FACTORIZATION DEPTH AS A REGULARIZATION DIMENSION FOR LORA

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing deep low-rank formulations largely treat depth as a fixed design choice. We investigate whether depth can instead serve as an another regularization dimension for Low-Rank Adaptation (LoRA), alongside trainable parameter count and structural-rank. To characterize this role, we analyze a mergeable L-factor parameterization, with L controlling factorization depth, through its representation cost, induced optimization geometry, and idealized singular-mode dynamics. These analyses motivate a finite-horizon mechanism hypothesis: deeper factorization may selectively delay weak singular modes, promoting spectral concentration at moderate depths but hindering task fitting when taken too far. We test this hypothesis through four complementary approaches: capacity controls, depth ablations, spectral diagnostics, and synthetic experiments. Across these experiments, the resulting accuracy–rank–loss patterns support the hypothesized trade-off between spectral concentration and task fit, rather than a monotonic benefit from increasing depth. Finally, to probe the mechanism behind this effect, we show that a nuclear-norm intervention recovers only part of the performance gain observed at moderate depths, suggesting that the depth effect is not fully explained by explicit nuclear-norm regularization in the tested settings. Overall, our results support viewing factorization depth as a promising, non-monotonic regularization dimension for LoRA, while motivating further study of whether this phenomenon generalizes across optimizers, backbone models, and tasks.

1 INTRODUCTION

Parameter-efficient fine-tuning (PEFT) adapts pretrained models through a small trainable parameter set (Houlsby et al., 2019; Aghajanyan et al., 2021). Low-Rank Adaptation (LoRA; Hu et al., 2022) is especially attractive because its update is mergeable at inference. Existing variants modify mainly rank allocation, scaling, initialization, decomposition, or optimization (Zhang et al., 2023; Kalajdzievski, 2023; Hayou et al., 2024; Liu et al., 2024b; Meng et al., 2024). The factorization depth is another degree of freedom: the same structural-rank bound can be represented by products with different numbers of trainable factors.

The closest prior work studies compressible dynamics in deep overparameterized low-rank learning and adaptation (Yaras et al., 2024), while deep-linear and matrix-factorization analyzes connect depth to singular-mode learning and parameterization-induced bias (Saxe et al., 2014; Gunasekar et al., 2017; Arora et al., 2019). What remains less clear is whether depth is useful as a controllable LoRA variable, how its effect changes beyond a single deep configuration, and whether its spectral behavior can be reduced to an explicit penalty.

To make this question concrete, we use Depth-Regularized LoRA (DR-LoRA) as an experimental instantiation of depth-controlled LoRA, rather than as a separate task-specific adaptation objective. DR-LoRA represents the update as a static product of L trainable factors, where L denotes the factorization depth, and the product can be merged into the frozen weight. Our analysis proceeds in three stages: we characterize the representation cost of a fixed merged update, derive the induced optimization geometry, and specialize the resulting matrix-level relation to idealized singular-mode dynamics under conditional assumptions. This analysis motivates a finite-horizon hypothesis: deeper factorization may delay weak singular modes and concentrate the learned spectrum, while excessive depth may also delay task-relevant directions and impair task fitting.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

We evaluate this hypothesis through controlled fine-tuning experiments and complementary diagnostics. In the main six-task comparison, the evaluated DR-LoRA configuration outperforms parameter-count-matched and structural-rank-matched LoRA controls under the aligned protocol, while producing updates with lower effective rank. We separately analyze the depth–accuracy relationship on CommonsenseQA (Talmor et al., 2019). Across the tested depths, effective rank decreases but accuracy does not improve monotonically: at larger depths, stronger spectral concentration is accompanied by higher training loss and weaker task fit. A nuclear-norm intervention recovers only part of the performance gain in the favorable depth regime, suggesting that the tested depth effect cannot be fully explained by explicit nuclear-norm regularization. These findings support interpreting factorization depth as a non-monotonic regularization dimension.

Our contributions are threefold: we characterize the representation and optimization geometry induced by factorization depth; demonstrate a non-monotonic depth–generalization trade-off under complementary capacity controls; and show that the tested nuclear-norm intervention does not reproduce the full spectral and accuracy behavior of depth.

2 Related Work

LoRA design space. LoRA research has explored rank allocation, scaling, initialization, weight decomposition, and optimization (Zhang et al., 2023; Kalajdzievski, 2023; Hayou et al., 2024; Liu et al., 2024b; Meng et al., 2024; Wang et al., 2025a). Composition and subspace methods provide additional flexibility (Xia et al., 2024; Lialin et al., 2024; Wu et al., 2024; Wang et al., 2025b; He et al., 2025). These methods establish a broad adapter-design space; our focus is the number of factors in a single static, mergeable product.

Deep factorization. The most closely related study is the compressible-dynamics analysis of deep overparameterized low-rank learning and adaptation by Yaras et al. (2024). Our emphasis is complementary: rather than evaluating depth only as a fixed adapter design, we vary it as a regularization dimension and characterize its non-monotonic behavior under capacity-aware comparisons and explicit spectral regularization. The analysis builds on deep-linear and matrix-factorization results for singular-mode dynamics and implicit bias (Saxe et al., 2014; Gunasekar et al., 2017; Arora et al., 2019; Lyu and Li, 2020), together with Schatten factorization identities (Shang et al., 2016) and cautions against reducing implicit bias to norm minimization (Razin and Cohen, 2020).

Spectral structure. Direction selection, spectral clipping, full-rank adaptation, and related subspace methods emphasize that performance depends on how update directions are used, not only on the hard rank bound (Xiao et al., 2026; Wi and Park, 2026; Albert et al., 2025). We use entropy-based effective rank (Roy and Vetterli, 2007) as a scale-invariant measure of the distribution of singular-value mass in the learned update. Unlike structural rank, which describes nominal representational capacity, effective rank captures how many spectral directions are actually used by the learned update. In this paper, it serves as a descriptive diagnostic for depth-dependent optimization and spectral concentration.

3 Factorization Depth and Optimization Geometry

We develop the analysis in three stages. First, we characterize the minimum factor energy required to represent a fixed merged update, which reveals how factorization depth changes the representation geometry. Second, we derive a general first-order identity that maps factor-wise gradient updates to motion of the merged matrix. Third, under aligned and balanced factors with squared loss, we specialize this matrix-level identity to mode-wise spectral dynamics, showing how depth can selectively delay sufficiently weak singular modes. This ordering distinguishes representation geometry from optimization dynamics and makes explicit the additional assumptions used in the spectral specialization. In particular, the variational result alone does *not* imply that unregularized training minimizes a spectral penalty.

Let ΔW denote the learned update and $s > 0$ its fixed scaling; set $X = \Delta W/s$. DR-LoRA parameterizes X as

$$X = W_1 \cdots W_L, \quad W_\ell \in \mathbb{R}^{r_{\ell-1} \times r_\ell}, \quad r_0 = d_{\text{out}},\ r_L = d_{\text{in}}, \tag{1}$$

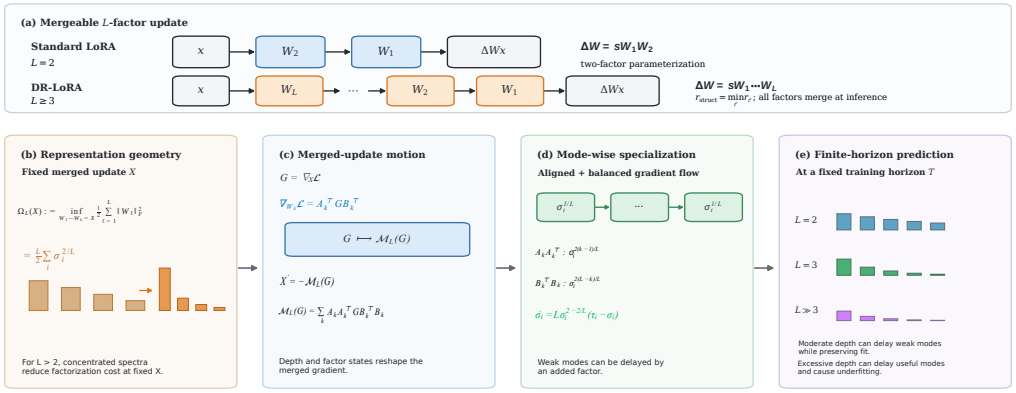


Figure 1: Overview of the factorization-depth analysis. (a) Standard LoRA is the L=2 case of the mergeable product $\Delta W = sW_1 \cdots W_L$; computation applies the factors from W_L to W_1, and all factors merge into one update at inference. (b) For a fixed merged update, depth changes the minimum factor energy and motivates studying spectral concentration, but this variational cost is not the task objective. (c) Factor-wise gradient descent induces the state-dependent merged-update operator $\mathcal{M}_L$, providing the direct optimization link. (d) Under aligned, balanced gradient flow, the operator decouples into singular-mode dynamics and can selectively delay sufficiently small modes. (e) At a fixed training horizon, moderate delay may concentrate the spectrum while preserving task fit, whereas excessive delay may also suppress useful modes and cause underfitting. The arrows summarize the logical chain; only panels (c)–(d) provide the direct dynamical derivation.

where L is the factorization depth, W_ℓ is the ℓ-th trainable factor, and d_{out} and d_{in} are the output and input dimensions. Standard LoRA is the case $L = 2$. The structural-rank bound and trainable parameter count are

$$\operatorname{rank}(X) \le r_{\text{struct}} := \min_{0 \le \ell \le L} r_\ell, \qquad P_L = \sum_{\ell=1}^{L} r_{\ell-1} r_\ell, \tag{2}$$

where r_{struct} is the nominal rank bound and P_L counts the trainable factor parameters. The factors merge into the frozen weight at inference; varying L therefore changes the training parameterization while keeping the merged-update form fixed. Thus r_{struct} describes a nominal representational constraint, whereas L can change how that space is parameterized and traversed during training.

3.1 Representation Cost of Deep Factorization

For a fixed update X, different factorizations can assign different Frobenius norms to the factors. Define the least factor energy

$$\Omega_L(X) := \inf_{W_1 \cdots W_L = X} \frac{1}{2} \sum_{\ell=1}^{L} \|W_\ell\|_F^2. \tag{3}$$

Here $\|\cdot\|_F$ is the Frobenius norm, and $\Omega_L(X)$ is a representation cost rather than the task loss: it measures the smallest total squared factor norm needed to represent an already fixed merged update X.

Proposition 1 (factorization cost). Assume $\min_{1 \le \ell \le L-1} r_\ell \ge \operatorname{rank}(X)$. Then

$$\Omega_L(X) = \frac{L}{2} \sum_i \sigma_i(X)^{2/L}. \tag{4}$$

Interpretation. For $L = 2$, Eq. 4 equals the nuclear norm $\|X\|_* := \sum_i \sigma_i(X)$. For $L > 2$, the concavity of $\sigma^{2/L}$ makes the variational cost more sensitive to spectral concentration, motivating the spectral analysis below. This result does not imply that unregularized AdamW minimizes Ω_L; the proof of Proposition 1 is provided in Appendix D.

3.2 Optimization Geometry: The Induced Update Operator

Let $G = \nabla_X \mathcal{L}$. To isolate the contribution of the k-th factor, write the merged update as $X = A_k W_k B_k$, where $A_k = W_1 \cdots W_{k-1}$ and $B_k = W_{k+1} \cdots W_L$, with empty products equal to identities. Perturbing all factors gives

$$dX = \sum_{k=1}^{L} A_k (dW_k) B_k. \tag{5}$$

Using $d\mathcal{L} = \langle G, dX \rangle_F$ and moving A_k and B_k across the Frobenius inner product yields

$$\nabla_{W_k} \mathcal{L} = A_k^\top G B_k^\top. \tag{6}$$

Thus the gradient of one factor is the merged gradient viewed through the products of all factors on its left and right. This dependence is the first place where depth enters the training dynamics directly.

Proposition 2 (induced update operator). For simultaneous factor gradient descent, $W_k^+ = W_k - \eta \nabla_{W_k} \mathcal{L}$,

$$X^+ - X = -\eta \mathcal{M}_L(G) + O(\eta^2), \quad \mathcal{M}_L(G) = \sum_{k=1}^{L} A_k A_k^\top G B_k^\top B_k. \tag{7}$$

Interpretation. The result follows by expanding the updated product and retaining terms with one factor update; the derivation is given in Appendix D. Unlike direct optimization of X, factorized optimization maps G through the depth- and state-dependent operator $\mathcal{M}_L$, so parameterizations with the same structural-rank bound can follow different merged-update trajectories.

This operator also clarifies why Proposition 1 should not be read as an optimizer-equivalence claim. An explicit nuclear-norm penalty changes the objective by adding a spectral subgradient, whereas deep factorization changes the map from factor gradients to merged-update motion. We now specialize $\mathcal{M}_L$ to a setting in which its action on each singular mode can be computed explicitly.

3.3 Spectral Dynamics: From Weak-Mode Delay to Effective Rank

Finite-horizon mechanism. Factorization depth may change which singular modes are learned within a fixed training budget even when different depths can represent the same target. The prediction therefore concerns transient mode selection: strong modes may be learned while weaker modes remain underdeveloped at the measurement horizon.

Proposition 2 allows both singular-value changes and subspace rotations. For a transparent mode-wise specialization, we follow deep-linear analyses (Saxe et al., 2014; Arora et al., 2019) and assume squared loss, shared singular vectors between the current and target matrices, and aligned, balanced factors. Under ordinary factor gradient flow, each active mode then follows

$$\dot{\sigma}_i = L \sigma_i^{2-2/L} (\tau_i - \sigma_i), \qquad g_L(\sigma) = L \sigma^{2-2/L}. \tag{8}$$

Here a *weak mode* has a small current singular value σ_i, not necessarily a small target τ_i. At the same positive state and target residual, adjacent depths have gain ratio

$$\frac{g_{L+1}(\sigma)}{g_L(\sigma)} = \frac{L+1}{L} \sigma^{2/[L(L+1)]}, \tag{9}$$

so an additional factor slows the mode when

$$\sigma < \theta_L := \left(\frac{L}{L+1} \right)^{L(L+1)/2}. \tag{10}$$

This is a local comparison and does not order complete trajectories with different initializations, scalings, or learning-rate schedules.

For $0 < \sigma_{i,0} < \sigma_i^\star \ll \tau_i$, approximating the early residual by τ_i gives, for $L > 2$,

$$t_i^\star \approx \frac{\sigma_{i,0}^{-(1-2/L)} - (\sigma_i^\star)^{-(1-2/L)}}{(L-2)\tau_i}. \tag{11}$$

For $L = 2$, $t_i^\star \approx (2\tau_i)^{-1} \log(\sigma_i^\star / \sigma_{i,0})$. Thus activation depends polynomially on inverse initialization size for $L > 2$, but only logarithmically for $L = 2$, providing a conditional finite-horizon mechanism for spectral concentration.

Connecting the spectrum to effective rank. For a nonzero update, use the entropy-based effective rank of Roy and Vetterli (2007):

$$p_i = \frac{\sigma_i}{\sum_j \sigma_j}, \quad H(p) = -\sum_i p_i \log p_i, \quad \operatorname{erank}(\Delta W) = \exp H(p) = \operatorname{erank}(X). \tag{12}$$

Unlike structural rank, this scale-invariant quantity measures the learned distribution of singular-value mass. Its dynamics satisfy

$$\frac{d}{dt} \log \operatorname{erank}(X) = -\operatorname{Cov}_p(a_i, \log p_i), \qquad a_i = L\sigma_i^{1-2/L}(\tau_i - \sigma_i). \tag{13}$$

Effective rank decreases when high-mass modes grow faster in relative terms than tail modes. Because growth also depends on the target residual, monotonic rank reduction is not guaranteed; the derivation and sufficient conditions are provided in Appendix D.

Depth–generalization trade-off and scope. Moderate depth may delay weak modes while preserving task fit, whereas excessive depth may also delay task-relevant directions, raise training loss, and reduce accuracy. This motivates the effective-rank diagnostics, depth sweep, synthetic experiment, and nuclear-norm intervention, but does not determine a universal optimal depth. The analysis assumes ordinary factor GD and, for the mode-wise specialization, aligned and balanced gradient flow with positive active modes. Our AdamW initialization is initially unbalanced; the analysis is therefore a conditional mechanism hypothesis rather than a Transformer training law.

4 Experimental Setup

Backbone and tasks. We evaluate on LLaMA-2-7B (Touvron et al., 2023) using MRPC, CoLA, and RTE from GLUE (Wang et al., 2019), plus ScienceQA (Lu et al., 2022), CommonsenseQA (Talmor et al., 2019), and OpenBookQA (Mihaylov et al., 2018). The mix spans sentence-level NLU and multiple-choice QA, allowing us to test whether the depth pattern is tied to one task format rather than to make an official leaderboard claim.

Adapter and training protocol. Adapters are inserted only into `q_proj` and `v_proj`. Unless noted otherwise, all methods use AdamW, cosine decay, learning rate 2×10^{-4}, batch size 4, and 4 gradient-accumulation steps. The six tasks are combined into one training mixture and jointly fine-tuned for 3 epochs. Backbone, data mixture, schedule, adapter placement, and evaluation are fixed; gradient clipping is not used.

Comparison philosophy. The default DR-LoRA uses $L = 3$, $r_1 = 12$, $r_2 = 4$, and $\alpha = 8$, giving $r_{\text{struct}} = 4$, $s = \alpha / r_{\text{struct}} = 2$, and 65,584 trainable parameters per adapted matrix. LoRA r8 ($\alpha = 16$, $s = 2$; 65,536 parameters) controls trainable parameter count, whereas LoRA r4 ($\alpha = 8$, $s = 2$) controls the structural-rank bound with a smaller budget. Thus the nominal scaling is aligned across the three main configurations. These are complementary capacity controls, not a complete causal isolation of depth, because the factorization structures, initialization dynamics, and factor-wise optimization still differ. The within-DR-LoRA rank comparison and the depth sweep separately test rank sensitivity and depth sensitivity. Additional PEFT variants are reported in Appendix B.

Evaluation and reporting. Candidates are selected by length-normalized log likelihood under fixed prompts and verbalizers (Appendix A). GLUE tasks use validation splits, and the QA tasks use the same held-out splits across methods. We report a unified accuracy-style metric; CoLA therefore uses accuracy rather than its official MCC. Main LoRA results, training dynamics, the depth sweep, and the nuclear-norm intervention average seeds 42, 1, and 123. The remaining comparisons use seed 42 and are presented as complementary exploratory analyses.

5 Main Results

5.1 DR-LoRA versus LoRA

Table 1 reports the three-seed six-task comparison. The default $L = 3$ configuration reaches 0.843, exceeding parameter-count-matched LoRA r8 by 3.4 points and structural-rank-matched LoRA r4

Table 1: Six-task results under the unified protocol (three-seed mean). *StructR*: structural-rank upper bound; *Acc*: macro-average accuracy. LoRA r8 controls parameter count, while LoRA r4 controls structural rank with a smaller budget. Best Acc in **bold**; macro-average standard deviations are ± 0.004, ± 0.003, and ± 0.002 for DR-LoRA r12c4, LoRA r8, and LoRA r4.

Method	Config	StructR	MRPC	CoLA	RTE	SciQA	CSQA	OBQA	Acc
LoRA	r8, α=16	8	0.865	0.850	0.836	0.819	0.734	0.747	0.809
LoRA	r4, α=8	4	0.869	0.850	0.841	0.818	0.761	0.790	0.821
DR-LoRA (default)	r_1=12, r_2=4	4	0.878	0.864	0.879	0.828	0.792	0.815	**0.843**
DR-LoRA	r_1=8, r_2=8, α=16	8	0.886	0.860	0.885	0.829	0.785	0.813	**0.843**

Table 2: Three-seed single-task fine-tuning results on CommonsenseQA. *eRank* denotes the post-training effective rank of the merged update; DR-LoRA yields stronger spectral concentration and smaller observed factor-gradient peaks.

Metric	DR-LoRA	LoRA r4	LoRA r8
Accuracy	**0.816**	0.776	0.769
eRank	**2.47**	3.83	7.61
Top σ	**1.39**	7.64	11.41
Grad. peak	**9**	36	107

by 2.2 points. The controls constrain complementary notions of capacity and use aligned nominal scaling, but, as noted in Section 4, they do not fully isolate depth from factor-wise optimization and other parameterization-specific effects. LoRA r4 also exceeds LoRA r8, while DR-LoRA improves over both; thus neither a larger rank bound nor parameter count alone explains the pattern.

5.2 Comparison with PEFT Variants

We also compare DR-LoRA with representative LoRA variants that modify scaling, optimization, initialization, or weight decomposition. The complete seed-42 comparison is reported in Appendix B.

5.3 Structural-Rank Ablation

The two DR-LoRA rows in Table 1 vary the structural-rank bound from 4 to 8 while keeping $L = 3$, approximately matching trainable parameters, and using $\alpha = 16$ for the rank-8 configuration to preserve $s = 2$. Both reach 0.843. This within-parameterization comparison suggests that the default gain is not explained by a larger structural-rank budget.

6 Mechanism Analysis

6.1 Training Dynamics

Table 2 and Figure 2 report three-seed single-task diagnostics on CommonsenseQA. Their CommonsenseQA accuracies are not expected to match the CommonsenseQA column of Table 1: the main benchmark uses joint fine-tuning over the six-task mixture, whereas the mechanism, depth, and nuclear-norm analyses use single-task CommonsenseQA fine-tuning. Relative to both LoRA controls, DR-LoRA attains higher accuracy with lower effective rank and a smaller leading singular value. It also exhibits smaller observed factor-gradient peaks, which we treat as a parameterization-dependent diagnostic rather than evidence of coordinate-invariant smoothness. These results are consistent with depth altering how the available update space is used; they do not establish effective rank as a causal mediator. The full singular-value spectra appear in Appendix C.

6.2 Synthetic Effective-Rank Dynamics

We fit a fixed rank-8 target with different depths while holding the target, initialization scale, optimizer, and training horizon fixed. This construction isolates the central finite-horizon prediction in a setting where all depths are trained toward the same matrix. Deeper products retain less tail mass and exhibit lower effective rank at the measured horizon (Appendix C.1), matching the predicted delay of

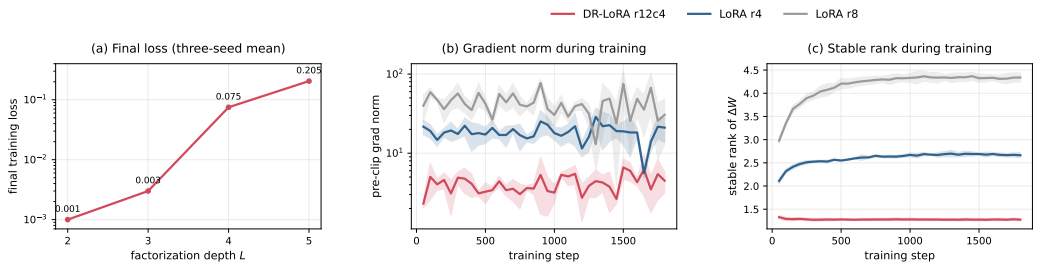


Figure 2: Optimization and spectral diagnostics. Left: final training loss across the depth sweep (three-seed mean, log scale), using the values in Table 3. Center: gradient norm (three-seed mean $\pm$ one standard deviation, log scale). Right: stable rank $\|\Delta W\|_F^2/\|\Delta W\|_2^2$, a complementary spectral diagnostic to the entropy-based eRank in Table 2.

Table 3: Depth ablation on CommonsenseQA, three-seed mean. Accuracy follows a non-monotonic pattern over the tested depths; effective rank falls monotonically; train loss rises sharply past L=3, indicating underfitting at excessive depth.

L	Acc	eRank	σ_1	Loss
2 (LoRA)	0.769	7.61	11.41	0.001
3	**0.816**	2.47	1.39	0.003
4	0.808	2.26	1.28	0.075
5	0.802	2.10	1.16	0.205

weak modes. Because exact recovery of the common target would produce the same singular-value spectrum, the observed cross-depth difference cannot be attributed to different target optima. Instead, it supports a transient learning bias: within the fixed optimization horizon, depth changes which portions of the target spectrum have already been recovered. This experiment provides a controlled test of the mode-selection component and supports its finite-horizon interpretation. Its quantitative relevance to Transformer fine-tuning remains to be established empirically.

7 Sensitivity and Intervention Analysis

We next test the predicted finite-horizon trade-off through a depth sensitivity analysis and then alter the objective with an explicit nuclear-norm intervention. Auxiliary structural diagnostics appear in Appendix B.2.

7.1 Depth Sensitivity

The depth sweep is a sensitivity analysis across the tested configurations, not a jointly parameter- and rank-matched isolation. Table 3 and Figure 3 reveal a clear trade-off: effective rank decreases monotonically with depth, whereas accuracy is highest at a moderate depth under this protocol, while training loss rises sharply for deeper configurations. A complementary learning-rate sensitivity check is reported in Appendix B.3; it shows that the favorable region remains associated with moderate depth, while the location and magnitude of the peak vary across the tested rates.

Lower effective rank is therefore not sufficient for better generalization. The moderate-depth regime is consistent with useful weak-mode delay, whereas the higher losses observed in the deeper tested configurations suggest that excessive depth can also impede directions needed for task fitting. This joint accuracy–rank–loss pattern supports, but does not causally establish, the predicted finite-horizon trade-off.

7.2 Nuclear-Norm Intervention

We next alter the objective directly by adding a nuclear-norm penalty to LoRA r8. Across the tested weights, Table 4 shows that explicit spectral shrinkage improves accuracy only partially: the best setting reaches 0.786, versus 0.769 for LoRA and 0.816 for DR-LoRA.

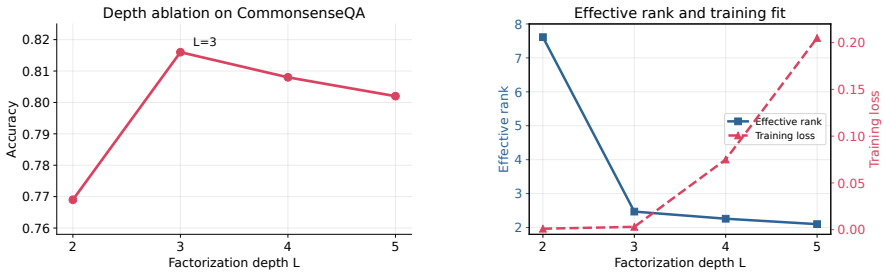


Figure 3: Depth ablation, three-seed mean. Left: accuracy vs. depth, with the highest accuracy at a moderate depth under this protocol. Right: increasing depth lowers effective rank but raises training loss in the deeper tested configurations.

Table 4: Nuclear-norm regularization sweep on LoRA r8 (CommonsenseQA, three-seed mean). DR-LoRA shown for reference (not regularized). Accuracy peaks at λ=5×10^{-4}, the best explicit penalty improves accuracy but leaves effective rank above DR-LoRA's.

λ	Acc	eRank	σ_1	Loss
0 (baseline)	0.769	7.61	11.41	0.001
10^{-4}	0.774	7.58	6.92	0.001
$\mathbf{5\times10^{-4}}$	**0.786**	7.38	4.61	0.001
10^{-3}	0.778	7.28	3.85	0.002
5×10^{-3}	0.774	6.93	2.15	0.006
DR-LoRA	0.816	2.47	1.39	0.003

Table 5: Single-seed depth bonus across base adapter structures (seed 42, six-task macro-average). Here L=2 denotes the two-matrix adapter and L=3 denotes the three-matrix adapter. The comparable gains suggest that the effect is not restricted to the plain LoRA base.

Base	L=2	L=3	Δ
Plain	0.811	0.844	+3.3
DoRA	0.803	0.838	+3.5

At its best-accuracy setting, the penalty changes effective rank only from 7.61 to 7.38; even the strongest tested penalty remains far above DR-LoRA at 2.47. Thus this finite sweep indicates that factorization depth and nuclear-norm regularization do not exhibit equivalent behavior under the tested optimization settings, while leaving other spectral penalties and schedules for future study.

8 Preliminary Transfer Probes

We conduct single-seed probes to assess whether the depth-dependent pattern extends across adapter structures, backbones, and modalities.

8.1 Across Base Adapters

Table 5 reports seed-42 gains of 3.3 and 3.5 points from increasing depth from L=2 to L=3 on plain LoRA and DoRA bases, respectively, suggesting that the effect may extend beyond plain LoRA.

8.2 Across Backbones and Modalities

To test whether the depth bonus can survive a change of backbone and modality, we replace LLaMA-2-7B with the vision-language model LLaVA-1.5-7B (Liu et al., 2023a; 2024a) and fine-tune on ScienceQA-IMG, the image-grounded split of ScienceQA (Lu et al., 2022). We keep the adapter configuration identical to the text experiments and change nothing else. As shown in Table 6,

Table 6: Preliminary transfer to a multimodal backbone: LLaVA-1.5-7B on ScienceQA-IMG (image-grounded subset, 2,017 test instances, seed 42). Adapter parameter counts are approximately matched across methods.

Method	Parameters	Accuracy
LoRA r8	4.19M	0.638
DR-LoRA r12c4	4.20M	**0.692**
Δ (depth bonus)		**+5.4 pp**

DR-LoRA improves accuracy by 5.4 points in this single-seed setting. This single-seed result is a preliminary probe rather than a multimodal benchmark. Together, the transfer probes suggest possible portability beyond the default text-only setting, but broader multi-seed evaluation is required.

9 Conclusion

We identify factorization depth as a non-monotonic design dimension for LoRA. Structural rank bounds representational capacity, whereas depth changes how the update space is traversed. Across complementary controls, moderate depth improves accuracy, while excessive depth increases training loss and degrades performance; effective rank is therefore a descriptive rather than causal indicator. Together with the conditional theory and nuclear-norm intervention, these results motivate selecting depth empirically and evaluating the effect across broader optimizers, backbones, and tasks.

10 Limitations

First, the comparisons provide complementary parameter-count and structural-rank controls rather than a complete causal isolation of depth. The nominal scaling is aligned across the default DR-LoRA, LoRA r8, and LoRA r4 settings ($s = 2$), but deeper products still change factor-wise optimization, initialization dynamics, and the factorization structure. A fully crossed depth–rank–scaling–initialization study would strengthen attribution.

Second, the theoretical results concern factorized gradient descent, and the singular-mode analysis additionally assumes aligned, balanced gradient flow. Our AdamW initialization is initially unbalanced, so the theory provides a mechanism hypothesis rather than a Transformer convergence law; effective rank is used as a diagnostic of the learned spectrum.

Third, the main analyses use three seeds, but the PEFT-variant, structural, and transfer probes require broader replication. Experiments cover LLaMA-2-7B and one LLaVA-1.5-7B setting, leaving larger backbones, generative tasks, and broader multimodal benchmarks open. The unified likelihood-based accuracy protocol improves cross-task comparability but differs from official metrics, most notably CoLA's MCC.

AI Use Statement

In this work, we used generative AI tools to provide critical ingredients for proving mathematical claims and to draft parts of the manuscript. We also used generative AI tools to assist with searching for and obtaining relevant information during the research and writing process.

The use of generative AI in the mathematical analysis was limited to providing potential intermediate ideas, derivation suggestions, and other critical ingredients related to the proofs. The authors independently checked the mathematical validity of these suggestions and remain responsible for the final statements, derivations, propositions, and proofs presented in the paper.

Generative AI tools were additionally used to draft selected portions of the paper and to assist with language, organization, and readability. They were also used as an auxiliary tool for searching for and obtaining research information. All AI-assisted text and information were reviewed by the authors, and relevant claims and sources were checked before inclusion in the manuscript.

The authors take full responsibility for the accuracy, originality, and integrity of this work, including all mathematical claims, proofs, experimental descriptions, interpretations, and AI-assisted text. No AI-generated material was included without author review and verification.

REFERENCES

Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. 2021. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. In *ACL-IJCNLP*.

Paul Albert, Frederic Z. Zhang, Hemanth Saratchandran, Cristian Rodriguez-Opazo, Anton van den Hengel, and Ehsan Abbasnejad. 2025. RandLoRA: Full-rank parameter-efficient fine-tuning of large models. In *ICLR*.

Sanjeev Arora, Nadav Cohen, Wei Hu, and Yuping Luo. 2019. Implicit regularization in deep matrix factorization. In *NeurIPS*.

Suriya Gunasekar, Blake Woodworth, Srinadh Bhojanapalli, Behnam Neyshabur, and Nathan Srebro. 2017. Implicit regularization in matrix factorization. In *NeurIPS*.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. 2024. LoRA+: Efficient low rank adaptation of large models. In *ICML*.

Zhiwei He, Zhaopeng Tu, Xing Wang, Xingyu Chen, Zhijie Wang, Jiahao Xu, Tian Liang, Wenxiang Jiao, Zhuosheng Zhang, and Rui Wang. 2025. RaSA: Rank-sharing low-rank adaptation. arXiv:2503.12576.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin de Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *ICML*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. LoRA: Low-rank adaptation of large language models. In *ICLR*.

Damjan Kalajdzievski. 2023. A rank stabilization scaling factor for fine-tuning with LoRA. *arXiv:2312.03732*.

Vladislav Lialin, Sherin Muckatira, Namrata Shivagunde, and Anna Rumshisky. 2024. ReLoRA: High-rank training through low-rank updates. In *ICLR*.

Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. 2023a. Visual instruction tuning. In *NeurIPS*.

Haotian Liu, Chunyuan Li, Yuheng Li, and Yong Jae Lee. 2024a. Improved baselines with visual instruction tuning. In *CVPR*.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. 2024b. DoRA: Weight-decomposed low-rank adaptation. In *ICML*.

Pan Lu, Swaroop Mishra, Tanglin Xia, Liang Qiu, Kai-Wei Chang, Song-Chun Zhu, Oyvind Tafjord, Peter Clark, and Ashwin Kalyan. 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. In *NeurIPS*.

Kaifeng Lyu and Jian Li. 2020. Gradient descent maximizes the margin of homogeneous neural networks. In *ICLR*.

Fanxu Meng, Zhaohui Wang, and Muhan Zhang. 2024. PiSSA: Principal singular values and singular vectors adaptation of large language models. In *NeurIPS*.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? A new dataset for open book question answering. In *EMNLP*.

Noam Razin and Nadav Cohen. 2020. Implicit regularization in deep learning may not be explainable by norms. In *NeurIPS*.

Olivier Roy and Martin Vetterli. 2007. The effective rank: A measure of effective dimensionality. In *15th European Signal Processing Conference (EUSIPCO)*, pp. 606–610.

Andrew M. Saxe, James L. McClelland, and Surya Ganguli. 2014. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. In *ICLR*.

Fanhua Shang, Yuanyuan Liu, and James Cheng. 2016. Unified scalable equivalent formulations for Schatten quasi-norms. arXiv:1606.00668.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2019. CommonsenseQA: A question answering challenge targeting commonsense knowledge. In *NAACL-HLT*.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, et al. 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv:2307.09288.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *ICLR*.

Zhengbo Wang, Jian Liang, Ran He, Zilei Wang, and Tieniu Tan. 2025a. LoRA-Pro: Are low-rank adapters properly optimized? In *ICLR*.

Sheng Wang, Liheng Chen, Pengan Chen, Jingwei Dong, Boyang Xue, Jiyue Jiang, Lingpeng Kong, and Chuan Wu. 2025b. MoS: Unleashing parameter efficiency of low-rank adaptation with mixture of shards. In *ICLR*.

Hyowon Wi and Noseong Park. 2026. Can spectral-clipping enable better learning while forgetting less for low-rank adaptation? In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 25708–25734. Association for Computational Linguistics.

Taiqiang Wu, Jiahao Wang, Zhe Zhao, and Ngai Wong. 2024. Mixture-of-subspaces in low-rank adaptation. In *EMNLP*, pages 7880–7899. Association for Computational Linguistics.

Wenhan Xia, Chengwei Qin, and Elad Hazan. 2024. Chain of LoRA: Efficient fine-tuning of language models via residual learning. *arXiv:2401.04151*.

Xi Xiao, Chenrui Ma, Yunbei Zhang, Chen Liu, Zhuxuanzi Wang, Yanshu Li, Lin Zhao, Guosheng Hu, Tianyang Wang, and Hao Xu. 2026. Not all directions matter: Towards structured and task-aware low-rank model adaptation. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2132–2154. Association for Computational Linguistics.

Can Yaras, Peng Wang, Laura Balzano, and Qing Qu. 2024. Compressible dynamics in deep overparameterized low-rank learning and adaptation. In *ICML*.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. 2023. Adaptive budget allocation for parameter-efficient fine-tuning. In *ICLR*.

A LIKELIHOOD SCORING

All methods are evaluated with the same fixed prompts and candidate completions. For a candidate answer $y = (y_1, \ldots, y_T)$ and prompt x, we score

$$\text{score}(y \mid x) = \frac{1}{T} \sum_{t=1}^{T} \log p(y_t \mid x, y_{<t}), \tag{14}$$

and select the candidate with the highest length-normalized log probability. Tables 7 and 8 report the exact prompt formats and candidate completions used for the two task families.

Table 7: Prompt templates and candidate completions for GLUE-style tasks.

Task	Prompt template	Candidate completions
MRPC	Sentence 1: `{sentence1}` Sentence 2: `{sentence2}` Are the two sentences equivalent? Answer:	yes, no
CoLA	Sentence: `{sentence}` Is this sentence grammatically acceptable? Answer:	acceptable, unacceptable
RTE	Premise: `{premise}` Hypothesis: `{hypothesis}` Does the premise entail the hypothesis? Answer:	yes, no

Table 8: Prompt format for multiple-choice QA tasks. Each listed answer choice is scored as a candidate completion.

Task	Prompt template
ScienceQA	Question: `{question}` Choices: (A) `{choice A}` ... Answer:
CommonsenseQA	Question: `{question}` Choices: (A) `{choice A}` ... Answer:
OpenBookQA	Question: `{question}` Choices: (A) `{choice A}` ... Answer:

B ADDITIONAL RESULTS

The following auxiliary results are omitted from the main text to keep the central argument concise.

B.1 COMPARISON WITH PEFT VARIANTS

Table 9 compares the default DR-LoRA r12c4 with representative LoRA variants under the same backbone, training recipe, and evaluation protocol. On the seed-42 run, the default DR-LoRA r12c4 obtains the numerically highest overall Acc at 0.841, while PiSSA is essentially tied at 0.840; we therefore treat this table as evidence of competitiveness rather than a statistically decisive win. More importantly, DR-LoRA r12c4 does so with the lowest structural-rank upper bound among the compared methods. While PiSSA leverages spectral initialization and LoRA+ improves optimization through asymmetric learning rates, DR-LoRA r12c4 modifies the factorization depth itself. These results suggest that depth-based factorization is a complementary design choice for PEFT.

B.2 ADDITIONAL SEED-42 DIAGNOSTICS

As a secondary seed-42 diagnostic, we test whether simple structural additions on top of DR-LoRA explain the observed gain. Table 10 evaluates magnitude modulation, multiplicative gating, and residual gating on the same DR-LoRA base. None improves the six-task macro-average, and one is harmful. This diagnostic suite was conducted in a separate experimental run from the PEFT comparison in Table 9; accordingly, the relevant comparison here is the within-run difference between

Table 9: Comparison with PEFT variants (seed 42). All methods use the same LLaMA-2-7B backbone, training recipe, and length-normalized likelihood evaluation. *StructR*: structural-rank upper bound; *Params*: trainable parameters per adapted weight matrix. *Acc*: overall accuracy under our unified protocol. Highest seed-42 overall in **bold**; PiSSA is within 0.1 point, so this table should be read as competitiveness evidence. DR-LoRA r12c4 (default) uses structural rank 4.

Method	StructR	Params	MRPC	CoLA	RTE	SciQA	CSQA	OBQA	Acc
LoRA r8	8	65,536	0.885	0.849	0.838	0.816	0.731	0.748	0.811
rsLoRA	8	65,536	0.846	0.847	0.827	0.819	0.743	0.750	0.805
LoRA+	8	65,536	0.880	0.857	0.848	0.824	0.754	0.776	0.823
DoRA	8	69,632	0.855	0.851	0.827	0.814	0.737	0.732	0.803
PiSSA	8	65,536	0.875	0.855	0.877	0.827	**0.800**	**0.806**	0.840
DR-LoRA (default)	4	65,584	**0.892**	**0.858**	**0.881**	**0.831**	0.783	0.800	**0.841**

each augmented variant and its corresponding base, rather than exact agreement with the baseline score in Table 9.

This auxiliary test supplies a third, negative control. Starting from the same DR-LoRA r12c4 base, adding magnitude modulation changes the macro-average from 0.840 to 0.838, multiplicative gating to 0.833, and residual gating to 0.838. None improves the base result, so the observed gain is not trivially recovered by adding gates or magnitude modulation to ΔW. These negative controls support focusing on factorization depth rather than these added mechanisms.

Table 10: Additional structural diagnostics on DR-LoRA r12c4 (seed 42). *Acc*: six-task macro-average; Δ in percentage points relative to the unaugmented base. These secondary variants do not improve over the base setting.

Variant	Acc	Δ
Base	**0.840**	—
+ DoRA mag.	0.838	-0.24
+ Mult. gate	0.833	-0.67
+ Residual gate	0.838	-0.20

B.3 Learning-Rate Sensitivity of Depth

To assess whether the depth–accuracy pattern depends on the default optimization scale, we repeat the CommonsenseQA depth comparison at three AdamW learning rates. This is a sensitivity analysis rather than a claim that one depth is universally optimal. The results show that moderate depth is favorable across the tested settings, but the location and magnitude of the peak vary across learning rates.

Table 11: CommonsenseQA report accuracy under three learning rates. The 5×10^{-5} and 8×10^{-4} rows use learning-rate sweep means; the 2×10^{-4} row reproduces the original paper depth-ablation values.

Learning rate	L=2	L=3	L=4	L=5
5×10^{-5}	0.7760	0.8002	0.7957	0.7851
2×10^{-4}	0.7690	0.8160	0.8080	0.8020
8×10^{-4}	0.7250	0.8092	0.8161	0.8135

C Additional Spectral Results

C.1 Synthetic factorization dynamics

These simulations isolate the optimization effect of factorization depth from task-specific modeling choices. With the target, initialization scale, step size, and optimization horizon held fixed, increasing the number of factors delays the growth of weaker singular modes and produces a more concentrated final spectrum. The effect is therefore a finite-horizon diagnostic: it explains why deeper parameterizations can impose stronger spectral selection, but it does not by itself imply improved task accuracy at every depth.

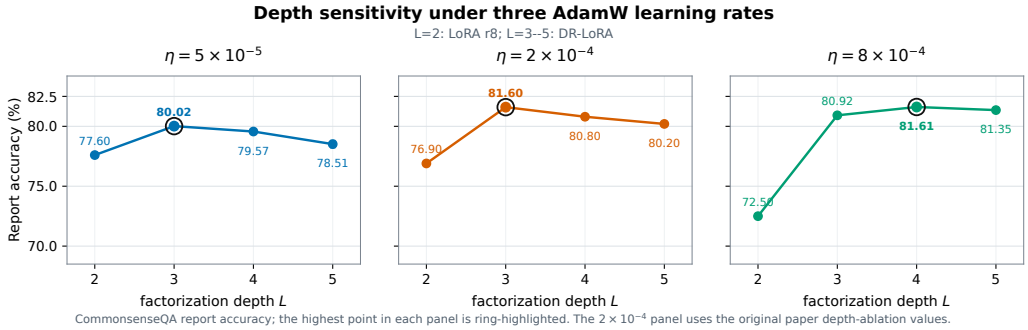


Figure 4: Depth sensitivity under three learning rates. Points show CommonsenseQA report accuracy for L=2–5; the highest point in each panel is ring-highlighted. The center panel uses the original paper depth-ablation protocol, while the two outer panels use the learning-rate sweep means.

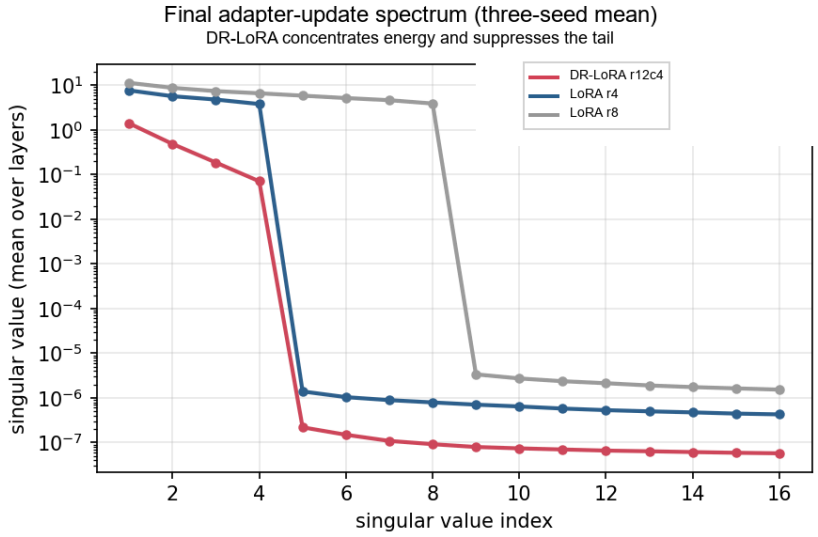


Figure 5: Post-training ΔW singular-value spectrum. DR-LoRA yields a steeper spectrum and a substantially smaller leading singular value than parameter-count-matched LoRA r8.

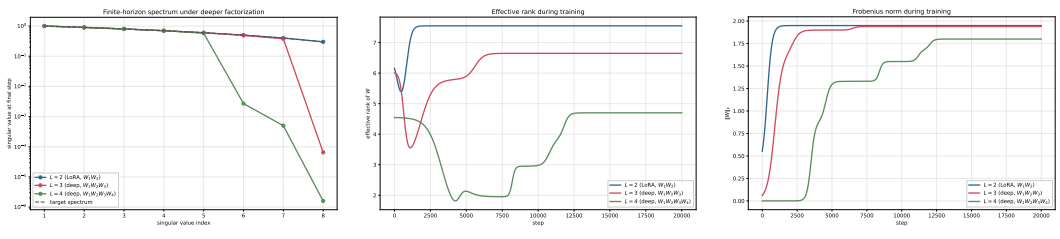


Figure 6: Simulation (L=2/3/4, identical GD). Left: spectra at the final optimization step. Center: eRank during training. Right: Frobenius norm during training.

C.2 Effective-rank mediation

The mediation plot relates post-training effective rank to task accuracy across the depth and explicit-penalty conditions. Accuracy is not monotone in effective rank: moderate spectral compression is compatible with the strongest performance, whereas further rank reduction can coincide with increased training loss and degraded task fit. Thus effective rank is informative about the observed depth effect, but it is not a sufficient standalone predictor of generalization.

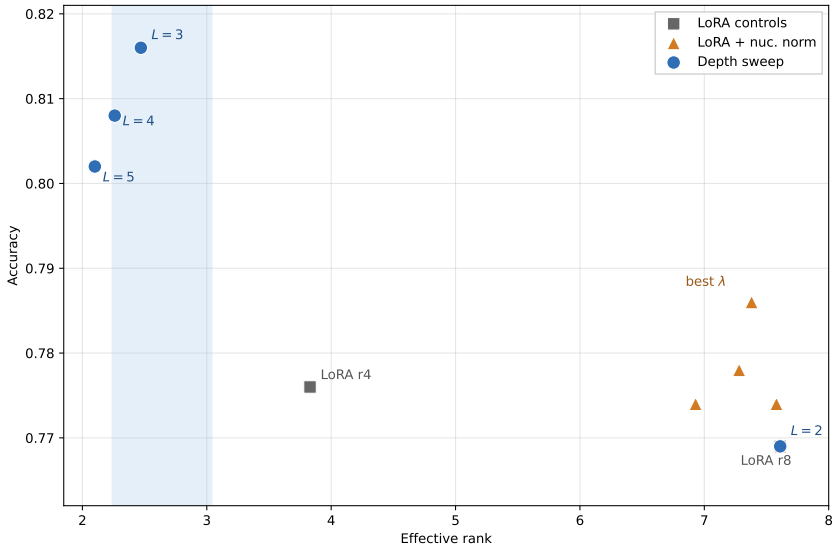


Figure 7: Effective-rank analysis, three-seed mean. Accuracy is highest at moderate effective rank across LoRA, LoRA+λ, and depth L=2/3/4/5, consistent with a useful but limited amount of spectral compression.

D Proofs and Effective-Rank Dynamics

D.1 Factorization cost for arbitrary factors

For $p > 0$, write $\|X\|_{S_p} = (\sum_i \sigma_i(X)^p)^{1/p}$. The generalized Schatten product inequality, including positive exponents below one, gives (Shang et al., 2016)

$$\|W_1 \cdots W_L\|_{S_{2/L}} \le \prod_{\ell=1}^{L} \|W_\ell\|_F. \tag{15}$$

Raising to the power $2/L$ and applying AM–GM proves

$$\sum_i \sigma_i(X)^{2/L} \le \prod_{\ell=1}^{L} \|W_\ell\|_F^{2/L} \le \frac{1}{L} \sum_{\ell=1}^{L} \|W_\ell\|_F^2. \tag{16}$$

This lower bound applies to arbitrary factors; it does not assume that their individual singular vectors align. To attain it, let $q = \operatorname{rank}(X) > 0$, $X = U\Sigma V^\top$ be a compact SVD, and choose matrices $Q_\ell \in \mathbb{R}^{r_\ell \times q}$ with orthonormal columns. Set $Q_0 = U$, $Q_L = V$, and

$$W_\ell = Q_{\ell-1} \Sigma^{1/L} Q_\ell^\top, \qquad \ell = 1, \dots, L. \tag{17}$$

The width assumption ensures these matrices exist. The product is X, and each squared factor norm is $\operatorname{tr}(\Sigma^{2/L})$. The case $X = 0$ follows by setting all factors to zero. This proves Proposition 1. For $p = 2/L < 1$ and fixed total singular-value mass, concavity of z^p means that transferring mass from a smaller singular value to a larger one reduces this spectral cost. It does not establish a general optimizer-equivalence result.

D.2 MERGED-UPDATE AND SINGULAR-MODE DYNAMICS

The differential of the product is $dX = \sum_k A_k(dW_k)B_k$, so the Frobenius inner product identity $d\mathcal{L} = \langle G, dX\rangle$ gives $\nabla_{W_k}\mathcal{L} = A_k^\top G B_k^\top$. Expanding the finite simultaneous step gives Proposition 2 with an $O(\eta^2)$ remainder for fixed factors as $\eta \to 0$. For ordinary factor gradient flow, this becomes the exact ODE $\dot{X} = -\mathcal{M}_L(G)$.

For the balanced aligned construction above, $A_k A_k^\top$ and $B_k^\top B_k$ act on active singular directions with gains $\sigma_i^{2(k-1)/L}$ and $\sigma_i^{2(L-k)/L}$. With squared loss, $G = U(\Sigma - \mathrm{diag}(\tau_i))V^\top$. Thus every term contributes $\sigma_i^{2-2/L}(\sigma_i - \tau_i)$ and

$$\dot{\sigma}_i = L\sigma_i^{2-2/L}(\tau_i - \sigma_i). \tag{18}$$

This aligned family remains invariant under this gradient flow. It is a conditional diagnostic, not a model of arbitrary initialization or AdamW. For positive σ, the adjacent-depth gain ratio and threshold are

$$\frac{g_{L+1}(\sigma)}{g_L(\sigma)} = \frac{L+1}{L}\sigma^{2/[L(L+1)]}, \tag{19}$$

$$g_{L+1}(\sigma) < g_L(\sigma) \quad \Longleftrightarrow \quad \sigma < \theta_L := \left(\frac{L}{L+1}\right)^{L(L+1)/2}. \tag{20}$$

These are same-scale, same-state comparisons, not an ordering of full trajectories with different initialization or learning-rate schedules.

D.3 FINITE-TIME WEAK-MODE DELAY

For $0 < \sigma_{i,0} < \sigma_i^\star \ll \tau_i$ and $L > 2$, approximating $\tau_i - \sigma_i$ by τ_i and integrating yields

$$t_i^\star \approx \frac{1}{L\tau_i}\int_{\sigma_{i,0}}^{\sigma_i^\star} \sigma^{-(2-2/L)}\, d\sigma = \frac{\sigma_{i,0}^{-(1-2/L)} - (\sigma_i^\star)^{-(1-2/L)}}{(L-2)\tau_i}. \tag{21}$$

At $L = 2$, $\sigma_i(t) \approx \sigma_{i,0}e^{2\tau_i t}$, giving $t_i^\star \approx (2\tau_i)^{-1}\log(\sigma_i^\star/\sigma_{i,0})$. The polynomial-versus-logarithmic distinction concerns the dependence on inverse initialization size. Exactly zero modes of the balanced diagnostic do not activate; the positive-initialization assumption is essential. In our practical initialization only the final factor is zero, while other factors are nonzero and unbalanced, so its initial motion is not governed by this balanced-mode simplification.

D.4 FROM RELATIVE MODE GROWTH TO EFFECTIVE RANK

Let $S = \sum_i \sigma_i > 0$, $p_i = \sigma_i/S$, and consider a fixed positive active support. Write $a_i = \dot{\sigma}_i/\sigma_i$ and $\bar{a} = \sum_i p_i a_i$. Then

$$\dot{p}_i = p_i(a_i - \bar{a}), \qquad \sum_i \dot{p}_i = 0. \tag{22}$$

Differentiating the entropy gives

$$\frac{d}{dt}\log \mathrm{erank}(X) = \dot{H}(p) = -\sum_i \dot{p}_i(\log p_i + 1) \tag{23}$$

$$= -\sum_i p_i(a_i - \bar{a})\log p_i = -\,\mathrm{Cov}_p(a_i, \log p_i). \tag{24}$$

Substituting the balanced dynamics yields Eq. 13. A sufficient condition for nonincreasing effective rank is $(a_i - a_j)(p_i - p_j) \geq 0$ for every pair of active modes, since

$$\mathrm{Cov}_p(a_i, \log p_i) = \frac{1}{2}\sum_{i,j} p_i p_j (a_i - a_j)(\log p_i - \log p_j) \geq 0. \tag{25}$$

Equality includes uniform relative growth, which scales the matrix without changing its effective rank. The residuals $\tau_i - \sigma_i$ may prevent this ordering, so the theory does not guarantee that effective rank

decreases monotonically over training, nor that greater factorization depth always produces a lower effective rank. Both behaviors depend on the evolving singular values and their target residuals at the finite training horizon. This is why the paper measures effective rank rather than inferring it from depth or leading singular value alone. For any nonzero matrix, $1 \leq \mathrm{erank}(X) \leq \mathrm{rank}(X) \leq r_{\mathrm{struct}}$; the upper entropy bound follows from comparison with the uniform distribution on its nonzero singular values. We do not assign an entropy-based effective rank to the exactly zero initial update.