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]LTCI, Télécom Paris, Institut Polytechnique de Paris, 19 place Marguerite Perey, 91120 Palaiseau, France

[2]L2TI, Université Sorbonne Paris Nord, 99 avenue Jean-Baptiste Clément, 93430 Villetaneuse, France

[3]L2S, CentraleSupélec, Université Paris-Saclay, 3 rue Joliot-Curie, 91192 Gif-sur-Yvette, France

[4]INRS-EMT, Université du Québec, 1650 boulevard Lionel-Boulet, Varennes, Québec, Canada

Résumé – Ce travail examine les relations entre la complexité de l'échantillon et la complexité du modèle dans l'apprentissage à faible rang des réseaux neuronaux quadratiques superficiels (QSNN). Nous proposons un nouveau cadre enseignant-étudiant à double corrélation qui intègre des corrélations entre paramètres pour mieux refléter les propriétés des données réelles. Ce cadre étend les théories existantes sur les QSNN en analysant l'influence de la taille d'échantillon sur l'erreur de généralisation pour des modèles appris à faible rang ou présentants un biais intrinsèque. Nos résultats révèlent un comportement à deux phases dans les lois d'échelle de la capacité de généralisation selon la taille d'échantillon, et démontrent que les corrélations paramétriques du modèle enseignant améliorent considérablement la généralisation des modèles à rang réduit. Des simulations numériques étendues confirment ces résultats et offrent des perspectives théoriques ainsi que des recommandations pratiques pour concevoir des architectures de réseaux neuronaux efficaces dans le contexte de l'apprentissage à faible rang.

Abstract – We investigate the interplay between sample complexity and model complexity in low-rank learning of quadratic shallow neural networks (QSNN), within a novel doubly-correlated teacher-student framework that incorporates parameter correlations to reflect real-world data properties. This framework generalizes existing theories for QSNN by analyzing the impact of sample size on generalization loss for models under low-rank learning or exhibiting inherent bias. We observe a two-regime behavior in the scaling law of generalization ability with respect to sample size and show that parameter correlations in the teacher model significantly enhance the generalization of rank-reduced models. Extensive numerical simulations confirm the results and offer theoretical insights and practical guidance for designing efficient neural network architectures under low-rank learning.

1 Introduction

Neural networks have achieved remarkable success in fields such as computer vision, natural language processing, and speech recognition. However, their over-parameterized nature—where the number of model parameters far exceeds what is needed to represent the target function—leads to increased computational costs and challenges for deployment on resource-constrained devices [15]. While model compression and low-rank approximation techniques can reduce over-parameterization, most remain heuristic and lack a unified theoretical foundation [1].

This work addresses fundamental questions for training efficient neural networks from scratch: How much can over-parameterization be reduced during training? What is the minimum sample size required for a good generalization performance? Can real-world data properties, such as parameter correlations, facilitate model reduction?

We study Quadratic Shallow Neural Networks (QSNN) in a teacher-student framework, where a fixed "teacher" model generates labels and a "student" model is trained to approximate the teacher. While matrix sensing provides theoretical insights into the interplay between approximation, data complexity, and parameterization [6, 17], its conclusions generally do not extend to neural networks with nonlinear activations. For QSNNs, recent works have established sample complexity results for full-rank students learning from random teachers [3, 16, 12]. Quadratic activations are popular in theoretical studies due to their convexity, enabling analytical results and robust validation. Additionally, it has been demonstrated that stacking multiple quadratic layers can approximate more complex neural networks [11].

We extend current QSNN theory by analyzing sample complexity for both full-rank and low-rank students and proposing a teacher model that better reflects parameter correlations in real-world data. This provides a more accurate and complete description of neural scaling laws in teacher-student scenarios with low-rank learning.

Our main contributions include:

— Analytical derivation of the optimal generalization loss for rank-reduced QSNN models;

— A doubly-correlated teacher-student framework closer to real-world data, where the teacher's parameters are generated using a correlated stochastic process

— More generalized extension of sample complexity theory to both full-rank and low-rank networks, revealing a two-regime neural scaling law;

— Empirical results demonstrating the impact of parameter correlations and rank constraints on generalization performance.

These findings bridge the gap between theory and practice for the scaling law of low-rank deep learning, offering new insights for training-time model compression.

2 Problem Formulation

Consider a teacher-student setup with single-hidden-layer NN structure. For input datum $x \in \mathbb{R}^d$ and weight matrix $W \in \mathbb{R}^{m\times d}$ given input dimension d and number of neurons m, the QSNN is defined as:

$$f(W;x) = \frac{1}{m} x^T W^T W x. \tag{1}$$

With a fixed weight matrix $W^* \in \mathbb{R}^{m^*\times d}$, the teacher network generates labels as $f(W^*;x)$. Given a dataset of n samples $\mathcal{X} = \{x_1,\dots,x_n\}$ drawn from a distribution X and corresponding labels $\{f(W^*;x_1),\dots,f(W^*;x_n)\}$, the training seeks a student model that minimizes the training loss $\hat{L}(W;\mathcal{X})$ based on Mean Squared Error (MSE) on the dataset:

$$\hat{L}(W;\mathcal{X}) = \frac{1}{n}\sum_{x\in\mathcal{X}} \left(f(W;x) - f(W^*;x)\right)^2. \tag{2}$$

The generalization loss $L(W)$ is defined as the expected MSE over the distribution X:

$$L(W) = \mathbb{E}_{x\sim X}\left[\left(f(W;x) - f(W^*;x)\right)^2\right]. \tag{3}$$

We present the state-of-the-art sample complexity theory for the teacher-student setup in single hidden layer QSNNs as defined in Equation 1. For full-rank students and teachers (i.e., W has rank $\min(m,d)$ and W^* has rank $\min(m^*,d)$), if $m \geq m^*$, there exists a threshold n_c for the training sample size n, determined by d and m^*. If $n > n_c$, the student model has a positive probability of generalizing, meaning it can learn the solution $\frac{1}{m}W^TW = \frac{1}{m^*}W^{*T}W^*$, achieving an optimal generalization loss $L(W) = 0$. Conversely, if $n \leq n_c$, the student model will overfit with probability 1. The formula for n_c is given in [12].

$$n_c = \begin{cases} d(m^*+1) - \dfrac{m^*(m^*+1)}{2} & \text{if } m^* < d \\ \dfrac{d(d+1)}{2} & \text{if } m^* \geq d. \end{cases} \tag{4}$$

The case of $m^* \geq d$ in Equation 4 is proved in [5]. For the case where $m^* < d$, [12] provides a rigorous proof for $m^* = 1$ and a heuristic explanation for $m^* > 1$, although the latter can be confirmed through numerical simulations. As shown in Figure 1, a notable decrease in validation losses to nearly zero is evident when the sample size exceeds n_c.

We extend QSNN sample complexity theory by studying rank-reduced student networks and parameter correlations in the teacher model. For low-rank training, we constrain the student network's rank to $r < \min(m,d)$ by factorizing the weight matrix W into $W = W_1W_2^T$ where $W_1 \in \mathbb{R}^{m\times r}$ and $W_2 \in \mathbb{R}^{d\times r}$. When $r < \frac{dm}{d+m}$, the number of parameters of the low-rank model is less than that of the original structure. This factorization is equivalent to inserting a linear hidden layer between the input and original hidden layer (Figure 2). Unlike full-rank students which can achieve zero error, we find that rank-reduced students exhibit persistent generalization bias across all hyperparameter settings, as demonstrated in Theorem 1. This indicates that the current sample complexity framework does not extend to the low-rank case.

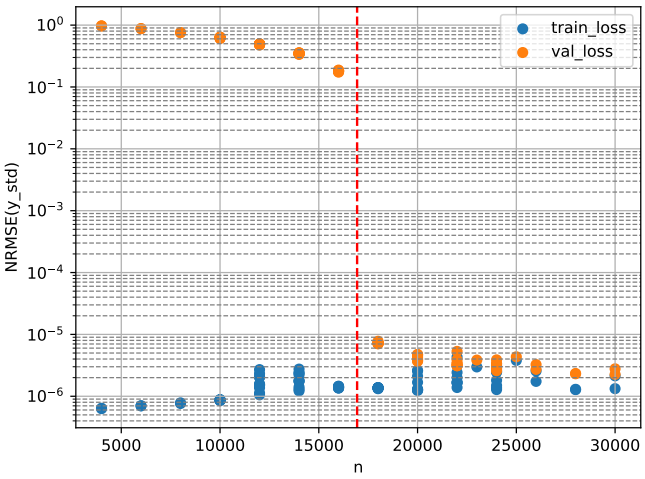
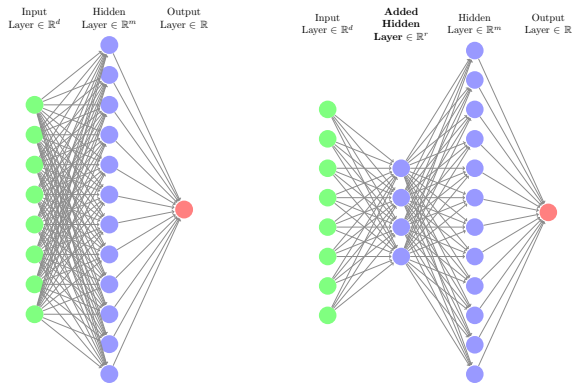


Figure 1 – Training losses and validation losses from multiple random teacher/student initialization and training with different n under the hyperparameters $d = 200$ and $m^* = 120$. The vertical dashed red line indicates the corresponding sample complexity $n_c = 16940$.



(a) NN without factorization (b) NN with factorization

Figure 2 – Structure of the QSNN with and without low-rank factorization.

Theorem 1. *If $X \sim \mathcal{N}(0, I_d)$, fix an arbitrary teacher model $W^* \in \mathbb{R}^{m^*\times d}$ with $W^{*T}W^* = U\,\mathrm{diag}(s_1^*,\dots,s_d^*)U^T$ where $s_1^* \geq \cdots \geq s_d^* \geq 0$ and U an orthogonal matrix. For any student $W \in \mathbb{R}^{m\times d}$ of rank at most $r < d$, the generalization loss-optimal W^o satisfies:*

$$W^{oT}W^o = U\ \mathrm{diag}\left(s_1^* + \tfrac{S}{r+2},\dots,s_r^* + \tfrac{S}{r+2}, 0,\dots,0\right)U^T, \tag{5}$$

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

Proof. With $A = \frac{1}{m}W^TW$ and $A^* = \frac{1}{m^*}W^{*T}W^*$, as $X \sim \mathcal{N}(0, I_d)$, the generalization loss becomes

$$L(W) = \mathbb{E}\left[(X^T(A^* - A)X)^2\right] = \mathrm{tr}(\Delta)^2 + 2||\Delta||_F^2, \tag{6}$$

where $\Delta = A^* - A$ is symmetric.

Assume $A = U\,\mathrm{diag}(b_1,\dots,b_d)U^T$ with $b_i \geq \dots b_d \geq 0$ and $\mathrm{rank}(A) \leq r$, and let $b_i = 0$ for $i > r$. Writing $\chi_i = s_i^* - b_i$ for each i, the objective becomes:

$$L(W) = \left(\sum_{i=1}^r \chi_i + S\right)^2 + 2\sum_{i=1}^r \chi_i^2 \tag{7}$$

with $S = \sum_{i=r+1}^d s_i^*$. Because $L(W)$ in Equation 7 is strictly convex and symmetric in the variables $\chi_1,\dots,\chi_r$, minimizing

$L(W)$ reduces to minimizing a single-variable function:

$$g(\chi) = (r\chi + S)^2 + 2r\chi^2, \tag{8}$$

which yields optimal $\chi^o = -\frac{S}{r+2}$ and thus $b_i^o = s_i^* + \frac{S}{r+2}$ for $i \leq r$.

This solution is feasible since all singular values s_i^* are non-negative. The result thus follows from unitary invariance. □

When both the teacher and the student are rank-1, the optimization problem reduces to the classical phase retrieval problem. Notably, the optimal solution in this case differs from the standard Eckart–Young–Mirsky theorem [4] due to the presence of the quadratic activation. For a full-rank W, the generalization loss $L(W)$ can be made zero by the trivial solution $W^T W = W^{*T} W^*$, which corresponds to the conditions in Equation 4. However, when W is constrained to be low-rank, the residual singular values introduce an unavoidable approximation error, implying that some loss must be tolerated in low-rank learning. Despite this, numerous studies on real-world regression tasks have shown that rank-reduced neural networks can often achieve generalization performance comparable to that of full-rank networks [2, 8, 14], indicating that the loss in precision due to residual singular values is often negligible in practice. Several works attribute the strong generalization ability of low-rank models to inherent correlations in the parameters of real-world mappings [10, 9]. Motivated by these observations, we introduce the analytical framework described in the next section.

3 Proposed Analytical Framework

We propose a doubly-correlated teacher-student framework, where the teacher's weight matrix W^* is generated by a stable 2D first-order autoregressive (AR(1)) process with correlation parameter ϕ ($0 \leq \phi < 1$). This induces correlations between adjacent elements, while mean and variance are preserved by scaling the noise with $\sqrt{1-\phi^2}$. Though being a simplified framework, such correlations in the parameters are observed in real-world regression networks. The presence of correlations between parameters strongly affect the singular value distribution of W^*. For uncorrelated weights, the singular values follow the Marchenko-Pastur law [13]. Introducing correlations deforms this distribution, concentrating singular values near the largest as ϕ increases [7] (see Figure 3). This makes low-rank approximation more effective as the least significant singular values are ignorable without much loss of precision.

Proposition: *In this doubly-correlated setup, the generalization loss as a function of sample size exhibits two regimes: an initial exponential decay, followed by a plateau determined by the model's rank constraint and the teacher's singular value distribution.*

This behavior can be understood via singular value alignment. In the low-sample regime, the model quickly learns the leading singular value directions as training samples increase, resulting in exponential decay of generalization loss ($L(W) \propto e^{-\alpha n}$). For teachers with more uniform singular values (low correlation), this decay is nearly vertical, regressing to the hard sample size threshold in Equation 4. However, as the sample size grows, depending on the inherent bias of the student, the additional information primarily pertains to the lower singular directions, which correspond to fine-grained variations that are not easily distinguishable. Further gains are now limited not by sample size but by the alignment between the model's limited singular modes and those of the data distribution. The generalization loss then plateaus, set by the model's rank and the teacher's singular spectrum (see Theorem 1). The correlation parameter ϕ fundamentally shapes this two-regime behavior. Higher correlation concentrates the teacher's singular values toward the largest components, not only making the exponential decay occur sooner as the model can more efficiently capture the dominant patterns, but also decreases the plateau level as the residual singular values become smaller. Thus, the trade-off in the decision of low-rank learning is between model complexity and tolerable error, shaped by data mapping correlations and sample size.

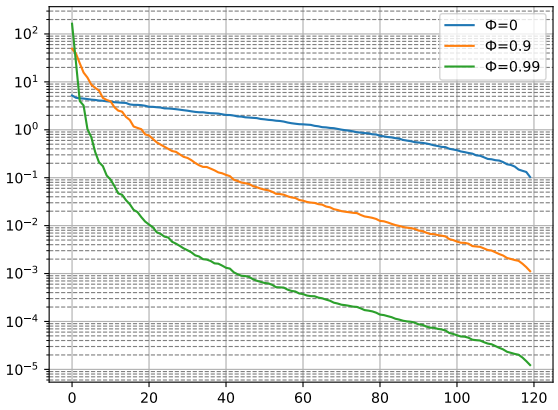


Figure 3 – Singular values of $A^* = \frac{1}{m^*} W^{*T} W^*$ to indices, where $d = 200$, $m^* = 120$. W^* generated by stable 2D AR(1) process with different ϕ.

4 Numerical Simulation

We empirically validate the two-regime generalization loss behavior using the doubly-correlated teacher-student setup with $m^* = 120$, $d = 200$, $m = 120$, and $\phi = 0.99$. Students are either two-layer full-rank networks or rank-constrained three-layer bottleneck networks. All models are trained with Adam optimizer (learning rate 10^{-3}) and batch learning (using a batch size equal to the sample size), for 10^5 epochs.

Experiments span various student ranks and random initializations. Figure 4 shows results for both uncorrelated and correlated teachers, with the teacher model fixed within each trial to ensure a stable singular value spectrum. Validation losses are computed on large enough test sets and reported as Normalized Root Mean Square Error (NRMSE) for comparability. In all cases, the low-sample regime shows exponential decay in generalization loss, with a rate determined by the teacher's singular values, not the student's capacity. In the high-sample regime, the loss saturates to a bias that depends on student rank, corresponding to the irreducible error in Theorem 1. For $\phi = 0$, the exponential decay closely matches the hard threshold n_c from Equation 4. We note that the gap is due to imperfect training. With perfect training, full-rank students shall reach zero loss beyond n_c.

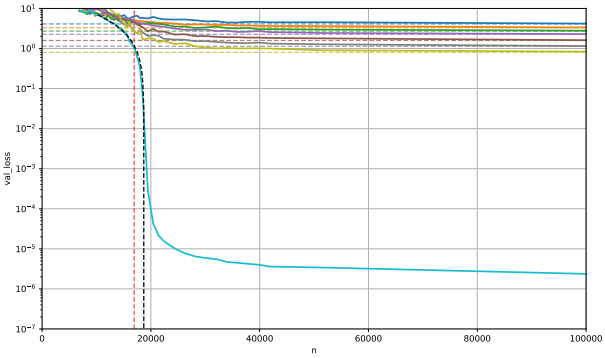
Parameter correlations in the teacher help students achieve lower generalization loss with less parameters, as seen by the reduced bias in rank-reduced models, even though a small loss

compared to the full-rank model is introduced due to the least significant singular values being difficult to represent in the training data.

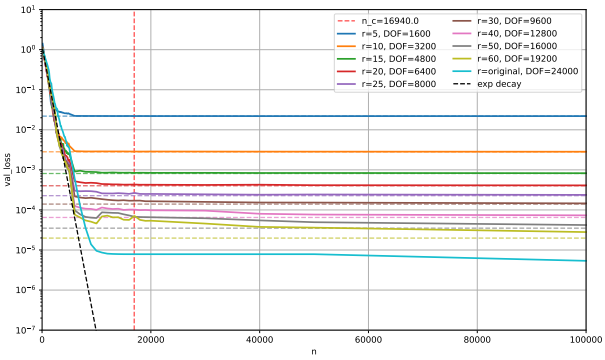
5 Discussion

This work explores the potential of reducing certain degrees of over-parameterization when training a neural network. From state-of-the-art parameterization theories on QSNN, we derive theoretical insights into the two-regime behavior of generalization loss to sample size and demonstrate the critical role of parameter correlations in enhancing the generalization performance of rank-reduced models. Our findings provide practical guidance for deciding and designing low-rank learning of efficient neural networks. We note that our results on sample complexity for low-rank QSNNs and correlated cases are primarily empirical and heuristic, and a full theoretical treatment is left for future work.

In addition, there are other future directions to explore, such as: (i) more empirical evidence on real-world deep regression tasks and the measurement of the correlation of parameters in real-world datasets ; (ii) classification tasks with different loss functions (e.g. cross-entropy loss) could be considered, as thresholding the output of the network could further decrease the precision tradeoff of rank reduction ; (iii) the interplay between parameter correlations and explicit regularizations to further improve training-time compression techniques.



(a) $\phi = 0$



(b) $\phi = 0.99$

Figure 4 – Average validation losses vs training sample number n of various rank-reduced and full-rank students approximating teachers generated by 2D AR(1) process of different ϕ values. Degree of freedom (DOF) for each r calculated from $(d+m)r$ or dm. The black dashed line is the exponential decay. The horizontal colored dashed lines represent the theoretically-optimal generalization loss from Equation 5 of each rank with corresponding color. n_c is calculated from Equation 4.

References

[1] P. V. Dantas, W. Sabino Da Silva, L. C. Cordeiro, and C. B. Carvalho. A Comprehensive Review of Model Compression Techniques in Machine Learning. *Applied Intelligence*, 54(22):11804–11844, Nov. 2024.

[2] M. Denil, B. Shakibi, L. Dinh, M. Ranzato, and N. de Freitas. Predicting Parameters in Deep Learning, Oct. 2014, arXiv:1306.0543.

[3] 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.

[4] C. Eckart and G. Young. The approximation of one matrix by another of lower rank. *Psychometrika*, 1(3):211–218, Sept. 1936.

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

[6] 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.

[7] M. Hisakado and T. Kaneko. Deformation of Marchenko-Pastur distribution for the correlated time series, Nov. 2024, arXiv:2305.12632.

[8] 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.

[9] G. Jin, X. Yi, L. Zhang, L. Zhang, S. Schewe, and X. Huang. How does Weight Correlation Affect the Generalisation Ability of Deep Neural Networks, arXiv:2010.05983.

[10] A. K. Lampinen and S. Ganguli. An Analytic Theory of Generalization Dynamics and Transfer Learning in Deep Linear Networks, arXiv:1809.10374.

[11] R. Livni, S. Shalev-Shwartz, and O. Shamir. On the Computational Efficiency of Training Neural Networks. In *NeurIPS*, volume 27, Montréal, Canada, 2014.

[12] 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.

[13] V. A. Marčenko and L. A. Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457–483, Apr. 1967.

[14] X. Ou, Z. Chen, C. Zhu, and Y. Liu. Low Rank Optimization for Efficient Deep Learning: Making A Balance between Compact Architecture and Fast Training, Mar. 2023, arXiv:2303.13635.

[15] W. Roth, G. Schindler, B. Klein, R. Peharz, S. Tschiatschek, H. Fröning, F. Pernkopf, and Z. Ghahramani. Resource-efficient neural networks for embedded systems. *Journal of Machine Learning Research*, 25(50):1–51, 2024.

[16] M. Soltani and C. Hegde. Towards Provable Learning of Polynomial Neural Networks Using Low-Rank Matrix Estimation. In *Proc. of the 21st AISTATS*. PMLR, Mar. 2018.

[17] 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.