sample size behaves like

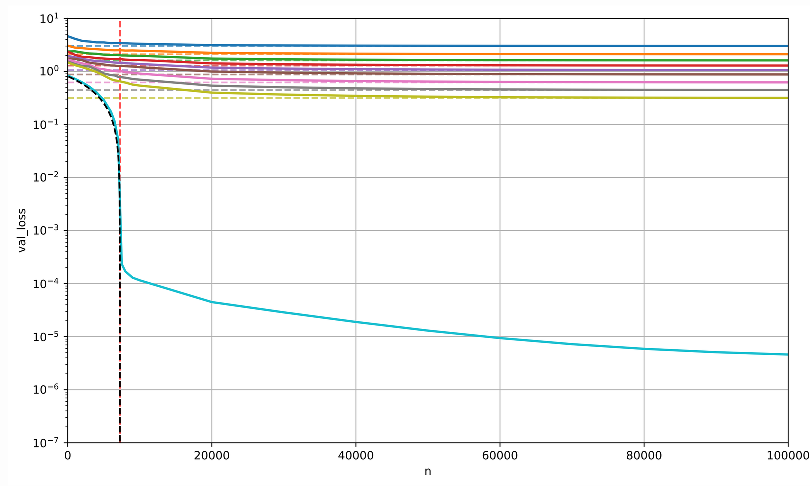
$$\bar{L}(W^*, n) \approx A(W^*) \exp(-\alpha n) + L(W^o) \qquad (3)$$

- **Low-sample regime:** Exponential decay $L(W) \propto e^{-\alpha n}$
- **High-sample regime:** Plateau determined by rank constraint and teacher's residual singular values s_i^*

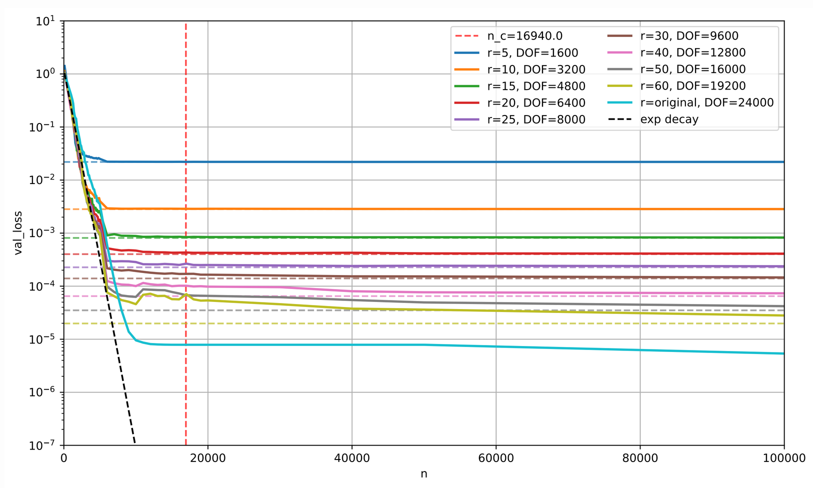
7. Experimental Results



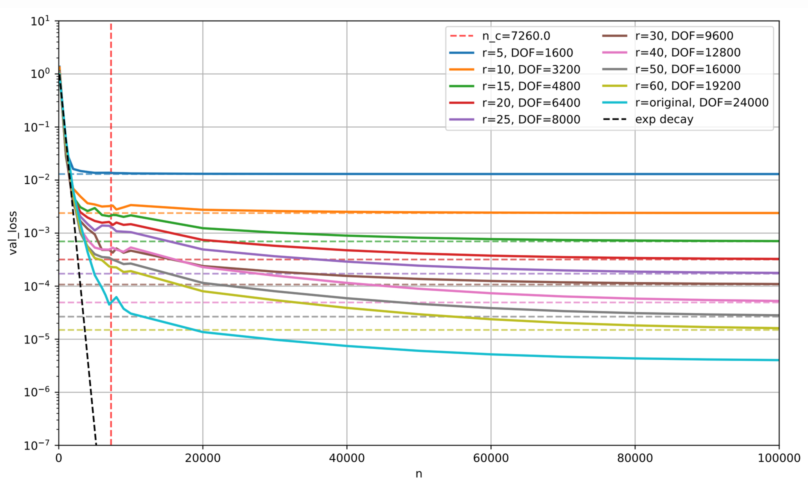
(a) $m^* = 120$, $d = 200$, $m = 120$, $\phi = 0$



(b) $m^* = 200$, $d = 120$, $m = 200$, $\phi = 0$



(c) $m^* = 120$, $d = 200$, $m = 120$, $\phi = 99\%$



(d) $m^* = 200$, $d = 120$, $m = 200$, $\phi = 99\%$

Findings:

- Exponential decay in low-sample regime determined by teacher
- Loss saturates to bias based on rank constraint
- Parameter correlations reduce sample complexity
- On low-correlation data and high-rank students, sample complexity reduces to hard threshold (vertical red-dashed line)

Parameter correlations make low-rank learning feasible!

References

[1] S. Du and J. Lee. On the Power of Over-parametrization in Neural Networks with Quadratic Activation. In Proceedings of the 35th ICML, pages 1329–1338. PMLR, July 2018.

[2] D. Gamarnik, E. C. Kızıldağ, and I. Zadik. Stationary Points of Shallow Neural Networks with Quadratic Activation Function, July 2020, arXiv:1912.01599.

[3] K. Geyer, A. Kyrillidis, and A. Kalev. Low-Rank Regularization and Solution Uniqueness in Over-parameterized Matrix Sensing. In Proc. of the 23rd AISTATS, pages 930–940. PMLR, June 2020.

[4] Y. Idelbayev and M. Á. Carreira-Perpiñán. Low-Rank Compression of Neural Nets: Learning the Rank of Each Layer. In 2020 IEEE / (CVPR), pages 8046–8056, June 2020.

[5] G. Jin, X. Yi, L. Zhang, et al. How does Weight Correlation Affect the Generalisation Ability of Deep Neural Networks, arXiv:2010.05983.

[6] S. S. Mannelli, E. Vanden-Eijnden, and L. Zdeborová. Optimization and Generalization of Shallow Neural Networks with Quadratic Activation Functions. In NeurIPS, volume 33, pages 13445–13455, Vancouver, Canada, 2020.

[7] W. Roth, G. Schindler, B. Klein, et al. Resource-efficient neural networks for embedded systems. Journal of Machine Learning Research, 25(50):1–51, 2024.

[8] N. Xiong, L. Ding, and S. S. Du. How Over-Parameterization Slows Down Gradient Descent in Matrix Sensing: The Curses of Symmetry and Initialization. In 12th ICLR, Oct. 2023.

contact : xiaolin.wang@telecom-paris.fr

https://xiaolin.wang