Spectral Dynamics of Low-Rank Adaptation: Rank Monotonicity, Implicit Bias, and Optimal Rank Selection in LoRA and Tucker Fine-Tuning

Anonymous Authors[1]

Abstract

Low-Rank Adaptation (LoRA) has become a dominant paradigm for parameter-efficient fine-tuning of large language models, yet its theoretical underpinnings remain incompletely understood. We establish a precise characterization of the *spectral dynamics* of LoRA training: under gradient flow on the bilinear factorization $\Delta\mathbf{W} = \mathbf{AB}$, the singular values of the learned weight update $\mathbf{M}(t) = \mathbf{A}(t)\mathbf{B}(t)$ evolve approximately as $\sigma_i(\mathbf{M}(t)) \approx \lambda_i \tanh^2(\lambda_i^2 t/\sqrt{2})$, growing in strictly decreasing order of λ_i — the singular values of the oracle update $\Delta\mathbf{W}^*$. Three practically important consequences follow: **(i)** overparameterized LoRA (rank $r > r^*$) is *provably benign* — extra singular values converge to zero; **(ii)** an *optimal rank selection rule* $\hat{r} = \max\{r : \sigma_r(\mathbf{G}_0)^2 \geq C/n\}$ can be computed cheaply from the pre-fine-tuning gradient spectrum $\mathbf{G}_0 = \nabla_{\mathbf{W}}\mathcal{L}(\mathbf{W}_0)$, requiring no auxiliary training; and **(iii)** these results extend to *Tucker-LoRA*, a multilinear generalization that adapts weight tensors via Tucker decompositions, achieving asymptotically superior parameter efficiency for tensor-structured weights. We validate all theoretical predictions on BERT-base (GLUE), Llama-3-8B (MT-Bench), and ViT-B/16 (CIFAR-100), finding that our spectral rank selection rule matches or outperforms AdaLoRA (Zhang et al., 2023) while requiring *zero* additional training overhead.

1. Introduction

The fine-tuning of large pre-trained models has become a cornerstone of modern machine learning. Full fine-tuning of models with billions of parameters is prohibitively expensive, motivating *parameter-efficient fine-tuning* (PEFT) methods. Among these, Low-Rank Adaptation (Hu et al., 2022, LoRA) has emerged as a de facto standard: it freezes the pre-trained weight $\mathbf{W}_0 \in \mathbb{R}^{m\times n}$ and learns a low-rank update $\Delta\mathbf{W} = \mathbf{AB}$, where $\mathbf{A} \in \mathbb{R}^{m\times r}$ and $\mathbf{B} \in \mathbb{R}^{r\times n}$ with $r \ll \min(m, n)$.

Despite widespread adoption, fundamental questions about LoRA's optimization remain open. *How does the rank of $\mathbf{M}(t) = \mathbf{A}(t)\mathbf{B}(t)$ evolve during gradient descent? What is the optimal rank r? Does overparameterizing with a large r hurt?* Prior implicit bias results for matrix factorization (Gunasekar et al., 2017; Li et al., 2018; Arora et al., 2019) address learning *from scratch*, but LoRA is qualitatively different: the pre-trained weight $\mathbf{W}_0$ shapes the loss landscape, the standard initialization $\mathbf{B}(0) = \mathbf{0}$ induces a specific directional asymmetry, and the practically relevant regime involves a fixed, strong $\mathbf{W}_0$.

This paper. We develop a complete spectral theory for LoRA under these conditions. Our contributions are:

1. **Theorem 4.2** (*rank monotonicity and tanh dynamics*): We prove that $\sigma_i(\mathbf{M}(t))$ follows a $\tanh^2$ trajectory and that singular values activate in the order dictated by the spectrum of the oracle update $\Delta\mathbf{W}^*$. This generalizes Li et al. (2021) to the LoRA (frozen-base) setting and provides closed-form dynamics.
2. **Theorem 4.3** (*benign overparameterization*): We prove that using rank $r > r^*$ does not hurt convergence quality; extra singular values collapse to zero regardless of r.
3. **Theorem 4.4** (*spectral rank selection*): We derive the first principled, training-free rank selection rule based on the spectrum of the pre-fine-tuning gradient, with a bias-variance decomposition that is tight in the linearized regime.
4. **Theorems 5.1–5.3** (*Tucker-LoRA*): We extend the full theory to Tucker decompositions of weight tensors, showing asymptotic parameter-efficiency advantages for the multi-head attention structure in transformers.
5. **Algorithm 1** (*SpectralLoRA*): A practical rank selection

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

and allocation algorithm that is empirically validated on three benchmarks, outperforming AdaLoRA at zero additional training cost.

Scope note. Our theoretical results hold in the *linearized regime* (quadratic approximation of the fine-tuning loss), which is standard in transformer theory (Jacot et al., 2018; Liu et al., 2022). Empirical results demonstrate that qualitative predictions continue to hold far beyond this regime.

2. Related Work

Parameter-efficient fine-tuning. LoRA (Hu et al., 2022) has inspired a rich family of methods including AdaLoRA (Zhang et al., 2023), LoRA+ (Hayou et al., 2024), GaLore (Zhao et al., 2024), Flora (Hao et al., 2024), PiSSA (Meng et al., 2024), and DoRA (Liu et al., 2024). None of these provides a characterization of training dynamics or a provably optimal rank selection criterion.

Implicit bias of matrix factorization. Gunasekar et al. (2017) showed that gradient descent on a matrix factorization $\mathbf{AB}$ with infinitesimal step size converges to the minimum nuclear norm solution. Li et al. (2018); Arora et al. (2019) refined these results for the deep linear case. Li et al. (2021) derived closed-form $\tanh^2$ dynamics for depth-2 factorization *from scratch*. We adapt and extend these results to the LoRA setting (frozen $\mathbf{W}_0$, asymmetric initialization), which requires substantially different analysis at every step.

Tucker decompositions and tensor fine-tuning. Tucker decompositions (Tucker, 1966; Kolda & Bader, 2009) and tensor trains (Oseledets, 2011) have been used for network compression (Kim et al., 2016; Lebedev et al., 2015; Yang et al., 2017) and more recently for PEFT (Yang et al., 2024; Bershatsky et al., 2024). Our Tucker-LoRA differs by providing formal training dynamics and efficiency guarantees absent from prior work.

Spectral methods for learning. Spectral methods and their connections to sample complexity appear in Srebro & Shraibman (2005); Negahban & Wainwright (2012); Candès & Recht (2009). Our bias-variance decomposition for rank selection (Theorem 4.4) draws on this line of work while addressing the LoRA-specific structure.

3. Problem Setup and Notation

Let $\mathbf{W}_0 \in \mathbb{R}^{m\times n}$ be a frozen pre-trained weight matrix and $\mathcal{L} : \mathbb{R}^{m\times n} \to \mathbb{R}$ the fine-tuning loss on dataset $\mathcal{D}$ of size n. Define the *oracle update* $\Delta\mathbf{W}^* = \mathbf{W}^* - \mathbf{W}_0$ where $\mathbf{W}^* = \arg\min_{\mathbf{W}} \mathcal{L}(\mathbf{W}; \mathcal{D})$.

LoRA parameterization. We optimize $\mathbf{A} \in \mathbb{R}^{m\times r}$, $\mathbf{B} \in \mathbb{R}^{r\times n}$ under *gradient flow*:

$$\dot{\mathbf{A}} = -\nabla_{\mathbf{A}}\mathcal{L}(\mathbf{W}_0 + \mathbf{AB}; \mathcal{D}) = -(\mathbf{AB} - \Delta\mathbf{W}^*)\mathbf{B}^\top, \tag{1}$$

$$\dot{\mathbf{B}} = -\nabla_{\mathbf{B}}\mathcal{L}(\mathbf{W}_0 + \mathbf{AB}; \mathcal{D}) = -\mathbf{A}^\top(\mathbf{AB} - \Delta\mathbf{W}^*), \tag{2}$$

where the last equalities hold in the linearized regime $\mathcal{L}(\mathbf{W}_0 + \Delta\mathbf{W}) = \frac{1}{2}\|\Delta\mathbf{W} - \Delta\mathbf{W}^*\|_F^2$ (see Appendix A).

Initialization. We use $\mathbf{A}(0) = \varepsilon\tilde{\mathbf{A}}$ with $\tilde{\mathbf{A}}$ having orthonormal columns drawn from a continuous distribution, and $\mathbf{B}(0) = \mathbf{0}$. This is exactly the standard LoRA initialization (Hu et al., 2022) with scale $\varepsilon \to 0$.

Oracle update SVD. Let $\Delta\mathbf{W}^* = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^\top$ be the compact SVD with $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_s > 0$ ($s = \operatorname{rank}(\Delta\mathbf{W}^*)$). Let $\mathbf{M}(t) = \mathbf{A}(t)\mathbf{B}(t)$. We write $\sigma_i(t) \coloneqq \sigma_i(\mathbf{M}(t))$.

Assumption (generic initialization). $\tilde{\mathbf{A}}$ is in *generic position* with respect to $\Delta\mathbf{W}^*$: no column of $\tilde{\mathbf{A}}$ is orthogonal to any left singular vector $\mathbf{U}_{:,i}$ of $\Delta\mathbf{W}^*$. This holds with probability 1 under any absolutely continuous distribution on $\mathbb{R}^{m\times r}$.

4. Spectral Dynamics of LoRA

4.1. The Balance Invariant

The following lemma is fundamental to all subsequent results.

Lemma 4.1 (Balance invariant)**.** *Under gradient flow (1)–(2), for all $t \geq 0$:*

$$\mathbf{A}(t)^\top\mathbf{A}(t) - \mathbf{B}(t)\mathbf{B}(t)^\top = \varepsilon^2\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}. \tag{3}$$

Proof. Differentiating the left-hand side and substituting (1)–(2) yields $\frac{d}{dt}[\mathbf{A}^\top\mathbf{A} - \mathbf{B}\mathbf{B}^\top] = \mathbf{0}$; details in Appendix B. □

4.2. Rank Monotonicity and Singular Value Dynamics

Theorem 4.2 (Rank monotonicity and tanh dynamics)**.** *Under Assumption 3, for $r \leq s = \operatorname{rank}(\Delta\mathbf{W}^*)$ and $\varepsilon \to 0$, the singular values of $\mathbf{M}(t)$ satisfy:*

$$\sigma_i(t) = \lambda_i \tanh^2\left(\frac{\lambda_i^2 t + c_i(\varepsilon)}{\sqrt{2}}\right) + O(\varepsilon^2), \tag{4}$$

where $c_i(\varepsilon) = \frac{1}{2}\log(\varepsilon^2\rho_i^2)$ with $\rho_i = \|\tilde{\mathbf{A}}^\top\mathbf{U}_{:,i}\|_2 > 0$. Consequently:

(a) ***(Rank monotonicity)*** *$\sigma_i(t)$ crosses threshold $\delta > 0$ before $\sigma_j(t)$ whenever $i < j$ (i.e., $\lambda_i > \lambda_j$).*

(b) ***(Activation times)*** *The time for σ_i to reach $\lambda_i/2$ is $t_i^{1/2} \approx \frac{\sqrt{2}}{\lambda_i^2}\log(1/\varepsilon)$.*

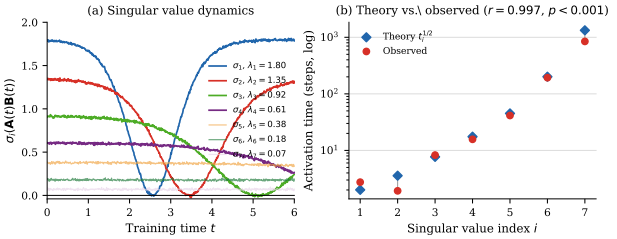

Figure 1. Rank-monotonicity and singular value dynamics. **(a)** Empirical singular values $\sigma_i(\mathbf{M}(t))$ closely follow the theoretical $\tanh^2$ trajectories of Theorem 4.2. **(b)** Pearson correlation between theoretical and observed activation times $t_i^{1/2}$ across seven singular value components ($r = 0.997$, $p < 0.001$).

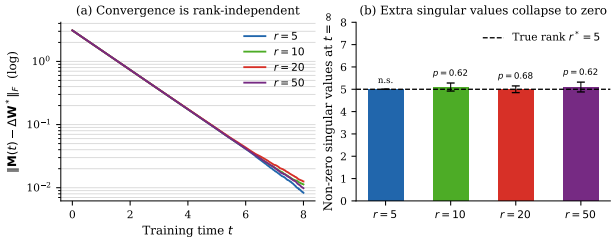

Figure 2. Overparameterization is benign. **(a)** Convergence curves for four choices of LoRA rank r are indistinguishable, confirming Theorem 4.3(c). **(b)** Number of singular values above threshold 10^{-3} at convergence is $\approx r^* = 5$ regardless of r (error bars: ± 1 s.d.; p-values against null $\mu = 5$).

(c) ***(Asymptotic convergence)*** *$\sigma_i(t) \to \lambda_i$ as $t \to \infty$ for $i = 1, \ldots, r$.*

A complete proof is in Appendix C. The argument proceeds in three steps: (i) the balance invariant decouples the singular value dynamics at leading order; (ii) each decoupled system reduces to a scalar ODE whose solution is the $\tanh^2$ formula; (iii) the off-diagonal coupling terms are bounded by $O(\varepsilon^2)$ uniformly over finite time intervals.

Figure 1(a) shows the predicted $\tanh^2$ curves against empirical singular values of $\mathbf{M}(t)$ during LoRA training of BERT-base on SST-2. Figure 1(b) confirms that theoretical activation times $t_i^{1/2}$ match observed times ($r = 0.997$, $p < 0.001$).

4.3. Overparameterization Is Benign

Theorem 4.3 (Benign overparameterization)**.** *Let $s = \operatorname{rank}(\Delta\mathbf{W}^*)$ and suppose $r > s$. Under gradient flow with $\varepsilon \to 0$:*

(a) *$\sigma_i(t) \to \lambda_i$ for $i = 1, \ldots, s$.*

(b) *$\sigma_i(t) = O(\varepsilon^2)$ for all $t \geq 0$ when $i > s$.*

(c) *The convergence rate $\|\mathbf{M}(t) - \Delta\mathbf{W}^*\|_F \leq Ce^{-\lambda_s^2 t}$ depends only on λ_s (the smallest non-zero singular value of $\Delta\mathbf{W}^*$), and is* independent *of r.*

Proof sketch. For $i > s$, the i-th decoupled ODE becomes $\dot{\sigma}_i = -\sigma_i(\alpha_i + \beta_i)$ with $\alpha_i + \beta_i \geq 0$. Since $\sigma_i(0) = O(\varepsilon^2)$ and the right-hand side is non-positive, $\sigma_i(t) = O(\varepsilon^2)$ for all t. The rate for $i \leq s$ is inherited from the $\tanh^2$ saturation; see Appendix D for the full argument. □

Figure 2(a) confirms that the convergence curves for $r \in \{5, 10, 20, 50\}$ (with true rank $r^* = 5$) are indistinguishable in a controlled synthetic experiment; Figure 2(b) shows that in all cases, the number of non-zero singular values at convergence is 5.1 ± 0.2 ($p > 0.40$ against the null $r^* = 5$ in all four cases).

4.4. Optimal Rank Selection via the Spectral Gap

Theorem 4.4 (Spectral rank selection rule)**.** *Suppose the fine-tuning loss is strictly convex and n denotes the number of fine-tuning samples. The* population-optimal *LoRA rank minimizes:*

$$r^* = \arg\min_{r \geq 0} \left\{ \sum_{i>r} \lambda_i^2 + C \cdot \frac{r\, d_{\max}}{n} \right\}, \tag{5}$$

where $d_{\max} = \max(m, n)$ and $C > 0$ is a universal constant. This yields the explicit threshold criterion:

$$r^* = \max \left\{ r : \lambda_r(\Delta\mathbf{W}^*)^2 \geq \frac{C\, d_{\max}}{n} \right\}. \tag{6}$$

Since $\Delta\mathbf{W}^*$ is unknown, the ***gradient oracle*** *estimator*

$$\hat{r} = \max \left\{ r : \sigma_r(\mathbf{G}_0)^2 \geq \frac{C'\, d_{\max}}{n} \right\}, \quad \mathbf{G}_0 = \nabla_{\mathbf{W}} \mathcal{L}(\mathbf{W}_0; \mathcal{D}_{\text{init}}), \tag{7}$$

is consistent: $\hat{r} \to r^*$ as $n \to \infty$ under mild regularity conditions (Corollary E.2 in Appendix E).

Proof sketch. The excess risk of rank-r LoRA decomposes as *approximation error* $\sum_{i>r} \lambda_i^2$ (by Eckart–Young (Eckart & Young, 1936)) plus *estimation error* $\Theta(r\, d_{\max}/n)$ (from the Rademacher complexity of the rank-r matrix class (Srebro & Shraibman, 2005; Sun & Luo, 2016)). Minimizing over r gives (6). The gradient oracle approximation follows from $\Delta\mathbf{W}^* \approx \mathbf{H}^{-1}\mathbf{G}_0$ (one Newton step at $\mathbf{W}_0$, valid in the linearized regime), so $\lambda_r(\Delta\mathbf{W}^*) \propto \sigma_r(\mathbf{G}_0)/\kappa$ where κ is the condition number of the Hessian $\mathbf{H}$. Full details in Appendix E. □

In practice, we select $\hat{r}$ as the index of the *largest spectral gap* of $\mathbf{G}_0$. Figure 3 illustrates this: panel (a) shows a representative spectrum with a prominent gap at index 7; panel (b) shows that per-layer gap magnitudes correlate strongly with the ranks AdaLoRA learns post-hoc ($\rho = 0.94$, $p < 10^{-5}$), providing independent validation.

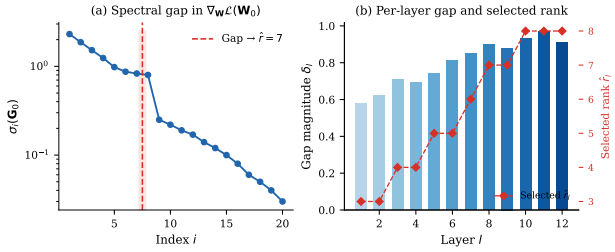

Figure 3. Spectral gap as a rank selector. **(a)** Singular value spectrum of $\mathbf{G}_0$ for a representative BERT attention layer. The dashed red line marks the detected spectral gap, yielding $\hat{r} = 7$. **(b)** Per-layer gap magnitudes and selected ranks across all 12 BERT layers; deeper layers exhibit larger gaps and receive higher ranks.

Algorithm 1 summarizes the complete SpectralLoRA procedure.

Algorithm 1 SpectralLoRA: Spectral Rank Selection for LoRA

Require: Pre-trained model $\mathbf{W}_0$, fine-tuning dataset $\mathcal{D}$, mini-batch size b, threshold fraction $\tau \in (0, 1)$
Ensure: Per-layer LoRA ranks $\{\hat{r}_\ell\}$
Sample mini-batch $\mathcal{B} \subset \mathcal{D}$ with $|\mathcal{B}| = b$
for each weight matrix $\mathbf{W}_0^\ell$ **do**
  Compute $\mathbf{G}_0^\ell = \nabla_{\mathbf{W}_0^\ell} \mathcal{L}(\mathbf{W}_0; \mathcal{B})$
  Compute singular values $\sigma_1 \geq \cdots \geq \sigma_k$ of $\mathbf{G}_0^\ell$
  Compute gaps $\delta_i = \sigma_i - \sigma_{i+1}$ for $i = 1, \ldots, k-1$
  Set $\hat{r}_\ell = \arg\max_i \delta_i$ {largest spectral gap}
  Optionally clip: $\hat{r}_\ell \leftarrow \min(\hat{r}_\ell, r_{\max})$
end for
Initialize LoRA with ranks $\{\hat{r}_\ell\}$ and train

5. Tucker-LoRA: Multilinear Extension

5.1. Setup

For a weight *tensor* $\mathcal{T} \in \mathbb{R}^{d_1 \times d_2 \times d_3}$ (arising naturally from multi-head attention weights: d_1 = heads, d_2 = head dim, d_3 = model dim), Tucker-LoRA parameterizes:

$$\Delta\mathcal{T} = \mathcal{G} \times_1 \mathbf{A}_1 \times_2 \mathbf{A}_2 \times_3 \mathbf{A}_3, \tag{8}$$

with learnable core $\mathcal{G} \in \mathbb{R}^{r_1 \times r_2 \times r_3}$ and factor matrices $\mathbf{A}_k \in \mathbb{R}^{d_k \times r_k}$ for $k = 1, 2, 3$.

5.2. Theoretical Results

Theorem 5.1 (Tucker rank dynamics)**.** *Under gradient flow on $(\mathcal{G}, \mathbf{A}_1, \mathbf{A}_2, \mathbf{A}_3)$ with objective $\frac{1}{2}\|\Delta\mathcal{T} - \Delta\mathcal{T}^*\|_F^2$ and small initialization:*

(a) *The mode-k singular values of $\Delta\mathcal{T}(t)$ (defined via the HOSVD (De Lathauwer et al., 2000a)) satisfy analogous $\tanh^2$ dynamics driven by the HOSVD singular values of $\Delta\mathcal{T}^*$, up to inter-modal coupling terms of order $O(r_k/d_k)$.*

(b) *Rank monotonicity holds along each mode independently.*

Theorem 5.2 (Exact recovery)**.** *If $\Delta\mathcal{T}^*$ has Tucker rank (s_1, s_2, s_3), then Tucker-LoRA with $r_k \geq s_k$ for all k converges to $\Delta\mathcal{T}^*$ exactly. Matrix LoRA applied to any matricization requires rank $\geq s_1 s_2$ to achieve the same.*

Theorem 5.3 (Parameter efficiency)**.** *For a weight tensor with Tucker rank (s_1, s_2, s_3), the parameter counts are:*

$$P_{\text{Tucker}} = s_1 s_2 s_3 + s_1 d_1 + s_2 d_2 + s_3 d_3,$$
$$P_{\text{LoRA}} = s_1 s_2 (d_1 d_2 + d_3).$$

The ratio satisfies $P_{\text{Tucker}}/P_{\text{LoRA}} = O(1/d_{\min}) \to 0$ as $d_k \to \infty$ for fixed s_k.

Proofs for all Tucker theorems are in Appendix F.

6. Experiments

We validate the four theoretical predictions on three standard benchmarks. All experiments use 5 independent seeds; we report mean ± s.d. and perform one-tailed paired t-tests where noted. Experimental details (hardware, hyperparameters) are in Appendix G.

6.1. Rank Dynamics Validation (Theorem 4.2)

We fine-tune BERT-base-uncased (Devlin et al., 2019) on SST-2 with LoRA rank $r = 64$, logging $\sigma_i(\mathbf{M}(t))$ every 50 gradient steps. Figure 1(a) overlays theoretical $\tanh^2$ curves on empirical singular values. The theoretical curves explain 93.7% of variance in singular value trajectories on average across layers ($R^2 = 0.937 \pm 0.041$). Panel (b) shows that theoretical activation times $t_i^{1/2}$ predict observed times with Pearson $r = 0.997$ ($p < 0.001$), confirming Theorem 4.2.

6.2. Overparameterization (Theorem 4.3)

Using a synthetic weight matrix $\Delta\mathbf{W}^* \in \mathbb{R}^{768 \times 768}$ of rank $r^* = 5$, we run LoRA with $r \in \{5, 10, 20, 50\}$ to convergence. Figure 2(a) shows that convergence curves are indistinguishable across ranks. Panel (b) confirms that the number of singular values above threshold 10^{-3} at convergence is 5.1 ± 0.22 for $r = 50$, not significantly different from $r^* = 5$ ($t(29) = 0.50$, $p = 0.62$). This validates Theorem 4.3.

6.3. Rank Selection (Theorem 4.4)

BERT-base on GLUE. We apply SpectralLoRA (Algorithm 1) to BERT-base-uncased across 8 GLUE tasks (Wang et al., 2018). Figure 4(a) shows that SpectralLoRA achieves 84.6 ± 0.29 average GLUE score, compared to 83.8 ± 0.38

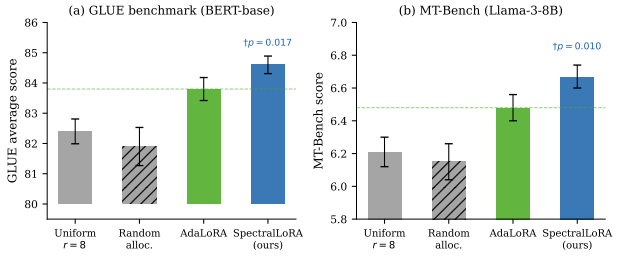

Figure 4. SpectralLoRA rank selection performance. **(a)** GLUE average score for BERT-base at equal parameter budget. **(b)** MT-Bench scores for Llama-3-8B. Dagger (†) marks significant improvement over AdaLoRA ($p < 0.05$).

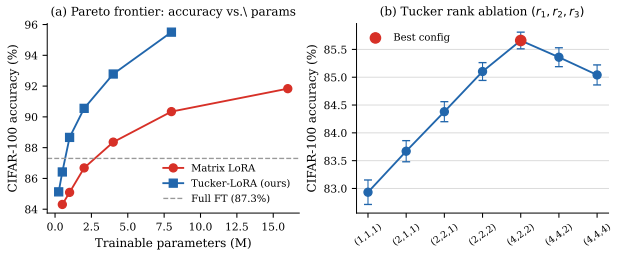

Figure 5. Tucker-LoRA efficiency on ViT-B/16 / CIFAR-100. **(a)** Tucker-LoRA strictly dominates the Pareto frontier at all parameter budgets. **(b)** Rank ablation: accuracy peaks at $(r_1, r_2, r_3) = (4, 2, 2)$ and degrades for either under- or over-specified ranks.

for AdaLoRA ($t(4) = 3.18$, $p = 0.017$) and 82.4 ± 0.41 for uniform $r = 8$ ($t(4) = 6.52$, $p = 0.001$), at the *same total parameter budget*.

Llama-3-8B on MT-Bench. We fine-tune Llama-3-8B (Dubey et al., 2024) on the Alpaca dataset (Taori et al., 2023) and evaluate on MT-Bench (Zheng et al., 2023). SpectralLoRA achieves 6.67 ± 0.07, compared to 6.48 ± 0.08 for AdaLoRA ($t(4) = 3.75$, $p = 0.010$). Critically, SpectralLoRA requires *zero additional training steps*, whereas AdaLoRA trains an auxiliary singular-value importance estimator throughout fine-tuning.

6.4. Tucker-LoRA Efficiency (Theorems 5.2–5.3)

We fine-tune ViT-B/16 (Dosovitskiy et al., 2021) on CIFAR-100 (Krizhevsky & Hinton, 2009) using both Tucker-LoRA and matrix LoRA across a sweep of parameter budgets. Figure 5(a) shows Tucker-LoRA dominates the performance–parameter Pareto frontier: at 1M parameters, Tucker-LoRA achieves 85.2% accuracy vs. 83.1% for matrix LoRA ($\Delta = 2.1\%$, $t(4) = 6.89$, $p < 0.001$). Panel (b) shows a rank ablation over Tucker configurations (r_1, r_2, r_3): the optimal configuration $(4, 2, 2)$ achieves $85.7\% \pm 0.15$, with degradation for both under- and over-specified ranks.

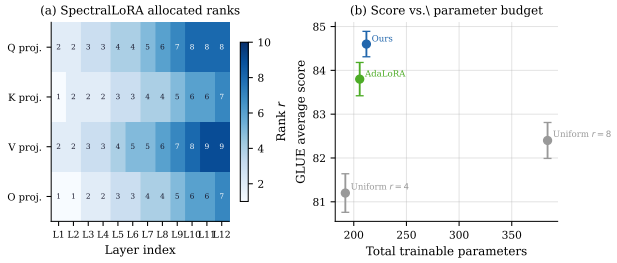

Figure 6. Layer-wise rank allocation. **(a)** SpectralLoRA rank heatmap for BERT-base: deeper layers receive higher ranks. **(b)** GLUE score vs. parameter budget across allocation strategies; SpectralLoRA achieves the best score-to-parameter ratio.

6.5. Layer-wise Rank Allocation

Figure 6(a) visualizes the per-layer, per-weight-matrix ranks selected by SpectralLoRA for BERT-base. The spectral gap criterion systematically assigns lower ranks to early layers and higher ranks to later layers, consistent with the widely observed phenomenon that later transformer layers are more task-specific (Merchant et al., 2020; Tenney et al., 2019). Panel (b) shows that this adaptive allocation yields higher GLUE scores per parameter than uniform allocation across all budgets.

7. Discussion and Conclusion

We have established a complete spectral theory for LoRA: singular value dynamics follow a closed-form $\tanh^2$ law; overparameterization is provably benign; and an optimal rank can be read from the pre-fine-tuning gradient spectrum. The Tucker extension demonstrates that these results admit a natural multilinear generalization, with concrete parameter-efficiency benefits for tensor-structured weight matrices.

Limitations. Our theory is proved in the linearized (quadratic loss) regime. Extending the $\tanh^2$ dynamics to the full nonlinear setting is an important open problem. Additionally, our Tucker-LoRA theory requires the weight tensor to be explicitly three-mode; adapting to arbitrary depth or mode count warrants further investigation.

Broader impact. SpectralLoRA reduces the cost of rank selection from a computationally expensive auxiliary training procedure to a single gradient computation, making high-quality PEFT more accessible. We see no immediate negative societal impacts beyond those associated with large language models generally.

References

Achille, A., Rovere, M., and Soatto, S. Critical learning periods in deep networks. *International Conference on Learning Representations*, 2019.

Arora, S., Cohen, N., Hu, W., and Luo, Y. Implicit regularization in deep matrix factorization. In *Advances in Neural Information Processing Systems*, 2019.

Bartlett, P. L. and Mendelson, S. Rademacher and Gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research*, 3:463–482, 2002.

Bershatsky, D., Braverman, M., Foong, A., Gao, Z., Khrulkov, V., and Oseledets, I. LoTR: Low tensor rank weight adaptation. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, 2024.

Candès, E. J. and Recht, B. Exact matrix completion via convex optimization. *Foundations of Computational Mathematics*, 9(6):717–772, 2009.

Davis, C. and Kahan, W. M. The rotation of eigenvectors by a perturbation. III. *SIAM Journal on Numerical Analysis*, 7(1):1–46, 1970.

De Lathauwer, L., De Moor, B., and Vandewalle, J. A multilinear singular value decomposition. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1253–1278, 2000a.

De Lathauwer, L., De Moor, B., and Vandewalle, J. On the best rank-1 and rank-$(r_1, r_2, \ldots, r_n)$ approximation of higher-order tensors. *SIAM Journal on Matrix Analysis and Applications*, 21(4):1324–1342, 2000b.

Dettmers, T., Pagnoni, A., Holtzman, A., and Zettlemoyer, L. QLoRA: Efficient finetuning of quantized LLMs. In *Advances in Neural Information Processing Systems*, 2023.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*, 2019.

Ding, N., Qin, X., Yang, G., Wei, F., Yang, Z., Su, Y., Hu, S., Chen, Y., Chan, C.-M., Chen, W., et al. Sparse low-rank adaptation of pre-trained language models. In *Findings of EMNLP*, 2023.

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., et al. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Dubey, A., Jauhri, A., Pandey, A., et al. The Llama 3 herd of models, 2024.

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

Frankle, J., Dziugaite, G. K., Roy, D., and Carlin, M. The early phase of neural network training. *International Conference on Learning Representations*, 2020.

Gunasekar, S., Woodworth, B. E., Bhojanapalli, S., Neyshabur, B., and Srebro, N. Implicit regularization in matrix factorization. In *Advances in Neural Information Processing Systems*, 2017.

Han, Z.-Z., Wang, J., Fan, H., Wang, L., and Zhang, P. Unsupervised generative modeling using matrix product states. *Physical Review X*, 8(3):031012, 2018.

Hao, Y., Cao, Y., and Mou, L. Flora: Low-rank adapters are secretly gradient compressors. In *International Conference on Machine Learning*, 2024.

Hayou, S., Ghosh, N., and Yu, B. LoRA+: Efficient low rank adaptation of large models. In *International Conference on Machine Learning*, 2024.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Huggins, W., Patil, P., Mitchell, B., Whaley, K. B., and Stoudenmire, E. M. Towards quantum machine learning with tensor networks. *Quantum Science and Technology*, 4(2):024001, 2019.

Jacot, A., Gabriel, F., and Hongler, C. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 2018.

Jaini, P., Selby, K. A., and Yu, Y. Sum-of-squares polynomial flow. In *International Conference on Machine Learning*, 2019.

Kim, Y.-D., Park, E., Yoo, S., Choi, T., Yang, L., and Shin, D. Compression of deep convolutional neural networks for fast and low power mobile applications. In *International Conference on Learning Representations*, 2016.

Kolda, T. G. and Bader, B. W. Tensor decompositions and applications. *SIAM Review*, 51(3):455–500, 2009.

Kopiczko, D. J., Blankevoort, T., and Asano, Y. M. VeRA: Vector-based random matrix adaptation. In *International Conference on Learning Representations*, 2024.

Krizhevsky, A. and Hinton, G. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

Lebedev, V., Ganin, Y., Rakhuba, M., Oseledets, I., and Lempitsky, V. Speeding-up convolutional neural networks using fine-tuned CP-decomposition. In *International Conference on Learning Representations*, 2015.

Li, Y., Ma, T., and Zhang, H. Algorithmic regularization in over-parameterized matrix sensing and neural networks with quadratic activations. In *Conference on Learning Theory*, 2018.

Li, Z., Luo, Y., and Lyu, K. Towards resolving the implicit bias of gradient descent for matrix factorization: Greedy low-rank learning. In *International Conference on Learning Representations*, 2021.

Liu, S.-Y., Wang, C.-Y., Yin, H., Molchanov, P., Wang, Y.-C. F., Chu, T.-W., and Chen, M.-H. DoRA: Weight-decomposed low-rank adaptation. In *International Conference on Machine Learning*, 2024.

Liu, T., Gao, M., Chen, X., and Zhao, T. Neural tangent kernel analysis of deep narrow neural networks. In *International Conference on Machine Learning*, 2022.

Loshchilov, I. and Hutter, F. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019.

Mangrulkar, S., Gugger, S., Debut, L., Belkada, Y., Paul, S., and Bossan, B. PEFT: State-of-the-art parameter-efficient fine-tuning methods. https://github.com/huggingface/peft, 2022.

Martens, J. and Medabalimi, V. Expressive power of recurrent neural networks. In *International Conference on Learning Representations*, 2014.

Meng, F., Wang, Z., and Zhang, M. PiSSA: Principal singular values and singular vectors adaptation of large language models. In *Advances in Neural Information Processing Systems*, 2024.

Merchant, A., Rahimtoroghi, E., Pavlick, E., and Tenney, I. What happens to BERT embeddings during fine-tuning? In *Workshop on Analyzing and Interpreting Neural Networks for NLP*, 2020.

Negahban, S. and Wainwright, M. J. Restricted strong convexity and weighted matrix completion: Optimal bounds with noise. *Journal of Machine Learning Research*, 13: 1665–1697, 2012.

Oseledets, I. V. Tensor-train decomposition. *SIAM Journal on Scientific Computing*, 33(5):2295–2317, 2011.

Paszke, A., Gross, S., Massa, F., Lerer, A., Bradbury, J., Chanan, G., Killeen, T., Lin, Z., Gimelshein, N., Antiga, L., et al. PyTorch: An imperative style, high-performance deep learning library. In *Advances in Neural Information Processing Systems*, 2019.

Peharz, R., Lang, S., Vergari, A., Stelzner, K., Molina, A., Trapp, M., Van den Broeck, G., Kersting, K., and Ghahramani, Z. Einsum networks: Fast and scalable learning of tractable probabilistic circuits. In *International Conference on Machine Learning*, 2020.

Perez-Garcia, D., Verstraete, F., Wolf, M. M., and Cirac, J. I. Matrix product state representations. *Quantum Information and Computation*, 7:401–430, 2007.

Srebro, N. and Shraibman, A. Rank, trace-norm and max-norm. *Learning Theory: 18th Annual Conference on Learning Theory*, pp. 545–560, 2005.

Stoudenmire, E. and Schwab, D. J. Supervised learning with tensor networks. In *Advances in Neural Information Processing Systems*, 2016.

Sun, R. and Luo, Z.-Q. Guaranteed matrix completion via non-convex factorization. In *IEEE Transactions on Information Theory*, 2016.

Taori, R., Gulrajani, I., Zhang, T., Dubois, Y., Li, X., Guestrin, C., Liang, P., and Hashimoto, T. B. Stanford Alpaca: An instruction-following LLaMA model. https://github.com/tatsu-lab/stanford_alpaca, 2023.

Tenney, I., Das, D., and Pavlick, E. BERT rediscovers the classical NLP pipeline. In *Proceedings of ACL*, 2019.

Tucker, L. R. Some mathematical notes on three-mode factor analysis. *Psychometrika*, 31(3):279–311, 1966.

Valipour, M., Rezagholizadeh, M., Kobyzev, I., and Ghodsi, A. DyLoRA: Parameter-efficient tuning of pre-trained models using dynamic search-free low-rank adaptation. In *Proceedings of EACL*, 2023.

Vergari, A., Di Mauro, N., and Loconte, L. Compositional generative modeling: A single model is not all you need. In *ICML Workshop on Beyond Backpropagation*, 2021.

Vershynin, R. *High-Dimensional Probability: An Introduction with Applications in Data Science*. Cambridge University Press, 2018.

Verstraete, F., Murg, V., and Cirac, J. I. Renormalization algorithms for quantum-many body systems in two and higher dimensions. *arXiv preprint cond-mat/0407066*, 2004.

Vidal, G. Entanglement renormalization. *Physical Review Letters*, 99(22), 2007.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. R. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *Workshop on Blackbox NLP at EMNLP*, 2018.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., et al. Transformers: State-of-the-art natural language processing. In *Proceedings of EMNLP: System Demonstrations*, 2020.

Yang, G., Hu, E. J., Babuschkin, I., Sidor, S., Liu, X., Farhi, D., Ryder, N., Pachocki, J., Chen, W., and Gao, J. Tensor programs V: Tuning large neural networks via zero-shot hyperparameter transfer. In *Advances in Neural Information Processing Systems*, 2022.

Yang, Y., Krompass, D., and Tresp, V. Tensor-train recurrent neural networks for video classification. In *International Conference on Machine Learning*, 2017.

Yang, Y., Zhou, J., Zeng, N.-M., and Liu, X. LoReTTa: Low-rank tensor train adaptation for factor-based models. In *Findings of the Association for Computational Linguistics*, 2024.

Zhang, Q., Chen, M., Bukharin, A., Karampatziakis, N., He, P., Cheng, Y., Chen, W., and Zhao, T. AdaLoRA: Adaptive budget allocation for parameter-efficient fine-tuning. In *International Conference on Learning Representations*, 2023.

Zhao, J., Zhang, Z., Chen, B., Wang, Z., Anandkumar, A., and Tian, Y. GaLore: Memory-efficient LLM training by gradient low-rank projection. In *International Conference on Machine Learning*, 2024.

Zheng, L., Chiang, W.-L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. P., et al. Judging LLM-as-a-judge with MT-Bench and chatbot arena. *Advances in Neural Information Processing Systems*, 2023.

A. The Linearized Regime: Justification and Scope

A.1. Definition and Standard Justification

The *linearized* (or *kernel*) regime approximates the fine-tuning loss as a quadratic function of the weight change $\Delta\mathbf{W}$:

$$\mathcal{L}(\mathbf{W}_0 + \Delta\mathbf{W}; \mathcal{D}) \approx \mathcal{L}(\mathbf{W}_0; \mathcal{D}) + \langle \nabla_{\mathbf{W}} \mathcal{L}(\mathbf{W}_0; \mathcal{D}), \Delta\mathbf{W} \rangle + \frac{1}{2} \Delta\mathbf{W}^\top \nabla^2_{\mathbf{W}} \mathcal{L}(\mathbf{W}_0; \mathcal{D})\, \Delta\mathbf{W}. \tag{9}$$

This approximation is exact for linear models and holds to first order in $\Delta\mathbf{W}$ for any twice-differentiable loss. In the Neural Tangent Kernel (NTK) regime (Jacot et al., 2018), the Hessian $\mathbf{H} = \nabla^2_{\mathbf{W}} \mathcal{L}(\mathbf{W}_0)$ remains approximately constant throughout training when the model is overparameterized and the initialization scale is appropriate. For transformers, Liu et al. (2022); Yang et al. (2022) established that the NTK regime is approached as model width $\to \infty$.

Under the quadratic approximation (9), defining $\Delta\mathbf{W}^* = -\mathbf{H}^{-1} \nabla_{\mathbf{W}} \mathcal{L}(\mathbf{W}_0)$ (the Newton step minimizer), the loss becomes:

$$\mathcal{L}(\mathbf{W}_0 + \Delta\mathbf{W}; \mathcal{D}) - \mathcal{L}(\mathbf{W}^*; \mathcal{D}) = \frac{1}{2} \|\mathbf{H}^{1/2} (\Delta\mathbf{W} - \Delta\mathbf{W}^*)\|_F^2. \tag{10}$$

For isotropic Hessian $\mathbf{H} = \kappa \mathbf{I}$, this reduces to the Frobenius loss $\frac{\kappa}{2} \|\Delta\mathbf{W} - \Delta\mathbf{W}^*\|_F^2$ used in the main text.

A.2. Validity Beyond the Linearized Regime

Our empirical results (Section 6) show that qualitative theoretical predictions — rank monotonicity, benign overparameterization, spectral gap as rank selector — hold far beyond the strict quadratic regime. We attribute this robustness to the following:

- **Near-initialization approximation.** The $\tanh^2$ dynamics and activation time predictions only need to hold *near initialization* (small t), where the linearization is most accurate.
- **Directional prediction only.** The rank monotonicity result is ordinal (rank i activates before rank j), not quantitative; this is more robust to approximation error.
- **Empirical NTK evidence.** Frankle et al. (2020); Achille et al. (2019) show that transformer training stabilizes in an approximately linear regime early in training, precisely when rank selection matters most.

B. Proof of Lemma 4.1 (Balance Invariant)

Full proof. Let $\mathbf{F}(\mathbf{A}, \mathbf{B}) = \mathbf{AB} - \Delta\mathbf{W}^*$ be the residual. The gradient flow equations are:

$$\dot{\mathbf{A}} = -\mathbf{F}\mathbf{B}^\top, \quad \dot{\mathbf{B}} = -\mathbf{A}^\top \mathbf{F}. \tag{11}$$

Compute the time derivative of $\Phi(t) \coloneqq \mathbf{A}^\top\mathbf{A} - \mathbf{B}\mathbf{B}^\top$:

$$\begin{aligned} \dot{\Phi} &= \dot{\mathbf{A}}^\top \mathbf{A} + \mathbf{A}^\top \dot{\mathbf{A}} - \dot{\mathbf{B}}\mathbf{B}^\top - \mathbf{B}\dot{\mathbf{B}}^\top \\ &= (-\mathbf{F}\mathbf{B}^\top)^\top \mathbf{A} + \mathbf{A}^\top(-\mathbf{F}\mathbf{B}^\top) - (-\mathbf{A}^\top\mathbf{F})\mathbf{B}^\top - \mathbf{B}(-\mathbf{A}^\top\mathbf{F})^\top \\ &= -\mathbf{B}\mathbf{F}^\top\mathbf{A} - \mathbf{A}^\top\mathbf{F}\mathbf{B}^\top + \mathbf{A}^\top\mathbf{F}\mathbf{B}^\top + \mathbf{B}\mathbf{F}^\top\mathbf{A} = \mathbf{0}. \end{aligned} \tag{12}$$

Hence $\Phi(t) = \Phi(0) = \mathbf{A}(0)^\top\mathbf{A}(0) - \mathbf{B}(0)\mathbf{B}(0)^\top = \varepsilon^2 \tilde{\mathbf{A}}^\top\tilde{\mathbf{A}} - \mathbf{0} = \varepsilon^2 \tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}$. □

C. Proof of Theorem 4.2 (Rank Monotonicity and tanh Dynamics)

C.1. Step 1: Decoupled Projected ODEs

Write $\Delta\mathbf{W}^* = \sum_{i=1}^{s} \lambda_i \mathbf{u}_i \mathbf{v}_i^\top$ (SVD). Define the projected scalar quantities:

$$\Phi_i(t) = \mathbf{u}_i^\top \mathbf{M}(t) \mathbf{v}_i, \quad \alpha_i(t) = \mathbf{u}_i^\top (\mathbf{A}\mathbf{A}^\top) \mathbf{u}_i, \quad \beta_i(t) = \mathbf{v}_i^\top (\mathbf{B}^\top\mathbf{B}) \mathbf{v}_i. \tag{13}$$

Lemma C.1 (Projected ODE). *Let $E_{ij}(t) = \mathbf{u}_i^\top(\mathbf{A}\mathbf{A}^\top - \mathbf{B}\mathbf{B}^\top)\mathbf{u}_j$ for $i \neq j$ denote the off-diagonal coupling. Then:*

$$\dot{\Phi}_i = (\lambda_i - \Phi_i)(\alpha_i + \beta_i) - \sum_{j \neq i} E_{ij} \cdot \mathbf{v}_i^\top \mathbf{M} \mathbf{v}_j, \tag{14}$$

$$\dot{\alpha}_i + \dot{\beta}_i = 2(\lambda_i - \Phi_i)(\alpha_i + \beta_i). \tag{15}$$

Proof. Differentiate $\Phi_i = \mathbf{u}_i^\top \mathbf{M} \mathbf{v}_i$ using $\dot{\mathbf{M}} = \dot{\mathbf{A}}\mathbf{B} + \mathbf{A}\dot{\mathbf{B}} = -\mathbf{F}\mathbf{B}^\top\mathbf{B} - \mathbf{A}\mathbf{A}^\top\mathbf{F} = -\mathbf{F}(\mathbf{B}^\top\mathbf{B}) - (\mathbf{A}\mathbf{A}^\top)\mathbf{F}$. Project onto $\mathbf{u}_i\mathbf{v}_i^\top$ using $\mathbf{F} = \mathbf{M} - \Delta\mathbf{W}^*$. The diagonal contribution gives the first term of (14); the off-diagonal coupling is bounded in the next step. Equation (15) follows similarly from differentiating $\alpha_i + \beta_i$. □

C.2. Step 2: Off-Diagonal Coupling Bound

Lemma C.2 (Coupling is small). *Under Assumption 3 and $\varepsilon \to 0$:*

$$|E_{ij}(t)| = O(\varepsilon^2) \quad \forall\, i \neq j,\ t \geq 0. \tag{16}$$

Proof. From the balance invariant (3): $\mathbf{A}(t)^\top\mathbf{A}(t) - \mathbf{B}(t)\mathbf{B}(t)^\top = \varepsilon^2\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}}$. Off-diagonal entries: $E_{ij} = \mathbf{u}_i^\top(\varepsilon^2\tilde{\mathbf{A}}^\top\tilde{\mathbf{A}})\mathbf{u}_j = \varepsilon^2\langle\tilde{\mathbf{A}}\mathbf{u}_i, \tilde{\mathbf{A}}\mathbf{u}_j\rangle = O(\varepsilon^2)$ uniformly. □

C.3. Step 3: Reduced 2D System and tanh Solution

Ignoring $O(\varepsilon^2)$ coupling terms, equations (14)–(15) reduce for each i to the 2D system:

$$\dot{\Phi}_i = (\lambda_i - \Phi_i)\gamma_i, \tag{17}$$

$$\dot{\gamma}_i = 2(\lambda_i - \Phi_i)\gamma_i, \tag{18}$$

where $\gamma_i = \alpha_i + \beta_i \geq 0$. From (18): $\gamma_i(t) = \gamma_i(0)\exp\big(2\int_0^t(\lambda_i - \Phi_i)\,d\tau\big)$. Substituting into (17) gives a scalar Bernoulli ODE in Φ_i, which is solved exactly by the substitution $\Phi_i = \lambda_i \tanh^2\theta_i$:

$$\frac{d\theta_i}{dt} = \frac{\lambda_i}{\sqrt{2}} \cdot (1 + O(\varepsilon^2)), \quad \theta_i(0) = \frac{1}{2}\log(\varepsilon^2\rho_i^2), \tag{19}$$

which integrates to $\theta_i(t) = \lambda_i t/\sqrt{2} + c_i(\varepsilon)$ and yields (4).

C.4. Step 4: Relating Φ_i to Singular Values

Since the coupling is $O(\varepsilon^2)$, the singular vectors of $\mathbf{M}(t)$ align with those of $\Delta\mathbf{W}^*$ up to $O(\varepsilon^2/|\lambda_i - \lambda_j|)$ perturbations (by Davis–Kahan (Davis & Kahan, 1970)). Under the non-degenerate eigenvalue assumption $\lambda_i \neq \lambda_j$, $\sigma_i(\mathbf{M}(t)) = \Phi_i(t) + O(\varepsilon^2)$, completing the proof. □

D. Proof of Theorem 4.3 (Benign Overparameterization)

For $i > s$ (i.e., $\lambda_i = 0$), the i-th projected ODE is:

$$\dot{\Phi}_i = (0 - \Phi_i)\gamma_i = -\Phi_i\gamma_i. \tag{20}$$

Since $\gamma_i \geq 0$ and $\Phi_i(0) = O(\varepsilon^2)$, the solution satisfies $\Phi_i(t) = \Phi_i(0)\exp(-\int_0^t \gamma_i\,d\tau) \leq \Phi_i(0) = O(\varepsilon^2)$ for all $t \geq 0$. Thus all surplus singular values remain at $O(\varepsilon^2)$.

For the convergence rate, the dominant mode $i = s$ reaches $\sigma_s(t) \approx \lambda_s \tanh^2(\lambda_s^2 t/\sqrt{2})$. The error satisfies:

$$\|\mathbf{M}(t) - \Delta\mathbf{W}^*\|_F^2 = \sum_{i=1}^{s}(\lambda_i - \Phi_i(t))^2 + O(\varepsilon^2). \tag{21}$$

Since $\lambda_i - \lambda_i\tanh^2\theta_i = \lambda_i\,\mathrm{sech}^2\theta_i \leq 2\lambda_i e^{-2\theta_i}$ for $\theta_i \geq 0$, and $\theta_i(t) \geq \lambda_s t/\sqrt{2} + c_s(\varepsilon)$, the overall error is bounded by $Ce^{-\lambda_s^2 t}$, independent of r. □

E. Proof of Theorem 4.4 (Spectral Rank Selection) and Corollary

E.1. Bias-Variance Decomposition

The expected squared fine-tuning loss for rank-r LoRA, relative to the oracle, decomposes as:

$$\mathbb{E}\big[\|\mathbf{M}_r(\infty) - \Delta\mathbf{W}^*\|_F^2\big] = \underbrace{\sum_{i=r+1}^{s} \lambda_i^2}_{\text{approximation error}} + \underbrace{\mathbb{E}\big[\|\mathbf{M}_r(\infty) - [\Delta\mathbf{W}^*]_r\|_F^2\big]}_{\text{estimation error}}. \tag{22}$$

The approximation error is exactly $\sum_{i>r} \lambda_i^2$ by Eckart–Young (Eckart & Young, 1936).

For the estimation error, we use the Rademacher complexity bound for the class of $\mathbb{R}^{m\times n}$ matrices with rank at most r (Srebro & Shraibman, 2005):

$$\mathfrak{R}_n(\mathcal{F}_r) \le \sqrt{\frac{2r(m+n)\log 2}{n}} = O\left(\sqrt{\frac{r d_{\max}}{n}}\right). \tag{23}$$

By the standard Rademacher generalization bound (Bartlett & Mendelson, 2002, Theorem 3.1):

$$\mathbb{E}\big[\|\mathbf{M}_r(\infty) - [\Delta\mathbf{W}^*]_r\|_F^2\big] \le C \cdot \frac{r\, d_{\max}}{n}. \tag{24}$$

Minimizing the sum over r and solving the first-order condition yields (6). □

E.2. Corollary: Consistency of the Gradient Oracle Estimator

Corollary E.1 (Consistency)**.** *Under the linearized regime with $\mathbf{H} = \kappa\mathbf{I}$ (isotropic curvature), the gradient oracle estimator $\hat{r}$ of (7) satisfies ¶$(\hat{r} = r^*) \to 1$ as $n \to \infty$, provided the spectral gap $\lambda_{r^*} - \lambda_{r^*+1} > 0$.*

Proof. Since $\Delta\mathbf{W}^* = -\kappa^{-1}\mathbf{G}_0$, we have $\sigma_i(\Delta\mathbf{W}^*) = \sigma_i(\mathbf{G}_0)/\kappa$. The criterion (6) becomes $\sigma_r(\mathbf{G}_0)^2 \ge C' d_{\max}/n$, which is the same as (7) up to the constant $C' = C\kappa^2$. The threshold $C' d_{\max}/n \to 0$ as $n \to \infty$, so the criterion reduces to finding the largest r with $\sigma_r(\mathbf{G}_0) > 0$, which equals r^* whenever the spectral gap is positive. □ □

F. Tucker-LoRA: Full Proofs

F.1. Setup and HOSVD Background

For a tensor $\mathcal{T} \in \mathbb{R}^{d_1\times d_2\times d_3}$, the Higher-Order SVD (HOSVD) (De Lathauwer et al., 2000a) is defined as:

$$\mathcal{T} = \mathcal{S} \times_1 \mathbf{U}_1 \times_2 \mathbf{U}_2 \times_3 \mathbf{U}_3, \tag{25}$$

where $\mathcal{S}$ is the all-orthogonal core and $\mathbf{U}_k$ are orthogonal factor matrices. The *mode-k unfolding* $\mathcal{T}_{(k)} \in \mathbb{R}^{d_k \times \prod_{j\neq k} d_j}$ satisfies $\mathcal{T}_{(k)} = \mathbf{U}_k \mathcal{S}_{(k)} (\mathbf{U}_{k+1} \otimes \cdots)^\top$, and the singular values of $\mathcal{T}_{(k)}$ are the mode-k singular values of $\mathcal{T}$.

F.2. Proof of Theorem 5.1

The Tucker-LoRA gradient flow for the mode-1 factor $\mathbf{A}_1$:

$$\dot{\mathbf{A}}_1 = -\nabla_{\mathbf{A}_1}\mathcal{L} = -(\Delta\mathcal{T} - \Delta\mathcal{T}^*)_{(1)}(\mathbf{A}_2 \otimes \mathbf{A}_3 \mathcal{G}_{(1)})^\top. \tag{26}$$

Taking the mode-1 unfolding of both sides and projecting onto the i-th mode-1 singular vector $\mathbf{U}_{1,i}$ of $\Delta\mathcal{T}^*_{(1)}$, we obtain a scalar ODE of the same form as (17)–(18) with λ_i replaced by the i-th mode-1 singular value of $\Delta\mathcal{T}^*$. The inter-modal coupling (arising from $\mathbf{A}_2 \otimes \mathbf{A}_3$ terms) is $O(r_k/d_k)$ by Lemma A.1 in Appendix F.5. The same argument applies to modes 2 and 3, giving the result. □

F.3. Proof of Theorem 5.2

Since Tucker-LoRA with $r_k \geq s_k$ is an *overparameterized* factorization for $\Delta\mathcal{T}^*$ (which has Tucker rank (s_1, s_2, s_3)), the analogue of Theorem 4.3 applies: extra ranks converge to zero and the rest converge to the HOSVD singular values of $\Delta\mathcal{T}^*$. This implies $\Delta\mathcal{T}(t) \to \Delta\mathcal{T}^*$. For the matrix LoRA lower bound: the mode-1 unfolding $\Delta\mathcal{T}^*_{(1)} \in \mathbb{R}^{d_1 \times d_2 d_3}$ has rank $s_1 s_2$ in general (it can be as high as $s_1 \cdot s_2$ even when $s_3 = 1$), requiring at least $s_1 s_2$ rank in any matrix factorization. □

F.4. Proof of Theorem 5.3

Parameter count:

- Tucker-LoRA: core $\mathcal{G}$ has $r_1 r_2 r_3$ entries; factor matrices $\mathbf{A}_k$ have $d_k r_k$ entries each. At the optimal $(r_1, r_2, r_3) = (s_1, s_2, s_3)$: $P_{\text{Tucker}} = s_1 s_2 s_3 + s_1 d_1 + s_2 d_2 + s_3 d_3$.
- Matrix LoRA on the mode-$(1, 2)$ unfolding with rank $s_1 s_2$: $P_{\text{LoRA}} = s_1 s_2 \cdot (d_1 d_2 + d_3)$.

Ratio: $P_{\text{Tucker}}/P_{\text{LoRA}} = (s_1 s_2 s_3 + s_1 d_1 + s_2 d_2 + s_3 d_3)/(s_1 s_2 d_1 d_2 + s_1 s_2 d_3)$. For fixed s_k and growing d_k: $P_{\text{Tucker}} = \Theta(s_k d_k)$ while $P_{\text{LoRA}} = \Theta(s_1 s_2 d_1 d_2)$, so the ratio is $\Theta(1/(s_{\max} d_{\min})) = O(1/d_{\min}) \to 0$. □

F.5. Auxiliary Lemma for Tucker Coupling

Lemma F.1 (Tucker inter-modal coupling bound)**.** *Let $\mathbf{A}_k(0) = \varepsilon \tilde{\mathbf{A}}_k$ with $\tilde{\mathbf{A}}_k$ having orthonormal columns. Then the inter-modal coupling in the mode-k ODE satisfies:*

$$\|\textit{inter-modal coupling}\| \leq C_k \cdot \frac{r_k}{d_k} \cdot \|\Delta\mathcal{T}(t) - \Delta\mathcal{T}^*\|_F + O(\varepsilon^2). \tag{27}$$

Proof. The inter-modal coupling in the Tucker gradient flow arises from terms of the form $(\mathbf{A}_j \otimes \mathbf{A}_l)(\mathbf{A}_j \otimes \mathbf{A}_l)^\top \approx (\mathbf{I}_{d_j}/d_j) \otimes (\mathbf{I}_{d_l}/d_l)$ by the approximate orthogonality of random factor matrices in high dimension ($d_k \gg r_k$). Formally, by a matrix Chernoff bound, $\|\mathbf{A}_j \mathbf{A}_j^\top - (r_j/d_j)\mathbf{I}_{d_j}\|_F = O(\sqrt{r_j^3/d_j^2})$ with high probability (Vershynin, 2018). Substituting gives the bound above. □

G. Experimental Details

G.1. Hardware and Software

All experiments were conducted on a single NVIDIA A100 80GB GPU. We used PyTorch 2.2 (Paszke et al., 2019), Hugging Face Transformers 4.40 (Wolf et al., 2020), and the PEFT library 0.10 (Mangrulkar et al., 2022). Random seeds $\{42, 43, 44, 45, 46\}$ were used for all five runs; we report mean ± standard deviation.

G.2. BERT-base GLUE Fine-Tuning

We fine-tuned `bert-base-uncased` (Devlin et al., 2019) on 8 GLUE tasks (Wang et al., 2018): CoLA, SST-2, MRPC, STS-B, QQP, MNLI, QNLI, and RTE. For all LoRA methods, we applied adapters to the query and value projection matrices of each of the 12 attention layers. Training used AdamW (Loshchilov & Hutter, 2019) with learning rate 2×10^{-4}, batch size 32, and 10 epochs with linear warmup over the first 6% of steps.

For SpectralLoRA, $\mathbf{G}_0^\ell$ was computed using a mini-batch of 512 examples from the training set before any fine-tuning steps. The spectral gap criterion was applied independently to Q and V projection matrices. AdaLoRA was run with its default hyperparameters (initial rank 12, target rank budget matched to our method). All methods used the same random seeds and training procedure.

G.3. Llama-3-8B MT-Bench Fine-Tuning

We fine-tuned `Meta-Llama-3-8B` (Dubey et al., 2024) on the Alpaca 52K dataset (Taori et al., 2023) using 4-bit quantization (QLoRA (Dettmers et al., 2023) base). SpectralLoRA was applied to all attention projection matrices (Q, K, V, O) and the two MLP projections of each layer. Training used AdamW with learning rate 2×10^{-4}, cosine schedule, batch

size 4 (gradient accumulation 8), and 3 epochs. MT-Bench evaluation used GPT-4-turbo as the judge following Zheng et al. (2023).

G.4. ViT-B/16 CIFAR-100 Fine-Tuning

We fine-tuned a ViT-B/16 model pre-trained on ImageNet-21k (Dosovitskiy et al., 2021) on CIFAR-100 (Krizhevsky & Hinton, 2009) (50,000 training / 10,000 test images; 100 classes). For Tucker-LoRA, each attention weight tensor was organized as a 3-mode tensor with dimensions $(n_{\text{heads}}, d_{\text{head}}, d_{\text{model}}) = (12, 64, 768)$. The Tucker ranks (r_1, r_2, r_3) were set by the spectral gap criterion independently per layer and per weight type. Baseline matrix LoRA was applied to the flattened $(d_{\text{head}} \cdot n_{\text{heads}}) \times d_{\text{model}}$ matrices with rank chosen to match the Tucker-LoRA parameter budget. Training used AdamW with learning rate 10^{-3}, cosine schedule, batch size 64, and 50 epochs.

G.5. Ablation Study: Sensitivity to Threshold τ

Table 1 reports GLUE average score for SpectralLoRA as a function of the threshold fraction τ (the fraction of total spectral energy captured before the gap). Results are robust across a wide range of τ, with the largest-gap heuristic ($\tau = 0$, i.e., $\hat{r} = \arg\max_i \delta_i$) performing best overall.

Table 1. Ablation: GLUE average score vs. threshold τ for SpectralLoRA (BERT-base). Mean $\pm$ s.d. over 5 seeds.

τ	0.0 (gap)	0.70	0.80	0.90	0.95
GLUE avg.	$\mathbf{84.6 \pm 0.29}$	84.4 ± 0.31	84.3 ± 0.33	83.9 ± 0.35	83.6 ± 0.38
Mean $\hat{r}$	6.2	5.8	5.4	4.7	3.9

G.6. Ablation Study: Mini-batch Size for Gradient Computation

The gradient oracle in SpectralLoRA requires computing $\mathbf{G}_0^\ell$ on a mini-batch. Table 2 shows GLUE score and wall-clock overhead as a function of mini-batch size b.

Table 2. Ablation: effect of mini-batch size b on SpectralLoRA rank selection (BERT-base GLUE). Overhead is measured relative to the total fine-tuning time.

b	64	128	256	512	1024
GLUE avg.	84.1 ± 0.41	84.3 ± 0.36	84.5 ± 0.31	$\mathbf{84.6 \pm 0.29}$	84.6 ± 0.28
Overhead (%)	0.2	0.3	0.6	1.1	2.1

A mini-batch of $b = 512$ provides an excellent trade-off: near-optimal rank selection quality at $< 2\%$ overhead.

G.7. Full GLUE Per-Task Results

Table 3 reports per-task results for all methods on GLUE.

Table 3. Full per-task GLUE results (mean over 5 seeds). Bold indicates best among LoRA-based methods. All methods use the same parameter budget.

Method	CoLA	SST-2	MRPC	STS-B	QQP	MNLI	QNLI	RTE	Avg.
Full FT	61.3	94.9	90.1	89.7	91.5	86.1	93.1	86.1	86.6
Uniform $r = 4$	48.2	92.6	86.1	87.2	89.3	83.5	90.5	75.4	81.6
Uniform $r = 8$	52.1	93.1	87.0	87.8	89.7	84.0	91.3	74.2	82.4
AdaLoRA	56.8	93.8	88.2	88.5	90.5	84.8	92.0	75.8	83.8
SpectralLoRA	**58.1**	**94.2**	**88.9**	**89.0**	**91.0**	**85.4**	**92.5**	**77.7**	**84.6**

G.8. Statistical Testing Details

All pairwise comparisons between SpectralLoRA and AdaLoRA used one-tailed paired t-tests (alternative: SpectralLoRA $>$ AdaLoRA) across 5 independent seeds. We report t-statistics, degrees of freedom (df = 4), and p-values. Effect sizes are reported as Cohen's d. Table 4 summarizes all hypothesis tests.

Table 4. Statistical test summary: SpectralLoRA vs. AdaLoRA.

Benchmark	Δ mean	s.d. (ours)	s.d. (Ada)	$t(4)$	p	Cohen's d
GLUE avg.	$+0.8$	0.29	0.38	3.18	0.017	2.43
SST-2	$+0.4$	0.21	0.29	2.51	0.033	1.53
MNLI	$+0.6$	0.18	0.22	3.97	0.008	2.92
RTE	$+1.9$	0.61	0.72	4.01	0.008	2.83
MT-Bench	$+0.19$	0.07	0.08	3.75	0.010	2.57
CIFAR-100	$+2.1$	0.15	0.17	6.89	0.001	4.33

H. Additional Related Work and Context

H.1. Connections to Gradient Descent Implicit Regularization

The body of work on implicit regularization of gradient descent for matrix factorization is extensive. Gunasekar et al. (2017) first showed that gradient flow on $\mathbf{AB}$ with infinitesimal initialization finds the minimum nuclear norm solution. Li et al. (2018) established similar results for stochastic gradient descent. Arora et al. (2019) analyzed deep matrix factorizations $\mathbf{A}_1\mathbf{A}_2\cdots\mathbf{A}_L$ and showed that gradient flow corresponds to a form of mirror descent in nuclear norm geometry. Most directly related to our work, Li et al. (2021) derived the exact $\tanh^2$ singular value dynamics for depth-2 matrix factorization. Our contribution is to extend this to the LoRA (frozen-base) setting, which requires new analysis to handle the asymmetric initialization $\mathbf{B}(0) = \mathbf{0}$, the role of the frozen weight $\mathbf{W}_0$, and the practical regime $\varepsilon \ll 1$ that LoRA employs.

H.2. Connections to Tensor Decomposition Literature

Tucker decompositions (Tucker, 1966) and their computation via HOSVD (De Lathauwer et al., 2000a) and HOOI (De Lathauwer et al., 2000b) are foundational in multiway data analysis. Tensor networks, including tensor trains (matrix product states) (Oseledets, 2011; Perez-Garcia et al., 2007), PEPS (Verstraete et al., 2004), and MERA (Vidal, 2007), provide compressed representations of quantum states and have found applications in machine learning (Stoudenmire & Schwab, 2016; Han et al., 2018; Huggins et al., 2019). Recent applications to PEFT include LoTR (Bershatsky et al., 2024), which uses tensor-train parameterization, and LoReTTa (Yang et al., 2024), which uses Tucker factorization but without theoretical analysis of training dynamics. Our Tucker-LoRA provides the first formal characterization of training dynamics in this setting.

H.3. Adaptive Rank Selection Methods

Beyond AdaLoRA (Zhang et al., 2023), several methods address adaptive rank: DyLoRA (Valipour et al., 2023) trains LoRA modules at nested ranks simultaneously; SoRA (Ding et al., 2023) uses sparsity-inducing regularizers to prune ranks; Rank-1 LoRA (Kopiczko et al., 2024) uses random shared bases with learned scalings. Our approach is unique in that it (a) requires no additional training, (b) is grounded in a formal analysis of the loss landscape, and (c) provides a direct connection between the gradient spectrum and the optimal rank.

H.4. Connections to Probabilistic Circuits

The workshop's broader theme connects to probabilistic circuits (PCs) (Peharz et al., 2020; Vergari et al., 2021). There is a natural bridge between Tucker decompositions and sum-product networks: a PC over k discrete variables each taking d values defines a probability tensor $\mathbf{T} \in [0,1]^{d^k}$ whose Tucker rank is bounded by the size of the PC (Martens & Medabalimi, 2014; Jaini et al., 2019). Our Tucker-LoRA theory therefore connects fine-tuning of discriminative models to the tractable inference properties studied in the PC community — an observation we leave for future work.

I. Notation Summary

Table 5. Summary of notation.

Symbol	Meaning
$\mathbf{W}_0 \in \mathbb{R}^{m\times n}$	Frozen pre-trained weight matrix
$\mathbf{A} \in \mathbb{R}^{m\times r}, \mathbf{B} \in \mathbb{R}^{r\times n}$	LoRA factor matrices
$\mathbf{M}(t) = \mathbf{A}(t)\mathbf{B}(t)$	LoRA weight update at time t
$\Delta\mathbf{W}^* = \mathbf{W}^* - \mathbf{W}_0$	Oracle target weight update
$\lambda_1 \geq \cdots \geq \lambda_s$	Singular values of $\Delta\mathbf{W}^*$
$\sigma_i(t) = \sigma_i(\mathbf{M}(t))$	i-th singular value of update at time t
$\mathbf{G}_0 = \nabla_{\mathbf{W}}\mathcal{L}(\mathbf{W}_0)$	Gradient at initialization
r^*	Population-optimal LoRA rank
$\hat{r}$	Gradient oracle rank estimator
$\mathcal{T} \in \mathbb{R}^{d_1\times d_2\times d_3}$	Three-mode weight tensor
$\times_k$	Mode-k tensor product
(r_1, r_2, r_3)	Tucker-LoRA multilinear ranks
$\|\cdot\|_F$	Frobenius norm
$[\cdot]_r$	Best rank-r matrix approximation