Towards Understanding Low-Rank Learning for Classification

Xiaolin Wang*, Olivier Rioul*, Anissa Mokraoui† and Pierre Duhamel‡
*LTCI, Télécom Paris, Institut Polytechnique de Paris, Palaiseau, France
Email: firstname.lastname@telecom-paris.fr
†L2TI, Université Sorbonne Paris Nord, Villetaneuse, France
Email: anissa.mokraoui@univ-paris13.fr
‡L2S, CentraleSupélec, Université Paris-Saclay, Gif-sur-Yvette, France
Email: pierre.duhamel@centralesupelec.fr

***Abstract*—Understanding the sampling complexity of a neural network—the number of training examples it needs to generalize—is a fundamentally unresolved question, especially for modern architectures that use low-rank constraints for compression and fine-tuning. Although low-rank methods such as LoRA are widely adopted in practice, their effect on sampling complexity for classification remains poorly understood. In this work, we investigate shallow ReLU networks with low-rank constrained weights in a teacher-student context. It is shown that the inherent structure of classification tasks is a prerequisite for successful low-rank learning. A distinctive three-stage sample complexity curve is observed : (i) initial phase of random guessing; (ii) abrupt transition to rapid improvement once a critical sample size is reached; and (iii) final plateau whose height depends on the network's rank. It is shown that the critical sample size is determined by the effective task structure rather than by the raw dimension of the input data, and that the rank constraint limits the maximum achievable accuracy. Simulations on synthetic data validate our predictions, revealing a rigorous trade-off between compression and accuracy. This can guide rank selection in low-rank learning and constitutes a crucial step toward understanding low-rank training in more complex architectures and real-world contexts.**

***Index Terms*—Sample complexity, low-rank learning, feature learning, scaling laws, neural networks.**

I. Introduction

Modern neural networks generalize remarkably well despite massive overparameterization—a behavior that classical measures such as Vapnik–Chervonenkis (VC) dimension fail to explain [1], [2]. This puzzle has motivated the study of neural scaling laws [3], [4], revealing surprising phenomena such as grokking [5], [6]. To explain the criteria of generalization, shallow regression networks with quadratic activations optimizing Mean Squared Error (MSE) have been favored by researchers as they facilitate calculations. References [7], [8] established sharp sample complexity thresholds separating memorization from generalization. In [9], we extended them to soft sample complexity curves, enabling the analysis of low-rank models and structured tasks. However, the assumed quadratic activation reduced learning to matrix sensing, missing the nonlinear dynamics of ReLU networks [10], which are more common in practice.

A key distinction for understanding generalization is between *lazy learning*, where the network behaves as a fixed random feature model [11], and *feature learning*, where the network adapts its internal representations to the task [12], [13]. Recently, [14] proposed a scaling argument framework predicting the onset of feature learning. By comparing the minimum energy cost of three candidate strategies (lazy regime, global feature rotation, and neuron specialization), they predicted sample complexity exponents without solving the full dynamics. However, it was developed for MSE on full-scale scalar networks and does not account for models with limited capacity.

In parallel, low-rank methods have become a cornerstone of practical deep learning. Techniques such as Low-Rank Adaptation (LoRA) [15], matrix factorization, and Singular Value Decomposition (SVD)-based weight compression have been proven effective [16], [17]. Theoretically, matrix factorization biases gradient descent toward low-rank solutions through implicit regularization [18], and generalization bounds tighten when weight matrices have bounded rank [19]. For ReLU networks, [20] showed that the optimal low-rank approximation differs fundamentally from standard SVD by Frobenius-norm truncation. Most low-rank methods operate in a *post-training* regime. By contrast, our work addresses *training-time* compression, where the low-rank constraint is imposed from initialization, a setting in which the interplay between rank, task structure, and sample efficiency is less understood.

Our present analysis connects both perspectives by studying how task structure and model compression jointly shape sample complexity in ReLU classification networks. We adopt a teacher-student framework where the teacher's weights have tunable spectral structure—modeled by auto-regressive (AR) correlation—and the student learns through a low-rank factorized weight matrix, isolating two sources of dimensional reduction: intrinsic task complexity and student's rank constraint. The central question is the following: When the model is deliberately compressed, with certain samples, how does structured data change the onset of learning and the best achievable accuracy? Our main contributions are as follows:

a) Three-Stage Phase Transition: We identify a robust three-stage sample complexity curve in low-rank ReLU classification networks: (i) a random guessing phase, (ii) rapid feature learning once a critical number of samples N_c is

exceeded, and (iii) a saturation phase where accuracy is capped by the student's rank. Unlike the sharp threshold observed in quadratic models [7], [8], the rank constraint imposes a soft ceiling on the accuracy that the student can ultimately achieve.

b) Scaling Analysis with Effective Dimension: Extending the scaling argument framework of [14] to low-rank networks on structured classification tasks, we derive that the critical sample size scales as $N_c \propto r_{\min} \cdot d$, where $r_{\min} = \min(r, r_{\text{eff}})$ is the smaller of the student rank r and the task's effective dimensionality r_{eff}. For structured tasks, r_{eff} is much smaller than the input dimension d, so learning begins with significantly fewer samples. We also derive an upper bound on the achievable accuracy determined by the spectral energy captured in the top modes.

c) Implicit Spectral Regularization: We show that the low-rank factorization steers gradient descent toward the most important features of the task via a positive feedback loop, explaining the sharp onset of feature learning observed in several experiments.

II. Problem Formulation

A *shallow neural network* (or single hidden-layer network) with ReLU activation computes a function $f : \mathbb{R}^d \to \mathbb{R}^K$ of the form:

$$f(\boldsymbol{x}; \boldsymbol{W}, \boldsymbol{A}) = \boldsymbol{A}\,\sigma(\boldsymbol{W}\boldsymbol{x}), \tag{1}$$

where $\boldsymbol{x} \in \mathbb{R}^d$ is the input vector, $\boldsymbol{W} \in \mathbb{R}^{m\times d}$ is the hidden layer weight matrix, $\boldsymbol{A} \in \mathbb{R}^{K\times m}$ is the output layer weight matrix, m is the number of hidden neurons, K is the output dimension (number of classes), and $\sigma(\cdot) = \max(0, \cdot)$ denotes the element-wise ReLU activation function. For multi-classification tasks, the network output is passed through a softmax layer to produce class probabilities.

In the teacher-student framework, a teacher network $f_T(\boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{W}^*, \boldsymbol{A}^*)$ with fixed, unknown weights $\boldsymbol{W}^* \in \mathbb{R}^{m^*\times d}$ and $\boldsymbol{A}^* \in \mathbb{R}^{K\times m^*}$ generates the labels, and a student network $f_S(\boldsymbol{x}) = f(\boldsymbol{x}; \boldsymbol{W}, \boldsymbol{A})$ is trained to recover this mapping from samples. The input data follows a standard Gaussian distribution. The student is given N training samples $\mathcal{D} = \{(\boldsymbol{x}_i, y_i)\}_{i=1}^N$, where y_i is generated by the teacher network. The student is trained by minimizing the cross-entropy loss. The test (generalization) accuracy is measured on fresh samples. Equivalently, the test accuracy is $1 - \epsilon_{\text{test}}$. The *saturation accuracy* $\text{Acc}_\infty = \lim_{N\to\infty} \text{Acc}(N)$ represents the best achievable performance given the student's structure.

Instead of learning the full weight matrix $\boldsymbol{W} \in \mathbb{R}^{m\times d}$ directly, we impose a low-rank constraint through factorization for the purpose of training-time compression [20]: $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{V}^\top$, where $\boldsymbol{U} \in \mathbb{R}^{m\times r}$ and $\boldsymbol{V} \in \mathbb{R}^{d\times r}$ are the learnable factor matrices, and $r < \min(m, d)$ is the rank constraint. This parametrization reduces the number of parameters in the first layer from md to $r(m+d)$ and induces an implicit bias toward learning the dominant spectral components first [18]. As a trade-off, it places the student in a capacity-limited regime where the alignment with the teacher is bounded away from being perfect, which is intentional in compression-oriented settings such as Principal Component Analysis (PCA), SVD, tensor decomposition, and LoRA.

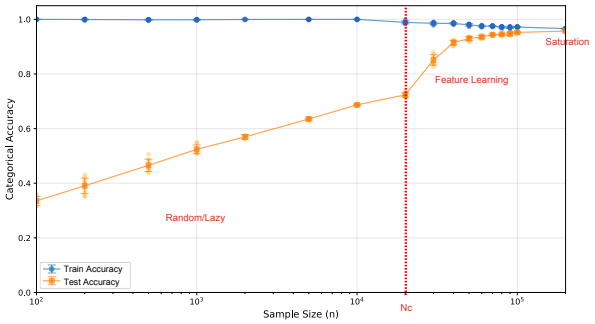

Fig. 1. Example of three-stage phase transition in test accuracy as a function of sample size N. The curve exhibits: (I) a random/lazy regime with near-chance accuracy, (II) a rapid feature learning regime following the critical sample size N_c, and (III) a saturation regime.

Natural data—images, audio, text—exhibits spectral structure where a small number of modes capture most of the energy, as reflected in the power-law eigenvalue distributions of empirical covariance matrices [4]. The AR(1) model is the simplest stationary process that produces geometrically decaying eigenvalues with a single parameter $\rho \in [0, 1)$, continuously interpolating between isotropic ($\rho = 0$) and highly structured ($\rho \to 1$) tasks. This makes it an interpretable testbed for studying the effect of task structure on sample complexity, compared for instance to ad-hoc spectrum shaping. We assume the teacher's weight matrix $\boldsymbol{W}^*$ has correlated entries following a separable 2D AR(1) process with parameter ρ controlling the correlation length. Each entry satisfies $(\text{Cov}(\boldsymbol{W}^*))_{ij,kl} = \rho^{|i-k|} \cdot \rho^{|j-l|}$. The spectral concentration of the teacher is quantified by the *effective rank* of the column covariance $\boldsymbol{\Sigma}_{\text{col}}$ (with entries $(\boldsymbol{\Sigma}_{\text{col}})_{jl} = \rho^{|j-l|}$), defined via the participation ratio $r_{\text{eff}} = (\operatorname{tr}\boldsymbol{\Sigma}_{\text{col}})^2 / \operatorname{tr}\boldsymbol{\Sigma}_{\text{col}}^2$. Since $\operatorname{tr}\boldsymbol{\Sigma}_{\text{col}} = d$ and $\operatorname{tr}\boldsymbol{\Sigma}_{\text{col}}^2 = \sum_{j,l} \rho^{2|j-l|} \approx d(1+\rho^2)/(1-\rho^2)$ for large d, we obtain the closed-form expression:

$$r_{\text{eff}}(\rho) \approx d\,\frac{1-\rho^2}{1+\rho^2}. \tag{2}$$

III. Analysis of Sample Complexity

We propose an analysis framework that explains the three-stage phase transition of the test error as a function of sample size. Our analysis focuses on the interplay between the structural constraints of the student network and the spectral properties of the teacher's weight configuration.

A. Empirical Observation: Three-Stage Phase Transition

Fig. 1 illustrates the distinctive three-stage phase transition observed in our experiments. As sample size N increases, the test accuracy evolves through: (i) a flat random guessing regime where the student struggles to extract signal, (ii) a sharp "turn" from a critical sample size N_c marking the onset of feature learning, and (iii) a saturation plateau determined by the student's rank constraint. This empirical observation motivates the analysis that follows.

B. Scaling Analysis of Critical Sample Size

To explain the observed phase transition, we adapt the scaling argument framework of [14]—originally developed for regression tasks with full-rank models and MSE loss—to our low-rank learning classification setting. The key idea is that the onset of feature learning corresponds to the sample size at which it becomes energetically favorable for the student to align its column space span($\boldsymbol{V}$) with the principal eigenspace of $\boldsymbol{W}^*$, rather than remaining in the random initialization regime. Following [14], the critical sample size $P_\star$ is proportional to the minimum variational energy $\tilde{E}_q(\alpha)$ required to achieve alignment α with the teacher: $P_\star \propto \min_q \tilde{E}_q(\alpha) = \min_q \sum_{l,i} \Delta_{l,i}(q) + \langle y, \mathcal{K}^{-1}, y \rangle$, where $\Delta_{l,i}$ measures the deviation of q from the initialization prior and $\mathcal{K}$ is the induced kernel at the penultimate layer. Compared with a lazy Gaussian-process regime (cost $\asymp d^2$) and unspecialized Gaussian feature learning (cost $\asymp (m^2 r^2 d^2)^{1/3}$), *neuron specialization*—where M student neurons each acquire a unit-magnitude feature direction inside span($\boldsymbol{V}$)—dominates whenever the saturation alignment $A_{\text{sat}}(r) \equiv \alpha_{\max}(r)$ is bounded away from zero. Each specialized neuron pays $\Delta = d/2$ (not discounted by the rank constraint), while M such neurons inject a kernel spike of size M/m yielding target overlap $(M/m)\, A_{\text{sat}}^2(r)$. Collecting the two competing terms,

$$N_c \propto \tilde{E}_{\text{sp}}^\star \asymp M \cdot d \;+\; \frac{m}{K\, M\, A_{\text{sat}}^2(r)}, \quad M \le r. \tag{3}$$

The unconstrained saddle sits at $M^\star \asymp \sqrt{m/(K\, d\, A_{\text{sat}}^2(r))}$; the rank cap binds whenever $r \lesssim M^\star$, in which case $M = r_{\min} = \min(r, r_{\text{eff}})$. Although this energy was derived for MSE regression, the cost terms count geometric degrees of freedom that are independent of the loss function; we therefore expect the scaling exponents to carry over to classification, which our experiments confirm. When the linear-in-rd term dominates ($K\, r_{\min}^2\, d\, A_{\text{sat}}^2(r) \gg m$), the headline scaling reads $N_c \approx C \cdot r_{\min} \cdot d$.

This scaling prediction reveals two distinct compression effects: **1) Task compression** ($r_{\text{eff}} < d$): structured tasks reduce $r_{\min}$ via spectral concentration (cf. Eq. (2)), lowering N_c even for full-rank students. **2) Model compression** ($r < r_{\text{eff}}$): the student reaches its attainable performance with fewer samples ($N_c \propto r \cdot d$), but $\alpha_{\max}$ is bounded below 1. The fraction of teacher variance captured by the top-r eigenmodes defines the maximum achievable alignment:

$$\alpha_{\max} \approx \frac{\sum_{k=1}^{r} \lambda_k^*}{\sum_{k=1}^{d} \lambda_k^*}, \tag{4}$$

where λ_k^* are the eigenvalues of $\boldsymbol{W}^{*\top}\boldsymbol{W}^*$ in decreasing order. For teachers with large ρ, the spectral energy concentrates in the top modes, so $\alpha_{\max}$ is close to 1 even for moderate r.

C. Mechanism: Online Spectral Alignment

The factorization $\boldsymbol{W} = \boldsymbol{U}\boldsymbol{V}^\top$ induces dynamics that naturally decompose the teacher's spectrum. Under gradient descent, the gradient of $\boldsymbol{V}$ satisfies $\dot{\boldsymbol{V}} \propto (\boldsymbol{U}^\top\boldsymbol{U})\nabla_{\boldsymbol{W}}\mathcal{L}$, where the prefactor $\boldsymbol{U}^\top\boldsymbol{U}$ amplifies directions where the student already has significant weight mass. This creates a positive feedback loop: directions aligned with the teacher's dominant eigenmodes receive stronger gradients, accelerating their learning. Combining the AR(1) teacher (signal concentrated in few modes) with the low-rank student (designed to find modes), the "turn" in the accuracy curve corresponds to the student's singular vectors locking onto the teacher's first principal component. The subsequent rapid rise represents sequential discovery of the remaining top-r components.

D. Saturation Analysis

We characterize the asymptotic error floor in the saturation regime by adopting the nonlinear low-rank approximation (NLRA) of [20]. Let $\boldsymbol{h} = \sigma(\boldsymbol{W}^*\boldsymbol{x})$ and $\hat{\boldsymbol{h}} = \sigma(\hat{\boldsymbol{W}}\boldsymbol{x})$ denote the hidden representations of the teacher and rank-r student, respectively. For Gaussian inputs, the hidden-layer MSE decomposes neuron-by-neuron via the first-order arc-cosine kernel. At the optimal rank-r approximation $\hat{\boldsymbol{W}}$, the error takes the form:

$$\mathcal{E}_{\text{hidden}} = \mathbb{E}\left[\|\boldsymbol{h} - \hat{\boldsymbol{h}}\|^2\right] = \tfrac{1}{2}\sum_{i=1}^{m^*} \|\boldsymbol{w}_i^*\|^2 \big(1 - h(p_i^*)\big), \tag{5}$$

where $\boldsymbol{w}_i^*$ is the i-th row of $\boldsymbol{W}^*$, $p_i^* \in [0,1]$ is the cosine similarity between $\boldsymbol{w}_i^*$ and its projection onto the optimal r-dimensional subspace selected by the NLRA, and $h(p) = \big[\frac{\sqrt{1-p^2} + (\pi - \arccos p)\, p}{\pi}\big]^2$ arises from the ReLU arc-cosine kernel [21]. To connect (5) to the alignment bound (4), consider the (suboptimal) choice of the top-r SVD directions of $\boldsymbol{W}^*$. The weighted average correlation satisfies $\sum_i \|\boldsymbol{w}_i^*\|^2 p_i^2 / \|\boldsymbol{W}^*\|_F^2 = \alpha_{\max}$, so $h(p_i) \ge p_i^2$ yields:

$$\mathcal{E}_{\text{hidden}} \le \tfrac{1}{2}(1 - \alpha_{\max})\|\boldsymbol{W}^*\|_F^2. \tag{6}$$

The logit error $\Delta\boldsymbol{z} = \boldsymbol{A}^*(\boldsymbol{h} - \hat{\boldsymbol{h}})$ then satisfies $\mathbb{E}[\|\Delta\boldsymbol{z}\|^2] \le \|\boldsymbol{A}^*\|_{\text{op}}^2 \cdot \mathcal{E}_{\text{hidden}}$, and since the softmax is locally Lipschitz, the classification error floor vanishes as $\alpha_{\max} \to 1$. For AR(1) teachers with large ρ, spectral energy concentrates in the top modes, making $\alpha_{\max}$ close to 1 even for moderate r and yielding a low error floor, as confirmed in Fig. 4.

E. Implications for Low-Rank Network Design

The scaling law (3) and alignment bound (4) together reveal a compression–accuracy trade-off that can guide rank selection. Reducing the student rank r lowers N_c (faster convergence) but caps $\alpha_{\max}$ (higher error floor). In structured tasks, the effective rank r_{eff} sets a practical target: once $r \gtrsim r_{\text{eff}}$, further increases in r no longer raise sample complexity (which is capped by $r_{\text{eff}} \cdot d$) and only add computational cost. This yields a clear operating point: small r trades accuracy for speed, while $r \approx r_{\text{eff}}$ achieves near-optimal accuracy without incurring a sample complexity penalty.

IV. Simulation

We validate our scaling predictions through numerical experiments on synthetic teacher-student tasks. We consider a multi-class ($K = 10$) classification task. The teacher has

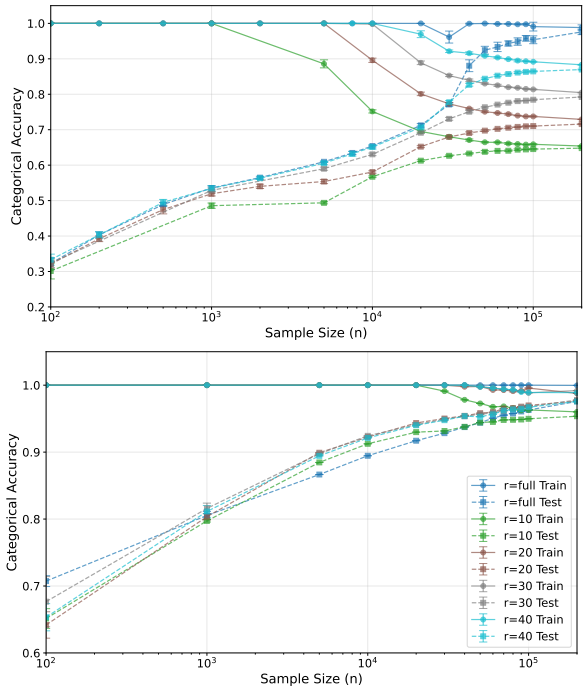

Fig. 2. Classification accuracy vs. sample size for unstructured teacher ($\rho = 0$, top) and structured teacher ($\rho = 0.9$, bottom). The structured teacher enables faster learning and higher saturation accuracy for low-rank students.

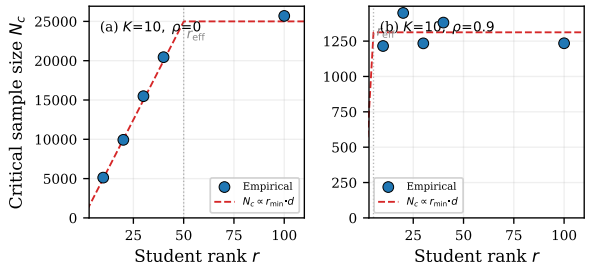

Fig. 3. Critical sample size N_c vs. student rank r. (a) Unstructured teacher ($\rho = 0$): N_c scales linearly for $r \leq r_{\text{eff}}$. (b) Structured teacher ($\rho = 0.9$): N_c is approximately constant since all $r > r_{\text{eff}}$.

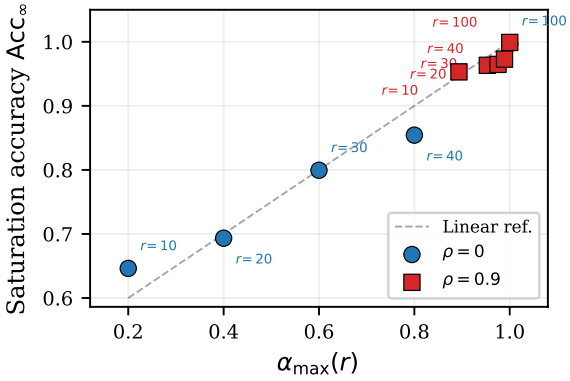

Fig. 4. Saturation accuracy Acc_∞ vs. theoretical alignment $\alpha_{\max}(r)$ for $\rho = 0$ (circles) and $\rho = 0.9$ (squares). Each point is labeled with the corresponding student rank r. The dashed line shows a linear reference.

$m^* = 50$ hidden neurons with input dimension $d = 100$. The teacher's weight matrix $\boldsymbol{W}^*$ is generated from the 2D AR(1) process with $\rho \in \{0, 0.9\}$. The factorized students use rank $r \in \{10, 20, 30, 40\}$. Training uses the Adam optimizer with learning rate 10^{-3} for 5×10^4 epochs, ensuring optimization convergence. Results are averaged over 10 independent runs with different random seeds on the same teacher.

Fig. 2 compares the sample complexity curves for unstructured ($\rho = 0$) and structured ($\rho = 0.9$) teachers in the 10-class setting. For both ρ values, lower-rank students saturate at lower accuracy levels, confirming that the asymptotic error floor $\epsilon_\infty \approx \sum_{k>r} \lambda_k^*$ depends on the spectral tail energy. With the AR-correlated teacher ($\rho = 0.9$), the critical sample size N_c is significantly reduced compared to the isotropic case ($\rho = 0$). Moreover, low-rank students achieve higher saturation accuracy when $\rho = 0.9$, because the AR spectrum concentrates energy in fewer modes that can be captured by the rank-r approximation. This is consistent with our scaling law (3): increasing ρ reduces r_{eff} and hence $r_{\min}$, lowering N_c. Note that for $\rho = 0.9$ the phase transition is less abrupt than for $\rho = 0$: the discrete nature of classification labels smooths the onset of feature learning compared to regression, as the margin-based tolerance allows partially aligned models to already classify correctly.

Fig. 3 directly validates the scaling prediction (3). For $\rho = 0$, the critical sample size N_c grows approximately linearly in r up to $r_{\text{eff}} = 50$, beyond which it saturates. For $\rho = 0.9$, all tested ranks exceed $r_{\text{eff}} \approx 5.2$, so N_c remains flat. Fig. 4 plots the empirical saturation accuracy against the theoretical alignment $\alpha_{\max}(r)$. For both ρ values, the points follow a monotonically increasing trend, confirming that $\alpha_{\max}$ is a reliable predictor of saturation performance. The $\rho = 0.9$ points cluster at high $\alpha_{\max}$ even for small r, reflecting the spectral concentration of the AR teacher. Overall, both scaling predictions—$N_c \propto r_{\min} \cdot d$ and $\text{Acc}_\infty \sim \alpha_{\max}(r)$—are consistently validated across different initializations and rank constraints, confirming that the interplay between task structure and model capacity fundamentally governs sample complexity in low-rank ReLU networks.

To test whether our findings generalize to non-Gaussian, real-world inputs, we replicate the experiment on MNIST ($d = 784$, $K = 10$) with a teacher of 50 hidden neurons (test accuracy 97.5%, $r_{\text{eff}} \approx 25$). We train low-rank students and full-rank baselines for varying sample sizes. Fig. 5(a) shows smooth learning curves consistent with the synthetic setting: the three-stage behavior is present, though the phase transition is softened by the classification margin's tolerance to partial spectral alignment. Fig. 5(b) validates the key quantitative prediction of Section III on real data: Acc_∞ increases near-linearly with $\alpha_{\max}(r)$, mirroring the synthetic results of Fig. 4. Notably, even $r = 5$ reaches 87% accuracy, well above what the spectral energy fraction alone would predict. At $r = 20$, the student achieves 94.2%, approaching the teacher's accuracy, consistent with $r \lesssim r_{\text{eff}} \approx 25$ being the practical operating point. This confirms the margin-based tolerance of classification and shows that $\alpha_{\max}$ serves as a quantitative

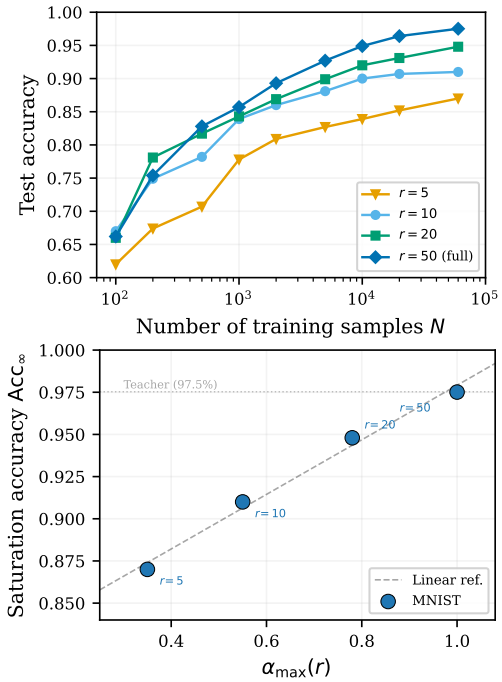

Fig. 5. MNIST results for rank-constrained student networks ($m = 50$), where $r = 50$ is effectively full-rank. (a) Test accuracy vs. training sample size N for student ranks $r \in \{5, 10, 20, 50\}$. (b) Saturation accuracy Acc_∞ vs. theoretical alignment $\alpha_{\max}(r)$. The dashed line shows the best linear fit, confirming that $\alpha_{\max}$ reliably predicts saturation performance on real data, consistent with the synthetic results.

predictor of achievable accuracy in low-rank learning on real data.

V. Conclusions

We have studied sample complexity in low-rank factorized ReLU shallow networks learning from teachers with AR-correlated weights. Our main findings are threefold. First, a robust three-stage phase transition (random guessing $\rightarrow$ feature learning $\rightarrow$ rank-limited saturation) governs test accuracy as a function of sample size. Second, the critical sample size scales as $N_c \propto r_{\min} \cdot d$, where $r_{\min} = \min(r, r_{\text{eff}})$ captures the bottleneck between model capacity and task complexity; AR correlation reduces r_{eff} and accelerates learning. Third, low-rank factorization acts as an implicit spectral regularizer, inducing online alignment with the teacher's dominant eigenmodes. These results provide an analytical lens for understanding why low-rank training remain effective: structured tasks have low effective dimension, and rank-constrained models automatically discover the dominant features with correspondingly fewer samples. An MNIST experiment further confirms that $\alpha_{\max}(r)$ reliably predicts saturation on real data, and that the margin-based tolerance of classification enables low-rank models to achieve near-full-rank performance even with limited spectral alignment. Extending the spectral alignment analysis to deep networks is non-trivial because per-layer alignment dynamics couple across depth through products of layer-wise Jacobians, and lazy and feature-learning regimes may coexist at different layers, each with its own effective rank and critical sample size. In standard post-training LoRA, by contrast, the rank constraint acts on residual updates around a frozen backbone, so the critical sample size is still governed by the effective dimension of the residual task while the saturation floor is set by how well the pre-trained features already span the teacher's target subspace rather than by the LoRA rank alone. Future work includes validating and extending these predictions on deeper architectures, other activation functions, and contemporary real-world benchmarks.

References

[1] A. Daniely and S. Shalev-Shwartz, "Optimal learners for multiclass problems," in *Proc. Conf. Learn. Theory*, May 2014.

[2] Y. Wang and C. Scott, "VC dimension of partially quantized neural networks in the overparametrized regime," in *Proc. ICLR*, Oct. 2021.

[3] J. Kaplan, S. McCandlish, T. Henighan, T. B. Brown, B. Chess, R. Child, S. Gray, A. Radford, J. Wu, and D. Amodei, "Scaling laws for neural language models," 2020, arXiv:2001.08361.

[4] Y. Bahri, E. Dyer, J. Kaplan, J. Lee, and U. Sharma, "Explaining neural scaling laws," *Proc. Nat. Acad. Sci.*, vol. 121, no. 27, p. e2311878121, Jul. 2024.

[5] A. Power, Y. Burda, H. Edwards, I. Babuschkin, and V. Misra, "Grokking: Generalization beyond overfitting on small algorithmic datasets," presented at the MATH-AI Workshop, ICLR, Jan. 2022.

[6] N. Rubin, I. Seroussi, and Z. Ringel, "Grokking as a first order phase transition in two layer networks," in *Proc. ICLR*, May 2024.

[7] S. S. Mannelli, E. Vanden-Eijnden, and L. Zdeborová, "Optimization and generalization of shallow neural networks with quadratic activation functions," in *Proc. NeurIPS*, vol. 33, Dec. 2020, pp. 13 445–13 455.

[8] D. Gamarnik, E. C. Kızıldağ, and I. Zadik, "Stationary points of a shallow neural network with quadratic activations and the global optimality of the gradient descent algorithm," *Math. Operations Res.*, vol. 50, no. 1, pp. 209–251, Feb. 2025.

[9] X. Wang, O. Rioul, A. Mokraoui, P. Duhamel, and J. Benesty, "Generalizability and sample complexity of quadratic shallow neural networks under low-rank learning," in *GRETSI*, Aug. 2025, pp. 941–944.

[10] S. Akiyama and T. Suzuki, "On learnability via gradient method for two-layer relu neural networks in teacher-student setting," in *Proc. ICML*, vol. 139, Jul. 2021, pp. 152–162.

[11] A. Jacot, F. Gabriel, and C. Hongler, "Neural tangent kernel: Convergence and generalization in neural networks," in *Proc. NeurIPS*, vol. 31, 2018.

[12] S. Goldt, M. Advani, A. M. Saxe, F. Krzakala, and L. Zdeborová, "Dynamics of stochastic gradient descent for two-layer neural networks in the teacher-student setup," in *Proc. NeurIPS*, vol. 32, 2019.

[13] D. Saad and S. Solla, "Dynamics of on-line gradient descent learning for multilayer neural networks," in *Proc. NIPS*, vol. 8, 1995.

[14] N. Rubin, O. Davidovich, and Z. Ringel, "Mitigating the curse of detail: Scaling arguments for feature learning and sample complexity," Dec. 2025, arXiv:2512.04165.

[15] E. J. Hu, Y. Shen, P. Wallis, Z. Allen-Zhu, Y. Li, S. Wang, L. Wang, and W. Chen, "LoRA: Low-rank adaptation of large language models," in *Proc. ICLR*, Oct. 2021.

[16] 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," *J. Syst. Eng. Electron.*, vol. 35, no. 3, pp. 509–531, Jun. 2024.

[17] L. Balzano, T. Ding, B. D. Haeffele, S. M. Kwon, Q. Qu, P. Wang, Z. Wang, and C. Yaras, "An overview of low-rank structures in the training and adaptation of large models," Feb. 2026, arXiv:2503.19859.

[18] S. Gunasekar, B. E. Woodworth, S. Bhojanapalli, B. Neyshabur, and N. Srebro, "Implicit regularization in matrix factorization," in *Proc. NIPS*, vol. 30, 2017.

[19] A. Pinto, A. Rangamani, and T. A. Poggio, "On generalization bounds for neural networks with low rank layers," in *Proc. 36th Int. Conf. Algorithmic Learn. Theory*, Feb. 2025, pp. 921–936.

[20] K. Vodrahalli, R. Shivanna, M. Sathiamoorthy, S. Jain, and E. H. Chi, "Nonlinear initialization methods for low-rank neural networks," May 2022, arXiv:2202.00834.

[21] Y. Cho and L. Saul, "Kernel methods for deep learning," in *Proc. NIPS*, vol. 22, 2009.