How Finite-Rank Bottlenecks Shape the Low-Rank Adaptation Landscape

Long Nguyen-Chi[⋆,1], **Quynh Nguyen**[1], **Thanh Nguyen-Cung**[1], **Binh T. Nguyen**[⋆,1,2]
[1]College of Engineering & Computer Science, VinUniversity
[2]Center for AI Research, VinUniversity, Hanoi, Vietnam.

Abstract

Low-rank adaptation (LoRA) has become the standard method for parameter-efficient fine-tuning of large pretrained models, yet theoretical explanations for why low-rank updates suffice remain incomplete. Most existing analyses rely on the Neural Tangent Kernel (NTK) regime, which linearizes the network around the pretrained weights. In this work, we take a different approach that requires no linearization: our key observation is that, once the backbone is frozen, the loss sees each adapted weight matrix only through a finite-rank linear map induced by the frozen activations. Trace-norm regularization then forces the optimizer to pick a low-rank matrix among all matrices that produce the same observed output. This yields global minimizers whose ranks are controlled by the rank of this map, which we call the *bottleneck rank*, rather than the sample size. For multi-head attention, we obtain tighter rank upper bounds by exploiting two additional invariances: row-wise softmax is insensitive to additive row constants in the score matrix, and value updates are observable only after attention weighting and frozen output projection. For the factorized LoRA objective with weight decay, we prove a complementary result on the nonconvex side: after an arbitrarily small generic positive semidefinite perturbation, every first-order stationary point is rank-deficient once the LoRA rank exceeds a threshold determined by the bottleneck rank. Finally, we show that global minimizers of the perturbed objective are near-optimal for the original trace-norm problem, and we provide uniform generalization bounds governed by the spectral structure of the frozen features.

1 Introduction

Fine-tuning large pretrained models on downstream tasks is effective but expensive: updating all parameters of a billion-scale model requires substantial memory and compute. Parameter-efficient fine-tuning (PEFT) solves this problem by adapting a large pretrained model while updating only a small subset of parameters. Some of the representative approaches include adapter modules (Houlsby et al., 2019), prefix tuning (Li and Liang, 2021), prompt tuning, and low-rank adaptation (LoRA, Hu et al. 2022). Among these, LoRA stands out as an attractive option because it keeps the pretrained backbone frozen and modifies selected linear maps by adding trainable low-rank updates. More formally, for a weight matrix $W_0 \in \mathbb{R}^{m\times n}$, LoRA parameterizes the adapted weight as

$$W = W_0 + T, \qquad T = UV^\top, \qquad U \in \mathbb{R}^{m\times r}, V \in \mathbb{R}^{n\times r},$$

The rank r is typically chosen by the practitioner as a hyperparameter rather than something derived from the fine-tuning objective itself (Hu et al., 2022; Schulman and Lab, 2025). A fundamental theoretical question is therefore:

⋆Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

When does the regularized fine-tuning objective itself admit a low-rank global minimizer, and what controls the rank?

Most existing theory addresses nearby questions through linearization or simplified training dynamics. In the NTK regime, Jang et al. (2024) shows that full fine-tuning admits a low-rank solution and that sufficiently overparameterized LoRA eliminates spurious local minima. Afzal et al. (2025) analyze fine-tuning through linearization under an explicit ℓ_2 bias toward the pretrained model. Beyond classical linearization, Kim et al. (2025) study LoRA dynamics and prove a dichotomy between convergence to a low-rank global minimizer and convergence to a qualitatively different high-rank solution. These works are complementary to ours, but they do not answer a more structural question: what property of the objective function makes low-rank optima available in the first place?

Contributions. Our answer starts with a simple observation: once the backbone is frozen, the loss does not observe an adapted matrix T directly, but only through a finite-rank linear map determined by the frozen activations. Concretely, if $H = [h_1, \ldots, h_N]$ is the matrix of frozen input features across the N training samples, the empirical loss depends on T only through the product TH. We call $\mathrm{rank}(\Phi)$, where $\Phi(T) = TH$, the *bottleneck rank* of the adapted block. The same principle applies blockwise to multiple adapted stages and attention heads. Trace-norm regularization then selects a low-rank representative among all matrices that induce the same observed output. This yields exact low-rank existence results without linearizing the network around the pretrained weights. For multi-head attention (Vaswani et al., 2017), the same principle becomes sharper because attention introduces additional invariances: row-wise softmax is unchanged by additive row constants in the score matrix, and value updates are observable only after attention weighting and the frozen output projection.

Table 1: Comparison with the closest related theoretical analyses of LoRA.

	This work	Jang et al. (2024)	Kim et al. (2025)
Requires NTK linearization	No	Yes	No
Multilayer multi-head attention bound	Yes (Thm. 4.4)	No	No
Spurious local minima excluded	No (Prop. 4.9)	Yes (NTK)	Partial
Additional loss assumptions	Differentiability and stagewise C^1 regularity	NTK-regime conditions	Restricted strong convexity and smoothness
Generalization bounds	Yes	Yes	No

Our contribution is summarized as follows.

- **Low-rank representer theorem for trace-norm fine-tuning.** For a frozen network in which an adapted block T is observed only through a finite-rank linear map $\Phi(T)$ of bottleneck rank d_Φ, we prove that the trace-norm-regularized problem has a global minimizer whose rank is controlled by d_Φ; see Theorem 4.1. The multi-stage extension is Theorem 4.3.
- **Rank bounds from attention invariances.** For multi-head attention, query and key updates are observable only through centered score subspaces because row-wise softmax is invariant to row-wise additive constants. Value updates are observable only after attention weighting and frozen output projection. Combining these invariances with projector compression gives the optimizer-dependent subspace rank bounds in Theorem 4.4.
- **Generic first-order rank deficiency for factorized LoRA.** For the weight-decayed LoRA factorization, we prove that after a generic positive semidefinite perturbation, full-column-rank first-order stationary points are excluded whenever the Pataki–Barvinok rank threshold exceeds the dimension of the associated family of symmetric KKT matrices; see Theorem 4.7. The multistage and attention consequence are Theorem C.9 and Corollary 4.8. This is a first-order exclusion theorem, not a no-spurious-local-minima theorem; Proposition 4.9 gives an explicit limitation.
- **Perturbation transfer and generalization.** We show that global minimizers of the perturbed factorized objective are near-optimal for the original trace-norm-regularized problem when the perturbation is small; see Theorem 4.10.

2 Related works

Linearized and kernel views of fine-tuning. The closest prior work to ours is Jang et al. (2024), which studies LoRA in the NTK regime. Using K_{out} for the number of output coordinates in the linearized training predictions, their low-rank existence theorem gives a solution satisfying $\frac{r(r+1)}{2} \leq K_{\text{out}}N$. Thus the rank scale in that analysis is $r = \lfloor \frac{\sqrt{8K_{\text{out}}N+1}-1}{2} \rfloor$, or $r = O(\sqrt{K_{\text{out}}N})$ in the scalar-output convention. Our results improve this comparison in two ways. First, our bottleneck reduction is exact for frozen upstream computation and does not linearize the network around the pretrained weights. Second, the optimizer's rank scale is $r \leq \min\{K_{\text{out}}, N\}$ in Attention, and this is better than Jang et al. (2024)'s bound. Afzal et al. (2025) also analyzed fine-tuning through linearization under an explicit ℓ_2 bias toward the pretrained model and quantified when nonlinear fine-tuning stays close to the linearized problem. In contrast, our reduction is exact once upstream computation is frozen; no approximation-by-linearization argument is needed.

Optimization theory beyond linearization. Recent works also study LoRA outside the lazy regime. Kim et al. (2025) analyzes the LoRA loss landscape and proves that training converges either to a low-rank global minimizer or to a qualitatively different high-rank solution. Xu et al. (2025) study gradient flow for LoRA in matrix factorization and show how initialization and singular-space alignment shape the final error. Liu et al. (2025) compares the landscapes of LoRA, GaLore, and full-rank adaptation through Polyak–Łojasiewicz-type regions and spurious local minima. Zhang and Pilanci (2024) analyzes a Riemannian preconditioner for LoRA in the infinite-width setting. These works ask how LoRA training behaves. Our question comes earlier: before one analyzes dynamics, when does the frozen regularized objective itself admit low-rank optima, and when must first-order stationary points of the factored problem be rank-deficient?

Attention and Transformer expressivity. Since transformers are built around attention (Vaswani et al., 2017), we instantiate our framework for $Q/K/V$ projections in attention modules. Orthogonally, a rich literature studies the expressive power and limitations of self-attention and transformers (Yun et al., 2020; Likhosherstov et al., 2021; Hahn, 2020; Zhou et al., 2024). Complementary work studies what functions low-rank adaptation can represent; Zeng and Lee (2024) characterized the expressive power of low-rank adaptation in several settings. For Transformer architectures, Hu et al. (2025) studied computational limitations of LoRA fine-tuning.

3 Problem setup

We study fine-tuning of a pretrained network in which all weights are frozen except for a specified collection of linear maps. The full notation tables are collected in Appendix A; here we introduce only what is needed to state the main results.

Single adapted block. For a frozen weight matrix $W_0 \in \mathbb{R}^{m\times n}$, we write the adapted weight as $W = W_0 + T$ with update $T \in \mathbb{R}^{m\times n}$. Given training data $S_{\text{data}} = \{(x_i, y_i)\}_{i=1}^N$, let $h_i \in \mathbb{R}^n$ denote the frozen input activation at this layer for sample x_i, and define the *frozen feature matrix*

$$H \coloneqq [h_1, \ldots, h_N] \in \mathbb{R}^{n\times N}.$$

When the input h_i is independent of the update T and all other parameters are fixed, the network output on sample x_i depends on T only through Th_i (Lemma B.1).

Stacking across samples, the empirical loss depends on T only through $\Phi(T) \coloneqq TH \in \mathbb{R}^{m\times N}$. The map Φ is linear, and its range has dimension exactly $m\,\text{rank}(H) \leq m\min\{n, N\}$.

Definition 3.1 (Bottleneck rank)**.** *For a linear map $\Phi : \mathbb{R}^{m\times n} \to \mathcal{E}$ through which the loss observes an adapted block T, the bottleneck rank is*

$$d_\Phi \coloneqq \text{rank}(\Phi) = \dim \mathcal{R}(\Phi).$$

When $\Phi(T) = TH$ as above, we have $d_\Phi = m\,\text{rank}(H)$. The bottleneck rank quantifies how much information the frozen network extracts from the update T: any two updates T, T' with $TH = T'H$ produce identical training predictions.

Trace-norm regularization and the LoRA factorization. We consider the nuclear-norm-regularized (trace-norm-regularized) objective

$$\min_{T\in\mathbb{R}^{m\times n}} \mathcal{L}^{\mathrm{vanilla}}(T) + \lambda\|T\|_\star, \qquad \lambda > 0, \tag{3.1}$$

where $\mathcal{L}^{\mathrm{vanilla}}(T)$ is the empirical loss with update T and $\|T\|_\star$ denotes the nuclear norm. The factorized LoRA objective with weight decay at rank r is

$$\mathcal{L}^{\mathrm{lora}}_\lambda(U,V) \coloneqq \mathcal{L}^{\mathrm{vanilla}}(UV^\top) + \frac{\lambda}{2}\big(\|U\|_F^2 + \|V\|_F^2\big), \quad U\in\mathbb{R}^{m\times r}, V\in\mathbb{R}^{n\times r}.$$

By the standard equivalence between nuclear-norm regularization and balanced weight decay (Recht et al., 2010), the factorized problem has the same global minimum as (3.1) restricted to rank $\le r$:

$$\min_{U,V} \mathcal{L}^{\mathrm{lora}}_\lambda(U,V) = \min_{\substack{T\in\mathbb{R}^{m\times n}\\ \mathrm{rank}(T)\le r}} \Big\{\mathcal{L}^{\mathrm{vanilla}}(T) + \lambda\|T\|_\star\Big\}.$$

Thus, whenever the unrestricted trace-norm problem (3.1) admits a global minimizer of rank $\le r$, the rank-r LoRA factorization incurs no loss of optimality.

Attention mechanism For multi-head attention with head set $\mathcal{A}$, head $a\in\mathcal{A}$ produces adapted queries, keys, and values on sample i:

$$\widetilde{Q}_i^{(a)} = (W_{Q,0}^{(a)} + T_Q^{(a)})H_{Q,i}, \quad \widetilde{K}_i^{(a)} = (W_{K,0}^{(a)} + T_K^{(a)})H_{K,i}, \quad \widetilde{V}_i^{(a)} = (W_{V,0}^{(a)} + T_V^{(a)})H_{V,i},$$

where $H_{S,i}$ is the frozen token matrix entering projection $S\in\{Q,K,V\}$ for sample i. Queries and keys share output dimension d_k; values have output dimension d_v. The attention matrix and block output are

$$A_i^{(a)} = \mathrm{softmax}_{\mathrm{row}}\left(\frac{(\widetilde{Q}_i^{(a)})^\top \widetilde{K}_i^{(a)}}{\sqrt{d_k}} + C_i^M\right), \qquad O_i = \sum_{a\in\mathcal{A}} W_O^{(a)}\widetilde{V}_i^{(a)}(A_i^{(a)})^\top,$$

where C_i^M is an attention mask and $W_O^{(a)}$ is the frozen output-projection block for head a. We write $H_S \coloneqq [H_{S,1},\dots,H_{S,N}]$ for the sample-stacked feature matrices. For sample i, let $\varrho_{Q,i}$ and $\varrho_{K,i}$ denote the numbers of query and key token positions, respectively, so that $A_i^{(a)}\in\mathbb{R}^{\varrho_{Q,i}\times\varrho_{K,i}}$. In the stage-indexed notation used below, at stage t, these quantities are written as $\varrho_{t,Q,i}$ and $\varrho_{t,K,i}$.

Multiple adapted layers. When L consecutive attention modules are adapted (indexed so that larger t denotes a more upstream stage), the frozen features at stage t may depend on the updates at stages $\tau > t$. This creates a stagewise bottleneck structure, formalized as follows.

Definition 3.2 (L-stage block-linear dependence)**.** *Let $L\in\mathbb{N}$ and let $\mathcal{S}_t$ be finite index sets for $t\in[L]$. Let $\mathcal{Z}_{L+1}$ be a singleton and let $\mathcal{Z}_t$ be finite-dimensional inner-product spaces for $t\in[L]$. We say that a loss map $\mathcal{L}(\{T_{t,S}\}_{t\in[L],S\in\mathcal{S}_t})$ has* L-stage block-linear dependence *if there exist maps*

$$\Phi_{t,S} : \mathbb{R}^{m_{t,S}\times n_{t,S}}\times\mathcal{Z}_{t+1}\to\mathcal{E}_{t,S}, \qquad g_t : \Big(\prod_{S\in\mathcal{S}_t}\mathcal{E}_{t,S}\Big)\times\mathcal{Z}_{t+1}\to\mathcal{Z}_t \quad (t\in[L]),$$

and $g_0 : \mathcal{Z}_1\to\mathbb{R}$ such that:

1. *for every fixed $z_{t+1}\in\mathcal{Z}_{t+1}$, the map $T\mapsto\Phi_{t,S}(T,z_{t+1})$ is linear;*
2. *if z_{L+1} denotes the unique point of $\mathcal{Z}_{L+1}$ and*

$$z_t \coloneqq g_t\Big(\{\Phi_{t,S}(T_{t,S},z_{t+1})\}_{S\in\mathcal{S}_t}, z_{t+1}\Big), \quad t = L, L-1,\dots,1,$$

then $\mathcal{L}(\{T_{t,S}\}) = g_0(z_1)$.

The key feature is that each block $T_{t,S}$ enters stage t through a linear map $\Phi_{t,S}$, but the stage transition maps g_t may couple the blocks in an arbitrary (possibly nonconvex) way. Stagewise affine shifts (e.g., the frozen term $W_{t,S,0}H_{t,S}(z_{t+1})$) are absorbed into the stage maps g_t. For attention, the concrete map is $\Phi_{t,(a,S)}(T,z) = TH_{t,S}(z)$, where $H_{t,S}(z)$ is the frozen token matrix at stage t. An output projection adapted alongside $Q/K/V$ receives inputs that depend on those same-stage updates.

Stagewise bottleneck ranks. For $z_{t+1} \in \mathcal{Z}_{t+1}^{\mathrm{reach}}$ (the set of reachable upstream states; see Definition A.2 in the appendix), define

$$d_{t,S}(z_{t+1}) \coloneqq \dim \mathcal{R}(\Phi_{t,S}(\cdot, z_{t+1})), \qquad d_{t,S}^{\max} \coloneqq \sup_{z_{t+1} \in \mathcal{Z}_{t+1}^{\mathrm{reach}}} d_{t,S}(z_{t+1}).$$

The recursive *upstream dimension budget* is

$$\beta_{L+1} \coloneqq 0, \qquad \beta_t \coloneqq \beta_{t+1} + \sum_{S \in \mathcal{S}_t} d_{t,S}^{\max}, \quad t = L, L-1, \ldots, 1.$$

Intuitively, β_t upper-bounds the intrinsic dimension of the stage-t reachable state family.

Standing assumptions. We use two mild assumptions throughout, which are exactly the standard evaluation-mode condition.

Assumption 1 (Well-posed objective)**.** *The loss function g (or g_0 in L-stage formulation, through which the empirical loss factors) is bounded below and lower semicontinuous, and all maps in the stagewise decomposition are continuous.*

Assumption 2 (Deterministic forward pass)**.** *The forward pass is deterministic per sample (no Dropout, no BatchNorm).*

4 Main results

All of our results follow from the same principle: freezing the backbone makes each adapted block observable only through a finite-rank linear map, and trace-norm regularization selects a low-rank representative among all updates that produce the same observations.

4.1 Low-rank minimizers for a single frozen block

Theorem 4.1 (Low-rank global minimizers from bottleneck rank)**.** *Let Assumptions 1 and 2 hold. Suppose the empirical loss depends on T only through a linear map $\Phi(T)$ of bottleneck rank d_Φ. Then the trace-norm-regularized problem* (3.1) *admits a global minimizer. Moreover, for every global minimizer $\tilde{T}$, there exists a global minimizer $T^\star$ with*

$$\Phi(T^\star) = \Phi(\tilde{T}) \qquad \textit{and} \qquad \frac{\mathrm{rank}(T^\star)\big(\mathrm{rank}(T^\star)+1\big)}{2} \le d_\Phi.$$

Consequently,

$$\mathrm{rank}(T^\star) \le \min\left\{m, n, \left\lfloor \frac{\sqrt{8d_\Phi + 1} - 1}{2} \right\rfloor \right\}.$$

Proof idea. Among all PSD lifts $Z \succeq 0$ with $\Phi(\bar{Z}) = \Phi(\tilde{T})$, choose one that minimizes $\mathrm{tr}(Z)$. Any extreme point of this spectrahedron satisfies the Barvinok–Pataki rank bound (Pataki, 1998; Barvinok, 2001): its rank r obeys $r(r+1)/2 \le d_\Phi$. The off-diagonal block $\bar{Z}$ recovers $T^\star$ with the same or smaller nuclear norm. The full proof is in Appendix C.1.

Remark 4.2 (Single-block rank scales)**.** *For a single frozen block with $\Phi(T) = TH$, we have $d_\Phi \le m\,\mathrm{rank}(H) \le mN$. When the observed output dimension is smaller than $\min\{K_{\mathrm{out}}, \mathrm{rank}(H)\} \le \min\{K_{\mathrm{out}}, N\}$, this has the same worst-case counting scale as $O(\sqrt{K_{\mathrm{out}} N})$ in Jang et al. (2024).*

Extension to L consecutive adapted layers. The next theorem concerns multiple adapted stages, each potentially containing several heads and projections. The same argument extends by stagewise induction: fix upstream stages, apply Theorem 4.1 to the conditional objective at each stage, and propagate downward.

Theorem 4.3 (Low-rank global minimizers for L stages)**.** *Under Assumptions 1 and 2, suppose the loss has L-stage block-linear dependence (Definition 3.2) and let all $\lambda_{t,S} > 0$. Then the trace-norm-regularized objective admits a global minimizer $\{T_{t,S}^\star\}$ satisfying, for every stage $t \in [L]$ and block $S \in \mathcal{S}_t$,*

$$\frac{r_{t,S}(r_{t,S}+1)}{2} \le d_{t,S}(z_{t+1}^\star) \le d_{t,S}^{\max}, \qquad r_{t,S} \coloneqq \mathrm{rank}(T_{t,S}^\star).$$

The case $L = 1$ recovers Theorem 4.1. For generic frozen linear layers, these bounds can still be loose because $d_{t,S}^{\max}$ may be large. The next subsection shows that multi-head attention *substantially tightens* the effective bottleneck.

4.2 Rank bounds for multi-head attention

For attention modules, the bottleneck rank from Theorem 4.3 can be refined using two invariances below that are specific to the attention mechanism.

Invariance 1: score centering. Row-wise softmax satisfies $\mathrm{softmax}_{\mathrm{row}}(M + C) = \mathrm{softmax}_{\mathrm{row}}(MP_{\mathbf{1}}^{\perp} + C)$, where $M, C \in \mathbb{R}^{q\times k}$ (allowing $-\infty$ entries in the mask C) and $P_{\mathbf{1}}^{\perp} = I_k - \frac{1}{k}\mathbf{1}_k\mathbf{1}_k^{\top} \in \mathbb{R}^{k\times k}$ (Lemma C.3). Thus, query/key scores are observed only after centering across key columns, reducing the key-token dimension from $\varrho_{K,i}$ to $\varrho_{K,i} - 1$. For queries and keys, the loss observes the update $T_Q^{(a)}$ only through the *centered key subspace*, and similarly for $T_K^{(a)}$ through the query subspace.

Invariance 2: value-projection compression. Value updates are observable to the loss only after two compressions: (i) attention-weighted pooling by $A_i^{(a)}$, which projects onto a subspace of dimension $\mathrm{rank}(A_i^{(a)})$; and (ii) the frozen output projection $W_O^{(a)}$.

Projector compression principle. Both invariances are instances of the following general lemma: if $\Phi(T) = \Phi(P_{\mathrm{L}}TP_{\mathrm{R}})$ for orthogonal projectors $P_{\mathrm{L}}, P_{\mathrm{R}}$, then $T^{\star} = P_{\mathrm{L}}\tilde{T}P_{\mathrm{R}}$ is an equally optimal global minimizer with $\mathrm{rank}(T^{\star}) \le \min\{\mathrm{rank}(P_{\mathrm{L}}), \mathrm{rank}(P_{\mathrm{R}})\}$ (Lemma C.1).

Combining these ingredients gives per-head, per-projection rank certificates:

Theorem 4.4 (Rank bounds for multi-head attention from invariances)**.** *Let Assumptions 1 and 2 hold. Consider L consecutive multi-head attention modules with adapted blocks $\mathcal{S}_t \subseteq \mathcal{A}_t \times \{Q, K, V\}$, fixed output-projection blocks $W_{t,O}^{(a)}$, and regularization weights $\lambda_{t,S}^{(a)} > 0$. Every attention row has at least one unmasked key. Then the nuclear-norm-regularized objective admits a global minimizer. Moreover, there exists a global minimizer $\{T_{t,S}^{(a)\star}\}$ that minimizes total rank $\sum_{t,a,S}\mathrm{rank}(T_{t,S}^{(a)})$ over the global minimizer set, and this minimizer satisfies:*

Query bound:

$$\mathrm{rank}(T_{t,Q}^{(a)\star}) \le \min\left\{d_{k,t}, \mathrm{rank}(H_{t,Q}^{\star}), \mathrm{rank}(H_{t,K}^{\star}), \sum_{i=1}^{N}(\varrho_{t,K,i} - 1)\right\} \quad \forall\,(a, Q) \in \mathcal{S}_t.$$

Key bound:

$$\mathrm{rank}(T_{t,K}^{(a)\star}) \le \min\left\{d_{k,t}, \mathrm{rank}(H_{t,Q}^{\star}), \mathrm{rank}(H_{t,K}^{\star}), \sum_{i=1}^{N}(\varrho_{t,K,i} - 1)\right\} \quad \forall\,(a, K) \in \mathcal{S}_t.$$

Value bound:

$$\mathrm{rank}(T_{t,V}^{(a)\star}) \le \min\left\{d_{v,t}, \mathrm{rank}(H_{t,V}^{\star}), \sum_{i=1}^{N}\mathrm{rank}(A_{t,i}^{(a)\star})\right\} \quad \forall\,(a, V) \in \mathcal{S}_t.$$

Here, $H_{t,S}^{\star}$ and $A_{t,i}^{(a)\star}$ are the frozen features and attention matrices evaluated at the same global minimizer whose rank is being bounded.

Interpretation. Our bounds have the following implications. First, query and key ranks are controlled by the centered key/query subspace. The term $\sum_{i=1}^{N}(\varrho_{t,K,i}-1)$ arises from the score-centering invariance. Second, value rank is controlled by attention sparsity. The term $\sum_{i=1}^{N}\mathrm{rank}(A_{t,i}^{(a)\star})$ reflects the effective number of distinct attention patterns. When attention is sparse or concentrated on a few dominant keys per query, $\mathrm{rank}(A_{t,i}^{(a)\star})$ is small, yielding a tight value-rank

bound. Third, the frozen output projection provides an unconditional value-rank bound. Since $\operatorname{rank}(T_{t,V}^{(a)\star}) \le \operatorname{rank}(W_O^{(a)\top}) \le d_{v,t}$, the value update can never exceed the per-head value dimension, regardless of the data.

Adapted output projections. The results aboved are extended for output projection in the next theorem. The preceding rank certificates assumed fixed output-projection weights. They extend structurally when one also adapts the full concatenated-head output projection, provided its input is treated as fixed only *conditionally* on the other adapted blocks.

Theorem 4.5 (Attention rank bounds with adapted output projections)**.** *In the setting of Theorem 4.4, suppose that at each selected stage t the full output projection $W_{t,O,0} \in \mathbb{R}^{d_{\mathrm{out},t} \times |\mathcal{A}_t| d_{v,t}}$ also receives an update $T_{t,O}$ of the same shape with nuclear-norm weight $\lambda_{t,O} > 0$. Then the nuclear-norm-regularized objective admits a global minimizer that minimizes the total rank of all adapted $Q/K/V/O$ blocks over the global minimizer set. At every such minimizer, the query and key bounds of Theorem 4.4 hold unchanged, and*

$$\operatorname{rank}(T_{t,V}^{(a)\star}) \le \min\left\{ d_{v,t}, \operatorname{rank}(H_{t,V}^{\star}), \sum_{i=1}^{N} \operatorname{rank}(A_{t,i}^{(a)\star}) \right\},$$

$$\operatorname{rank}(T_{t,O}^{\star}) \le \min\left\{ d_{\mathrm{out},t}, |\mathcal{A}_t| d_{v,t}, \sum_{i=1}^{N} \varrho_{t,Q,i} \right\},$$

where all features and attention matrices are evaluated at this minimizer.

A value update $T_{t,V}^{(a)}$ enters the stage output only after left multiplication by the head-a column block $\widetilde{W}_{t,O}^{(a)} \in \mathbb{R}^{d_{\mathrm{out},t} \times d_{v,t}}$ of the adapted output projection $W_{t,O,0} + T_{t,O}$. Adapting the output projection changes which directions of the $d_{v,t}$-dimensional head output are retained, but it cannot enlarge that output. The update $T_{t,O}$ differs from the $Q/K/V$ updates because its input, the attention-weighted values, depends on the same-stage $Q/K/V$ updates. The proof therefore compresses each block conditionally on the other blocks instead of using a joint linear map, and for the same reason Corollary 4.8 does not apply to $T_{t,O}$.

4.3 Generic rank deficiency of first-order stationary points

The results in Sections 4.1–4.2 concern *global* minimizers of the convex trace-norm objective. In practice, LoRA solves the *nonconvex* factorized problem $\mathcal{L}_{\lambda}^{\mathrm{lora}}(U, V)$ via gradient descent, which may converge to stationary points that are not global minimizers. We now show that the bottleneck structure also constrains the rank at stationary points of the factorized problem.

Assumption 3 (Differentiability for stationarity)**.** *The induced empirical loss $\mathcal{L}$ is differentiable with respect to every adapted block.*

Assumption 4 (Stagewise C^1 regularity for stagewise linear maps)**.** *In Definition 3.2, for every stage $t \in [L]$, every block $S \in \mathcal{S}_t$, every basis vector e_j of $\mathcal{E}_{t,S}$, and every basis matrix B_k of $\mathbb{R}^{m_{t,S} \times n_{t,S}}$, the maps $z_{t+1} \mapsto \Phi_{t,S}(\cdot, z_{t+1})^{\star}(e_j)$, $z_{t+1} \mapsto \Phi_{t,S}(B_k, z_{t+1})$ are C^1 on an open neighborhood of $\mathcal{Z}_{t+1}^{\mathrm{reach}}$. Here, $\Phi_{t,S}^{\star}$ is the adjoint map of $\Phi_{t,S}$. Moreover, each transition map g_t is C^1 on an open neighborhood of the reachable set of its arguments.*

Assumption 4 is a regularity assumption on how the moving linear observation maps depend on upstream states, and common evaluation-mode attention blocks satisfy this condition on bounded reachable sets.

Why a perturbation is needed. The KKT conditions at stationarity gives $(\mathcal{B}_{m,n}(G) + \lambda I_{m+n})R = 0$ with $G = \nabla_T \mathcal{L}$ and $\mathcal{B}_{m,n}(G) = \left[\begin{smallmatrix} 0 & G \\ G^{\top} & 0 \end{smallmatrix}\right]$. Full column rank of R forces an r-dimensional nullspace, whereas bottleneck structure restricts G to the low-dimensional range $\mathcal{R}(\Phi^{\star})$ (Lemma C.7). A generic PSD perturbation makes this low-dimensional symmetric-matrix family avoid such a high-dimensional nullspace.

Definition 4.6 (Perturbed LoRA objective)**.** *For perturbation $P \in \mathbb{S}_{+}^{m+n}$, define*

$$\mathcal{L}_{\lambda,P}^{\mathrm{lora}}(U, V) := \mathcal{L}^{\mathrm{vanilla}}(UV^{\top}) + \frac{\lambda}{2}\left(\|U\|_F^2 + \|V\|_F^2\right) + \frac{1}{2}\langle P, RR^{\top} \rangle.$$

Theorem 4.7 (Generic exclusion of full-rank stationary points)**.** *Fix $d \geq 1$, $r \geq 1$, $\lambda > 0$, and let $\mathcal{G} \subseteq \mathbb{S}^d$ be contained in a finite union of C^1 images of dimension at most p. If $\frac{r(r+1)}{2} > p$, then for Lebesgue-a.e. $P \in \mathbb{S}^d_+$, every matrix $R \in \mathbb{R}^{d\times r}$ satisfying $(G + \lambda I + P)R = 0$ for some $G \in \mathcal{G}$ has* $\operatorname{rank}(R) < r$.

Proof idea. For each $s \geq r$, the set of symmetric matrices with nullspace dimension exactly s is a smooth embedded submanifold of $\mathbb{S}^d$ of codimension $s(s+1)/2$. Therefore, the set of symmetric matrices with nullspace dimension at least r is a finite union of such manifolds, not generally a single submanifold. The family $\mathcal{G}+\lambda I$ has dimension at most p. When $r(r+1)/2 > p$, the translated family $\mathcal{G} + \lambda I + P$ generically misses this submanifold by a dimension-counting argument (Lemma C.6).

Application to attention. Using the stagewise gradient-range characterization (Lemma C.8), the family $\mathcal{G}_{t,S}$ for each attention block is C^1-parameterizable with dimension at most $\beta_{t+1} + d^{\max}_{t,S}$. This yields:

Corollary 4.8 (Generic rank deficiency for L attention modules)**.** *Under Assumptions 1,2, and mild differentiability conditions (Assumptions 3 and 4), consider L consecutive attention modules. For each adapted block $\alpha = (a, S) \in \mathcal{S}_t$, let $r^{\max}_{H,t,\alpha} \coloneqq \sup_{z_{t+1}} \operatorname{rank}(H_{t,S}(z_{t+1}))$. If*

$$\frac{r_{t,\alpha}(r_{t,\alpha}+1)}{2} > \beta_{t+1} + m_{t,\alpha} r^{\max}_{H,t,\alpha} \qquad \forall\, t \in [L], \alpha \in \mathcal{S}_t,$$

then for Lebesgue-a.e. PSD perturbations $\{P_{t,\alpha}\}$, every first-order stationary point of the perturbed LoRA objective satisfies $\operatorname{rank}(R_{t,\alpha}) < r_{t,\alpha}$.

Limitation: spurious local minima persist. This is a first-order *exclusion* theorem, not a no-spurious-local-minima result. Indeed, spurious local minima can occur even for arbitrarily overparameterized LoRA:

Proposition 4.9 (Spurious local minima for overparameterized LoRA)**.** *For any rank $r \geq 2$, any $\lambda > 0$, and any $\varepsilon \in (0, 1)$, there exists a smooth nonconvex loss g such that for every $P \in \mathbb{S}^2_+$ with $\|P\|_F \leq \varepsilon$, the perturbed rank-r LoRA objective has a strict local minimum at $R = 0$ that is not global.*

This clarifies the scope of the generic rank-deficiency result: it guarantees that stationary points are rank-deficient (the LoRA factorization does not waste its rank budget at stationary points), but it does not guarantee that all such points are globally optimal.

4.4 Perturbation transfer and generalization

Near-optimality of the perturbed factorization. The PSD perturbation in Theorem 4.7 can be taken arbitrarily small. The following result shows that global minimizers of the perturbed factorized objective remain near-optimal for the original trace-norm problem:

Theorem 4.10 (Perturbation-to-unperturbed transfer)**.** *Let $\mathcal{I}$ be any finite index set of adapted blocks. For each $j \in \mathcal{I}$, fix a factorization rank r_j and a regularization parameter $\lambda_j > 0$. Assume that the trace-norm-regularized problem admits a global minimizer $\{T^\star_j\}$ satisfying* $\operatorname{rank}(T^\star_j) \leq r_j \forall\, j \in \mathcal{I}$, *and let $R^\star_j \coloneqq \left[\begin{smallmatrix} U^\star_j \\ V^\star_j \end{smallmatrix}\right]$ for a balanced factorization of $T^\star_j$. Let $\{(\widehat{U}_j, \widehat{V}_j)\}$ be a global minimizer of the perturbed LoRA objective with parameters λ_j, P_j (Definition 4.6). Then*

$$\mathcal{L}(\{\widehat{T}_j\}) + \sum_j \lambda_j \|\widehat{T}_j\|_\star \leq \min_{\{T_j\}} \left\{ \mathcal{L}(\{T_j\}) + \sum_j \lambda_j \|T_j\|_\star \right\} + \frac{1}{2} \sum_j \langle P_j, R^\star_j R^{\star\top}_j \rangle.$$

In particular, if $\|P_j\|_2 \leq \varepsilon_j$, the gap is at most $\sum_j \varepsilon_j \|T^\star_j\|_\star$.

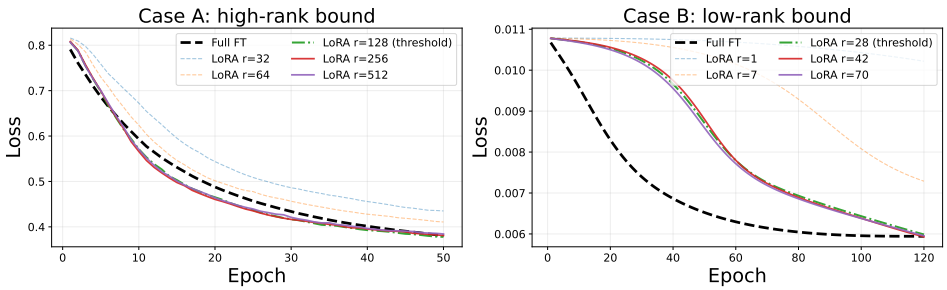
Our assumptions satisfy that the trace-norm-regularized problem admits a global minimizer, and this theorem can be applied to a single block index set $\mathcal{S}$, a stagewise index set $\mathcal{S}_t$, and the flattened attention index set $\widetilde{\mathcal{S}} \coloneqq \{(t, a, S) : t \in [L], (a, S) \in \mathcal{S}_t\}$.

Generalization. Uniform generalization bounds for trace-norm-bounded updates in frozen networks are governed by $\mathbb{E}[\|EH^\top\|_2]$, where E is a Rademacher sign matrix. The key identity is the trace/spectral duality:

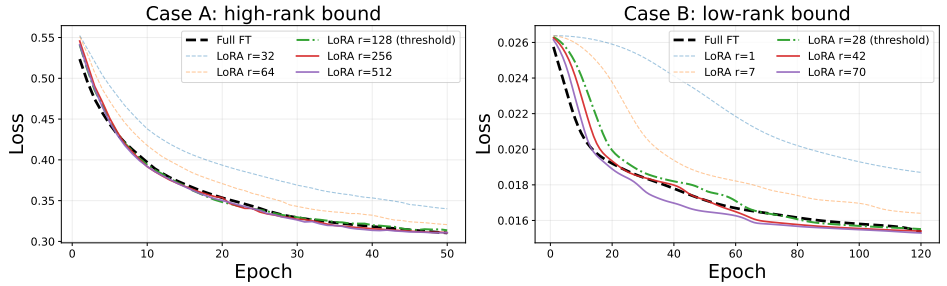
$$\sup_{\|T\|_\star \le D} \langle T, EH^\top \rangle_F = D\|EH^\top\|_2,$$

which reduces the Rademacher complexity of the loss class to the spectral norm of random-sign-weighted frozen features. The full single-block bound (Theorem B.5) and the L-stage attention extension (Theorem B.8) are stated and proved in Appendix B.

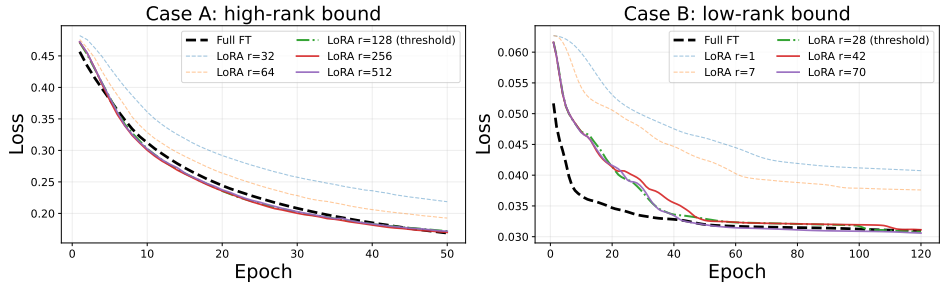
5 Experiment



(a) MNLI



(b) SST-2



(c) RTE

Figure 1: Training loss trajectories on GLUE classification tasks (MNLI, SST-2, and RTE) using Qwen2.5-1.5B across various LoRA ranks. The results demonstrate a clear threshold effect: ranks below the estimated threshold exhibit distinct, suboptimal convergence trajectories, whereas ranks at or above the threshold tightly cluster and converge to nearly indistinguishable final losses, closely matching the convergence point of full fine-tuning.

We test the predicted rank threshold on SST-2, MNLI, and RTE (Wang et al., 2018) using Qwen2.5-1.5B (Yang et al., 2025). Only one query-projection block in one head of layers 7 and 8 is adapted; after a classification-head warm-up, both the backbone and classifier are frozen. The threshold is the smallest rank at which the upstream and downstream rank proxies both saturate.

Two rank-bound regimes. For the selected query blocks, Theorem 4.4 gives the upper bound

$$\operatorname{rank}(T_{t,Q}^{(a)\star}) \le \min\left\{ d_{k,t}, \operatorname{rank}(H_{t,Q}^{\star}), \operatorname{rank}(H_{t,K}^{\star}), \sum_{i=1}^{N} (\varrho_{t,K,i} - 1) \right\} \quad \forall\, (a, Q) \in \mathcal{S}_t.$$

We evaluate in two regimes. We test a many-example regime where this estimate reaches the head cap $d_{k,t}$ and a few-example regime where it is strictly smaller. For Qwen2.5-1.5B, $d_{k,t} = 128$.

Results. Figure 1 shows a consistent threshold effect: ranks below the estimate follow distinct, worse loss trajectories, while ranks at or above it nearly overlap and approach the full fine-tuning loss. This is consistent with the high-rank LoRA/full-fine-tuning alignment observed by Jang et al. (2024). We also observe that high-rank LoRA runs converge to almost the same point as the full fine-tuning baseline, although their training dynamics differ, consistent with the related phenomenon observed by Jang et al. (2024).

6 Conclusion

In this paper, we identify that LoRA operates via a finite-rank mechanism where frozen layers project updates into low-dimensional subspaces. Through this, trace-norm regularization selects a low-rank representative among all updates producing the same observed outputs. This yields global low-rank minimizers without NTK linearization and gives rank bounds controlled by bottleneck dimensions rather than by the ambient number of trainable parameters.

For attention modules, we showed that the observable subspaces are smaller than the naive frozen-feature image, row-wise softmax removes one key-column direction per sample, and value updates are observable only after attention-weighted pooling and the frozen output projection. These invariances give per-head rank certificates for query, key, and value updates at rank-minimal global minimizers. We also prove that small perturbations exclude full-rank stationary points in nonconvex LoRA while maintaining near-optimality.

Limitations. First, our rank theorems certify existence for trace-norm objectives, not convergence of gradient-based LoRA. Second, we can't explicitly calculate bounds in Theorem 4.4, which are based on global solutions. Finally, the attention bounds are a posteriori, and the generic rank-deficiency theorem is only a first-order statement for perturbed objectives; Proposition 4.9 shows that spurious local minima can remain.

Acknowledgments

The authors would like to thank anonymous reviewers and the Area Chair for their useful feedback to improve our work. This work is supported by grant VUNI.2526.CAIR.04 and Seed Fund 10000225 CECS.

References

Afzal, Z. R., Esmaeilbeig, T., Soltanalian, M., and Ohannessian, M. I. (2025). Linearization explains fine-tuning in large language models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

Barvinok, A. (2001). A remark on the rank of positive semidefinite matrices subject to affine constraints. *Discrete & Computational Geometry*, 25(1):23–31.

Hahn, M. (2020). Theoretical limitations of self-attention in neural sequence models. *Transactions of the Association for Computational Linguistics*, 8:156–171.

Houlsby, N., Giurgiu, A., Jastrzebski, S., Morrone, B., De Laroussilhe, Q., Gesmundo, A., Attariyan, M., and Gelly, S. (2019). Parameter-efficient transfer learning for nlp. In *International conference on machine learning*, pages 2790–2799. PMLR.

Hu, E. J., Shen, Y., Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., Chen, W., et al. (2022). Lora: Low-rank adaptation of large language models. *ICLR*, 1(2):3.

Hu, J. Y. C., Su, M., Kuo, E. J., Song, Z., and Liu, H. (2025). Computational limits of low-rank adaptation (lora) fine-tuning for transformer models. In *13th International Conference on Learning Representations, ICLR 2025*, pages 54606–54645. International Conference on Learning Representations, ICLR.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31.

Jang, U., Lee, J. D., and Ryu, E. K. (2024). Lora training in the ntk regime has no spurious local minima. In *Proceedings of the 41st International Conference on Machine Learning*, pages 21306–21328.

Kim, J., Kim, J., and Ryu, E. K. (2025). Lora training provably converges to a low-rank global minimum or it fails loudly (but it probably won't fail). In *Forty-second International Conference on Machine Learning*.

Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. (2019). Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32.

Li, X. L. and Liang, P. (2021). Prefix-tuning: Optimizing continuous prompts for generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4582–4597.

Likhosherstov, V., Choromanski, K., and Weller, A. (2021). On the expressive power of self-attention matrices. *arXiv preprint arXiv:2106.03764*.

Liu, X.-H., Du, Y., Wang, J., and Yu, Y. (2025). On the optimization landscape of low rank adaptation methods for large language models. In *International Conference on Learning Representations*. Poster.

Maurer, A. (2016). A vector-contraction inequality for rademacher complexities. In *International Conference on Algorithmic Learning Theory*, pages 3–17. Springer.

McShane, E. J. (1934). Extension of range of functions. *Bulletin of the American Mathematical Society*, 40:837–842.

Pataki, G. (1998). On the rank of extreme matrices in semidefinite programs and the multiplicity of optimal eigenvalues. *Mathematics of operations research*, 23(2):339–358.

Recht, B., Fazel, M., and Parrilo, P. A. (2010). Guaranteed minimum-rank solutions of linear matrix equations via nuclear norm minimization. *SIAM review*, 52(3):471–501.

Schulman, J. and Lab, T. M. (2025). Lora without regret. *Thinking Machines Lab: Connectionism*. https://thinkingmachines.ai/blog/lora/.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, 30.

Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., and Bowman, S. (2018). Glue: A multi-task benchmark and analysis platform for natural language understanding. In *Proceedings of the 2018 EMNLP workshop BlackboxNLP: Analyzing and interpreting neural networks for NLP*, pages 353–355.

Xu, Z., Min, H., MacDonald, L. E., Luo, J., Tarmoun, S., Mallada, E., and Vidal, R. (2025). Understanding the learning dynamics of LoRA: A gradient flow perspective on low-rank adaptation in matrix factorization. In *Proceedings of The 28th International Conference on Artificial Intelligence and Statistics*, volume 258 of *Proceedings of Machine Learning Research*, pages 4636–4644.

Yang, A., Li, A., Yang, B., Zhang, B., Hui, B., Zheng, B., Yu, B., Gao, C., Huang, C., Lv, C., et al. (2025). Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Yun, C., Bhojanapalli, S., Rawat, A. S., Reddi, S. J., and Kumar, S. (2020). Are transformers universal approximators of sequence-to-sequence functions? In *8th International Conference on Learning Representations, ICLR 2020*.

Zeng, Y. and Lee, K. (2024). The expressive power of low-rank adaptation. In *The Twelfth International Conference on Learning Representations*.

Zhang, F. and Pilanci, M. (2024). Riemannian preconditioned LoRA for fine-tuning foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pages 59641–59669.

Zhou, H., Bradley, A., Littwin, E., Razin, N., Saremi, O., Susskind, J. M., Bengio, S., and Nakkiran, P. (2024). What algorithms can transformers learn? a study in length generalization. In *The Twelfth International Conference on Learning Representations*.

A Full list of notations and auxiliary definitions

We first summarize the key notation used throughout the paper. Table 2 summarizes the basic symbols for the LoRA fine-tuning setup, and Table 3 lists the additional notation used for attention layers and multi-stage linear mappings.

Symbol	Type / Shape	Description
n	scalar	Input dimension
m	scalar	Output dimension
W_0	$\mathbb{R}^{m\times n}$	Pretrained (frozen) weight matrix
W	$\mathbb{R}^{m\times n}$	Adapted weight matrix ($W = W_0 + T$)
T	$\mathbb{R}^{m\times n}$	Low-rank update matrix
U	$\mathbb{R}^{m\times r}$	Left low-rank factor
V	$\mathbb{R}^{n\times r}$	Right low-rank factor
r	scalar	Rank of LoRA update ($r \ll \min(m, n)$)
f_W	function	Network with frozen weights except adapted blocks
S_{data}	$\{(x_i, y_i)\}_{i=1}^N$	Training dataset
N	scalar	Number of training samples
x_i	input	Input sample
y_i	output	Target label
$h(x)$	$\mathbb{R}^n$	Frozen input activation to adapted layer
h_i	$\mathbb{R}^n$	Activation for sample x_i ($h_i = h(x_i)$)
H	$\mathbb{R}^{n\times N}$	Stacked feature matrix $[h_1, \dots, h_N]$
$\mathcal{S}$	$\{Q, K, V\}$	Attention projection blocks
$H_S(x)$	matrix	Frozen token matrix for block S
$H_{S,i}$	matrix	Token matrix for sample x_i
W_S	matrix	Adapted weight for block S ($W_S = W_{S,0} + T_S$)
$W_{t,S}$	matrix	Weight at stage t for block S
$T_{t,S}$	matrix	Update at stage t
$H_{t,S}(x;\cdot)$	matrix	Token matrix entering block S at stage t
$H_{t,S,i}$	matrix	Same for sample x_i
$\mathcal{L}$	function	Loss function
Φ, Φ_S	operator	Linear maps
g	function	Scalar function in loss decomposition
s, s_S	vector	Affine shift(s)

Table 2: Summary of notation and parameters for LoRA fine-tuning setup

For a single block, we also write

$$F_{\text{primal}}(T) \coloneqq g(\Phi(T) + s) + \lambda\|T\|_\star, F_{\text{lift}}(Z) \coloneqq g(\Phi(\bar{Z}) + s) + \frac{\lambda}{2}\operatorname{tr}(Z), Z \succeq 0.$$

NTK linearization as a special case. Our setting also contains the NTK/lazy regime as a special case (Jacot et al., 2018; Lee et al., 2019). Let $\Delta\theta \in \mathbb{R}^p$ denote the trainable parameter update; if the adapted parameter is a matrix, one may take $\Delta\theta = \operatorname{vec}(T)$. The first-order Taylor model around θ_0 is

$$f(x; \theta_0 + \Delta\theta) \approx f(x; \theta_0) + J_{\theta_0}(x)\Delta\theta,$$

where $J_{\theta_0}(x) \in \mathbb{R}^{K\times p}$ is the Jacobian of the logits with respect to the trainable parameters at θ_0. Given training data $\{(x_i, y_i)\}_{i=1}^N$, define

$$\Phi_{\text{NTK}}(\Delta\theta) \coloneqq \big(J_{\theta_0}(x_1)\Delta\theta, \dots, J_{\theta_0}(x_N)\Delta\theta\big) \in (\mathbb{R}^K)^N$$

and

$$s_{\text{NTK}} \coloneqq \big(f(x_1; \theta_0), \dots, f(x_N; \theta_0)\big) \in (\mathbb{R}^K)^N.$$

Then any empirical loss depending only on the vector of training predictions can be written as

$$\mathcal{L}_{\text{NTK}}(\Delta\theta) = g\big(\Phi_{\text{NTK}}(\Delta\theta) + s_{\text{NTK}}\big)$$

Symbol	Type / Shape	Description
$H_{Q,i}, H_{K,i}, H_{V,i}$	matrices	Token matrices for query, key, value (sample i)
H_Q, H_K, H_V	stacked matrices	$[H_{S,1}, \ldots, H_{S,N}]$ for $S \in \{Q, K, V\}$
d_k	scalar	Query/key head dimension
d_v	scalar	Value head dimension
m_Q, m_K	scalar	Output dims of Q, K projections ($= d_k$)
$W_{S,0}$	matrix	Frozen projection for $S \in \{Q, K, V\}$
T_S	matrix	Trainable LoRA update for projection S
$\widetilde{Q}_i$	matrix	$(W_{Q,0} + T_Q)H_{Q,i}$ (adapted queries)
$\widetilde{K}_i$	matrix	$(W_{K,0} + T_K)H_{K,i}$ (adapted keys)
$\widetilde{V}_i$	matrix	$(W_{V,0} + T_V)H_{V,i}$ (adapted values)
A_i	$\mathbb{R}^{\varrho_{Q,i} \times \varrho_{K,i}}$	Attention matrix (row-softmax)
C_i^M	matrix	Attention mask (with $-\infty$ entries allowed)
O_i	matrix	Attention output: $W_O \widetilde{V}_i A_i^\top$
$\varrho_{Q,i}, \varrho_{K,i}$	scalars	Number of query/key tokens for sample i
$\varrho_{t,Q,i}, \varrho_{t,K,i}$	scalars	Number of query/key token positions at stage t for sample i
$\mathcal{A}$	set	Set of attention heads
$a \in \mathcal{A}$	index	Head index
$\widetilde{Q}_i^{(a)}, \widetilde{K}_i^{(a)}, \widetilde{V}_i^{(a)}$	matrices	Head-specific projections
$A_i^{(a)}$	matrix	Head-specific attention matrix
$W_O^{(a)}$	matrix	Output projection block for head a
O_i	matrix	Multi-head output: $\sum_a W_O^{(a)} \widetilde{V}_i^{(a)} (A_i^{(a)})^\top$
$\mathcal{A}_t$	set	Head set at stage t
$d_{k,t}, d_{v,t}$	scalars	Head dimensions at stage t
$H_{t,S,i}$	matrix	Token matrix at stage t, block S, sample i
$H_{t,S}$	stacked matrix	$[H_{t,S,1}, \ldots, H_{t,S,N}]$
z_t	element of $\mathcal{Z}_t$	Stage-t state in model
$\alpha = (a, S)$	index	Block index: head a, projection $S \in \{Q, K, V\}$
$\Phi_\alpha(T_\alpha)$	map	$T_\alpha H_S$
$\Phi_{t,\alpha}(T, z)$	map	Stagewise linear map: $T_{t,\alpha} H_{t,S}(z)$

Table 3: Summary of attention notation and multi-stage linear mapping

for a suitable map $g : (\mathbb{R}^K)^N \to \mathbb{R}$. For a sample-wise loss, one may take

$$g(z_1, \ldots, z_N) \coloneqq \frac{1}{N} \sum_{i=1}^{N} \ell(z_i, y_i).$$

In the scalar-output case $K = 1$, this reduces to a map into $\mathbb{R}^N$. We now introduce several auxiliary concepts used later in the analysis.

Definition A.1 (C^1-parameterizable subsets)**.** *Let $\mathcal{V}$ be a finite-dimensional real inner-product space and let $\mathcal{G} \subseteq \mathcal{V}$. We say that $\mathcal{G}$ is C^1-parameterizable with dimension at most p if there exist finitely many open sets $U_j \subseteq \mathbb{R}^{p_j}$ with $p_j \leq p$ and C^1 maps $f_j : U_j \to \mathcal{V}$ such that*

$$\mathcal{G} \subseteq \bigcup_{j=1}^{J} f_j(U_j).$$

Definition A.2 (Reachable sets)**.** *For L-stage block-linear dependence, set $\mathcal{Z}_{L+1}^{\mathrm{reach}} \coloneqq \mathcal{Z}_{L+1}$. For $t = L, L-1, \ldots, 1$, define the reachable stage-t linear-output set*

$$\mathcal{U}_t^{\mathrm{reach}} \coloneqq \left\{ \{\Phi_{t,S}(T_{t,S}, z_{t+1})\}_{S \in \mathcal{S}_t} : z_{t+1} \in \mathcal{Z}_{t+1}^{\mathrm{reach}}, T_{t,S} \in \mathbb{R}^{m_{t,S} \times n_{t,S}} \right\} \subseteq \prod_{S \in \mathcal{S}_t} \mathcal{E}_{t,S},$$

and the reachable stage-t state set

$$\mathcal{Z}_t^{\mathrm{reach}} \coloneqq \left\{ g_t\Big(\{u_{t,S}\}_{S \in \mathcal{S}_t}, z_{t+1} \Big) : z_{t+1} \in \mathcal{Z}_{t+1}^{\mathrm{reach}}, u_{t,S} \in \mathcal{R}(\Phi_{t,S}(\cdot, z_{t+1})) \right\}.$$

Population and empirical risk. For a single update $T \in \mathbb{R}^{m\times n}$, define

$$L(T) \coloneqq \mathbb{E}_{(x,y)\sim\mathcal{P}}[\ell(f_{W_0+T}(x), y)], \quad \mathcal{L}^{\text{vanilla}}_{S_{\text{data}}}(T) \coloneqq \frac{1}{N}\sum_{i=1}^{N} \ell(f_{W_0+T}(x_i), y_i).$$

For multiple adapted matrices $\{T_S\}$, we use the same notation blockwise. When the sample is fixed, we write $\mathcal{L}^{\text{vanilla}}$ for $\mathcal{L}^{\text{vanilla}}_{S_{\text{data}}}$.

For $D \geq 0$, define the deterministic activation ball

$$\mathcal{Z}_{\text{act}}(D, R_h) \coloneqq \{z \in \mathbb{R}^m : \|z\|_2 \leq (\|W_0\|_2 + D)R_h\}.$$

To state our generalization bounds, we also recall the empirical Rademacher complexity.

Empirical Rademacher complexity. Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \times \mathcal{Y} \to \mathbb{R}$. The (conditional) empirical Rademacher complexity of $\mathcal{F}$ on S_{data} is

$$\mathfrak{R}_{S_{\text{data}}}(\mathcal{F}) \coloneqq \mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^{N} \varepsilon_i f(x_i, y_i) \mid S_{\text{data}}\right],$$

where $\varepsilon_1, \ldots, \varepsilon_N$ are i.i.d. uniform signs in $\{-1, 1\}$. For $D \geq 0$, define the trace-norm ball

$$\mathbb{B}_\star(D) \coloneqq \{T \in \mathbb{R}^{m\times n} : \|T\|_\star \leq D\},$$

and define the loss class induced by trace-norm-bounded updates

$$\mathcal{F}_D \coloneqq \{(x, y) \mapsto \ell(f_{W_0+T}(x), y) : T \in \mathbb{B}_\star(D)\}.$$

Miscellaneous. For $Z \in \mathbb{S}^{m+n}_+$ write $Z = \begin{bmatrix} Z_{11} & Z_{12} \\ Z_{21} & Z_{22}\end{bmatrix}$ with $Z_{12} \in \mathbb{R}^{m\times n}$ and denote $\bar{Z} \coloneqq Z_{12}$. For sets A, B in a vector space, the Minkowski sum is $A + B \coloneqq \{a + b : a \in A, b \in B\}$. We use blkdiag for a block diagonal matrix. For a matrix M, let P_M denote the orthogonal projector onto the column space $\mathcal{R}(M)$. If $M = 0$, then $P_M = 0$.

B Generalization bounds

We now state the assumptions used to derive the generalization bounds in this section.

Assumption 5 (Bounded and Lipschitz loss)**.** *The loss satisfies $0 \leq \ell(\cdot, y) \leq B$ and is L_ℓ-Lipschitz ($L_\ell > 0$) with respect to $\|\cdot\|_2$ for every y.*

Assumption 6 (Almost-sure bounded frozen activations)**.** *For $(x, y) \sim \mathcal{P}$, the frozen input activation to the single adapted block satisfies $\|h(x)\|_2 \leq R_h$ almost surely. In the attention setting, for each $S \in \{Q, K, V\}$, the frozen token matrix satisfies $\|H_S(x)\|_F \leq R_{H,S}$ almost surely. Consequently, on every realized sample,*

$$\|h_i\|_2 \leq R_h, \qquad \|H_{S,i}\|_F \leq R_{H,S} \ (i \in [N],\ S \in \{Q, K, V\}).$$

Assumption 7 (Inter-stage token Lipschitzness for L consecutive attention modules)**.** *Fix trace-norm radii $(D_{t,S})_{t\in[L],S\in\mathcal{S}_t}$. For each $t \in \{1, \ldots, L-1\}$, assume there exists a finite constant $L_{H,t}(D_{>t})$ such that for all $i \in [N]$ and all two upstream update collections $\{T_{\tau,S}\}_{\tau>t}$ and $\{T'_{\tau,S}\}_{\tau>t}$ satisfying $\|T_{\tau,S}\|_\star, \|T'_{\tau,S}\|_\star \leq D_{\tau,S}(\tau > t)$, we have*

$$\left(\sum_{S\in\mathcal{S}_t} \|H_{t,S,i}(\{T_{\tau,S'}\}_{\tau>t}) - H_{t,S,i}(\{T'_{\tau,S'}\}_{\tau>t})\|_F^2\right)^{1/2}$$
$$\leq L_{H,t}(D_{>t})\left(\sum_{\tau=t+1}^{L}\sum_{S\in\mathcal{S}_\tau} \|T_{\tau,S}H_{\tau,S,i}(\{T_{\nu,S'}\}_{\nu>\tau}) - T'_{\tau,S}H_{\tau,S,i}(\{T'_{\nu,S'}\}_{\nu>\tau})\|_F^2\right)^{1/2}.$$

Assumption 8 (Regular frozen suffix on reachable activation sets for L attention modules)**.** *Fix trace-norm radii $(D_{t,S})$. Assume there exists a finite constant $L_{\text{Lattn}}(D)$ such that for all $i \in [N]$ and all pairs of reachable collections $\{Z_{t,S,i}\}_{t,S}, \{Z'_{t,S,i}\}_{t,S}$, one has*

$$\left\|\psi_{\text{Lattn},i}(\{Z_{t,S,i}\}_{t,S}) - \psi_{\text{Lattn},i}(\{Z'_{t,S,i}\}_{t,S})\right\|_2 \leq L_{\text{Lattn}}(D)\left(\sum_{t=1}^{L}\sum_{S\in\mathcal{S}_t} \|Z_{t,S,i} - Z'_{t,S,i}\|_F^2\right)^{1/2}.$$

For the single-block case, let $\sigma_i \in \{-1,1\}^m$ be an m-dimensional vector with independent Rademacher entries. Let $E \in \{-1,1\}^{m\times N}$ be the matrix whose i-th column is σ_i. Then $\sum_{i=1}^N \sigma_i h_i^\top = EH^\top$.

We also define

$$L_\psi(D,R_h) \coloneqq \max_{i\in[N]} \sup_{z\in\mathcal{Z}_{\mathrm{act}}(D,R_h)} \|D\psi_i(z)\|_2.$$

Assumption 9 (Regular single-block frozen suffix on reachable activation sets)**.** *Fix $D \geq 0$. Assume that for each $i \in [N]$, the map ψ_i is C^1 on an open neighborhood of $\mathcal{Z}_{\mathrm{act}}(D,R_h)$.*

The set $\mathcal{Z}_{\mathrm{act}}(D,R_h)$ is closed and bounded, hence compact. Since each $D\psi_i$ is continuous, the supremum $L_\psi(D,R_h)$ is finite.

The next two lemmas formalize the two "frozen suffix" decompositions used in the analysis. The first applies to a single adapted module, and the second extends the idea to L consecutive adapted attention modules.

Lemma B.1 (Single-cut suffix representation for one adapted module)**.** *Assume Assumption 2 holds. Let $\mathcal{S}$ index adapted linear maps that lie behind one computation-graph cut, in the following sense: for every training sample $i \in [N]$, the input tensor $H_{S,i}$ to each adapted map S is computed before any map in $\mathcal{S}$ is applied and is independent of the whole update tuple $\{T_{S'}\}$. Then there exist deterministic maps*

$$\psi_i : \prod_{S\in\mathcal{S}} \mathbb{R}^{m_S\times\varrho_{S,i}} \to \mathbb{R}^K$$

depending only on x_i and the frozen weights such that, for all updates $\{T_S\}$,

$$f_{\{W_{S,0}+T_S\}}(x_i) = \psi_i\Big(\{(W_{S,0}+T_S)H_{S,i}\}\Big) = \psi_i\Big(\{W_{S,0}H_{S,i}+T_SH_{S,i}\}\Big).$$

Consequently, for one adapted module, the empirical loss depends on each block T_S only through the bottleneck outputs $\{T_SH_{S,i}\}_{i=1}^N$, or equivalently through the stacked linear operator $\Phi_S(T_S) = T_SH_S$.

Proof of Lemma B.1. Fix $i \in [N]$. By the single-cut assumption, the tensors $\{H_{S,i}\}$ and all cached quantities produced before the adapted module are fixed functions of x_i and the frozen weights. Cut the computation graph at the outputs $Z_{S,i} \coloneqq (W_{S,0}+T_S)H_{S,i}, S \in \mathcal{S}$. After this cut, all remaining operations, residual connections, attention operations, nonlinearities, and output layers are frozen and deterministic. Therefore, the downstream computation defines a deterministic suffix map

$$\psi_i : \prod_{S\in\mathcal{S}} \mathbb{R}^{m_S\times\varrho_{S,i}} \to \mathbb{R}^K$$

from the collection $\{Z_{S,i}\}$ to the logits. Substituting $Z_{S,i} = W_{S,0}H_{S,i} + T_SH_{S,i}$ gives the displayed representation. Thus, changing T_S while keeping the rest of the module fixed can affect the empirical loss only through the finite collection $\{T_SH_{S,i}\}_{i=1}^N$. □

Lemma B.2 (Multi-cut decomposition for L consecutive adapted attention modules)**.** *Assume Assumption 2 holds. Consider L consecutive attention modules, indexed so that larger t is more upstream. For each stage t, the token matrix $H_{t,S,i}(\{T_{\tau,S'}\}_{\tau>t})$ entering block S is allowed to depend on upstream updates, but not on the current block update $T_{t,S}$. Then there exist deterministic maps*

$$\psi_{L\mathrm{attn},i} : \prod_{t=1}^{L} \prod_{S\in\mathcal{S}_t} \mathbb{R}^{m_{t,S}\times\varrho_{t,S,i}} \to \mathbb{R}^K$$

such that, for all update collections $\{T_{t,S}\}$,

$$f_{\{W_{t,S,0}+T_{t,S}\}}(x_i) = \psi_{L\mathrm{attn},i}\Big(\{(W_{t,S,0}+T_{t,S})H_{t,S,i}(\{T_{\tau,S'}\}_{\tau>t})\}_{t\in[L],S\in\mathcal{S}_t}\Big).$$

Proof of Lemma B.2. Fix $i \in [N]$. We cut the computation graph once at each adapted attention module. At the most upstream adapted stage L, the inputs to the adapted projections are fixed by the frozen prefix, so the stage outputs have the form

$$Z_{L,S,i} = (W_{L,S,0}+T_{L,S})H_{L,S,i}.$$

Because the computation between stage L and stage $L-1$ is deterministic, the token matrices entering stage $L-1$ are deterministic functions of the stage-L cut outputs. Thus, the stage-$L-1$ adapted outputs have the form

$$Z_{L-1,S,i} = (W_{L-1,S,0} + T_{L-1,S})H_{L-1,S,i}(\{T_{\tau,S'}\}_{\tau>L-1}).$$

Repeating this argument down to stage 1 shows that every adapted output can be written as

$$Z_{t,S,i} = (W_{t,S,0} + T_{t,S})H_{t,S,i}(\{T_{\tau,S'}\}_{\tau>t}).$$

After all these cuts, the remaining suffix from the collection $\{Z_{t,S,i}\}_{t,S}$ to the logits is frozen and deterministic. This suffix is the map $\psi_{L\mathrm{attn},i}$. The displayed formula follows by substituting the expressions for the cut outputs. □

The Rademacher-complexity proof below requires applying vector contraction to functions whose natural domains are not all of Euclidean space, but only the reachable activation sets induced by trace-norm-bounded updates. To avoid introducing any extension assumption, we use the classical real-valued Lipschitz extension formula of McShane (1934). The lemma is stated in the metric-space form needed later.

Lemma B.3 (McShane extension theorem)**.** *Let (X, d) be a metric space, let $\emptyset \neq A \subseteq X$, and let $f : A \to \mathbb{R}$ be L-Lipschitz. Define*

$$\bar{f}(x) \coloneqq \inf_{a\in A}\big(f(a) + Ld(x, a)\big), x \in X.$$

Then $\bar{f} : X \to \mathbb{R}$ is an L-Lipschitz extension of f.

Proof of Lemma B.3. Fix $x \in A$. For every $a \in A$, the L-Lipschitz property of f gives

$$f(a) + Ld(x, a) \geq f(x).$$

Taking $a = x$ shows equality is attained, hence $\bar{f}(x) = f(x)$. Thus $\bar{f}$ extends f.

Now fix $x, y \in X$. For every $a \in A$,

$$\bar{f}(x) \leq f(a) + Ld(x, a) \leq f(a) + Ld(y, a) + Ld(x, y).$$

Taking the infimum over $a \in A$ yields $\bar{f}(x) \leq \bar{f}(y) + Ld(x, y)$. Exchanging the roles of x and y gives $\bar{f}(y) \leq \bar{f}(x) + Ld(x, y)$. Therefore $|\bar{f}(x) - \bar{f}(y)| \leq Ld(x, y)$, so $\bar{f}$ is L-Lipschitz. □

We next show the standard Rademacher deviation bound that converts a complexity estimate into a high-probability generalization guarantee.

Lemma B.4 (Uniform deviation from conditional Rademacher complexity)**.** *Let $\mathcal{F}$ be a class of functions $f : \mathcal{X} \times \mathcal{Y} \to [0, B]$. Let $S_{\mathrm{data}} = \{(x_i, y_i)\}_{i=1}^N$ be i.i.d. from $\mathcal{P}$ and write*

$$\widehat{\mathbb{E}}_S f \coloneqq \frac{1}{N}\sum_{i=1}^N f(x_i, y_i), \qquad \mathbb{E}f \coloneqq \mathbb{E}_{(x,y)\sim\mathcal{P}}[f(x, y)].$$

Define the (conditional) empirical Rademacher complexity

$$\mathfrak{R}_{S_{\mathrm{data}}}(\mathcal{F}) \coloneqq \mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}}\frac{1}{N}\sum_{i=1}^N \varepsilon_i f(x_i, y_i) \mid S_{\mathrm{data}}\right],$$

where $\varepsilon_1, \ldots, \varepsilon_N$ are i.i.d. uniform signs. Then for every $\eta \in (0, 1)$, with probability at least $1-\eta$ over $S \sim \mathcal{P}^N$,

$$\sup_{f\in\mathcal{F}}\Big(\mathbb{E}f - \widehat{\mathbb{E}}_S f\Big) \leq 2\mathfrak{R}_{S_{\mathrm{data}}}(\mathcal{F}) + 3B\sqrt{\frac{\log(2/\eta)}{2N}}.$$

Proof of Lemma B.4. Define

$$g(S) \coloneqq \sup_{f\in\mathcal{F}}\big(\mathbb{E}f - \widehat{\mathbb{E}}_S f\big).$$

If we replace (x_i, y_i) by an independent copy (x_i', y_i') while keeping all other samples fixed, then for any fixed f,

$$\left|\widehat{\mathbb{E}}_S f - \widehat{\mathbb{E}}_{S^{(i)}_{\mathrm{data}}} f\right| = \frac{1}{N}\left|f(x_i, y_i) - f(x_i', y_i')\right| \le \frac{B}{N}.$$

Taking the supremum over f implies $|g(S) - g(S^{(i)}_{\mathrm{data}})| \le B/N$. McDiarmid's inequality therefore yields: with probability at least $1 - \eta/2$,

$$g(S) \le \mathbb{E}[g(S)] + B\sqrt{\frac{\log(2/\eta)}{2N}}.$$

Let $S' = \{(x_i', y_i')\}_{i=1}^N$ be an i.i.d. ghost sample from $\mathcal{P}^N$, independent of S. Then

$$\mathbb{E}[g(S)] = \mathbb{E}_S\left[\sup_{f\in\mathcal{F}}\Big(\mathbb{E}_{S'}\widehat{\mathbb{E}}_{S'} f - \widehat{\mathbb{E}}_S f\Big)\right] \le \mathbb{E}_{S,S'}\left[\sup_{f\in\mathcal{F}}\Big(\widehat{\mathbb{E}}_{S'} f - \widehat{\mathbb{E}}_S f\Big)\right].$$

Introduce i.i.d. signs $\varepsilon_1, \dots, \varepsilon_N$, independent of (S, S'), and use that (S, S') and (S', S) have the same distribution to obtain

$$\mathbb{E}_{S,S'}\left[\sup_f \frac{1}{N}\sum_{i=1}^N \big(f(x_i', y_i') - f(x_i, y_i)\big)\right] = \mathbb{E}_{S,S',\varepsilon}\left[\sup_f \frac{1}{N}\sum_{i=1}^N \varepsilon_i\big(f(x_i', y_i') - f(x_i, y_i)\big)\right].$$

By the triangle inequality and symmetry,

$$\mathbb{E}[g(S)] \le 2\mathbb{E}_S\left[\mathbb{E}_\varepsilon\left[\sup_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^N \varepsilon_i f(x_i, y_i) \mid S_{\mathrm{data}}\right]\right] = 2\mathbb{E}_S[\mathfrak{R}_{S_{\mathrm{data}}}(\mathcal{F})].$$

Fix ε and define

$$h_\varepsilon(S) \coloneqq \sup_{f\in\mathcal{F}} \frac{1}{N}\sum_{i=1}^N \varepsilon_i f(x_i, y_i).$$

Replacing one sample changes $h_\varepsilon(S)$ by at most B/N, hence $\mathfrak{R}_{S_{\mathrm{data}}}(\mathcal{F}) = \mathbb{E}_\varepsilon[h_\varepsilon(S)|S]$ also satisfies bounded differences with constant B/N. McDiarmid implies: with probability at least $1 - \eta/2$,

$$\mathbb{E}_S[\mathfrak{R}_{S_{\mathrm{data}}}(\mathcal{F})] \le \mathfrak{R}_{S_{\mathrm{data}}}(\mathcal{F}) + B\sqrt{\frac{\log(2/\eta)}{2N}}.$$

On the intersection of the two $1 - \eta/2$ events and using $\mathbb{E}[g(S)] \le 2\mathbb{E}[\mathfrak{R}_{S_{\mathrm{data}}}(\mathcal{F})]$, we get

$$g(S) \le 2\mathfrak{R}_{S_{\mathrm{data}}}(\mathcal{F}) + 3B\sqrt{\frac{\log(2/\eta)}{2N}}.$$

as desired. □

Combining these results gives the main single-block generalization result. The bound reduces uniform generalization over the trace-norm ball to the empirical Rademacher complexity of the induced loss class, and then to the spectral norm of a signed feature matrix.

Theorem B.5 (Uniform generalization for trace-norm-bounded updates in frozen networks)**.** *Assume Assumptions 2,5,6, and 9 hold. Then with probability at least* $1 - \eta$ *over* $S_{\mathrm{data}} \sim \mathcal{P}^N$*,*

$$\sup_{T\in\mathbb{B}_\star(D)}\Big(L(T) - \mathcal{L}^{\mathrm{vanilla}}_{S_{\mathrm{data}}}(T)\Big) \le 2\mathfrak{R}_{S_{\mathrm{data}}}(\mathcal{F}_D) + 3B\sqrt{\frac{\log(2/\eta)}{2N}}.$$

Moreover,

$$\mathfrak{R}_{S_{\mathrm{data}}}(\mathcal{F}_D) \le \frac{\sqrt{2}L_\ell L_\psi(D, R_h)D}{N}\mathbb{E}_\sigma\big[\|EH^\top\|_2 \mid S_{\mathrm{data}}\big], \tag{B.1}$$

Theorem B.5 reduces uniform generalization over the trace-norm ball $\mathbb{B}_\star(D)$ to the spectral norm of a random signed feature matrix $EH^\top$.

Proof of Theorem B.5. The first display follows immediately from Lemma B.4 applied to the bounded class $\mathcal{F}_D$. It remains to prove (B.1). For each $i \in [N]$, define

$$K_i(D) \coloneqq \{Th_i : \|T\|_\star \le D\} \subseteq \mathbb{R}^m, \phi_i(z) \coloneqq \ell\big(\psi_i(W_0 h_i + z), y_i\big), z \in K_i(D).$$

Whenever $z = Th_i$ and $z' = T'h_i$ with $\|T\|_\star, \|T'\|_\star \le D, |\phi_i(z) - \phi_i(z')| \le L_\ell L_\psi(D, R_h)\|z - z'\|_2$. Define $\widetilde{\phi}_i(z) \coloneqq \phi_i(z) - \phi_i(0)$ on $K_i(D)$. Then $\widetilde{\phi}_i(0) = 0$, and $\widetilde{\phi}_i$ is $L_\ell L_\psi(D, R_h)$-Lipschitz on $K_i(D)$.

Moreover, for every realization of ε,

$$\sup_{\|T\|_\star \le D} \frac{1}{N} \sum_{i=1}^N \varepsilon_i \phi_i(Th_i) = \sup_{\|T\|_\star \le D} \frac{1}{N} \sum_{i=1}^N \varepsilon_i \widetilde{\phi}_i(Th_i) + \frac{1}{N} \sum_{i=1}^N \varepsilon_i \phi_i(0).$$

Taking the conditional expectation and using

$$\mathbb{E}_\varepsilon \left[\sum_{i=1}^N \varepsilon_i \phi_i(0) \mid S_{\text{data}} \right] = 0$$

gives

$$\mathfrak{R}_{S_{\text{data}}}(\mathcal{F}_D) = \mathbb{E}_\varepsilon \left[\sup_{\|T\|_\star \le D} \frac{1}{N} \sum_{i=1}^N \varepsilon_i \widetilde{\phi}_i(Th_i) \mid S_{\text{data}} \right].$$

By Lemma B.3, each $\widetilde{\phi}_i$ extends from $K_i(D)$ to a function $\bar{\phi}_i : \mathbb{R}^m \to \mathbb{R}$ with the same Lipschitz constant $L_\ell L_\psi(D, R_h)$. Since $Th_i \in K_i(D)$ for every feasible T, we may write

$$\mathfrak{R}_{S_{\text{data}}}(\mathcal{F}_D) = \mathbb{E}_\varepsilon \left[\sup_{\|T\|_\star \le D} \frac{1}{N} \sum_{i=1}^N \varepsilon_i \bar{\phi}_i(Th_i) \mid S_{\text{data}} \right].$$

Applying Maurer's vector contraction inequality Maurer (2016) gives

$$\mathfrak{R}_{S_{\text{data}}}(\mathcal{F}_D) \le \frac{\sqrt{2} L_\ell L_\psi(D, R_h)}{N} \mathbb{E}_\sigma \left[\sup_{\|T\|_\star \le D} \sum_{i=1}^N \langle \sigma_i, Th_i \rangle \mid S_{\text{data}} \right].$$

Stacking the σ_i as columns of E gives

$$\sum_{i=1}^N \langle \sigma_i, Th_i \rangle = \left\langle \sum_{i=1}^N \sigma_i h_i^\top, T \right\rangle_F = \langle EH^\top, T \rangle_F.$$

Therefore, by trace/spectral duality,

$$\sup_{\|T\|_\star \le D} \sum_{i=1}^N \langle \sigma_i, Th_i \rangle = D\|EH^\top\|_2.$$

Substituting this into the previous display yields (B.1). □

Why generalization depends on $\|EH^\top\|_2$. The trace norm is the dual of the operator norm: $\sup_{\|T\|_\star \le D} \langle T, M \rangle = D\|M\|_2$. After symmetrization and vector contraction, the complexity of the frozen-network loss class reduces to such a dual pairing with $M = \sum_i \sigma_i h_i^\top = EH^\top$. Thus, the effective capacity is governed by the spectral geometry of frozen features, not by the ambient parameter dimension mn.

From here, for each (t, S), let $E_{t,S}$ be an independent Rademacher matrix in $\{-1, 1\}^{m_{t,S} \times \sum_{i=1}^N \varrho_{t,S,i}}$.

Standing notation for the L-stage generalization argument. Fix a realized sample $S_{\mathrm{data}} = \{(x_i, y_i)\}_{i=1}^N$ and radii $D = (D_{t,S})_{t\in[L],S\in\mathcal{S}_t}$. Define

$$\Theta_D \coloneqq \{\{T_{t,S}\}_{t\in[L],S\in\mathcal{S}_t} : \|T_{t,S}\|_\star \le D_{t,S}\ \forall t \in [L], S \in \mathcal{S}_t\}.$$

For each (t, S) let $\varrho_{t,S}^{\max} \coloneqq \max_{i\in[N]} \varrho_{t,S,i}$, and define the Hilbert product space

$$\mathcal{H} \coloneqq \bigoplus_{t=1}^{L} \bigoplus_{S\in\mathcal{S}_t} \mathbb{R}^{m_{t,S}\times\varrho_{t,S}^{\max}},$$

equipped with the product Frobenius norm, where each $\mathbb{R}^{m_{t,S}\times\varrho_{t,S,i}}$ is identified with its zero-padded copy inside $\mathbb{R}^{m_{t,S}\times\varrho_{t,S}^{\max}}$.

For each $i \in [N]$, define the reachable observed-output set

$$K_i(D) \coloneqq \left\{ \left(\{T_{t,S} H_{t,S,i}(\{T_{\tau,S'}\}_{\tau>t})\}_{S\in\mathcal{S}_t} \right)_{t=1}^{L} : \{T_{t,S}\} \in \Theta_D \right\} \subseteq \mathcal{H}.$$

Since the all-zero update tuple belongs to Θ_D, we have $0 \in K_i(D)$.

For $z \in K_i(D)$, choose any feasible tuple $\{T_{t,S}\} \in \Theta_D$ producing z and define

$$\phi_i(z) \coloneqq \ell(f_{\{W_{t,S,0}+T_{t,S}\}}(x_i), y_i), \widetilde{\phi}_i(z) \coloneqq \phi_i(z) - \phi_i(0).$$

This is well-defined by Lemma B.2, because the observed-output tuple determines all full adapted outputs stage by stage and therefore determines the logits uniquely.

Finally, define

$$\Psi_i(\{T_{t,S}\}) \coloneqq \left(\{T_{t,S} H_{t,S,i}(\{T_{\tau,S'}\}_{\tau>t})\}_{S\in\mathcal{S}_t} \right)_{t=1}^{L} \in \mathcal{H}.$$

Define the corresponding loss class

$$\mathcal{F}_D^{(L)} \coloneqq \left\{ (x, y) \mapsto \ell(f_{\{W_{t,S,0}+T_{t,S}\}}(x), y) : \{T_{t,S}\} \in \Theta_D \right\}.$$

The multi-stage attention bound has one additional difficulty compared with the single-block bound: perturbing an upstream observed output can change the token matrices seen by downstream adapted blocks. The next deterministic lemma isolates this recursive amplification.

Lemma B.6 (Recursive triangular amplification)**.** *Fix nonnegative numbers $\rho_1, \dots, \rho_{L-1}$. Set*

$$\Gamma_L \coloneqq 1, \qquad \Gamma_t \coloneqq \left\| \begin{bmatrix} 1 & \rho_t \\ 0 & \Gamma_{t+1} \end{bmatrix} \right\|_2 (t = L-1, \dots, 1).$$

If nonnegative numbers a_t, b_t satisfy

$$b_L \le a_L, b_t \le a_t + \rho_t \left(\sum_{\tau=t+1}^{L} a_\tau^2 \right)^{1/2} (t = 1, \dots, L-1),$$

then for every $t \in [L]$,

$$\left(\sum_{\tau=t}^{L} b_\tau^2 \right)^{1/2} \le \Gamma_t \left(\sum_{\tau=t}^{L} a_\tau^2 \right)^{1/2}.$$

Proof of Lemma B.6. We argue by backward induction on t. The case $t = L$ is immediate. Assume the claim holds for $t+1$ and set $c \coloneqq \left(\sum_{\tau=t+1}^{L} a_\tau^2 \right)^{1/2}$. Then

$$\left(\sum_{\tau=t}^{L} b_\tau^2 \right)^{1/2} \le \sqrt{(a_t + \rho_t c)^2 + \Gamma_{t+1}^2 c^2} = \left\| \begin{bmatrix} 1 & \rho_t \\ 0 & \Gamma_{t+1} \end{bmatrix} \begin{pmatrix} a_t \\ c \end{pmatrix} \right\|_2 \le \Gamma_t \sqrt{a_t^2 + c^2},$$

which is exactly the desired bound. □

The previous lemma provides the recursive norm-growth factor that controls how perturbations accumulate across consecutive stages. Combining it with the observed output Lipschitz bound yields the following estimate for the full L-module attention.

Lemma B.7 (Observed-output Lipschitzness for L consecutive attention modules)**.** *Assume Assumptions 2,5,7, and 8 hold. For $t = 1, \dots, L-1$, define*

$$\|W_{t,0}\|_{\mathcal{S}_t} \coloneqq \left(\sum_{S\in\mathcal{S}_t} \|W_{t,S,0}\|_2^2\right)^{1/2}, \rho_t(D) \coloneqq \|W_{t,0}\|_{\mathcal{S}_t} L_{H,t}(D_{>t}),$$

and let $\Gamma_t(D)$ be the constants from Lemma B.6. Then for any two update collections in the product trace-norm ball and every $i \in [N]$,

$$\begin{aligned}&\left|\ell\big(f_{\{W_{t,S,0}+T_{t,S}\}}(x_i), y_i\big) - \ell\big(f_{\{W_{t,S,0}+T'_{t,S}\}}(x_i), y_i\big)\right| \\ &\le L_\ell L_{L\mathrm{attn}}(D)\Gamma_1(D)\left(\sum_{t=1}^{L}\sum_{S\in\mathcal{S}_t}\left\|T_{t,S}H_{t,S,i}(\{T_{\tau,S'}\}_{\tau>t}) - T'_{t,S}H_{t,S,i}(\{T'_{\tau,S'}\}_{\tau>t})\right\|_F^2\right)^{1/2}.\end{aligned}$$

Proof of Lemma B.7. Fix i. Define

$$a_t \coloneqq \left(\sum_{S\in\mathcal{S}_t}\left\|T_{t,S}H_{t,S,i}(\{T_{\tau,S'}\}_{\tau>t}) - T'_{t,S}H_{t,S,i}(\{T'_{\tau,S'}\}_{\tau>t})\right\|_F^2\right)^{1/2}.$$

Also, define the actual adapted outputs

$$\begin{aligned}Z_{t,S,i} &\coloneqq (W_{t,S,0} + T_{t,S})H_{t,S,i}(\{T_{\tau,S'}\}_{\tau>t}),\\ Z'_{t,S,i} &\coloneqq (W_{t,S,0} + T'_{t,S})H_{t,S,i}(\{T'_{\tau,S'}\}_{\tau>t}),\end{aligned}$$

and

$$b_t \coloneqq \left(\sum_{S\in\mathcal{S}_t}\|Z_{t,S,i} - Z'_{t,S,i}\|_F^2\right)^{1/2}.$$

For the top stage L, the token matrices are frozen, hence $b_L \le a_L$. For $t < L$, write

$$\Delta H_{t,S,i} \coloneqq H_{t,S,i}(\{T_{\tau,S'}\}_{\tau>t}) - H_{t,S,i}(\{T'_{\tau,S'}\}_{\tau>t}),$$

so that

$$Z_{t,S,i} - Z'_{t,S,i} = W_{t,S,0}\Delta H_{t,S,i} + \Big(T_{t,S}H_{t,S,i}(\{T_{\tau,S'}\}_{\tau>t}) - T'_{t,S}H_{t,S,i}(\{T'_{\tau,S'}\}_{\tau>t})\Big).$$

Therefore

$$b_t \le a_t + \left(\sum_{S\in\mathcal{S}_t}\|W_{t,S,0}\Delta H_{t,S,i}\|_F^2\right)^{1/2}.$$

Using $\|AB\|_F \le \|A\|_2\|B\|_F$ and Cauchy-Schwarz across S,

$$b_t \le a_t + \|W_{t,0}\|_{\mathcal{S}_t}\left(\sum_{S\in\mathcal{S}_t}\|\Delta H_{t,S,i}\|_F^2\right)^{1/2}.$$

Assumption 7 gives

$$b_t \le a_t + \rho_t(D)\left(\sum_{\tau=t+1}^{L} a_\tau^2\right)^{1/2}.$$

Applying Lemma B.6,

$$\left(\sum_{t=1}^{L} b_t^2\right)^{1/2} \le \Gamma_1(D)\left(\sum_{t=1}^{L} a_t^2\right)^{1/2}.$$

Finally, Assumption 8 and the L_ℓ-Lipschitz property of $\ell(\cdot, y_i)$ imply

$$|\ell(f_T(x_i), y_i) - \ell(f_{T'}(x_i), y_i)| \le L_\ell L_{L\mathrm{attn}}(D)\left(\sum_{t=1}^{L} b_t^2\right)^{1/2},$$

which, combined with the previous display, proves the claim. □

With this estimate, we can state the multi-stage generalization bound. The result is the attention analogue of the single-block theorem above, with the random signed matrices appearing stage by stage.

Theorem B.8 (Uniform generalization for trace-norm-bounded updates in L consecutive attention modules)**.** *Assume Assumptions 2,5,7, and 8 hold. Fix radii $(D_{t,S})_{t\in[L],S\in\mathcal{S}_t}$. For each (t,S), let $E_{t,S}$ be an independent Rademacher matrix of the same shape as the stacked stage-t, block-S feature matrix. Then with probability at least $1-\eta$ over $S_{\mathrm{data}}\sim\mathcal{P}^N$,*

$$\sup_{\|T_{t,S}\|_\star\le D_{t,S}}\Big(L(\{T_{t,S}\})-\mathcal{L}^{\mathrm{vanilla}}_{S_{\mathrm{data}}}(\{T_{t,S}\})\Big)$$
$$\le\frac{2\sqrt{2}L_\ell L_{L\mathrm{attn}}(D)\Gamma_1(D)}{N}\sum_{t=1}^{L}\sum_{S\in\mathcal{S}_t}D_{t,S}\mathbb{E}\Bigg[\sup_{\substack{\|T_{\tau,S'}\|_\star\\ \le D_{\tau,S'}(\tau>t)}}\big\|E_{t,S}H_{t,S}(\{T_{\tau,S'}\}_{\tau>t})^\top\big\|_2\Big|S_{\mathrm{data}}\Bigg]$$
$$+3B\sqrt{\frac{\log(2/\eta)}{2N}}.$$

Here, the supremum over $(\tau>t)$ is interpreted as vacuous when $t=L$.

The same statement applies verbatim to multi-head attention by taking the block family at stage t to be a head-indexed collection such as $\mathcal{A}_t\times\{Q,K,V\}$, or the corresponding shared-block collection in grouped-query or multi-query attention.

Proof of Theorem B.8. By Lemma B.4, it suffices to bound $\mathfrak{R}_{S_{\mathrm{data}}}(\mathcal{F}_D^{(L)})$. Fix a realized sample S_{data}. By Lemma B.7,

$$|\widetilde{\phi}_i(z)-\widetilde{\phi}_i(z')|\le L_\ell L_{L\mathrm{attn}}(D)\Gamma_1(D)\|z-z'\|_{\mathcal{H}}\quad\forall z,z'\in K_i(D).$$

Because $0\in K_i(D)$ and $\widetilde{\phi}_i(0)=0$, Lemma B.3 yields an extension $\bar{\phi}_i:\mathcal{H}\to\mathbb{R}$ such that $\bar{\phi}_i|_{K_i(D)}=\widetilde{\phi}_i$ and

$$|\bar{\phi}_i(z)-\bar{\phi}_i(z')|\le L_\ell L_{L\mathrm{attn}}(D)\Gamma_1(D)\|z-z'\|_{\mathcal{H}}\quad\forall z,z'\in\mathcal{H}.$$

Therefore

$$\mathfrak{R}_{S_{\mathrm{data}}}(\mathcal{F}_D^{(L)})=\mathbb{E}_\varepsilon\left[\sup_{\{T_{t,S}\}\in\Theta_D}\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i\widetilde{\phi}_i\big(\Psi_i(\{T_{t,S}\})\big)\mid S_{\mathrm{data}}\right]$$
$$=\mathbb{E}_\varepsilon\left[\sup_{\{T_{t,S}\}\in\Theta_D}\frac{1}{N}\sum_{i=1}^{N}\varepsilon_i\bar{\phi}_i\big(\Psi_i(\{T_{t,S}\})\big)\mid S_{\mathrm{data}}\right].$$

Applying Maurer's vector contraction inequality Maurer (2016) to the class $\{i\mapsto\Psi_i(\{T_{t,S}\}):\{T_{t,S}\}\in\Theta_D\}$ gives

$$\mathfrak{R}_{S_{\mathrm{data}}}(\mathcal{F}_D^{(L)})\le C\mathbb{E}\left[\sup_{\{T_{t,S}\}\in\Theta_D}\sum_{t=1}^{L}\sum_{S\in\mathcal{S}_t}\sum_{i=1}^{N}\langle\Sigma_{t,S,i},T_{t,S}H_{t,S,i}(\{T_{\tau,S'}\}_{\tau>t})\rangle_F\mid S_{\mathrm{data}}\right],$$

where the $\Sigma_{t,S,i}$ are independent Rademacher sign matrices of matching size and $C:=\dfrac{\sqrt{2}L_\ell L_{L\mathrm{attn}}(D)\Gamma_1(D)}{N}$.

Stacking the $\Sigma_{t,S,i}$ across i gives

$$\sum_{i=1}^{N}\langle\Sigma_{t,S,i},T_{t,S}H_{t,S,i}(\{T_{\tau,S'}\}_{\tau>t})\rangle_F=\langle E_{t,S}H_{t,S}(\{T_{\tau,S'}\}_{\tau>t})^\top,T_{t,S}\rangle_F.$$

Therefore, using $\sup_x \sum_j a_j(x) \le \sum_j \sup_x a_j(x)$ and trace/spectral duality,

$$\mathfrak{R}_{S_{\mathrm{data}}}(\mathcal{F}_D^{(L)}) \le C \sum_{t=1}^{L} \sum_{S \in \mathcal{S}_t} \mathbb{E}\left[\sup_{\substack{\|T_{t,S}\|_\star \le D_{t,S} \\ \|T_{\tau,S'}\|_\star \le D_{\tau,S'}\ (\tau>t)}} \langle E_{t,S} H_{t,S}(\{T_{\tau,S'}\}_{\tau>t})^\top, T_{t,S} \rangle_F \right]$$

$$= C \sum_{t=1}^{L} \sum_{S \in \mathcal{S}_t} D_{t,S} \mathbb{E}\left[\sup_{\|T_{\tau,S'}\|_\star \le D_{\tau,S'}(\tau>t)} \left\| E_{t,S} H_{t,S}(\{T_{\tau,S'}\}_{\tau>t})^\top \right\|_2 \mid S_{\mathrm{data}} \right].$$

Applying Lemma B.4 to the bounded class $\mathcal{F}_D^{(L)}$ yields the stated uniform generalization bound. □

C Proofs for Main results

We begin with two reductions that are used repeatedly. The first reduction says that if a loss is invariant under projecting an update onto fixed left and right subspaces, then nuclear-norm regularization never prefers the unprojected representative. The second reduction rewrites the nuclear norm via a positive-semidefinite lift, which we can apply the Barvinok–Pataki rank bound (Pataki, 1998; Barvinok, 2001).

C.1 Proofs for Section 4.1

Lemma C.1 (Projector compression for nuclear-norm regularization)**.** *Let $P_{\mathrm{L}} \in \mathbb{R}^{m\times m}$ and $P_{\mathrm{R}} \in \mathbb{R}^{n\times n}$ be orthogonal projectors. Let $\Phi : \mathbb{R}^{m\times n} \to \mathcal{E}$ be any map such that $\Phi(T) = \Phi(P_{\mathrm{L}} T P_{\mathrm{R}})\ \forall T \in \mathbb{R}^{m\times n}$. Then $\|P_{\mathrm{L}} T P_{\mathrm{R}}\|_\star \le \|T\|_\star\ \forall T \in \mathbb{R}^{m\times n}$. Consequently, if*

$$T \mapsto g(\Phi(T) + s) + \lambda \|T\|_\star, \lambda > 0,$$

admits a global minimizer, then for every global minimizer $\tilde{T}$, the compressed matrix $T^\star \coloneqq P_{\mathrm{L}} \tilde{T} P_{\mathrm{R}}$ is also a global minimizer and satisfies

$$\mathrm{rank}(T^\star) \le \min\{\mathrm{rank}(P_{\mathrm{L}}), \mathrm{rank}(P_{\mathrm{R}})\}.$$

Proof of Lemma C.1. Since P_{L} and P_{R} are orthogonal projectors, we have $\|P_{\mathrm{L}}\|_2 \le 1, \|P_{\mathrm{R}}\|_2 \le 1$. By Schatten-norm submultiplicativity,

$$\|P_{\mathrm{L}} T P_{\mathrm{R}}\|_\star \le \|P_{\mathrm{L}}\|_2 \|T\|_\star \|P_{\mathrm{R}}\|_2 \le \|T\|_\star.$$

Let $\widetilde{T}$ be any global minimizer and define $T^\star \coloneqq P_{\mathrm{L}} \widetilde{T} P_{\mathrm{R}}$. By hypothesis, $\Phi(T^\star) = \Phi(\widetilde{T})$, so the loss term is unchanged. The nuclear norm does not increase, so

$$g(\Phi(T^\star) + s) + \lambda \|T^\star\|_\star \le g(\Phi(\widetilde{T}) + s) + \lambda \|\widetilde{T}\|_\star.$$

Since $\widetilde{T}$ is globally optimal, equality must hold, and $T^\star$ is also globally optimal. We also have

$$\mathrm{rank}(T^\star) \le \min\{\mathrm{rank}(P_{\mathrm{L}}), \mathrm{rank}(P_{\mathrm{R}})\}.$$

□

Lemma C.2 (Nuclear-norm lifting: equivalence of infima)**.** *Let $g : \mathcal{E} \to \mathbb{R}$ be any function and fix $s \in \mathcal{E}$. Then*

$$\min_{T \in \mathbb{R}^{m\times n}} F_{\mathrm{primal}}(T) = \min_{Z \succeq 0} F_{\mathrm{lift}}(Z).$$

Moreover:

1. *For every T there exists $Z \succeq 0$ with $\bar{Z} = T$ and $F_{\mathrm{lift}}(Z) = F_{\mathrm{primal}}(T)$.*
2. *For every $Z \succeq 0$, $F_{\mathrm{primal}}(\bar{Z}) \le F_{\mathrm{lift}}(Z)$.*
3. *If $Z^\star$ minimizes F_{lift}, then $T^\star = \bar{Z}^\star$ minimizes F_{primal}.*

Proof of Lemma C.2. Fix $Z \succeq 0$ and factor $Z = RR^\top$ with $R = \begin{bmatrix} U \\ V \end{bmatrix}$. Then $\bar{Z} = UV^\top$ and $\|UV^\top\|_\star \le \frac{1}{2}(\|U\|_F^2 + \|V\|_F^2) = \frac{1}{2}\operatorname{tr}(Z)$, so $F_{\mathrm{primal}}(\bar{Z}) \le F_{\mathrm{lift}}(Z)$.

Conversely, fix $T \in \mathbb{R}^{m\times n}$ and let $T = U_T \Sigma V_T^\top$ be a compact SVD. Set $U \coloneqq U_T\Sigma^{1/2} \in \mathbb{R}^{m\times r}$ and $V \coloneqq V_T\Sigma^{1/2} \in \mathbb{R}^{n\times r}$, where $r = \operatorname{rank}(T)$. Then $UV^\top = T$ and $\|U\|_F^2 = \|V\|_F^2 = \operatorname{tr}(\Sigma) = \|T\|_\star$. Define

$$Z \coloneqq \begin{bmatrix} UU^\top & UV^\top \\ VU^\top & VV^\top \end{bmatrix} = \begin{bmatrix} U \\ V \end{bmatrix}\begin{bmatrix} U \\ V \end{bmatrix}^\top \succeq 0,$$

so $\bar{Z} = UV^\top = T$ and $\operatorname{tr}(Z) = \|U\|_F^2 + \|V\|_F^2 = 2\|T\|_\star$. Hence $F_{\mathrm{lift}}(Z) = F_{\mathrm{primal}}(T)$. Taking infima yields equality, and the minimizer mapping follows. □

Proof of Theorem 4.1. Define

$$F(T) \coloneqq g(\Phi(T) + s) + \lambda\|T\|_\star.$$

Since $\lambda > 0$, the map $T \mapsto \lambda\|T\|_\star$ is coercive. Since g is bounded below and lower semicontinuous, and since Φ is continuous, the objective F is bounded below, lower semicontinuous, and coercive on the finite-dimensional space $\mathbb{R}^{m\times n}$. Therefore, F attains a global minimum.

Fix a global minimizer $\widetilde{T}$ and let $\widetilde{T} = AB^\top$ be a balanced factorization obtained from a compact singular value decomposition, so that

$$\frac{1}{2}\left(\|A\|_F^2 + \|B\|_F^2\right) = \|\widetilde{T}\|_\star.$$

Define

$$\widetilde{Z} \coloneqq \begin{bmatrix} A \\ B \end{bmatrix}\begin{bmatrix} A \\ B \end{bmatrix}^\top.$$

Then $\widetilde{Z} \succeq 0$, $\overline{\widetilde{Z}} = \widetilde{T}$, and $\operatorname{tr}(\widetilde{Z}) = 2\|\widetilde{T}\|_\star$.

Consider the conditional trace-minimization problem

$$\min_{Z\succeq 0} \operatorname{tr}(Z) \text{ subject to } \Phi(\bar{Z}) = \Phi(\widetilde{T}).$$

This feasible set is nonempty because it contains $\widetilde{Z}$. The objective is bounded below by 0. Moreover, every trace sublevel set is compact, because $Z \succeq 0$ implies $\|Z\|_F \le \operatorname{tr}(Z)$. Hence, the conditional trace-minimization problem attains its minimum. Let $v^\star$ denote its minimum value.

The set $\mathcal{C}^\star \coloneqq \left\{Z \succeq 0 : \Phi(\bar{Z}) = \Phi(\widetilde{T}), \operatorname{tr}(Z) = v^\star\right\}$ is nonempty and compact. Choose an extreme point $Z^\star$ of $\mathcal{C}^\star$. Since $\mathcal{C}^\star$ is the set of minimizers of the linear functional $Z \mapsto \operatorname{tr}(Z)$ over the spectrahedron $\mathcal{C} \coloneqq \{Z \succeq 0 : \Phi(\bar{Z}) = \Phi(\widetilde{T})\}$, the set $\mathcal{C}^\star$ is a face of $\mathcal{C}$. Therefore, $Z^\star$ is also an extreme point of $\mathcal{C}$.

Let $e_1, \dots, e_{d_\Phi}$ be an orthonormal basis of $\mathcal{R}(\Phi)$. For $\ell = 1, \dots, d_\Phi$, define

$$A_\ell \coloneqq \frac{1}{2}\begin{bmatrix} \mathbf{0} & \Phi^\star(e_\ell) \\ \Phi^\star(e_\ell)^\top & \mathbf{0} \end{bmatrix}.$$

Then the constraint $\Phi(\bar{Z}) = \Phi(\widetilde{T})$ is equivalent to the independent affine constraints

$$\langle A_\ell, Z\rangle = \langle e_\ell, \Phi(\widetilde{T})\rangle_{\mathcal{E}}, \ell = 1, \dots, d_\Phi.$$

The independence follows because $\Phi^\star$ is injective on $\mathcal{R}(\Phi)$.

By the Barvinok-Pataki rank bound for extreme points of spectrahedra with d_Φ independent affine constraints (Pataki, 1998; Barvinok, 2001), $Z^\star$ satisfies

$$\frac{\operatorname{rank}(Z^\star)\big(\operatorname{rank}(Z^\star)+1\big)}{2} \le d_\Phi.$$

Set $T^\star := \overline{Z^\star}$. Then $\Phi(T^\star) = \Phi(\overline{Z^\star}) = \Phi(\widetilde{T})$. Furthermore, Lemma C.2 gives

$$F(T^\star) \le g(\Phi(\widetilde{T}) + s) + \frac{\lambda}{2}\operatorname{tr}(Z^\star) \le g(\Phi(\widetilde{T}) + s) + \frac{\lambda}{2}\operatorname{tr}(\widetilde{Z}) = F(\widetilde{T}).$$

Since $\widetilde{T}$ is globally optimal, $T^\star$ is also globally optimal. Finally, $\operatorname{rank}(T^\star) \le \operatorname{rank}(Z^\star)$, so

$$\frac{\operatorname{rank}(T^\star)\big(\operatorname{rank}(T^\star)+1\big)}{2} \le d_\Phi.$$

Solving this quadratic inequality gives the displayed rank bound. □

Proof of Theorem 4.3. Let

$$F(\{T_{t,S}\}) := \mathcal{L}(\{T_{t,S}\}) + \sum_{t=1}^{L}\sum_{S\in\mathcal{S}_t} \lambda_{t,S}\|T_{t,S}\|_\star.$$

Because all $\lambda_{t,S} > 0$, the regularizer is coercive on the finite-dimensional product space $\prod_{t,S}\mathbb{R}^{m_{t,S}\times n_{t,S}}$. By Assumption 1, each stage map in the recursion of Definition 3.2 is continuous, hence the map from the parameter tuple to z_1 is continuous; since g_0 is bounded below and lower semicontinuous, $\mathcal{L} = g_0(z_1)$ is bounded below and lower semicontinuous. Therefore, F attains a global minimum.

Fix any global minimizer of F. We now process the stages from upstream to downstream, i.e. in the order $t = L, L-1, \dots, 1$.

Assume that all stages $u > t$ have already been fixed once and for all. Then the corresponding state $z^\star_{t+1}$ is fixed and will not change in the subsequent steps. Now fix a block $S \in \mathcal{S}_t$ and hold every block other than $T_{t,S}$ fixed. The resulting conditional objective has the form

$$T \mapsto \tilde{g}_{t,S}\big(\Phi_{t,S}(T, z^\star_{t+1})\big) + \lambda_{t,S}\|T\|_\star,$$

where $\tilde{g}_{t,S}$ is obtained by fixing all companion inputs and composing the remaining stage maps with g_0. Because the companion blocks and $z^\star_{t+1}$ are fixed, all maps below stage t are continuous functions of the single argument, and g_0 is lower semicontinuous and bounded below. Thus $\tilde{g}_{t,S}$ is bounded below and lower semicontinuous on the relevant finite-dimensional space.

Since the current value of $T_{t,S}$ is already globally optimal for this conditional problem with all companion blocks fixed, Theorem 4.1 allows us to replace it by another conditional minimizer $T^\star_{t,S}$ with the same $\Phi_{t,S}$ satisfying

$$\frac{r_{t,S}(r_{t,S}+1)}{2} \le d_{t,S}(z^\star_{t+1}), r_{t,S} := \operatorname{rank}(T^\star_{t,S}).$$

The equality of the resulting values implies that the stage-t state z_t and all downstream states $z_{t-1}, \dots, z_1$ are unchanged by the replacement. Thus, the global objective value is preserved, and all blocks already processed remain valid.

Proceeding inductively through $t = L, L-1, \dots, 1$ yields a global minimizer $\{T^\star_{t,S}\}$ such that for every $t \in [L]$ and $S \in \mathcal{S}_t$,

$$\frac{r_{t,S}(r_{t,S}+1)}{2} \le d_{t,S}(z^\star_{t+1}) \le d^{\max}_{t,S}.$$

The final scalar bound follows from the implication

$$\frac{r(r+1)}{2} \le d \Longrightarrow r \le \left\lfloor \frac{\sqrt{8d+1}-1}{2} \right\rfloor.$$

□

C.2 Proofs for Section 4.2

The proof of Theorem 4.4 has three ingredients: row-wise score centering for softmax, a blockwise projector-compression principle at a rank-minimal global minimizer, and explicit identification of the query, key, and value subspaces that remain observable after the frozen attention computation.

Lemma C.3 (Row-wise softmax depends only on the centered score matrix)**.** *Let $L' \in \mathbb{N}$ and let $P^{\perp}_{\mathbf{1}_{L'}} = I_{L'} - \frac{1}{L'}\mathbf{1}_{L'}\mathbf{1}_{L'}^{\top}$. Let $C \in (\mathbb{R} \cup \{-\infty\})^{a \times L'}$ be such that each row of C has at least one finite entry. Then for every $M \in \mathbb{R}^{a \times L'}$,*

$$\mathrm{softmax}_{\mathrm{row}}(M + C) = \mathrm{softmax}_{\mathrm{row}}(MP^{\perp}_{\mathbf{1}_{L'}} + C).$$

In particular, when $C = 0$, $\mathrm{softmax}_{\mathrm{row}}(M) = \mathrm{softmax}_{\mathrm{row}}(MP^{\perp}_{\mathbf{1}_{L'}})$.

Proof of Lemma C.3. Fix a row index $i \in [a]$ and define $J_i \coloneqq \{j \in [L'] : C_{ij} > -\infty\}$. By assumption, $J_i \neq \emptyset$. Set $c_i \coloneqq \frac{1}{L'}\sum_{j=1}^{L'} M_{ij}$. Then, for every $j \in J_i$, $(MP^{\perp}_{\mathbf{1}_{L'}} + C)_{ij} = (M + C)_{ij} - c_i$. Hence

$$\frac{\exp((MP^{\perp}_{\mathbf{1}_{L'}} + C)_{ij})}{\sum_{k \in J_i}\exp((MP^{\perp}_{\mathbf{1}_{L'}} + C)_{ik})} = \frac{e^{-c_i}\exp((M + C)_{ij})}{e^{-c_i}\sum_{k \in J_i}\exp((M + C)_{ik})} = \frac{\exp((M + C)_{ij})}{\sum_{k \in J_i}\exp((M + C)_{ik})}.$$

For $j \notin J_i$, both $(M + C)_{ij}$ and $(MP^{\perp}_{\mathbf{1}_{L'}} + C)_{ij}$ are equal to $-\infty$, so both rows assign zero softmax mass to coordinate j. Therefore, the i-th row of the two row-wise softmax matrices coincide. Since i was arbitrary, the matrices are equal. The special case $C = 0$ is immediate. □

Theorem C.4 (Self-consistent projector compression)**.** *Let $\lambda_S > 0$ for every $S \in \mathcal{S}$, and define*

$$F(\{T_S\}_{S \in \mathcal{S}}) \coloneqq \mathcal{L}(\{T_S\}_{S \in \mathcal{S}}) + \sum_{S \in \mathcal{S}} \lambda_S \|T_S\|_{\star}.$$

Assume that F admits a global minimizer. Suppose that for each block S and every fixed companion tuple $\{T_{S'}\}_{S' \neq S}$ there exist orthogonal projectors $P^{\mathrm{L}}_S(\{T_{S'}\}_{S' \neq S}) \in \mathbb{R}^{m_S \times m_S}$, $P^{\mathrm{R}}_S(\{T_{S'}\}_{S' \neq S}) \in \mathbb{R}^{n_S \times n_S}$, such that

$$\mathcal{L}(\{T_{S'}\}_{S' \neq S}, T_S) = \mathcal{L}\Big(\{T_{S'}\}_{S' \neq S}, P^{\mathrm{L}}_S(\{T_{S'}\}_{S' \neq S}) T_S P^{\mathrm{R}}_S(\{T_{S'}\}_{S' \neq S})\Big)\ \forall T_S.$$

Then there exists a global minimizer $\{T^{\star}_S\}$ minimizing

$$\sum_{S \in \mathcal{S}} \mathrm{rank}(T_S)$$

over the global minimizer set, and every such minimizer satisfies

$$\mathrm{rank}(T^{\star}_S) \leq \min\Big\{\mathrm{rank}\big(P^{\mathrm{L}}_S(\{T^{\star}_{S'}\}_{S' \neq S})\big), \mathrm{rank}\big(P^{\mathrm{R}}_S(\{T^{\star}_{S'}\}_{S' \neq S})\big)\Big\}.$$

Proof of Theorem C.4. Since $\mathrm{rank}(T_S) \in \{0, 1, \ldots, \min(m_S, n_S)\}$ for each block S, the set of possible values of $\sum_S \mathrm{rank}(T_S)$ is finite, and there exists a global minimizer $\{T^{\star}_S\} \in \mathrm{argmin}\, F$ for which $\sum_S \mathrm{rank}(T_S)$ is minimal.

Fix a block $S \in \mathcal{S}$ and abbreviate $P^{\mathrm{L},\star}_S \coloneqq P^{\mathrm{L}}_S(\{T^{\star}_{S'}\}_{S' \neq S}), P^{\mathrm{R},\star}_S \coloneqq P^{\mathrm{R}}_S(\{T^{\star}_{S'}\}_{S' \neq S})$. Define $\widehat{T}_S \coloneqq P^{\mathrm{L},\star}_S T^{\star}_S P^{\mathrm{R},\star}_S$. By the assumed invariance of the loss,

$$\mathcal{L}(\{T^{\star}_{S'}\}_{S' \neq S}, \widehat{T}_S) = \mathcal{L}(\{T^{\star}_{S'}\}_{S' \neq S}, T^{\star}_S).$$

By Lemma C.1, $\|\widehat{T}_S\|_{\star} \leq \|T^{\star}_S\|_{\star}$. Hence $F(\{T^{\star}_{S'}\}_{S' \neq S}, \widehat{T}_S) \leq F(\{T^{\star}_S\})$. Since $\{T^{\star}_S\}$ is already globally optimal, the new tuple is also a global minimizer.

If $\mathrm{rank}(T^{\star}_S) > \min\{\mathrm{rank}(P^{\mathrm{L},\star}_S), \mathrm{rank}(P^{\mathrm{R},\star}_S)\}$, then

$$\mathrm{rank}(\widehat{T}_S) \leq \min\{\mathrm{rank}(P^{\mathrm{L},\star}_S), \mathrm{rank}(P^{\mathrm{R},\star}_S)\} < \mathrm{rank}(T^{\star}_S),$$

so, the new global minimizer would have strictly smaller total rank, contradicting the choice of $\{T^{\star}_S\}$. Therefore, the displayed rank bound must hold for every block S. □

Proof of Theorem 4.4. By Lemma B.2, for each sample i the network output is a deterministic function of the collection

$$\left\{(W_{t,S,0}^{(a)} + T_{t,S}^{(a)})H_{t,S,i}(\{T_{\tau,S'}^{(a')}\}_{\tau>t})\right\}_{(t,a,S)\in\widetilde{\mathcal{S}}}.$$

Therefore the empirical loss exhibits L-stage blockwise linear transformation dependence in the sense of Definition 3.2, with linear mappings $\Phi_{t,(a,S)}(T,z) = TH_{t,S}(z)$. Assumption 1 lets us invoke Theorem 4.3, so the corresponding trace-norm regularized objective admits a global minimizer. We now verify the projector-invariance hypothesis of Theorem C.4 for the same objective, viewed as a multi-block problem indexed by $\widetilde{\mathcal{S}}$.

Fix an adapted block $(t,a,S) \in \widetilde{\mathcal{S}}$ and hold all companion blocks fixed. These companions determine a fixed upstream state z_{t+1} and therefore fixed samplewise input matrices $H_{t,Q,i}, H_{t,K,i}, H_{t,V,i} (i \in [N])$.

First, consider the case $S = Q$. Hold all blocks other than $T_{t,Q}^{(a)}$ fixed. Define

$$\widetilde{K}_{t,i}^{(a)} \coloneqq (W_{t,K,0}^{(a)} + T_{t,K}^{(a)})H_{t,K,i}, \quad \widetilde{V}_{t,i}^{(a)} \coloneqq (W_{t,V,0}^{(a)} + T_{t,V}^{(a)})H_{t,V,i},$$

and for $T \in \mathbb{R}^{d_{k,t}\times n_{t,Q}}$ set

$$A_{t,Q,i}^{(a)}(T) \coloneqq \mathrm{softmax}_{\mathrm{row}}\left(\frac{((W_{t,Q,0}^{(a)} + T)H_{t,Q,i})^\top \widetilde{K}_{t,i}^{(a)}}{\sqrt{d_{k,t}}} + C_{t,i}^M\right).$$

Since every other head, every residual contribution, and every downstream stage is fixed, the full stage-t output on sample i is equal to a fixed matrix plus

$$W_{t,O}^{(a)}\widetilde{V}_{t,i}^{(a)}A_{t,Q,i}^{(a)}(T)^\top.$$

Hence, the conditional loss depends on T only through the collection

$$\left\{W_{t,O}^{(a)}\widetilde{V}_{t,i}^{(a)}A_{t,Q,i}^{(a)}(T)^\top\right\}_{i=1}^N.$$

Define

$$P_{t,i}^\perp \coloneqq I_{\varrho_{t,K,i}} - \frac{1}{\varrho_{t,K,i}}\mathbf{1}_{\varrho_{t,K,i}}\mathbf{1}_{\varrho_{t,K,i}}^\top; K_{t,\mathrm{ctr}}^{(a)} \coloneqq [\widetilde{K}_{t,1}^{(a)}P_{t,1}^\perp, \ldots, \widetilde{K}_{t,N}^{(a)}P_{t,N}^\perp]; \widehat{T}_{t,Q}^{(a)} \coloneqq P_{K_{t,\mathrm{ctr}}^{(a)}}T_{t,Q}^{(a)}P_{H_{t,Q}}.$$

By Lemma C.3, for every i, the matrix $A_{t,Q,i}^{(a)}(T)$ depends only on

$$\left(\frac{((W_{t,Q,0}^{(a)} + T)H_{t,Q,i})^\top \widetilde{K}_{t,i}^{(a)}}{\sqrt{d_{k,t}}}\right)P_{t,i}^\perp.$$

Moreover, $(H_{t,Q,i})^\top(\widehat{T}_{t,Q}^{(a)})^\top\widetilde{K}_{t,i}^{(a)}P_{t,i}^\perp = (H_{t,Q,i})^\top(T_{t,Q}^{(a)})^\top\widetilde{K}_{t,i}^{(a)}P_{t,i}^\perp$, because $(H_{t,Q,i})^\top P_{H_{t,Q}} = (H_{t,Q,i})^\top$ and $P_{K_{t,\mathrm{ctr}}^{(a)}}\widetilde{K}_{t,i}^{(a)}P_{t,i}^\perp = \widetilde{K}_{t,i}^{(a)}P_{t,i}^\perp$. Therefore $A_{t,Q,i}^{(a)}(\widehat{T}_{t,Q}^{(a)}) = A_{t,Q,i}^{(a)}(T_{t,Q}^{(a)})\ \forall i$, so the conditional loss is invariant under

$$T_{t,Q}^{(a)} \mapsto P_{K_{t,\mathrm{ctr}}^{(a)}}T_{t,Q}^{(a)}P_{H_{t,Q}}.$$

Next, consider the case $S = K$. Hold all blocks other than $T_{t,K}^{(a)}$ fixed and define

$$\widetilde{Q}_{t,i}^{(a)} \coloneqq (W_{t,Q,0}^{(a)} + T_{t,Q}^{(a)})H_{t,Q,i}, \quad \widetilde{V}_{t,i}^{(a)} \coloneqq (W_{t,V,0}^{(a)} + T_{t,V}^{(a)})H_{t,V,i}.$$

For $T \in \mathbb{R}^{d_{k,t}\times n_{t,K}}$, set

$$A_{t,K,i}^{(a)}(T) \coloneqq \mathrm{softmax}_{\mathrm{row}}\left(\frac{(\widetilde{Q}_{t,i}^{(a)})^\top(W_{t,K,0}^{(a)} + T)H_{t,K,i}}{\sqrt{d_{k,t}}} + C_{t,i}^M\right).$$

As above, the conditional loss depends on T only through

$$\left\{W_{t,O}^{(a)}\widetilde{V}_{t,i}^{(a)}A_{t,K,i}^{(a)}(T)^\top\right\}_{i=1}^N.$$

By Lemma C.3, $A^{(a)}_{t,K,i}(T)$ depends only on

$$\frac{(\widetilde{Q}^{(a)}_{t,i})^\top W^{(a)}_{t,K,0} H_{t,K,i} P^\perp_{t,i} + (\widetilde{Q}^{(a)}_{t,i})^\top T H_{t,K,i} P^\perp_{t,i}}{\sqrt{d_{k,t}}}.$$

Define

$Q^{(a)}_{t,\mathrm{all}} \coloneqq [\widetilde{Q}^{(a)}_{t,1}, \dots, \widetilde{Q}^{(a)}_{t,N}]$, $H_{t,K,\mathrm{ctr}} \coloneqq [H_{t,K,1}P^\perp_{t,1}, \dots, H_{t,K,N}P^\perp_{t,N}]$, $\widehat{T}^{(a)}_{t,K} \coloneqq P_{Q^{(a)}_{t,\mathrm{all}}} T^{(a)}_{t,K} P_{H_{t,K,\mathrm{ctr}}}$.

Then $(\widetilde{Q}^{(a)}_{t,i})^\top \widehat{T}^{(a)}_{t,K} H_{t,K,i} P^\perp_{t,i} = (\widetilde{Q}^{(a)}_{t,i})^\top T^{(a)}_{t,K} H_{t,K,i} P^\perp_{t,i}$, because $(\widetilde{Q}^{(a)}_{t,i})^\top P_{Q^{(a)}_{t,\mathrm{all}}} = (\widetilde{Q}^{(a)}_{t,i})^\top$, $P_{H_{t,K,\mathrm{ctr}}} H_{t,K,i} P^\perp_{t,i} = H_{t,K,i} P^\perp_{t,i}$. Thus $A^{(a)}_{t,K,i}(\widehat{T}^{(a)}_{t,K}) = A^{(a)}_{t,K,i}(T^{(a)}_{t,K})$ $\forall i$, and the conditional loss is invariant under

$$T^{(a)}_{t,K} \mapsto P_{Q^{(a)}_{t,\mathrm{all}}} T^{(a)}_{t,K} P_{H_{t,K,\mathrm{ctr}}}.$$

Finally, consider the case $S = V$. Hold all blocks other than $T^{(a)}_{t,V}$ fixed. Then all attention matrices

$$A^{(a)}_{t,i} \coloneqq \mathrm{softmax}_{\mathrm{row}}\left(\frac{(\widetilde{Q}^{(a)}_{t,i})^\top \widetilde{K}^{(a)}_{t,i}}{\sqrt{d_{k,t}}} + C^M_{t,i}\right)$$

are fixed, and the conditional loss depends on T only through

$$\left\{W^{(a)}_{t,O}(W^{(a)}_{t,V,0} + T) H_{t,V,i} (A^{(a)}_{t,i})^\top\right\}_{i=1}^N.$$

Define $H^{(a)}_{t,V,A} \coloneqq [H_{t,V,1}(A^{(a)}_{t,1})^\top, \dots, H_{t,V,N}(A^{(a)}_{t,N})^\top]$ and set $\widehat{T}^{(a)}_{t,V} \coloneqq P_{W^{(a)\top}_{t,O}} T^{(a)}_{t,V} P_{H^{(a)}_{t,V,A}}$. Then for every i, $W^{(a)}_{t,O}\widehat{T}^{(a)}_{t,V} H_{t,V,i}(A^{(a)}_{t,i})^\top = W^{(a)}_{t,O} T^{(a)}_{t,V} H_{t,V,i}(A^{(a)}_{t,i})^\top$, because $W^{(a)}_{t,O} P_{W^{(a)\top}_{t,O}} = W^{(a)}_{t,O}$, $P_{H^{(a)}_{t,V,A}} H_{t,V,i}(A^{(a)}_{t,i})^\top = H_{t,V,i}(A^{(a)}_{t,i})^\top$. Hence, the conditional loss is invariant under

$$T^{(a)}_{t,V} \mapsto P_{W^{(a)\top}_{t,O}} T^{(a)}_{t,V} P_{H^{(a)}_{t,V,A}}.$$

We have therefore verified the hypotheses of Theorem C.4 for the flattened block family $\widetilde{\mathcal{S}}$. Hence, there exists a global minimizer minimizing total rank over the global minimizer set, and for that minimizer, the rank of each adapted block is bounded by the ranks of the corresponding projector pairs.

Evaluating the projector ranks at this rank-minimal global minimizer gives the following intermediate bounds:

$$\mathrm{rank}(T^{(a)\star}_{t,Q}) \le \min\{\mathrm{rank}(K^{(a)\star}_{t,\mathrm{ctr}}), \mathrm{rank}(H^\star_{t,Q})\}$$
$$\mathrm{rank}(T^{(a)\star}_{t,K}) \le \min\{\mathrm{rank}(Q^{(a)\star}_{t,\mathrm{all}}), \mathrm{rank}(H^\star_{t,K,\mathrm{ctr}})\},$$

and

$$\mathrm{rank}(T^{(a)\star}_{t,V}) \le \min\{\mathrm{rank}(W^{(a)}_{t,O}), \mathrm{rank}(H^{(a)\star}_{t,V,A})\}.$$

Here

$$K^{(a)\star}_{t,\mathrm{ctr}} \coloneqq [\widetilde{K}^{(a)\star}_{t,1} P^\perp_{t,1}, \dots, \widetilde{K}^{(a)\star}_{t,N} P^\perp_{t,N}], \quad Q^{(a)\star}_{t,\mathrm{all}} \coloneqq [\widetilde{Q}^{(a)\star}_{t,1}, \dots, \widetilde{Q}^{(a)\star}_{t,N}],$$

and

$$H^{(a)\star}_{t,V,A} \coloneqq [H^\star_{t,V,1}(A^{(a)\star}_{t,1})^\top, \dots, H^\star_{t,V,N}(A^{(a)\star}_{t,N})^\top], \quad H^\star_{t,K,\mathrm{ctr}} \coloneqq [H^\star_{t,K,1}P^\perp_{t,1}, \dots, H^\star_{t,K,N}P^\perp_{t,N}],$$

We now bound the ranks of these matrices. Since

$$K^{(a)\star}_{t,\mathrm{ctr}} = [\widetilde{K}^{(a)\star}_{t,1}, \dots, \widetilde{K}^{(a)\star}_{t,N}]\,\mathrm{blkdiag}(P^\perp_{t,1}, \dots, P^\perp_{t,N}),$$

we have

$$\mathrm{rank}(K^{(a)\star}_{t,\mathrm{ctr}}) \le \min\left\{\mathrm{rank}([\widetilde{K}^{(a)\star}_{t,1}, \dots, \widetilde{K}^{(a)\star}_{t,N}]), \sum_{i=1}^N \mathrm{rank}(P^\perp_{t,i})\right\}.$$

Because $\operatorname{rank}(P_{t,i}^{\perp}) = \varrho_{t,K,i} - 1$ and $[\widetilde{K}_{t,1}^{(a)\star}, \dots, \widetilde{K}_{t,N}^{(a)\star}] = (W_{t,K,0}^{(a)} + T_{t,K}^{(a)\star})H_{t,K}^{\star}$, it follows that

$$\operatorname{rank}(K_{t,\mathrm{ctr}}^{(a)\star}) \le \min\left\{ d_{k,t}, \operatorname{rank}(H_{t,K}^{\star}), \sum_{i=1}^{N} (\varrho_{t,K,i} - 1) \right\}.$$

Combining this with

$$\operatorname{rank}(T_{t,Q}^{(a)\star}) \le \min\{\operatorname{rank}(K_{t,\mathrm{ctr}}^{(a)\star}), \operatorname{rank}(H_{t,Q}^{\star})\}$$

gives the stated query bound.

For the key bound, first observe that $\operatorname{rank}(Q_{t,\mathrm{all}}^{(a)\star}) \le \min\{d_{k,t}, \operatorname{rank}(H_{t,Q}^{\star})\}$, because $Q_{t,\mathrm{all}}^{(a)\star} = (W_{t,Q,0}^{(a)} + T_{t,Q}^{(a)\star})H_{t,Q}^{\star}$. Also, $H_{t,K,\mathrm{ctr}}^{\star} = H_{t,K}^{\star} \operatorname{blkdiag}(P_{t,1}^{\perp}, \dots, P_{t,N}^{\perp})$, so

$$\operatorname{rank}(H_{t,K,\mathrm{ctr}}^{\star}) \le \min\left\{ \operatorname{rank}(H_{t,K}^{\star}), \sum_{i=1}^{N} (\varrho_{t,K,i} - 1) \right\}.$$

Combining these two estimates with

$$\operatorname{rank}(T_{t,K}^{(a)\star}) \le \min\{\operatorname{rank}(Q_{t,\mathrm{all}}^{(a)\star}), \operatorname{rank}(H_{t,K,\mathrm{ctr}}^{\star})\}$$

gives the stated key bound.

For the value bound, we have $H_{t,V,A}^{(a)\star} = H_{t,V}^{\star} \operatorname{blkdiag}\big((A_{t,1}^{(a)\star})^{\top}, \dots, (A_{t,N}^{(a)\star})^{\top}\big)$, and hence

$$\operatorname{rank}(H_{t,V,A}^{(a)\star}) \le \min\left\{ \operatorname{rank}(H_{t,V}^{\star}), \sum_{i=1}^{N} \operatorname{rank}(A_{t,i}^{(a)\star}) \right\}.$$

Since $W_{t,O}^{(a)}$ has at most $d_{v,t}$ columns, $\operatorname{rank}(W_{t,O}^{(a)}) \le d_{v,t}$ so combining this with

$$\operatorname{rank}(T_{t,V}^{(a)\star}) \le \min\{\operatorname{rank}(W_{t,O}^{(a)}), \operatorname{rank}(H_{t,V,A}^{(a)\star})\}$$

gives the stated value bound. □

The adapted-output extension follows from the same blockwise compression, but not from a simultaneous linear observation map for the full $Q/K/V/O$ tuple.

Proof of Theorem 4.5. Let $\mathcal{I}$ index all selected query, key, value, and full output-projection updates. By the assumed continuity of the forward computation, the empirical loss is lower semicontinuous and bounded below. Since every adapted block has a strictly positive nuclear-norm coefficient, the objective is coercive on the finite-dimensional product space $\prod_{j\in\mathcal{I}} \mathbb{R}^{m_j \times n_j}$ and therefore attains a global minimum.

Since the sum of matrix ranks is integer valued and bounded above by $\sum_{j\in\mathcal{I}} \min\{m_j, n_j\}$, there is a global minimizer $\{T_j^{\star}\}$ minimizing that sum over all global minimizers. We establish the hypotheses of Theorem C.4 by fixing all companion updates each time a block is compressed.

For any query update, the input matrices $H_{t,Q,i}$ and adapted keys $\widetilde{K}_{t,i}^{(a)}$ do not depend on that query update. Row-wise score centering and the argument in the proof of Theorem 4.4 therefore show that replacing $T_{t,Q}^{(a)}$ by $P_{K_{t,\mathrm{ctr}}^{(a)}} T_{t,Q}^{(a)} P_{H_{t,Q}}$ preserves every corresponding attention matrix. The adapted output projection is held fixed during this replacement, so it does not change that invariance. Similarly, for a key update, replacing $T_{t,K}^{(a)}$ by

$$P_{Q_{t,\mathrm{all}}^{(a)}} T_{t,K}^{(a)} P_{H_{t,K,\mathrm{ctr}}}$$

preserves the centered score products and the attention matrices. Thus the query and key projector bounds from Theorem 4.4 remain valid.

For an adapted value block, all attention matrices and the companion output update $T_{t,O}$ are fixed. Write the adapted output projection in head blocks as

$$\widetilde{W}_{t,O} \coloneqq W_{t,O,0} + T_{t,O} = [\widetilde{W}^{(a)}_{t,O}]_{a\in\mathcal{A}_t}, \qquad \widetilde{W}^{(a)}_{t,O} \in \mathbb{R}^{d_{\mathrm{out},t}\times d_{v,t}},$$

and put

$$C^{(a)}_{t,i} = H_{t,V,i}(A^{(a)}_{t,i})^\top, \quad C^{(a)}_t = [C^{(a)}_{t,1}, \dots, C^{(a)}_{t,N}].$$

Here $C^{(a)}_{t,i} \in \mathbb{R}^{n_{t,V}\times \varrho_{t,Q,i}}$. Set

$$P^{(a)}_O = P_{(\widetilde{W}^{(a)}_{t,O})^\top}, \qquad P^{(a)}_C = P_{C^{(a)}_t}.$$

For any $T \in \mathbb{R}^{d_{v,t}\times n_{t,V}}$ and every sample i,

$$\widetilde{W}^{(a)}_{t,O} P^{(a)}_O T P^{(a)}_C C^{(a)}_{t,i} = \widetilde{W}^{(a)}_{t,O} T C^{(a)}_{t,i}.$$

Thus the stage output and the downstream loss are unchanged by $T \mapsto P^{(a)}_O T P^{(a)}_C$. The corresponding nuclear norm does not increase by Lemma C.1.

For the full output projection, hold all $Q/K/V$ updates and all other blocks fixed. For each training sample i, define

$$B^{(a)}_{t,i} \coloneqq \widetilde{V}^{(a)}_{t,i}(A^{(a)}_{t,i})^\top \in \mathbb{R}^{d_{v,t}\times \varrho_{t,Q,i}}, \quad B_{t,i} \coloneqq \begin{bmatrix} B^{(a_1)}_{t,i} \\ \vdots \\ B^{(a_{|\mathcal{A}_t|})}_{t,i} \end{bmatrix}, \quad H_{t,O} \coloneqq [B_{t,1}, \dots, B_{t,N}].$$

Every $B_{t,i}$ is then fixed, and the varying part of the stage output equals $T_{t,O}B_{t,i}$. Since $P_{H_{t,O}}B_{t,i} = B_{t,i}$,

$$(T_{t,O}P_{H_{t,O}})B_{t,i} = T_{t,O}B_{t,i} \qquad (i \in [N]).$$

The replacement $T_{t,O} \mapsto T_{t,O}P_{H_{t,O}}$ therefore preserves the objective's loss term and does not increase the nuclear norm.

All conditional compression projectors depend only on companion updates. Theorem C.4 applies to the finite block family $\mathcal{I}$ and gives, at the chosen rank-minimal global minimizer,

$$\mathrm{rank}(T^{(a)\star}_{t,V}) \le \min\{\mathrm{rank}(\widetilde{W}^{(a)\star}_{t,O}), \mathrm{rank}(C^{(a)\star}_t)\}, \quad \mathrm{rank}(T^\star_{t,O}) \le \min\{d_{\mathrm{out},t}, \mathrm{rank}(H^\star_{t,O})\}. \tag{C.1}$$

It remains to bound the three ranks on the right-hand side of (C.1). First, $\widetilde{W}^{(a)\star}_{t,O}$ has $d_{\mathrm{out},t}$ rows and $d_{v,t}$ columns, so

$$\mathrm{rank}(\widetilde{W}^{(a)\star}_{t,O}) \le \min\{d_{\mathrm{out},t}, d_{v,t}\} \le d_{v,t}.$$

Second, the identity

$$C^{(a)\star}_t = H^\star_{t,V}\,\mathrm{blkdiag}((A^{(a)\star}_{t,1})^\top, \dots, (A^{(a)\star}_{t,N})^\top)$$

implies $\mathrm{rank}(C^{(a)\star}_t) \le \min\{\mathrm{rank}(H^\star_{t,V}), \sum_{i=1}^N \mathrm{rank}(A^{(a)\star}_{t,i})\}$. Third, $H^\star_{t,O} \in \mathbb{R}^{(|\mathcal{A}_t|d_{v,t})\times\sum_{i=1}^N \varrho_{t,Q,i}}$ by construction, so $\mathrm{rank}(H^\star_{t,O}) \le \min\{|\mathcal{A}_t|d_{v,t}, \sum_{i=1}^N \varrho_{t,Q,i}\}$. Substituting these three estimates into (C.1) yields the two bounds stated in the theorem and completes the proof. □

C.3 Proofs for Section 4.3

For the positive rank-deficiency statement, the first step is to write the first-order stationarity equations in a symmetric block form.

Lemma C.5 (Stationarity conditions for the perturbed block LoRA objective)**.** *In the setting of Definition 4.6, suppose $\mathcal{L}$ is differentiable and $\{(U_S, V_S)\}$ is a first-order stationary point of $\mathcal{L}^{\mathrm{lora}}_{\{\lambda_S\},\{P_S\}}$. Let $T_S \coloneqq U_S V_S^\top$ and $G_S \coloneqq \nabla_{T_S}\mathcal{L}(\{T_{S'}\})$. Then for every $S \in \mathcal{S}$,*

$$\big(\mathcal{B}_{m_S,n_S}(G_S) + \lambda_S I_{m_S+n_S} + P_S\big)R_S = 0, R_S = \begin{bmatrix} U_S \\ V_S \end{bmatrix}.$$

Proof of Lemma C.5. Fix S and differentiate w.r.t. U_S and V_S using the chain rule: $\nabla_{U_S}\mathcal{L}(\{U_{S'}V_{S'}^\top\}) = G_S V_S$ and $\nabla_{V_S}\mathcal{L}(\{U_{S'}V_{S'}^\top\}) = G_S^\top U_S$. The regularizer contributes $\lambda_S U_S$ and $\lambda_S V_S$. The perturbation term satisfies $\nabla_{R_S}\frac{1}{2}\langle P_S, R_S R_S^\top\rangle = P_S R_S$. Stacking the two stationarity equations yields the displayed KKT condition. □

The genericity argument is a dimension-counting statement. We will need the following elementary measure-zero lemma:

Lemma C.6 (Measure-zero Minkowski sums with C^1 images)**.** *Let $\mathcal{M}$ be an m-dimensional embedded C^1 submanifold of $\mathbb{R}^d$. Let $U \subseteq \mathbb{R}^p$ be open, and let $f : U \to \mathbb{R}^d$ be C^1. If $m + p < d$, then*

$$\mathcal{M} + f(U) \coloneqq \{x + f(u) : x \in \mathcal{M}, u \in U\}$$

has Lebesgue measure zero in $\mathbb{R}^d$.

Proof of Lemma C.6. Since $\mathcal{M} \subseteq \mathbb{R}^d$ is an embedded submanifold, it is second countable and admits a countable C^1 atlas. It is therefore enough to prove the claim on one coordinate patch.

Fix a chart $\varphi : V \to \mathcal{M}$, where $V \subseteq \mathbb{R}^m$ is open. On this patch, $\{\varphi(v) + f(u) : v \in V, u \in U\}$ is the image of the C^1 map

$$F : V \times U \to \mathbb{R}^d, \qquad F(v, u) = \varphi(v) + f(u).$$

The domain $V \times U$ is open in $\mathbb{R}^{m+p}$. Set $k \coloneqq m + p$. By assumption, $k < d$.

Cover $V \times U$ by countably many compact cubes Q_ℓ whose interiors cover $V \times U$ and whose closures are contained in $V \times U$. Fix one such cube Q. Since F is C^1, its derivative is bounded on Q. Hence there exists $L_Q < \infty$ such that

$$\|F(x) - F(y)\|_2 \le L_Q \|x - y\|_2 \ \forall x, y \in Q.$$

Subdivide Q into N^k subcubes of side length proportional to N^{-1}. The image of each subcube is contained in a Euclidean ball in $\mathbb{R}^d$ of radius at most $C_Q N^{-1}$, where C_Q is independent of N. Therefore the d-dimensional Lebesgue outer measure of $F(Q)$ is bounded by

$$N^k \omega_d \left(\frac{C_Q}{N}\right)^d = \omega_d C_Q^d N^{k-d},$$

where ω_d is the volume of the unit ball in $\mathbb{R}^d$. Since $k - d < 0$, this upper bound converges to 0 as $N \to \infty$. Thus $F(Q)$ has Lebesgue measure zero.

Taking the countable union over all cubes Q_ℓ and then over the countable atlas of $\mathcal{M}$ shows that $\mathcal{M} + f(U)$ has Lebesgue measure zero in $\mathbb{R}^d$. □

Proof of Theorem 4.7. If $r > d$, then every $R \in \mathbb{R}^{d\times r}$ satisfies $\operatorname{rank}(R) \le d < r$, so the conclusion is immediate. Assume from now on that $r \le d$.

Let

$$D_{\mathrm{sym}} \coloneqq \dim \mathbb{S}^d = \frac{d(d+1)}{2}.$$

For $s \in \{r, r+1, \dots, d\}$, define $\mathcal{A}_s \coloneqq \{M \in \mathbb{S}^d : \dim \mathcal{N}(M) = s\}$. The set $\mathcal{A}_s$ is an embedded C^∞ submanifold of $\mathbb{S}^d$ with

$$\dim(\mathcal{A}_s) = D_{\mathrm{sym}} - \frac{s(s+1)}{2}.$$

Indeed, $\mathcal{A}_s$ is the set of symmetric matrices of rank $d - s$; its codimension in $\mathbb{S}^d$ is the dimension of $\mathbb{S}^s$, namely $s(s+1)/2$.

Suppose that for some $P \in \mathbb{S}^d$ there exist $G \in \mathcal{G}$ and $R \in \mathbb{R}^{d\times r}$ such that $\operatorname{rank}(R) = r$ and

$$(G + \lambda I_d + P)R = 0.$$

Set $M \coloneqq G + \lambda I_d + P$. Since the columns of R are linearly independent and lie in $\mathcal{N}(M)$, we have $\dim \mathcal{N}(M) \ge r$. Therefore $M \in \bigcup_{s=r}^d \mathcal{A}_s$, and hence

$$P = M - \lambda I_d - G \in \bigcup_{s=r}^{d} \big((\mathcal{A}_s - \lambda I_d) - \mathcal{G}\big).$$

It remains to show that each set $(\mathcal{A}_s - \lambda I_d) - \mathcal{G}$ has Lebesgue measure zero in $\mathbb{S}^d$. Since $-\mathcal{G}$ is contained in a finite union of C^1 images of dimension at most p, there exist open sets $U_j \subseteq \mathbb{R}^{p_j}$ with $p_j \leq p$ and C^1 maps $f_j : U_j \to \mathbb{S}^d, j = 1, \ldots, J$, such that

$$-\mathcal{G} \subseteq \bigcup_{j=1}^{J} f_j(U_j).$$

Fix $s \in \{r, \ldots, d\}$ and $j \in [J]$. The translated manifold $\mathcal{A}_s - \lambda I_d$ is an embedded C^∞ submanifold of $\mathbb{S}^d$ with dimension $D_{\mathrm{sym}} - \frac{s(s+1)}{2}$. Because $s \geq r$ and $p_j \leq p$,

$$\left(D_{\mathrm{sym}} - \frac{s(s+1)}{2}\right) + p_j \leq D_{\mathrm{sym}} - \frac{r(r+1)}{2} + p < D_{\mathrm{sym}}.$$

Lemma C.6, applied inside the Euclidean space $\mathbb{S}^d$, implies that $(\mathcal{A}_s - \lambda I_d) + f_j(U_j)$ has Lebesgue measure zero in $\mathbb{S}^d$. Taking the finite union over j gives that $(\mathcal{A}_s - \lambda I_d) - \mathcal{G}$ has Lebesgue measure zero in $\mathbb{S}^d$.

Finally, taking the finite union over $s = r, \ldots, d$ gives a Lebesgue-null exceptional set. Thus, for every P outside this exceptional set, there do not exist $G \in \mathcal{G}$ and $R \in \mathbb{R}^{d \times r}$ with $\mathrm{rank}(R) = r$ satisfying

$$(G + \lambda I_d + P)R = 0.$$

Intersecting this full-measure set with $\mathbb{S}^d_+$ proves the conclusion for Lebesgue-a.e. positive semidefinite perturbation. □

The next lemma connects the stagewise observation model to the KKT matrices above.

Lemma C.7 (Stagewise gradient range under L-stage linear map dependence)**.** *Assume $\mathcal{L}$ from Definition 3.2 is differentiable. Fix a parameter tuple and the corresponding states z_t. Then for every $t \in [L]$ and $S \in \mathcal{S}_t$, we have $\nabla_{T_{t,S}} \mathcal{L} \in \mathcal{R}(\Phi_{t,S}(\cdot, z_{t+1})^\star)$.*

Proof of Lemma C.7. Fix $t \in [L]$ and $S \in \mathcal{S}_t$, and fix all blocks except $T_{t,S}$. Because all upstream blocks are fixed, the state z_{t+1} is fixed. Define the conditional loss

$$F_{t,S}(T) \coloneqq \mathcal{L}\big(\{T_{u,A}\}_{(u,A)\neq(t,S)}, T_{t,S} = T\big).$$

By Definition 3.2, the function $F_{t,S}$ depends on T only through the linear map $T \mapsto \Phi_{t,S}(T, z_{t+1})$. Therefore, for every $\Delta \in \mathcal{N}(\Phi_{t,S}(\cdot, z_{t+1}))$ and every T, $F_{t,S}(T + \Delta) = F_{t,S}(T)$. Since F is differentiable, this implies

$$0 = DF_{t,S}(T)[\Delta] = \langle \nabla_{T_{t,S}} \mathcal{L}, \Delta \rangle_F \forall \Delta \in \mathcal{N}(\Phi_{t,S}(\cdot, z_{t+1})).$$

Hence $\nabla_{T_{t,S}} \mathcal{L} \in \mathcal{N}(\Phi_{t,S}(\cdot, z_{t+1}))^\perp$. In finite dimensions, $\mathcal{N}(\Phi_{t,S}(\cdot, z_{t+1}))^\perp = \mathcal{R}(\Phi_{t,S}(\cdot, z_{t+1})^\star)$, so $\nabla_{T_{t,S}} \mathcal{L} \in \mathcal{R}(\Phi_{t,S}(\cdot, z_{t+1})^\star)$. □

To apply Theorem 4.7, we must bound the dimension of the union of these moving ranges over all reachable upstream states. The next lemma provides this bound by recursively parameterizing the reachable observed-output sets.

Lemma C.8 (Recursive parameterization of reachable states and moving adjoint ranges)**.** *Assume Assumption 4 holds. Then the following statements hold:*

1. *For each $t \in [L+1]$, the reachable state set $\mathcal{Z}_t^{\mathrm{reach}}$ is C^1-parameterizable with dimension at most β_t.*
2. *For each $t \in [L]$ and $S \in \mathcal{S}_t$, define*

$$\mathcal{G}_{t,S} \coloneqq \bigcup_{z_{t+1} \in \mathcal{Z}_{t+1}^{\mathrm{reach}}} \big\{\mathcal{B}_{m_{t,S}, n_{t,S}}(G) : G \in \mathcal{R}\big(\Phi_{t,S}(\cdot, z_{t+1})^\star\big)\big\} \subseteq \mathbb{S}^{m_{t,S}+n_{t,S}}.$$

Then $\mathcal{G}_{t,S}$ is C^1-parameterizable with dimension at most $\beta_{t+1} + d_{t,S}^{\max}$.

Proof of Lemma C.8. For each $t \in [L]$, let $\mathcal{O}_t$ be an open neighborhood of the reachable argument set

$$\mathcal{A}_t^{\arg} \coloneqq \left\{ (\{u_S\}_{S\in\mathcal{S}_t}, z_{t+1}) : z_{t+1} \in \mathcal{Z}_{t+1}^{\mathrm{reach}}, u_S \in \mathcal{R}\big(\Phi_{t,S}(\cdot, z_{t+1})\big) \right\}$$

on which g_t is C^1. For each pair (t, S), let $\mathcal{O}_{t,S}^{\Phi}$ be an open neighborhood of $\mathcal{Z}_{t+1}^{\mathrm{reach}}$ on which the coordinate maps

$$z_{t+1} \mapsto \Phi_{t,S}(B, z_{t+1}) \text{ and } z_{t+1} \mapsto \Phi_{t,S}(\cdot, z_{t+1})^\star(e)$$

are C^1 for every basis matrix B of $\mathbb{R}^{m_{t,S}\times n_{t,S}}$ and every basis vector e of $\mathcal{E}_{t,S}$. These open sets exist by Assumption 4.

We first prove the reachable-state claim by backward induction on t. For $t = L+1$, the set $\mathcal{Z}_{L+1}^{\mathrm{reach}}$ is a singleton, hence it is C^1-parameterizable with dimension $0 = \beta_{L+1}$.

Assume that $\mathcal{Z}_{t+1}^{\mathrm{reach}}$ is C^1-parameterizable with dimension at most β_{t+1}. Choose finitely many C^1 maps $z_{t+1}^{(j)} : \Theta_j \to \mathcal{Z}_{t+1}^{\mathrm{reach}}, j = 1, \dots, J$, whose images cover $\mathcal{Z}_{t+1}^{\mathrm{reach}}$, where each Θ_j is open and has dimension at most β_{t+1}. For each block $S \in \mathcal{S}_t$, choose a basis $B_{S,1}, \dots, B_{S,p_S}$ of $\mathbb{R}^{m_{t,S}\times n_{t,S}}$, and set $k_S \coloneqq d_{t,S}^{\max}$. For every j and every tuple of index sets $I = (I_S)_{S\in\mathcal{S}_t}, I_S = \{i_{S,1}, \dots, i_{S,q_S}\} \subseteq [p_S], q_S \le k_S$, define the preliminary argument map

$$Q_{j,I}\left(\theta, \{\alpha_S\}_{S\in\mathcal{S}_t}\right) \coloneqq \left(\left\{ \sum_{\ell=1}^{q_S} \alpha_{S,\ell} \Phi_{t,S}\big(B_{S,i_{S,\ell}}, z_{t+1}^{(j)}(\theta)\big) \right\}_{S\in\mathcal{S}_t}, z_{t+1}^{(j)}(\theta) \right).$$

Here $\alpha_S \in \mathbb{R}^{q_S}$. Restrict the domain to the open set $\Omega_{j,I} \coloneqq Q_{j,I}^{-1}(\mathcal{O}_t)$. On $\Omega_{j,I}$, define

$$F_{j,I}\left(\theta, \{\alpha_S\}_{S\in\mathcal{S}_t}\right) \coloneqq g_t \left(\left\{ \sum_{\ell=1}^{q_S} \alpha_{S,\ell} \Phi_{t,S}\big(B_{S,i_{S,\ell}}, z_{t+1}^{(j)}(\theta)\big) \right\}_{S\in\mathcal{S}_t}, z_{t+1}^{(j)}(\theta) \right).$$

The restriction to $\Omega_{j,I}$ is the important point: it ensures that the argument of g_t lies in the neighborhood on which g_t is assumed to be C^1. Since each $z_{t+1}^{(j)}(\theta)$ lies in $\mathcal{Z}_{t+1}^{\mathrm{reach}}$, and since the coordinate maps of $\Phi_{t,S}$ are C^1 on $\mathcal{O}_{t,S}^{\Phi}$, the map $F_{j,I}$ is C^1 on $\Omega_{j,I}$.

The dimension of the domain of $F_{j,I}$ is at most

$$\beta_{t+1} + \sum_{S\in\mathcal{S}_t} q_S \le \beta_{t+1} + \sum_{S\in\mathcal{S}_t} d_{t,S}^{\max} = \beta_t.$$

It remains to check that these finitely many maps cover $\mathcal{Z}_t^{\mathrm{reach}}$. Take any $z_t \in \mathcal{Z}_t^{\mathrm{reach}}$. By definition, there exist $z_{t+1} \in \mathcal{Z}_{t+1}^{\mathrm{reach}}, u_S \in \mathcal{R}\big(\Phi_{t,S}(\cdot, z_{t+1})\big)$ such that

$$z_t = g_t(\{u_S\}_{S\in\mathcal{S}_t}, z_{t+1}).$$

Choose j and $\theta \in \Theta_j$ with $z_{t+1} = z_{t+1}^{(j)}(\theta)$. For each S, the range $\mathcal{R}\big(\Phi_{t,S}(\cdot, z_{t+1})\big)$ has dimension at most k_S and is spanned by the finite list

$$\{\Phi_{t,S}(B_{S,a}, z_{t+1})\}_{a=1}^{p_S}.$$

Therefore, u_S can be expressed as a linear combination of at most k_S vectors from this list. This gives some tuple I and coefficients $\{\alpha_S\}$ such that $z_t = F_{j,I}\left(\theta, \{\alpha_S\}_{S\in\mathcal{S}_t}\right)$. Moreover, the corresponding argument belongs to $\mathcal{A}_t^{\arg} \subseteq \mathcal{O}_t$, so the parameter point lies in $\Omega_{j,I}$. Thus $\mathcal{Z}_t^{\mathrm{reach}}$ is covered by finitely many C^1 images of dimension at most β_t. This completes the induction and proves the first claim.

We now prove the state-dependent adjoint ranges claim. Fix $t \in [L]$ and $S \in \mathcal{S}_t$. By the first claim, $\mathcal{Z}_{t+1}^{\mathrm{reach}}$ is covered by finitely many C^1 parameterizations $z_{t+1}^{(j)} : \Theta_j \to \mathcal{Z}_{t+1}^{\mathrm{reach}}$, with domain dimension at most β_{t+1}. Let $e_1, \dots, e_{q_{t,S}}$ be a basis of $\mathcal{E}_{t,S}$, and set $k \coloneqq d_{t,S}^{\max}$. For every j and every index set $I = \{i_1, \dots, i_q\} \subseteq [q_{t,S}], q \le k$, define

$$M_{j,I}(\theta, \alpha) \coloneqq \sum_{\ell=1}^{q} \alpha_\ell \mathcal{B}_{m_{t,S},n_{t,S}} \left(\Phi_{t,S}(\cdot, z_{t+1}^{(j)}(\theta))^\star(e_{i_\ell}) \right),$$

where $\alpha \in \mathbb{R}^q$. As before, one may restrict the domain to the open preimage

$$\Omega_{j,I}^{\Phi} := \left\{(\theta, \alpha) : z_{t+1}^{(j)}(\theta) \in \mathcal{O}_{t,S}^{\Phi}\right\}.$$

This restriction contains all parameter points needed to cover the reachable set, because $z_{t+1}^{(j)}(\Theta_j) \subseteq \mathcal{Z}_{t+1}^{\mathrm{reach}} \subseteq \mathcal{O}_{t,S}^{\Phi}$. By Assumption 4, each map $M_{j,I}$ is C^1 on $\Omega_{j,I}^{\Phi}$. Its domain dimension is at most $\beta_{t+1} + q \leq \beta_{t+1} + d_{t,S}^{\max}$.

It remains to verify coverage. Take any element of $\mathcal{G}_{t,S}$. It has the form $\mathcal{B}_{m_{t,S},n_{t,S}}(G), G \in \mathcal{R}\big(\Phi_{t,S}(\cdot, z_{t+1})^\star\big)$ for some $z_{t+1} \in \mathcal{Z}_{t+1}^{\mathrm{reach}}$. Choose j and θ such that $z_{t+1} = z_{t+1}^{(j)}(\theta)$. The space $\mathcal{R}\big(\Phi_{t,S}(\cdot, z_{t+1})^\star\big)$ has dimension at most $d_{t,S}^{\max} = k$ and is spanned by

$$\{\Phi_{t,S}(\cdot, z_{t+1})^\star(e_a)\}_{a=1}^{q_{t,S}}.$$

Hence G can be written as a linear combination of at most k vectors from this list. Therefore $\mathcal{B}_{m_{t,S},n_{t,S}}(G)$ belongs to one of the images of $M_{j,I}$. Thus $\mathcal{G}_{t,S}$ is C^1-parameterizable with dimension at most $\beta_{t+1} + d_{t,S}^{\max}$. □

We now combine the KKT identity, the gradient-range lemma, and the parameterization lemma. The result is the stagewise version of the master genericity theorem:

Theorem C.9 (Generic rank-deficiency for L-stage perturbed LoRA)**.** *Assume the objective has L-stage blockwise linear map dependence, is differentiable, and Assumption 4 holds. Fix factor ranks $r_{t,S}$ and regularization parameters $\lambda_{t,S} > 0$. If*

$$\frac{r_{t,S}(r_{t,S}+1)}{2} > \beta_{t+1} + d_{t,S}^{\max} \ \forall t \in [L], S \in \mathcal{S}_t,$$

then for Lebesgue-a.e. choice of PSD perturbations $\{P_{t,S}\}$, every first-order stationary point of the perturbed LoRA objective satisfies

$$\mathrm{rank}(R_{t,S}) < r_{t,S} \forall t \in [L], S \in \mathcal{S}_t.$$

Proof of Theorem C.9. Fix a stage/block (t, S). At any first-order stationary point, Lemma C.5 gives

$$\big(\mathcal{B}_{m_{t,S},n_{t,S}}(G_{t,S}) + \lambda_{t,S} I + P_{t,S}\big) R_{t,S} = 0,$$

where $G_{t,S} = \nabla_{T_{t,S}} \mathcal{L}$. By Lemma C.7, $G_{t,S} \in \mathcal{R}(\Phi_{t,S}(\cdot, z_{t+1})^\star)$ for the reachable state z_{t+1} induced by the stationary point. Hence $\mathcal{B}_{m_{t,S},n_{t,S}}(G_{t,S}) \in \mathcal{G}_{t,S}$. By Lemma C.8, the family $\mathcal{G}_{t,S}$ is C^1-parameterizable with dimension at most $\beta_{t+1} + d_{t,S}^{\max}$. For the block (t, S), define $d_{t,S}^{\mathrm{blk}} := m_{t,S} + n_{t,S}$. By Lemma C.8, the family $\mathcal{G}_{t,S}$ is contained in a finite union of C^1 images of dimension at most $\beta_{t+1} + d_{t,S}^{\max}$. Since

$$\frac{r_{t,S}(r_{t,S}+1)}{2} > \beta_{t+1} + d_{t,S}^{\max},$$

Theorem 4.7, applied with $d = d_{t,S}^{\mathrm{blk}}$, $r = r_{t,S}$, $\lambda = \lambda_{t,S}$, and $\mathcal{G} = \mathcal{G}_{t,S}$, gives a full-measure set $\Omega_{t,S} \subseteq \mathbb{S}_+^{m_{t,S}+n_{t,S}}$ such that, whenever $P_{t,S} \in \Omega_{t,S}$, no matrix $R_{t,S} \in \mathbb{R}^{(m_{t,S}+n_{t,S}) \times r_{t,S}}$ with $\mathrm{rank}(R_{t,S}) = r_{t,S}$ can satisfy the block KKT equation

$$\big(\mathcal{B}_{m_{t,S},n_{t,S}}(G) + \lambda_{t,S} I + P_{t,S}\big) R_{t,S} = 0$$

for any $\mathcal{B}_{m_{t,S},n_{t,S}}(G) \in \mathcal{G}_{t,S}$.

Taking the product of the full-measure sets $\Omega_{t,S}$ over the finite block family preserves full measure. For any perturbation tuple in this product set, every first-order stationary point satisfies the displayed KKT equation for each block by Lemma C.5, with its gradient block belonging to $\mathcal{G}_{t,S}$ by Lemma C.7. Therefore every first-order stationary point satisfies $\mathrm{rank}(R_{t,S}) < r_{t,S}$ for every $t \in [L]$ and $S \in \mathcal{S}_t$. □

The attention corollary is obtained by evaluating the bottleneck dimension for the concrete map $T \mapsto TH_{t,S}(z_{t+1})$.

Proof of Corollary 4.8. For an adapted attention projection block $\alpha = (a, S)$, the stagewise linear map has the form $\Phi_{t,\alpha}(T, z_{t+1}) = TH_{t,S}(z_{t+1})$. Therefore

$$d_{t,\alpha}(z_{t+1}) = \dim \mathcal{R}(\Phi_{t,\alpha}(\cdot, z_{t+1})) = m_{t,\alpha} \operatorname{rank}(H_{t,S}(z_{t+1})).$$

Taking the supremum over $z_{t+1} \in \mathcal{Z}_{t+1}^{\mathrm{reach}}$ gives $d_{t,\alpha}^{\max} = m_{t,\alpha} r_{H,t,\alpha}^{\max}$. Substituting this identity into Theorem C.9 gives the stated result. □

Now we prove the limitation result from Section 4.3.

Proof of Proposition 4.9. Choose any constant $a > \lambda + \varepsilon$ and set

$$g(t) \coloneqq a(t^2 - 1)^2.$$

Then g is smooth and nonconvex, $g(0) = a$, $g(1) = 0$, and $g'(0) = 0$. Write $R = \begin{bmatrix} U \\ V \end{bmatrix}$ with $U = (u_1, \ldots, u_r)$ and $V = (v_1, \ldots, v_r)$. Let

$$t(R) \coloneqq UV^\top = \sum_{k=1}^{r} u_k v_k.$$

Under the column-wise vectorization $r' = \operatorname{vec}(R) \in \mathbb{R}^{2r}$, the chain rule gives

$$\nabla_{r'}^2 (g \circ t)(0) = g''(t(0)) \nabla t(0) \nabla t(0)^\top + g'(t(0)) \nabla^2 t(0) = 0,$$

because $t(0) = 0, \nabla t(0) = 0$, and $g'(0) = 0$. Moreover,

$$\nabla_{r'}^2 \left(\frac{\lambda}{2} \|R\|_F^2 \right) = \lambda I_{2r}, \nabla_{r'}^2 \left(\frac{1}{2} \langle P, RR^\top \rangle \right) = I_r \otimes P.$$

The perturbed objective is

$$F_P(R) \coloneqq g(t(R)) + \frac{\lambda}{2} \|R\|_F^2 + \frac{1}{2} \langle P, RR^\top \rangle.$$

Therefore $\nabla_{r'}^2 F_P(0) = \lambda I_{2r} + I_r \otimes P \succeq \lambda I_{2r} \succ 0$, so $R = 0$ is a strict local minimum. Now set

$$R_1 \coloneqq \begin{bmatrix} 1 & 0 & \cdots & 0 \\ 1 & 0 & \cdots & 0 \end{bmatrix}.$$

Then

$$t(R_1) = 1, \|R_1\|_F^2 = 2, R_1 R_1^\top = \begin{bmatrix} 1 & 1 \\ 1 & 1 \end{bmatrix}, \|R_1 R_1^\top\|_F = 2.$$

Hence

$$\begin{aligned} F_P(R_1) &= g(1) + \frac{\lambda}{2} \|R_1\|_F^2 + \frac{1}{2} \langle P, R_1 R_1^\top \rangle \\ &= \lambda + \frac{1}{2} \langle P, R_1 R_1^\top \rangle \\ &\leq \lambda + \frac{1}{2} \|P\|_F \|R_1 R_1^\top\|_F \\ &= \lambda + \|P\|_F \\ &\leq \lambda + \varepsilon < a = F_P(0). \end{aligned}$$

Thus, $R = 0$ is not global. □

C.4 Proofs for Section 4.4

Lemma C.10 ((Recht et al., 2010, Lemma 5.1))**.** *For any $T \in \mathbb{R}^{m \times n}$ and any integer $r \geq \operatorname{rank}(T)$,*

$$\|T\|_\star = \frac{1}{2} \min_{\substack{U \in \mathbb{R}^{m \times r}, V \in \mathbb{R}^{n \times r} \\ UV^\top = T}} \left\{ \|U\|_F^2 + \|V\|_F^2 \right\}.$$

Equivalently,

$$\|T\|_\star = \frac{1}{2} \min_{\substack{U \in \mathbb{R}^{m \times r}, V \in \mathbb{R}^{n \times r} \\ r \in \mathbb{N}, UV^\top = T}} \left\{ \|U\|_F^2 + \|V\|_F^2 \right\}.$$

We finally prove that small PSD perturbations do not substantially change the trace-norm objective value attained by a global factorized solution.

Proof of Theorem 4.10. The perturbed LoRA objective is lower semicontinuous because $\mathcal{L}$ is lower semicontinuous and the maps $(U_j, V_j) \mapsto U_j V_j^\top$ are continuous. It is coercive because

$$\sum_{j \in \mathcal{I}} \frac{\lambda_j}{2} \big(\|U_j\|_F^2 + \|V_j\|_F^2 \big)$$

is coercive, and each perturbation term is nonnegative. Since $\mathcal{L}$ is bounded below, the perturbed LoRA objective attains a global minimum.

For each S, write

$$\widehat{R}_j \coloneqq \begin{bmatrix} \widehat{U}_j \\ \widehat{V}_j \end{bmatrix}.$$

By Lemma C.10, $\|\widehat{T}_j\|_\star \le \frac{1}{2}\big(\|\widehat{U}_j\|_F^2 + \|\widehat{V}_j\|_F^2\big)$. Since $P_j \succeq 0$, $\langle P_j, \widehat{R}_j \widehat{R}_j^\top \rangle \ge 0$. Therefore

$$\mathcal{L}(\{\widehat{T}_j\}) + \sum_{j \in \mathcal{I}} \lambda_j \|\widehat{T}_j\|_\star \le \mathcal{L}^{\mathrm{lora}}_{\{\lambda_j\},\{P_j\}}(\{\widehat{U}_j, \widehat{V}_j\}).$$

For each S, since $\mathrm{rank}(T_j^\star) \le r_j$, choose a balanced SVD factorization padded with zero columns if necessary, so that

$$T_j^\star = U_j^\star V_j^{\star\top} \text{ and } \|T_j^\star\|_\star = \frac{1}{2}\big(\|U_j^\star\|_F^2 + \|V_j^\star\|_F^2\big).$$

By global optimality of $\{(\widehat{U}_j, \widehat{V}_j)\}$ for the perturbed factorized objective,

$$\mathcal{L}^{\mathrm{lora}}_{\{\lambda_j\},\{P_j\}}(\{\widehat{U}_j, \widehat{V}_j\}) \le \mathcal{L}^{\mathrm{lora}}_{\{\lambda_j\},\{P_j\}}(\{U_j^\star, V_j^\star\}).$$

The right-hand side equals

$$\mathcal{L}(\{T_j^\star\}) + \sum_{j \in \mathcal{I}} \lambda_j \|T_j^\star\|_\star + \frac{1}{2} \sum_{j \in \mathcal{I}} \langle P_j, R_j^\star R_j^{\star\top} \rangle.$$

Since $\{T_j^\star\}$ is a global minimizer of the nuclear-norm-regularized problem, combining the preceding displays proves the first inequality.

For the final bound, use $P_j \succeq 0$, $\|P_j\|_2 \le \varepsilon_j$, and

$$\mathrm{tr}(R_j^\star R_j^{\star\top}) = \|U_j^\star\|_F^2 + \|V_j^\star\|_F^2 = 2\|T_j^\star\|_\star.$$

Then

$$\frac{1}{2}\langle P_j, R_j^\star R_j^{\star\top} \rangle \le \frac{1}{2}\|P_j\|_2 \,\mathrm{tr}(R_j^\star R_j^{\star\top}) \le \varepsilon_j \|T_j^\star\|_\star.$$

Summing over $j \in \mathcal{I}$ gives the claimed perturbation bound. □

D Experiment details

Proxy estimation. The proxy rank is chosen large enough to exactly represent the selected head block; for Qwen2.5-1.5B, this is 128, matching the query-head dimension d_k.

The proxy is trained with the same regularized objective as the LoRA sweep, using $\lambda = 10^{-4}$, AdamW learning rate 5×10^{-4}, and best-checkpoint selection by the lowest training $\mathcal{L}_\lambda^{\mathrm{lora}}$. After training, the LoRA update is merged into the base linear weights, and the model is run in evaluation mode over the training split. We then estimate the downstream activation rank using $\mathrm{rank}(H) = \mathrm{rank}(HH^\top)$ to lower the computation by considering the rank of a smaller matrix. The numerical rank is computed by the number of eigenvalues of $HH^\top$ that are larger than the eigenvalue threshold 10^{-4}.

Let $\hat{r}_{\mathrm{up},S}$ denote the numerical rank of the upstream frozen-feature subspace for block S, and let $\hat{r}_{\mathrm{down},S}$ denote the numerical rank of the downstream observable subspace estimated after the trace-norm proxy optimization. The final predicted threshold is

$$r_{\mathrm{threshold}} = \max_S \min\{\hat{r}_{\mathrm{up},S}, \hat{r}_{\mathrm{down},S}\}.$$

Table 4: Hyperparameters for LoRA fine-tuning and full fine-tuning on NLP tasks.

LoRA fine-tuning				
Task/Dataset		MNLI	RTE	SST-2
Batch size		8	8	8
Learning rate	Head	5e-4	1e-4	1e-4
	Attention	8e-5	1e-3	5e-4
LoRA regularization λ		1e-4	1e-4	1e-4

Full fine-tuning			
Task/Dataset	MNLI	RTE	SST-2
Batch size	8	8	8
Learning rate	2.32e-5	3.375e-4	2e-4
Trace regularization λ	1e-4	1e-4	1e-4

This proxy approximates the optimizer-dependent quantity appearing in Theorem 4.4.

Before the LoRA sweep, the classification head is warmed up for 10 epochs, then frozen. For the high-rank bound case, we use 100 samples, and for the low-rank bound case, we use 4 samples for training.

NeurIPS Paper Checklist

The checklist is designed to encourage best practices for responsible machine learning research, addressing issues of reproducibility, transparency, research ethics, and societal impact. Do not remove the checklist: **The papers not including the checklist will be desk rejected.** The checklist should follow the references and follow the (optional) supplemental material. The checklist does NOT count towards the page limit.

Please read the checklist guidelines carefully for information on how to answer these questions. For each question in the checklist:

- You should answer [Yes], [No], or [N/A].
- [N/A] means either that the question is Not Applicable for that particular paper or the relevant information is Not Available.
- Please provide a short (1–2 sentence) justification right after your answer (even for [N/A]).

The checklist answers are an integral part of your paper submission. They are visible to the reviewers, area chairs, senior area chairs, and ethics reviewers. You will also be asked to include it (after eventual revisions) with the final version of your paper, and its final version will be published with the paper.

The reviewers of your paper will be asked to use the checklist as one of the factors in their evaluation. While [Yes] is generally preferable to [No], it is perfectly acceptable to answer [No] provided a proper justification is given (e.g., error bars are not reported because it would be too computationally expensive" or "we were unable to find the license for the dataset we used"). In general, answering [No] or [N/A] is not grounds for rejection. While the questions are phrased in a binary way, we acknowledge that the true answer is often more nuanced, so please just use your best judgment and write a justification to elaborate. All supporting evidence can appear either in the main paper or the supplemental material, provided in appendix. If you answer [Yes] to a question, in the justification please point to the section(s) where related material for the question can be found.

IMPORTANT, please:

- **Delete this instruction block, but keep the section heading "NeurIPS Paper Checklist"**,
- **Keep the checklist subsection headings, questions/answers and guidelines below.**
- **Do not modify the questions and only use the provided macros for your answers**.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: We provided technical proof and experimental evidence to support our claims.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation, as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss the limitations in Section 6.

Guidelines:

- The answer [N/A] means that the paper has no limitation, while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes]

   Justification: The paper provided the full set of assumptions and complete proofs in Section 3, Section 4, and Appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [Yes]

   Justification: We provided experiment descriptions in the main paper and appendix.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [No]

   Justification: The experiments use GLUE tasks and an open-source Qwen2.5 checkpoint, but the current paper does not yet include released code.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).

- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: We provided details for experiments in the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [No]

   Justification: The current figures show single optimization trajectories without error bars.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.
   - It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
   - For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
   - If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [No]

   Justification: The draft does not report GPU type, memory, runtime per run, and total compute.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.

- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: Our paper follows the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: This paper identifies several phenomena in general parameter-efficient fine-tuning methods that deepen our understanding of LoRA and may inform the development of more capable small models. Advances in these methods could have both positive and negative societal impacts.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

Justification: Our work does not pose such risks.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: We have properly cited papers and resources used in our experiment.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [N/A]

Justification: Our paper does not release new assets.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: We do not have experiments involving crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [N/A]

Justification: We do not involve crowdsourcing or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs have been employed solely for writing, editing, and formatting purposes. Their use does not affect the core methodology, scientific rigor, or originality of the research.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.