LoRA-Lens: Training Induces Spectral Compression in Low-Rank Adapters

Zhiyuan Gao
Department of Biostatistics & Bioinformatics
Duke University
Durham, NC 27705
zhiyuan.gao@duke.edu

Abstract

Low-Rank Adaptation (LoRA) is widely used for parameter-efficient fine-tuning, but how training reshapes adapter spectral structure—and what this implies for rank allocation—remains poorly understood. We introduce **LoRA-Lens**, a diagnostic that tracks *capacity utilization* (CU), the fraction of rank budget carrying meaningful spectral energy. Across five models from three families, training compresses adapter spectra (CU: 1.0 $\rightarrow$ 0.43–0.83), with V-projection compressing 2–3$\times$ more than Q. This compression does not reliably indicate redundancy: on three representative models, V `lora_B` parameters are approximately 30–350$\times$ more Fisher-sensitive than Q across training stages, consistent with the OV circuit's role as the content channel that writes directly to the residual stream. Symmetric rank sweeps show that Q-rank reduction is safe in the tested models while V-rank reduction can be harmful. We propose **FisherLoRA**—preserve V, reduce Q—reducing Q/V adapter parameters by 38% relative to uniform rank allocation with worst-case degradation of 0.006 across five models, vs. 0.067 for spectral-guided allocation.

1. Introduction

Low-Rank Adaptation [7] decomposes weight updates as $\Delta W = BA$ with $r \ll d$. Despite its ubiquity, practitioners select r and target modules largely by convention. Recent work has identified spectral under-utilization in LoRA [10, 16] and proposed various strategies—including adaptive importance-based pruning [20] and pretrained-spectrum-guided rank allocation [19]—but whether spectral compression of trained adapters implies functional redundancy has not been directly tested.

We introduce capacity utilization CU $= \hat{r}/r$, where $\hat{r} = \exp(-\sum_i \tilde{\sigma}_i \log \tilde{\sigma}_i)$ is the effective rank [11] and $\tilde{\sigma}_i = \sigma_i / \sum_j \sigma_j$ (CU=0 for $\Delta W = 0$). Our experiments reveal a dissociation between spectral structure and functional importance: V is spectrally more compressed, yet symmetric Q/V rank sweeps show that Q-rank reduction is broadly safe while V-rank reduction is model-specific. Fisher sensitivity analysis suggests an explanation grounded in attention mechanics [5]: V parameters sit in steeper loss regions because the OV circuit ($W_O W_V$) writes content into the residual stream more directly than the QK circuit influences it. This motivates **FisherLoRA**—preserve V, reduce Q—the opposite of what spectral analysis alone would suggest.

2. Spectral Compression

Setup. Qwen2.5-0.5B/1.5B/7B [15], Mistral-7B [8], and Yi-1.5-9B [17]; LoRA on Alpaca [14] (α=2r, cosine LR, 3 epochs, 1k samples); cross-dataset validation on Dolly [4] and GSM8K [3].

Eval: ARC-C [2], HellaSwag [18], TruthfulQA [9], GSM8K, IFEval subset (450 items); 3 seeds for key experiments. Data-size sensitivity (1k/5k/10k on two models) confirms all patterns hold.

2.1. Compression Is Training-Induced

We compare three conditions at each $r \in \{4, 8, 16, 32, 64\}$: **(i)** PEFT-default ($B$=0): CU≈0. **(ii)** Random ($B, A \sim \mathcal{N}(0, 1/\sqrt{r})$): CU≈1.0, no Q/V asymmetry. **(iii)** Trained: CU = 0.43–0.83. The trained condition falls strictly between the baselines, confirming that compression is an effect of optimization rather than initialization.

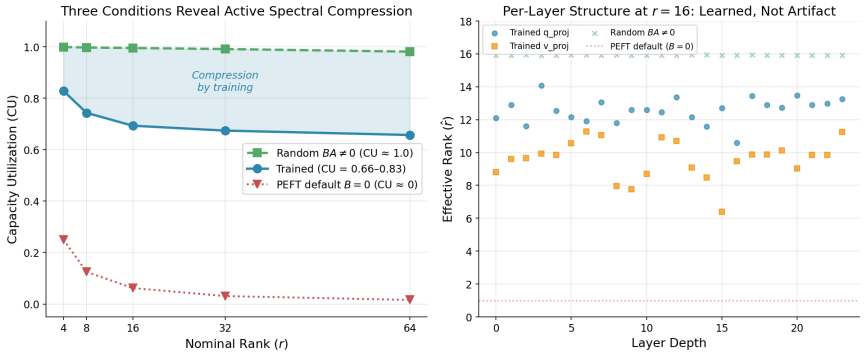

Figure 1: **Left:** CU across three conditions (0.5B). **Right:** Per-layer effective rank at r=16. Trained Q and V separate into distinct bands; random shows no asymmetry.

CU is reproducible across seeds (std $\leq$ 0.007), an order of magnitude more stable than downstream eval scores (std 0.009–0.027). Pretrained Q weights have 2.8–5.6× higher effective rank than V across all five models. This relationship extends beyond Q and V: correlating pretrained with adapter effective rank across all 7 linear module types (168 layer-module pairs) yields Pearson r=0.54 across 168 layer-module pairs (nominally $p < 10^{-13}$, though layer-wise dependencies make this descriptive) (Figure 2). Within the Qwen family, effective rank is well fit by $\hat{r} = a \cdot r^{\beta}$ over the tested range, with β decreasing: 0.948 (0.5B), 0.913 (1.5B), 0.896 (7B); all $R^2 > 0.999$.

2.2. Two-Phase Q–V Divergence

Tracking CU during training reveals two phases. In the first ∼10 steps, both Q and V compress rapidly; then Q partially recovers while V continues compressing, doubling the Q–V gap (Figure 3). The spectral structure is largely determined before the model has seen the full training set.

3. Compression and Redundancy Are Dissociated

V compresses more, so one might expect V-rank to be safer to reduce. We test this—and the converse—with symmetric rank sweeps on both modules.

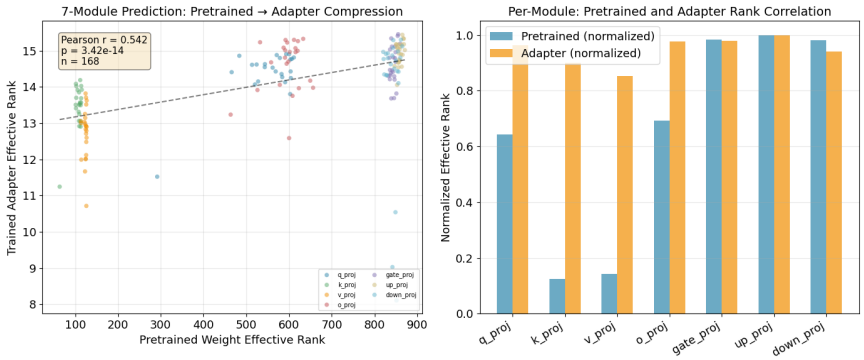

Figure 2: Pretrained vs. adapter effective rank across 7 module types (168 points, r=0.54). Modules with lower pretrained rank develop more compressed adapters.

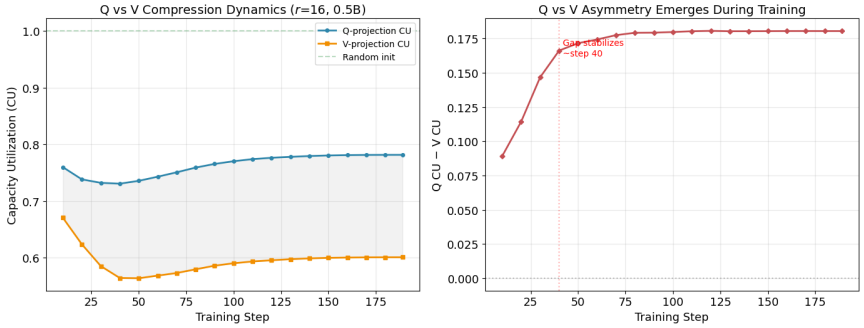

Figure 3: **Left:** Q and V CU trajectories (r=16, 0.5B). Both compress in ∼10 steps, then diverge. **Right:** Q–V gap doubles from 0.089 to 0.180, stabilizing around step 40.

3.1. Symmetric Rank Sweeps

The less compressed module (Q) tolerates r_Q=2 across all models; the more compressed module (V) is sensitive in Qwen-7B—the opposite of what spectral analysis predicts. We use p-values descriptively to flag consistent degradation patterns rather than as standalone confirmatory tests.

3.2. Fisher Sensitivity Provides a Mechanistic Account

We measure empirical Fisher information ($\mathbb{E}[\mathbf{g}^2]$, diagonal) for Q and V `lora_B` parameters, averaged over 16 Alpaca examples, at initialization, mid-training (step 50), and after convergence.

This persistent asymmetry is consistent with the QK/OV circuit decomposition of Elhage et al. [5]. Q participates in the QK circuit—a routing mechanism that determines attention patterns, where gradients are attenuated through the softmax. V participates in the OV circuit—the composed value-output pathway that writes source-token content into the residual stream [5], plausibly producing steeper loss curvature. We present this as a structural account consistent with our observations, not a complete causal explanation.

Table 1: Q-rank sweep (r_V=16, 3 seeds, Alpaca). Q reduction does not significantly degrade any model, including Qwen-7B.

r_Q	Qwen-0.5B	Qwen-7B	Mistral-7B
16	.361±.006	.612±.019	.536±.027
4	.368±.005	.608±.017	.530±.013
2	.376±.004	.607±.005	.522±.015

Table 2: V-rank sweep (r_Q=16, 3 seeds). Most models tolerate r_V=2; Qwen-7B degrades at $r_V \leq 4$ ($^*p < 0.05$), consistently across Alpaca, GSM8K, and Dolly (all $p < 0.025$). Baselines differ slightly across tables due to separate training runs with different rank configurations.

r_V	Qwen-0.5B	Qwen-1.5B	Yi-9B	Mistral-7B	**Qwen-7B**
16	$.371_{\pm.003}$	$.505_{\pm.003}$	$.498_{\pm.006}$	$.532_{\pm.022}$	$.610_{\pm.016}$
4	$.374_{\pm.010}$	$.501_{\pm.008}$	$.503_{\pm.009}$	$.542_{\pm.008}$	$.545_{\pm.018^*}$
2	$.374_{\pm.004}$	$.505_{\pm.009}$	$.508_{\pm.007}$	$.541_{\pm.012}$	$.529_{\pm.027^*}$

4. FisherLoRA

The dissociation between compression and sensitivity motivates a simple allocation rule: preserve V (high Fisher), reduce Q (low Fisher). A ∼50-step probe confirms the V ≫ Q asymmetry; then set r_Q=4, r_V=16 (62% of uniform parameters) and train normally—no additional overhead during the training run itself.

Under matched budgets (0.62×), FisherLoRA's worst-case degradation is 0.006—11× smaller than spectral-guided allocation (0.067), which uses the same parameter count but the opposite allocation direction. AdaLoRA uses sensitivity-based importance scores (smoothed $|w \cdot \nabla w|$) to dynamically prune singular values, and in principle could discover the Q/V asymmetry. However, with target_r=4 it converges to a much smaller budget (≈0.25×) and degrades on 3/5 models (worst Δ=−0.051), suggesting that this particular budget target is too aggressive for our setting. FisherLoRA's advantage is simplicity and safety: a fixed asymmetric allocation at a moderate budget, guided by a single Fisher probe, avoids both the risk of wrong-direction allocation (spectral) and over-aggressive pruning (AdaLoRA). FisherLoRA also slightly improves on 3/5 models (Qwen-0.5B, 1.5B, Yi-9B), consistent with Q rank being non-limiting under these budgets, though these small gains should not be over-interpreted.

5. Related Work

Shuttleworth et al. [12] identify intruder dimensions in LoRA updates; Si et al. [13] show fine-tuning amplifies top singular values; Tian et al. [16] propose post-hoc singular-value reweighting; Lion et al. [10] show stable rank falls below algebraic rank. Our work tests the *functional* consequences of these spectral observations via symmetric rank sweeps and Fisher analysis. AdaLoRA [20] uses sensitivity-based importance scores to adaptively allocate budget, while SR-LoRA [19] uses the stable rank of pretrained weights as a rank-allocation prior. Our results caution that spectral compression of trained adapters should not be interpreted directly as functional redundancy. The

Table 3: V/Q Fisher ratio at three training stages. V is consistently 30–350× more sensitive than Q, confirming the asymmetry is not an initialization artifact.

Model	Step 0	Step 50	Final
Qwen-0.5B	65.8×	94.7×	82.1×
Qwen-7B	32.5×	103.0×	93.5×
Mistral-7B	353.0×	59.3×	51.4×

Table 4: Method comparison across five models (3 seeds each). FisherLoRA and spectral use matched budgets (0.62×). AdaLoRA (init_r=16, target_r=4) dynamically prunes to ≈0.25×; its degradation suggests the pruning is too aggressive.

Method	Par.	Qwen 0.5B	Qwen 1.5B	Qwen 7B	Mistral 7B	Yi 9B	Worst Δ
Uniform	1.0×	.361	.501	.612	.536	.498	—
FisherLoRA	.62×	.368	.507	**.608**	.530	.503	**−.006**
Spectral	.62×	.374	.501	.545	.542	.503	−.067
AdaLoRA	dyn	.371	.462	.561	.546	.451	−.051

two-phase dynamics are reminiscent of early training phenomena documented by Frankle et al. [6], and the compression–sensitivity dissociation relates to intrinsic dimensionality [1]—we conjecture that lower intrinsic dimension implies higher per-direction importance.

6. Discussion

Training consistently compresses LoRA adapter spectra, with V compressing 2–3× more than Q across all five models. But low CU does not by itself imply redundancy: in our experiments, the more compressed V adapters are also substantially more Fisher-sensitive: V is 30–350× more Fisher-sensitive than Q throughout training, and Q-rank reduction is broadly safe while V-rank reduction is model-specific. FisherLoRA exploits this dissociation, saving 38% of parameters with worst-case degradation of 0.006.

The QK/OV circuit decomposition [5] offers a plausible structural explanation: the OV circuit's direct contribution to the residual stream makes V parameters more influential, while Q's routing role is more tolerant of rank reduction. More broadly, our results suggest that spectral diagnostics like CU are valuable for understanding training dynamics but should not be used directly to guide rank allocation without also considering functional sensitivity.

Limitations. FisherLoRA is validated on five models from three families; broader testing across additional architectures and larger training scales is needed. The mechanistic account based on QK/OV circuits is consistent with our observations but has not been verified through targeted interventions. All experiments use 1k training samples by default, with sensitivity analysis up to 10k confirming stability.

References

[1] Armen Aghajanyan, Sonal Gupta, and Luke Zettlemoyer. Intrinsic dimensionality explains the effectiveness of language model fine-tuning. *arXiv preprint arXiv:2012.13255*, 2021.

[2] Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

[3] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[4] Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. Free Dolly: Introducing the world's first truly open instruction-tuned LLM. 2023. Databricks Blog.

[5] Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 2021.

[6] Jonathan Frankle, David J Schwab, and Ari S Morcos. The early phase of neural network training. *arXiv preprint arXiv:2002.10365*, 2020.

[7] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

[8] Albert Q Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, et al. Mistral 7B. *arXiv preprint arXiv:2310.06825*, 2023.

[9] Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022.

[10] Kai Lion, Liang Zhang, Bingcong Li, and Niao He. PoLAR: Polar-decomposed low-rank adapter representation. *arXiv preprint arXiv:2506.03133*, 2025.

[11] Olivier Roy and Martin Vetterli. The effective rank: A measure of effective dimensionality. *European Signal Processing Conference (EUSIPCO)*, pages 606–610, 2007.

[12] Reece Shuttleworth, Jacob Andreas, Antonio Torralba, and Pratyusha Sharma. LoRA vs full fine-tuning: An illusion of equivalence. In *Advances in Neural Information Processing Systems*, 2025.

[13] Chongjie Si, Xuankun Yang, Muqing Liu, Yadao Wang, Xiaokang Yang, Wenbo Su, Bo Zheng, and Wei Shen. Weight spectra induced efficient model adaptation. *arXiv preprint arXiv:2505.23099*, 2025.

[14] Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B Hashimoto. Stanford Alpaca: An instruction-following LLaMA model. 2023. GitHub repository.

[15] Qwen Team. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024.

[16] Zailong Tian, Yanzhe Chen, Zhuoheng Han, and Lizi Liao. Spectral surgery: Training-free refinement of LoRA via gradient-guided singular value reweighting. *arXiv preprint arXiv:2603.03995*, 2026.

[17] Alex Young, Bei Chen, Chao Li, Chengen Huang, Ge Zhang, Guanwei Zhang, Heng Li, Jiangcheng Zhu, Jianqun Chen, Jing Chang, et al. Yi: Open foundation models by 01.AI. *arXiv preprint arXiv:2403.04652*, 2024.

[18] Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

[19] Chuyan Zhang, Kefan Wang, and Yun Gu. Beyond low-rank tuning: Model prior-guided rank allocation for effective transfer in low-data and large-gap regimes. In *ICCV*, 2025.

[20] Qingru Zhang, Minshuo Chen, Alexander Bukharin, Nikos Karampatziakis, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning. *arXiv preprint arXiv:2303.10512*, 2023.