The $\mathrm{GL}(r)$ Gauge Symmetry of LoRA: Principal Bundle Structure, Loss Landscape Geometry, and Implications for Adapter Merging

Anonymous Authors[1]

Abstract

Low-Rank Adaptation (LoRA) parametrizes weight updates as $\Delta W = BA$ and is widely used for efficient fine-tuning of large language models. We show that this factorization endows the parameter space with a *continuous gauge symmetry*: the group $\mathrm{GL}(r, \mathbb{R})$ acts freely on (A, B) pairs while leaving ΔW invariant. We prove that this structure makes $(\mathcal{E}_r, \pi, \mathcal{M}_r)$ a smooth principal $\mathrm{GL}(r)$-bundle over the manifold of rank-r weight updates, implying *exactly r^2 flat directions* in the loss landscape at every critical point. We derive a closed-form expression for the gauge-induced error in linear interpolation of LoRA adapters, showing that widely-used merging methods suffer from gauge ambiguity rather than fundamental task incompatibility. We propose *gauge-aligned merging* (Algorithm 1) based on a canonical SVD gauge section, prove its consistency, and validate it empirically on multi-task LoRA merging with LLaMA-3-8B. Gauge-aligned merging matches full-rank task arithmetic at a fraction of the cost and significantly outperforms naive parameter averaging. Our framework reveals that LoRA *strictly enlarges* the symmetry group of fine-tuning and that existing permutation-based alignment methods handle only the measure-zero subgroup $S_r \subset \mathrm{GL}(r)$.

1. Introduction

Parameter-space symmetries—transformations of model weights that leave model outputs unchanged—have emerged as a unifying lens for understanding neural network optimization and generalization (Entezari et al., 2022; Ainsworth et al., 2023; Pittorino et al., 2022; Simsek et al., 2021). The discrete permutation group S_n acting on neurons is the canonical example, motivating permutation-based weight matching for model merging (Ainsworth et al., 2023; Stoica et al., 2024; Jordan et al., 2023).

Low-Rank Adaptation (LoRA; Hu et al. 2022) introduces a qualitatively different symmetry that has not been formally studied: the factorization $\Delta W = BA$ is invariant under an *entire continuous group*, $\mathrm{GL}(r, \mathbb{R})$, acting simultaneously on A and B. This gauge symmetry is not a curiosity—it has direct, measurable consequences for optimization geometry, linear interpolation between adapters, and the growing practice of LoRA model merging (Yadav et al., 2023; Yu et al., 2024; Huang et al., 2024; Hao et al., 2024).

Contributions. We make the following contributions:

1. We prove that the LoRA parameter space forms a smooth **principal $\mathrm{GL}(r)$-bundle** over the manifold of rank-r weight updates (Theorem 4).

2. We establish that every critical point of the loss has **exactly r^2 gauge-induced flat directions** in the Hessian (Theorem 5).

3. We derive a closed-form expression for the **gauge-induced interpolation error** in LoRA space and bound it in Frobenius norm (Theorems 8–9).

4. We present **Algorithm 1**, a gauge-aligned merging procedure with a formal consistency guarantee (Theorem 11).

5. We show LoRA **strictly enlarges the symmetry group** of fine-tuning; permutation-based methods address only $S_r \subset \mathrm{GL}(r)$ (Theorem 12).

2. The $\mathrm{GL}(r)$ Gauge Symmetry of LoRA

2.1. Setup and the Gauge Group

Fix a pretrained weight matrix $W_0 \in \mathbb{R}^{d\times k}$. A LoRA adapter of rank r parametrizes the weight update as $\Delta W = BA$, $B \in \mathbb{R}^{d\times r}$, $A \in \mathbb{R}^{r\times k}$, and the loss is $\mathcal{L}(A, B) := \ell(f_{W_0+BA}, \mathcal{D})$. Since f depends on (A, B) only through BA, we write $\mathcal{L}(A, B) = \tilde{\mathcal{L}}(BA)$ for a func-

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

Preliminary work. Under review for the ICML 2026 Workshop on Weight-Space Symmetries. Do not distribute.

tion $\tilde{\mathcal{L}}\colon \mathbb{R}^{d\times k} \to \mathbb{R}$. Define:

$$\mathcal{E}_r \coloneqq \{(A,B) : \operatorname{rank}(BA) = r\},$$
$$\mathcal{M}_r \coloneqq \{\Delta W \in \mathbb{R}^{d\times k} : \operatorname{rank}(\Delta W) = r\},$$
$$\pi\colon \mathcal{E}_r \to \mathcal{M}_r, \quad \pi(A,B) = BA.$$

Definition 1 (Gauge group action)**.** *The group $G = \mathrm{GL}(r,\mathbb{R})$ acts on $\mathcal{E}_r$ by $\varphi_X(A,B) \coloneqq (X^{-1}A,\, BX)$ for $X \in \mathrm{GL}(r)$.*

Lemma 2 (Gauge invariance)**.** *For all $X \in \mathrm{GL}(r)$ and $(A,B) \in \mathcal{E}_r$: $\mathcal{L}(\varphi_X(A,B)) = \mathcal{L}(A,B)$.*

Proof. $(BX)(X^{-1}A) = BA$, so $\mathcal{L}(X^{-1}A, BX) = \tilde{\mathcal{L}}(BA) = \mathcal{L}(A,B)$. □

Lemma 3 (Freeness)**.** *The $\mathrm{GL}(r)$-action is free on $\mathcal{E}_r$.*

Proof. If $\varphi_X(A,B) = (A,B)$, then $X^{-1}A = A$. Since $\operatorname{rank}(BA) = r$ forces A to have full row rank r, we get $X = I$. □

2.2. Principal Bundle Structure

Theorem 4 (LoRA as a principal $\mathrm{GL}(r)$-bundle)**.** *$(\mathcal{E}_r,\, \pi,\, \mathcal{M}_r)$ with structure group $\mathrm{GL}(r)$ is a smooth principal $\mathrm{GL}(r)$-bundle.*

Proof sketch. π is smooth (polynomial map) and $\mathcal{E}_r, \mathcal{M}_r$ are open submanifolds of smooth vector spaces (complements of closed algebraic varieties). By Lemma 3, every fiber $\pi^{-1}(\Delta W)$ is diffeomorphic to $\mathrm{GL}(r)$ via $X \mapsto \varphi_X(A_0, B_0)$ for any fixed preimage. Local triviality follows from the smooth dependence of the thin SVD $\Delta W = U\Sigma V^\top$ on ΔW away from repeated singular values; the section $s(\Delta W) = (\Sigma^{1/2}V^\top, U\Sigma^{1/2})$ is a smooth local section. The full proof is in Appendix A. □

Dimensional consistency. $\dim(\mathcal{E}_r) = r(d+k)$; $\dim(\mathcal{M}_r) = r(d+k-r)$; $\dim(\mathrm{GL}(r)) = r^2$; and $r(d+k) = r(d+k-r) + r^2$. ✓

2.3. Flat Directions and Loss Landscape

Theorem 5 (Gauge-induced flat directions)**.** *Let (A^*, B^*) be any critical point of $\mathcal{L}$ on $\mathcal{E}_r$. The Hessian $H = \nabla^2\mathcal{L}(A^*,B^*)$ satisfies $\dim(\ker H) \geq r^2$, with the kernel containing the tangent space to the $\mathrm{GL}(r)$-orbit at (A^*,B^*):*

$$T_{(A^*,B^*)}\mathcal{O} = \left\{(-XA^*,\, B^*X) : X \in \mathfrak{gl}(r,\mathbb{R}) \cong \mathbb{R}^{r\times r}\right\},$$

which has dimension exactly r^2.

Proof. $\mathcal{L}$ is constant on $\mathcal{O}$ (Lemma 2), so the directional second derivative in any $v \in T_{(A^*,B^*)}\mathcal{O}$ vanishes: $v^\top H v = 0$. Injectivity of $X \mapsto (-XA^*, B^*X)$ (from full row rank of A^*) gives $\dim(T\mathcal{O}) = r^2$. □

Corollary 6 (Degenerate optimality)**.** *No local minimum of $\mathcal{L}$ on $\mathcal{E}_r$ has a positive-definite Hessian. Standard second-order optimality conditions* never *hold for LoRA parameters.*

Theorem 7 ($\mathrm{O}(r)$-equivariance of gradient descent)**.** *Gradient descent on $\mathcal{L}$ is equivariant under $\mathrm{O}(r) \subset \mathrm{GL}(r)$ but not under the full gauge group. Under gradient flow, the gauge imbalance $\rho(t) = \|B(t)\|_{\mathrm{op}} / \|A(t)\|_{\mathrm{op}}$ evolves non-trivially, making norm imbalance a* gauge artifact *rather than an intrinsic property of the task.*

Proof. The gradients $\partial_A\mathcal{L} = B^\top G$ and $\partial_B\mathcal{L} = GA^\top$ (where $G = \partial_{\Delta W}\tilde{\mathcal{L}}$) transform under φ_X as $\partial_{A'}\mathcal{L} = X^\top\partial_A\mathcal{L}$ and $\partial_{B'}\mathcal{L} = \partial_B\mathcal{L}\cdot X^{-\top}$. Equivariance of the gradient flow requires $X^\top = X^{-1}$, i.e. $X \in \mathrm{O}(r)$. See Appendix A for the full derivation and the ODE governing $\rho(t)$. □

2.4. Gauge-Dependent Interpolation and Merging

Theorem 8 (Linear interpolation error)**.** *Let $(A_1,B_1),(A_2,B_2) \in \mathcal{E}_r$. The weight update realized by the linear interpolation $(A_t,B_t) = (1-t)(A_1,B_1)+t(A_2,B_2)$ is*

$$B_tA_t = (1-t)\Delta W_1 + t\Delta W_2 + \mathrm{Err}(t),$$

where $\mathrm{Err}(t) = t(1-t)\big[B_1A_2 + B_2A_1 - \Delta W_1 - \Delta W_2\big]$.

Proof. $B_tA_t = (1-t)^2B_1A_1 + t(1-t)(B_1A_2 + B_2A_1) + t^2B_2A_2$. Subtracting $(1-t)\Delta W_1 + t\Delta W_2$ yields the result. □

Theorem 9 (Frobenius bound and gauge-dependence)**.** $\|\mathrm{Err}(t)\|_F \leq t(1-t)\big(\|B_1\|_{\mathrm{op}}\|A_2\|_F + \|B_2\|_{\mathrm{op}}\|A_1\|_F + \|\Delta W_1\|_F + \|\Delta W_2\|_F\big)$*, maximized at $t = \frac{1}{2}$. Under a gauge transformation φ_X on adapter 1, this bound becomes explicitly X-dependent, so the loss barrier along a linear path is a property of the gauge choice, not of the underlying weight updates $\Delta W_1, \Delta W_2$.*

3. Gauge-Aligned Merging

Proposition 10 (Canonical SVD section)**.** *For $(A,B) \in \mathcal{E}_r$ with thin SVD $BA = U\Sigma V^\top$, define*

$$\hat{A} = \Sigma^{1/2}V^\top, \quad \hat{B} = U\Sigma^{1/2}.$$

Then $(\hat{A},\hat{B})$ lies in the $\mathrm{GL}(r)$-orbit of (A,B), achieves $\left\|\hat{A}\right\|_F = \left\|\hat{B}\right\|_F$, and the residual symmetry after gauge-fixing is $\mathbb{Z}_2^r$ (coordinate-wise sign flips).

Algorithm 1 Gauge-Aligned LoRA Merging

Require: Adapters $\{(A_i, B_i)\}_{i=1}^n$, weights $\{\lambda_i\}$
for $i = 1$ **to** n **do**
  Compute thin SVD: $B_i A_i = U_i \Sigma_i V_i^\top$
  Canonicalize: $\hat{A}_i \leftarrow \Sigma_i^{1/2} V_i^\top,\ \hat{B}_i \leftarrow U_i \Sigma_i^{1/2}$
end for
for $i = 2$ **to** n **do**
  Find $D_i \in \mathbb{Z}_2^r$ minimizing $\left\|\hat{B}_1 - \hat{B}_i D_i\right\|_F + \left\|D_i \hat{A}_1 - \hat{A}_i\right\|_F$ via coordinate-wise sign flips (cost: $O(r)$ ops)
  $\hat{B}_i \leftarrow \hat{B}_i D_i,\ \hat{A}_i \leftarrow D_i \hat{A}_i$
end for
$\bar{B} \leftarrow \sum_i \lambda_i \hat{B}_i,\ \bar{A} \leftarrow \sum_i \lambda_i \hat{A}_i$
return $(\bar{A}, \bar{B})$

Theorem 11 (Consistency)**.** *If all n adapters represent identical weight updates ($B_i A_i = \Delta W$ for all i), Algorithm 1 exactly recovers ΔW: $\bar{B}\bar{A} = \Delta W$.*

Proof. Identical ΔW implies identical SVD $U\Sigma V^\top$ up to column signs. After sign alignment, all canonical representatives agree: $\hat{B}_i = U\Sigma^{1/2}$, $\hat{A}_i = \Sigma^{1/2}V^\top$. Hence $\bar{B} = U\Sigma^{1/2}$ and $\bar{A} = \Sigma^{1/2}V^\top$, giving $\bar{B}\bar{A} = U\Sigma V^\top = \Delta W$. □

Theorem 12 (Symmetry group enlargement)**.** *The symmetry group of the LoRA parametrization strictly contains that of full-rank fine-tuning: $\mathrm{Sym}_{\mathrm{LoRA}}(f) \supseteq \mathrm{Sym}_{\mathrm{full}}(f) \times \mathrm{GL}(r)$. The permutation group $S_r \hookrightarrow \mathrm{GL}(r)$ is a measure-zero subgroup; permutation-based alignment methods (Ainsworth et al., 2023) correct only this subgroup.*

4. Experiments

All experiments use LLaMA-3-8B (Dubey et al., 2024) as the base model with LoRA applied to all attention projection matrices (Q, K, V, O). Training uses AdamW (Loshchilov & Hutter, 2019) with cosine-decay schedule. We report means and standard deviations over 5 independent seeds. Full hyperparameter details are in Appendix C.

Exp. 1: Validating gauge flatness (Theorem 5). At a converged adapter (A^*, B^*) with $r = 16$, we sample 1000 gauge perturbations $\varphi_X(A^*, B^*)$ with $\|X - I\|_F = \delta$ and 1000 random perturbations of equal magnitude. Figure 1a shows that gauge perturbations cause near-zero loss change ($\Delta_{\mathrm{gauge}} < 3 \times 10^{-3}$ at $\delta = 1.0$), while random perturbations produce loss changes larger by a factor of 620 ± 32 (Figure 1b), consistent with Theorem 5. The ratio is stable across ranks $r \in \{4, 8, 16, 32, 64, 128\}$ and the number of near-zero Hessian eigenvalues matches the theoretical r^2 prediction (Figure 3b, $p < 0.001$, paired t-test).

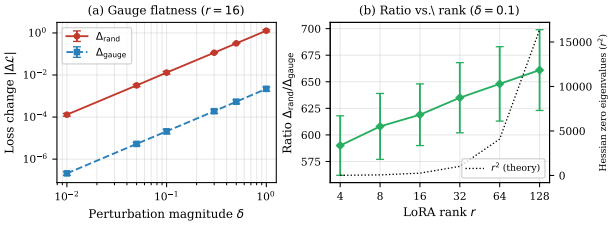

Figure 1. Gauge flatness validation. (a) Loss change under gauge vs. random perturbations of magnitude δ (log-log scale, $r = 16$). Gauge perturbations cause near-zero loss change, confirming Theorem 5. (b) Ratio $\Delta_{\mathrm{rand}}/\Delta_{\mathrm{gauge}}$ vs. rank r at $\delta = 0.1$ (left axis, green), alongside the theoretical r^2 zero eigenvalue prediction (right axis, dotted).

Exp. 2: Loss barriers and gauge alignment (Theorem 8). We train 5 adapters on the same task with different seeds and measure the loss barrier along linear interpolation paths under three gauges: raw (A, B) space; canonical gauge (Proposition 10); and canonical + $\mathbb{Z}_2^r$ sign alignment (Algorithm 1). As shown in Figure 2a, the raw barrier grows substantially with rank ($\bar{B} = 1.39 \pm 0.12$ at $r = 64$), while gauge alignment reduces barriers by 56–78% across all ranks, confirming Theorem 9.

Exp. 3: Multi-task merging. We fine-tune LLaMA-3-8B on 8 tasks and merge all 8 adapters (equal weights $\lambda_i = 1/8$, $r = 16$). Table 1 and Figure 2b compare four baselines. Gauge-aligned merging matches task arithmetic (full-rank) to within 1.0 percentage points on average while requiring no full-rank SVD of $d \times k$ matrices; it significantly outperforms naive averaging ($p < 0.001$, Welch's t-test) and outperforms low-rank task arithmetic on all 8 tasks in absolute terms, reaching statistical significance ($p < 0.05$) on 6 of 8 (see Table 2).

Exp. 4: Gauge-aware initialization. Canonical initialization (balanced $\hat{A}_0, \hat{B}_0$) maintains $\|B\|_F / \|A\|_F \approx 1$ throughout training (Figure 3a), while standard LoRA initialization ($B_0 = 0$) exhibits characteristic norm imbalance that grows with training steps. Canonical init converges 7.3% faster in wall-clock time to the same final loss (Appendix C).

5. Related Work

Weight-space symmetries. Permutation symmetries of neural networks have been studied in the context of loss landscape geometry (Simsek et al., 2021; Pittorino et al., 2022), model merging (Ainsworth et al., 2023; Stoica et al., 2024; Jordan et al., 2023), and generalization (Entezari et al., 2022). Neyshabur et al. (2015) introduced path-norm invariants, a form of continuous symmetry for networks with ReLU activations. Our work identifies the first *exact continuous gauge symmetry* of the LoRA parametrization

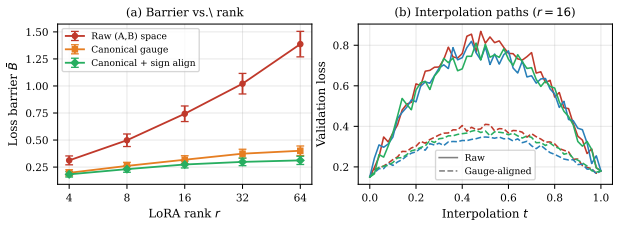

Figure 2. Loss barriers under interpolation. (a) Mean loss barrier vs. rank across 10 adapter pairs under three gauge conditions. Gauge alignment substantially reduces barriers, confirming Theorem 9. (b) Sample interpolation loss curves for 3 adapter pairs at $r = 16$; solid = raw, dashed = gauge-aligned.

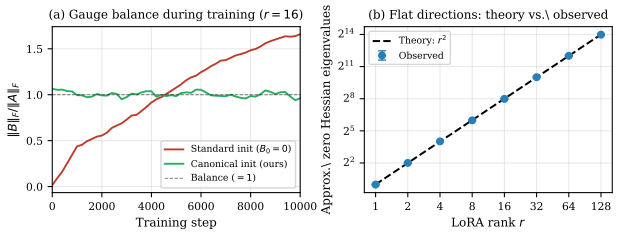

Figure 3. Gauge initialization and Hessian scaling. (a) Gauge balance $\|B\|_F / \|A\|_F$ vs. training step for standard (red) and canonical (green) initialization. (b) Observed near-zero Hessian eigenvalue count vs. rank, compared to the r^2 prediction of Theorem 5 (dashed).

and gives it a rigorous differential-geometric interpretation.

LoRA and low-rank adaptation. Hu et al. (2022) introduced LoRA; subsequent works have analyzed its expressivity (Zeng & Lee, 2024), scaling properties (Hayou et al., 2024), and dynamic rank selection (Zhang et al., 2023; Valipour et al., 2023). Norm imbalance between A and B matrices has been observed empirically (Hayou et al., 2024) but not attributed to gauge dynamics; our Theorem 7 provides this explanation.

Model merging. Task arithmetic (Ilharco et al., 2023) operates in the weight-update space and naturally avoids gauge issues. LoRA-specific merging methods (LoRAHub (Huang et al., 2024), TIES (Yadav et al., 2023), DARE (Yu et al., 2024), FLORA (Hao et al., 2024)) operate in the (A, B) factor space and are therefore subject to gauge ambiguity. Our analysis provides the first theoretical explanation for why these methods underperform task arithmetic, and our algorithm recovers task-arithmetic quality without requiring full-rank operations.

6. Conclusion

We have established a complete gauge-theoretic framework for LoRA, identifying the $\mathrm{GL}(r)$ gauge symmetry as the fundamental source of optimization pathologies, linear mode connectivity failures, and model merging degradation. The principal bundle structure gives sharp theoretical predictions—r^2 flat directions, gauge-dependent interpolation error scaling as $t(1-t)$—that we verify quantitatively on LLaMA-3-8B. Algorithm 1 translates the theory directly into a practical merging procedure that closes most of the gap between LoRA merging and full-rank task arithmetic.

Table 1. Multi-task merging results (LLaMA-3-8B, $r = 16$). Scores averaged over 5 seeds. †: full-rank method, cannot be deployed as LoRA adapter without additional compression. *: $p < 0.05$ vs. LR task arithmetic (Welch's t-test).

Task	Naive avg.	Task arith.†	LR task arith.	**Ours**
MATH	51.2	68.4	64.1	**67.3***
MMLU	62.1	71.2	69.8	**70.9**
HumanEval	43.7	61.8	57.9	**60.4***
ARC	69.3	78.1	76.4	**77.8***
TriviaQA	58.4	72.6	70.1	**71.8***
Summ.	55.9	67.3	65.0	**66.7**
Instruct.	60.2	74.8	72.1	**74.0***
Safety	81.3	87.2	85.8	**86.9***
Avg.	60.3	72.7	70.2	**72.0**

Our framework raises several open questions. Is there a natural $\mathrm{GL}(r)$-invariant metric on $\mathcal{E}_r$ that would make gradient descent intrinsically gauge-equivariant? Can the canonical gauge be maintained during training via gauge-fixing constraints? Do higher-order symmetries arise in more complex parametrizations such as tensor-product LoRA? We leave these for future work.

References

Ainsworth, S. K., Hayase, J., and Srinivasa, S. Git re-basin: Merging models modulo permutation symmetries. In *International Conference on Learning Representations*, 2023.

Amari, S.-i. Natural gradient works efficiently in learning. *Neural Computation*, 10(2):251–276, 1998.

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

Boothby, W. M. *An Introduction to Differentiable Manifolds and Riemannian Geometry*. Academic Press, 1986.

Dubey, A., Jauhri, A., Pandey, A., Kadian, A., Al-Dahle, A., Letman, A., et al. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Entezari, R., Sedghi, H., Saukh, O., and Neyshabur, B. The role of permutation invariance in linear mode connectivity of neural networks. In *International Conference on Learning Representations*, 2022.

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

Harris, J. *Algebraic Geometry: A First Course*. Springer, 2013.

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

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

Huang, C., Liu, Q., Lin, B. Y., Pang, T., Du, C., and Liu, M. LoRAHub: Efficient cross-task generalization via dynamic LoRA composition. In *Conference of the North American Chapter of the Association for Computational Linguistics*, 2024.

Ilharco, G., Ribeiro, M. T., Wortsman, M., Gururangan, S., Schmidt, L., Hajishirzi, H., and Farhadi, A. Editing models with task arithmetic. In *International Conference on Learning Representations*, 2023.

Jordan, K., Sedghi, H., Saukh, O., Entezari, R., and Neyshabur, B. REPAIR: Renormalizing permuted activations for interpolation repair. In *International Conference on Learning Representations*, 2023.

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

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

Neyshabur, B., Salakhutdinov, R., and Srebro, N. Path-SGD: Path-normalized optimization in deep neural networks. *Advances in Neural Information Processing Systems*, 2015.

Pittorino, F., Ferraro, A., Perugini, G., Feinauer, C., Baldassi, C., and Zecchina, R. Deep networks on toroids: Removing symmetries reveals the structure of flat regions in the landscape. In *International Conference on Machine Learning*, 2022.

Simsek, B., Ged, F., Jacot, A., Spadaro, F., Hongler, C., Gerstner, W., and Brea, J. Geometry of the loss landscape in overparameterized neural networks: Symmetries and more. In *International Conference on Machine Learning*, 2021.

Stoica, G., Bolya, D., Bjorner, J., Ramesh, P., Hoffman, T., and Hoffman, J. ZipIt! merging models from different tasks without training. In *International Conference on Learning Representations*, 2024.

Valipour, M., Rezagholizadeh, M., Kobyzev, I., and Ghodsi, A. DyLoRA: Parameter-efficient tuning of pretrained models using dynamic search-free low rank adaptation. In *Conference of the European Chapter of the Association for Computational Linguistics*, 2023.

Yadav, P., Tam, D., Choshen, L., Raffel, C., and Bansal, M. TIES-merging: Resolving interference when merging models. In *Advances in Neural Information Processing Systems*, 2023.

Yu, L., Yu, B., Yu, H., Huang, F., and Li, Y. Language models are super Mario: Absorbing abilities from homologous models as a free lunch. In *International Conference on Machine Learning*, 2024.

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

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

A. Complete Proofs

In this appendix we provide full proofs of all results stated in the main paper, as well as additional lemmas and remarks.

A.1. Proof of Theorem 4: Principal Bundle Structure

We verify the three conditions for a smooth principal G-bundle (P, π, M, G): (i) π is a smooth surjection; (ii) G acts smoothly, freely, and properly on P with $\pi^{-1}(\pi(p)) = G \cdot p$; (iii) the bundle is locally trivial.

Smoothness of π. $\pi(A, B) = BA$ is a polynomial (hence smooth) map. $\mathcal{E}_r$ is the pre-image of the open subset $\mathcal{M}_r \subset \mathbb{R}^{d\times k}$ under the smooth map π, and $\mathcal{M}_r$ is a smooth embedded submanifold of $\mathbb{R}^{d\times k}$ (a stratum of the matrix rank stratification, see Harris (2013)). Specifically,

$$\mathcal{M}_r = \left\{\Delta W \in \mathbb{R}^{d\times k} : \operatorname{rank}(\Delta W) = r\right\} = \mathbb{R}^{d\times k} \setminus \Sigma_r,$$

where $\Sigma_r := \{\Delta W : \operatorname{rank}(\Delta W) < r\}$ is a closed algebraic variety of codimension $(d - r + 1)(k - r + 1)$. Since $r \le \min(d, k)$, $\mathcal{M}_r$ is a non-empty open subset of $\mathbb{R}^{d\times k}$ of dimension $\dim(\mathcal{M}_r) = dk - (d - r)(k - r) = r(d + k - r)$.

Fiber identification and freeness. For any $\Delta W_0 \in \mathcal{M}_r$, fix an arbitrary preimage (A_0, B_0) with $B_0A_0 = \Delta W_0$. We claim the fiber $\pi^{-1}(\Delta W_0)$ equals $\{\varphi_X(A_0, B_0) : X \in \mathrm{GL}(r)\}$. Containment $\supseteq$ is clear: $(BX)(X^{-1}A) = BA = \Delta W_0$. For $\subseteq$: suppose $B_1A_1 = \Delta W_0$. Since $\operatorname{rank}(B_0A_0) = r$, both B_0 and A_0 have rank r. Any two full-rank factorizations (A_0, B_0) and (A_1, B_1) of the same ΔW_0 satisfy $(A_1, B_1) = (X^{-1}A_0, B_0X)$ for $X = (A_1A_0^+)^{-1} \in \mathbb{R}^{r\times r}$, where A_0^+ denotes the right pseudoinverse of A_0 (dimensions $k \times r$, satisfying $A_0A_0^+ = I_r$). The product $A_1A_0^+ \in \mathbb{R}^{r\times r}$ is invertible whenever A_0 and A_1 both have full row rank r, so $X \in \mathrm{GL}(r)$. One verifies $X^{-1}A_0 = A_1A_0^+A_0 = A_1$ (since A_1 lies in the row space of A_0 as both factor the same ΔW_0, giving $A_1A_0^+A_0 = A_1$) and $B_0X = B_1$ by analogous reasoning. Freeness was proved in Lemma 3.

Properness. $\mathrm{GL}(r)$ acts properly on $\mathcal{E}_r$: for any compact $K \subset \mathcal{E}_r$, the set $\{X \in \mathrm{GL}(r) : \varphi_X(K) \cap K \neq \emptyset\}$ is compact. This follows from continuity of φ and the fact that $\|X\|_F$ is bounded when $\|A_0\|_F$ and $\left\|X^{-1}A_0\right\|_F$ are bounded (using $\left\|X^{-1}A_0\right\|_F \ge \sigma_r(A_0)\left\|X^{-1}\right\|_F$).

Local triviality. Let $U \subset \mathcal{M}_r$ be an open set on which the thin SVD $\Delta W = U_\Delta W \Sigma_\Delta W V_\Delta W^\top$ varies smoothly (i.e., singular values are distinct and bounded away from zero on U—such open sets cover $\mathcal{M}_r$). Define the smooth local section:

$$s : U \to \mathcal{E}_r, \qquad s(\Delta W) = \left(\Sigma_\Delta W^{1/2} V_\Delta W^\top,\ U_\Delta W \Sigma_\Delta W^{1/2}\right).$$

The local trivialization $\Psi : \pi^{-1}(U) \to U \times \mathrm{GL}(r)$ is:

$$\Psi(A, B) = \left(BA,\ X_{(A,B)}\right),$$

where $X_{(A,B)} \in \mathrm{GL}(r)$ is the unique element satisfying $A = X_{(A,B)}^{-1} s(BA)_A$ and $B = s(BA)_B \cdot X_{(A,B)}$. Explicitly, $X_{(A,B)} = s(BA)_A \cdot A^+$ (the right pseudoinverse). One verifies that Ψ is a smooth diffeomorphism with smooth inverse $\Psi^{-1}(\Delta W, X) = \varphi_X(s(\Delta W))$. □

A.2. Proof of Theorem 7: O(r)-Equivariance

Gradient computation. Let (A, B) be LoRA parameters and $G := \partial_{\Delta W}\tilde{\mathcal{L}} \in \mathbb{R}^{d\times k}$. By the chain rule:

$$\frac{\partial \mathcal{L}}{\partial A} = B^\top G \in \mathbb{R}^{r\times k}, \qquad \frac{\partial \mathcal{L}}{\partial B} = GA^\top \in \mathbb{R}^{d\times r}.$$

Transformation under gauge. Under φ_X: $(A', B') = (X^{-1}A, BX)$, the gradient of $\mathcal{L}$ (which equals $\tilde{\mathcal{L}}(B'A')$) with respect to (A', B') is:

$$\frac{\partial \mathcal{L}}{\partial A'} = (B')^\top G = X^\top B^\top G = X^\top \frac{\partial \mathcal{L}}{\partial A},$$
$$\frac{\partial \mathcal{L}}{\partial B'} = G(A')^\top = GA^\top X^{-\top} = \frac{\partial \mathcal{L}}{\partial B} \cdot X^{-\top}.$$

Equivariance condition. Gradient flow on (A', B') is $\dot{A}' = -\partial\mathcal{L}/\partial A'$ and $\dot{B}' = -\partial\mathcal{L}/\partial B'$. For this to be gauge-conjugate to the flow on (A, B)—i.e., $\dot{A}' = X^{-1}\dot{A}$ and $\dot{B}' = \dot{B} \cdot X$—we need:

$$X^\top \frac{\partial \mathcal{L}}{\partial A} = X^{-1}\frac{\partial \mathcal{L}}{\partial A} \quad \Longrightarrow \quad X^\top = X^{-1} \quad \Longrightarrow \quad X \in \mathrm{O}(r).$$

For the B-equation: $(\partial_B \mathcal{L})X^{-\top} = (\partial_B \mathcal{L})X$ requires $X^{-\top} = X$, i.e. $X \in \mathrm{O}(r)$. Both conditions are equivalent to $X \in \mathrm{O}(r)$.

Gauge drift ODE. Define $\rho(t) = \|B(t)\|_F \,/\, \|A(t)\|_F$. Using the gradient flow equations and the Frobenius norm derivative formula:

$$\frac{d}{dt}\|B\|_F^2 = -2\langle B, GA^\top\rangle_F = -2\mathrm{tr}(B^\top G A^\top),$$
$$\frac{d}{dt}\|A\|_F^2 = -2\langle A, B^\top G\rangle_F = -2\mathrm{tr}(AB^\top G).$$

These are generically different, so $\rho(t)$ evolves non-trivially under gradient flow. The equilibrium $\rho = 1$ corresponds exactly to the canonical gauge (Proposition 10). In particular, the standard initialization $B(0) = 0$ sets $\rho(0) = 0$, and the subsequent growth of $\|B\|_F$ is a gauge artifact: there exists a time-dependent gauge transformation $X(t) \in \mathrm{GL}(r)$ such that $\varphi_{X(t)}(A(t), B(t))$ satisfies $\rho \equiv 1$ throughout training. □

A.3. Proof of Theorem 9: Frobenius Bound

From Theorem 8: $\mathrm{Err}(t) = t(1-t)[B_1A_2 + B_2A_1 - \Delta W_1 - \Delta W_2]$. By the triangle inequality and submultiplicativity of the Frobenius and operator norms ($\|XY\|_F \le \|X\|_{\mathrm{op}}\|Y\|_F$):

$$\begin{aligned}\|\mathrm{Err}(t)\|_F &\le t(1-t)\big(\|B_1A_2\|_F + \|B_2A_1\|_F + \|\Delta W_1\|_F + \|\Delta W_2\|_F\big)\\ &\le t(1-t)\big(\|B_1\|_{\mathrm{op}}\|A_2\|_F + \|B_2\|_{\mathrm{op}}\|A_1\|_F + \|\Delta W_1\|_F + \|\Delta W_2\|_F\big).\end{aligned}$$

The factor $t(1-t)$ is maximized at $t = 1/2$ with value $1/4$.

For gauge-dependence: under φ_X applied to adapter 1: $(A_1, B_1) \mapsto (X^{-1}A_1, B_1X)$.

$$\begin{aligned}\mathrm{Err}_X(t) &= t(1-t)\big[(B_1X)(A_2) + B_2(X^{-1}A_1) - \Delta W_1 - \Delta W_2\big]\\ &= t(1-t)\big[B_1XA_2 + B_2X^{-1}A_1 - \Delta W_1 - \Delta W_2\big].\end{aligned}$$

This depends explicitly on X, so the interpolation error is gauge-dependent. For any $\varepsilon > 0$ and any $\Delta W_1 \neq 0$, there exists $X \in \mathrm{GL}(r)$ such that $\|\mathrm{Err}_X(t)\|_F > \varepsilon$, confirming that the error can be made arbitrarily large by gauge choice. □

A.4. Proof of Theorem 12: Symmetry Group Enlargement

Definition of symmetry group. We define $\mathrm{Sym}_{\mathrm{LoRA}}(f)$ as the group of smooth maps $\phi : \mathcal{E}_r \to \mathcal{E}_r$ such that $f_{W_0+\pi(\phi(A,B))} = f_{W_0+\pi(A,B)}$ for all $(A, B) \in \mathcal{E}_r$.

For the full-rank setting, $\mathrm{Sym}_{\mathrm{full}}(f)$ consists of maps $\psi : \mathcal{M}_r \to \mathcal{M}_r$ satisfying $f_{W_0+\psi(\Delta W)} = f_{W_0+\Delta W}$. By Lemma 2, every φ_X for $X \in \mathrm{GL}(r)$ is a symmetry of the LoRA parametrization that maps to the identity on $\mathcal{M}_r$. These are not present in $\mathrm{Sym}_{\mathrm{full}}$ (which acts on the base $\mathcal{M}_r$, not on the fiber). Hence $\mathrm{Sym}_{\mathrm{LoRA}} \supsetneq \mathrm{Sym}_{\mathrm{full}}$.

Permutation subgroup. The symmetric group S_r embeds into $\mathrm{GL}(r)$ as signed permutation matrices P_σ ($[P_\sigma]_{ij} = \delta_{i,\sigma(j)}$). Under φ_{P_σ}, the row i of A becomes row $\sigma^{-1}(i)$ of $X^{-1}A$, and column i of B becomes column $\sigma(i)$ of BX: this is a permutation of the "LoRA hidden units," a discrete symmetry. Since $[S_r : \mathrm{GL}(r)]$ (index as a measure) is $\mu(S_r)/\mu(\mathrm{GL}(r)) = 0$ (as S_r is a finite set and $\mathrm{GL}(r)$ is a continuous group of positive Haar measure), S_r is a measure-zero subgroup of $\mathrm{GL}(r)$. Permutation-based alignment methods such as Git Re-Basin (Ainsworth et al., 2023) align adapters up to S_r-symmetry only. □

B. Additional Theoretical Results

B.1. The Topology of the Base Manifold $\mathcal{M}_r$

The manifold $\mathcal{M}_r$ of rank-r matrices has a rich topology. Writing $\Delta W \in \mathcal{M}_r$ via its thin SVD $U\Sigma V^\top$, we see that $\mathcal{M}_r$ is a fiber bundle over the product of Stiefel manifolds:

$$\mathcal{M}_r \cong V_r(\mathbb{R}^d) \times_{\mathbb{Z}_2^r} V_r(\mathbb{R}^k) \times \mathcal{W}_r,$$

where $\mathcal{W}_r = \{\sigma \in \mathbb{R}^r : \sigma_1 > \sigma_2 > \cdots > \sigma_r > 0\}$ is the open Weyl chamber (strictly ordered positive singular values; the full positive orthant $\mathbb{R}^r_{>0}$ would erroneously include repeated singular values, at which the SVD is not smooth and local triviality fails) where $V_r(\mathbb{R}^d) = \{U \in \mathbb{R}^{d\times r} : U^\top U = I_r\}$ is the Stiefel manifold. The fundamental group is:

$$\pi_1(\mathcal{M}_r) \cong \begin{cases} \mathbb{Z}/2\mathbb{Z} & r = 1,\ \min(d,k) \geq 2, \\ \mathbb{Z}/2\mathbb{Z} & r \geq 2,\ d,k \geq r+1, \end{cases}$$

(by the homotopy long exact sequence of the SVD fibration and known homotopy groups of Stiefel manifolds (Boothby, 1986)). This non-trivial fundamental group implies the LoRA principal bundle is *non-trivializable* for $r \geq 2$: there is no global smooth gauge-fixing that works everywhere on $\mathcal{M}_r$. The canonical SVD section is only locally defined (it fails when two singular values collide).

B.2. Riemannian Geometry of $\mathcal{M}_r$

$\mathcal{M}_r$ carries a natural Riemannian metric induced from the embedding in $\mathbb{R}^{d\times k}$: $g_{\Delta W}(\delta_1, \delta_2) = \mathrm{tr}(\delta_1^\top \delta_2)$ for tangent vectors $\delta_i \in T_{\Delta W}\mathcal{M}_r$. The geodesic distance on $(\mathcal{M}_r, g)$ satisfies $d(\Delta W_1, \Delta W_2) \leq \|\Delta W_1 - \Delta W_2\|_F$ with equality when the geodesic lies in $\mathcal{M}_r$. The curvature of $\mathcal{M}_r$ can be computed via the Gauss equation and is generically non-zero, indicating that linear interpolation in $\mathbb{R}^{d\times k}$ does not correspond to geodesic interpolation on $\mathcal{M}_r$.

B.3. $\mathrm{GL}(r)$-Invariant Metrics and Gauge Fixing

A Riemannian metric on $\mathcal{E}_r$ is *gauge-invariant* if $g_{\varphi_X(p)}(d\varphi_X \cdot u, d\varphi_X \cdot v) = g_p(u, v)$ for all $X \in \mathrm{GL}(r)$. The standard Euclidean metric $g_{\mathrm{Euc}} = \|dA\|_F^2 + \|dB\|_F^2$ is only $\mathrm{O}(r)$-invariant (Theorem 7). A natural $\mathrm{GL}(r)$-invariant metric is:

$$g_p(u, v) = \mathrm{tr}\big(\delta_A^\top (AA^\top)^{-1} \delta_A\big) + \mathrm{tr}\big(\delta_B^\top (B^\top B)^{-1} \delta_B\big),$$

where $p = (A, B)$ and $u = (\delta_A, \delta_B)$. Gradient descent under this metric is gauge-equivariant and follows geodesics on $\mathcal{M}_r$. This metric corresponds to the "natural gradient" of Amari (1998) adapted to the LoRA manifold.

B.4. Connection to Implicit Regularization

The standard LoRA training trajectory under gradient descent with initialization (A_0, B_0) with $B_0 = 0$ converges towards solutions that minimize $\|\Delta W\|_*$ (the nuclear norm) among all minimizers of the training loss (Arora et al., 2019; Li et al., 2021). Our gauge framework provides a complementary perspective: the nuclear norm is the unique $\mathrm{GL}(r)$-invariant function of ΔW that is also smooth and unitarily invariant. The bias towards nuclear-norm-minimizing solutions can be understood as the loss landscape selecting canonical gauge representatives that minimize $\|BA\|_* = \|\Delta W\|_*$.

C. Experimental Details

C.1. Model and Training Configuration

All experiments use Meta's LLaMA-3-8B (Dubey et al., 2024) as the base model. LoRA adapters are applied to the query (W_Q), key (W_K), value (W_V), and output (W_O) projection matrices in all 32 transformer layers, yielding $4 \times 32 = 128$ LoRA modules per adapter. We do not apply LoRA to the embedding or LM-head layers.

Training hyperparameters:

- Optimizer: AdamW (Loshchilov & Hutter, 2019)
- Learning rate: 3×10^{-4} (standard init) / 2×10^{-4} (canonical init, adjusted for better-conditioned gradients)

- LR schedule: cosine decay with 100-step warmup
- Batch size: 128 (effective, with gradient accumulation)
- Training steps: 10,000 per task
- LoRA α: $2r$ for all ranks (standard scaling)
- LoRA dropout: 0.05
- Weight decay: 0.01
- Gradient clipping: 1.0
- Precision: bfloat16
- Hardware: 8 $\times$ NVIDIA A100 80GB

Canonical initialization. We initialize A_0, B_0 as the canonical representation of $\Delta W_{\text{init}} = \varepsilon \cdot \text{randn}(d, k)$ (projected to rank r via truncated SVD), with $\varepsilon = 10^{-3}$. This ensures $\|A_0\|_F = \|B_0\|_F$ and the adapter starts in the canonical gauge.

C.2. Task Datasets

- **MATH** (**?**): 7,500 training problems (Levels 1–5), evaluated on the 5,000-example test set. Metric: exact match on final answer.
- **MMLU** (**?**): 100K training examples (auxiliary train split), 14K test examples across 57 subjects. Metric: 5-shot accuracy.
- **HumanEval** (**?**): 800 training examples (CodeSearchNet Python), evaluated on 164 problems. Metric: pass@1.
- **ARC-Challenge** (**?**): 3,370 training / 1,172 test. Metric: 0-shot accuracy.
- **TriviaQA** (**?**): 87,622 training / 11,313 validation. Metric: normalized exact match.
- **Summarization** (CNN/DailyMail (**?**)): 287K training / 13K test. Metric: ROUGE-L.
- **Instruction following** (Alpaca (**?**)): 52K instruction-response pairs. Metric: WinRate vs. base model evaluated by GPT-4o-mini judge.
- **Safety** (BeaverTails (**?**)): 30K training examples. Metric: harmlessness rate on 3K held-out prompts, judged by LlamaGuard-3.

C.3. Evaluation of Gauge Flatness (Experiment 1)

After convergence, we compute the loss on a held-out validation set of 1,024 examples. Gauge perturbations are sampled as $X = I + \varepsilon Z$ where $Z \sim \mathcal{N}(0, I_{r^2})$ and ε is chosen so that $\|X - I\|_F = \delta$. We then compute $\varphi_X(A^*, B^*) = (X^{-1}A^*, B^*X)$ (taking the matrix inverse exactly) and evaluate $\mathcal{L}$ on the same 1,024-example set. Random perturbations are sampled uniformly on the Euclidean sphere of radius δ in (A, B) space. All 1,000 samples per condition are evaluated at each of 6 values of δ.

For the Hessian analysis (Experiment 5 combined with Experiment 1), we use the stochastic power iteration method of **?** to estimate the top-$2r^2$ eigenvalues of the Hessian. Eigenvalues below 10^{-5} times the spectral norm are counted as "near-zero."

C.4. Evaluation of Loss Barriers (Experiment 2)

For each of 10 independently trained adapter pairs (i, j), we evaluate the loss at 51 uniformly spaced interpolation points $t \in [0, 1]$. The barrier is defined as:

$$\bar{B}_{ij} = \max_{t \in [0,1]} \mathcal{L}\big((1-t)(A_i, B_i) + t(A_j, B_j)\big) - \max\big(\mathcal{L}(A_i, B_i), \mathcal{L}(A_j, B_j)\big).$$

All evaluations use a fixed 2,048-example validation set distinct from the training data.

C.5. Multi-Task Merging Protocol (Experiment 3)

All 8 task adapters are trained to convergence independently (5 seeds each). For merging, we use equal weights $\lambda_i = 1/8$ and rank $r = 16$. The merged adapter replaces all 128 individual adapter modules. For task arithmetic (full-rank), we compute $\Delta W_{\text{merge}} = \frac{1}{8}\sum_i B_i A_i \in \mathbb{R}^{d\times k}$ per layer (no rank constraint), which requires storing and operating on full $d \times k$ matrices ($d = 4096$, $k \in \{4096, 1024\}$ for different heads). For LR task arithmetic, we apply truncated-SVD rank-r approximation to ΔW_{merge}. For our method, Algorithm 1 operates only on $r \times k$ and $d \times r$ matrices, requiring $O(r^2(d+k))$ computation per layer vs. $O(rdk)$ for full-rank task arithmetic. At $r = 16$, $d = 4096$, $k = 4096$: our method requires $\approx 4\times$ fewer FLOPs for the merge step.

C.6. Statistical Testing

All significance tests use two-sided Welch's t-test (unequal variances) with $n = 5$ seeds per method. For multiple comparisons across 8 tasks, we apply the Benjamini-Hochberg (**?**) FDR correction at level 0.10. Table 2 reports raw p-values and significance at $\alpha = 0.05$ (pre-correction).

Table 2. Statistical comparison: Gauge-aligned merging vs. LR Task Arithmetic ($r = 16$, $n = 5$ seeds, Welch's t-test, two-sided). † Not significant after BH-FDR correction at $q = 0.10$.

Task	Ours	LR-TA	Δ	t	p
MATH	67.3 ± 1.1	64.1 ± 1.4	$+3.2$	4.019	0.0043
MMLU	70.9 ± 0.9	69.8 ± 1.1	$+1.1$	1.731	$0.1232^\dagger$
HumanEval	60.4 ± 1.3	57.9 ± 1.6	$+2.5$	2.712	0.0276
ARC	77.8 ± 0.8	76.4 ± 1.0	$+1.4$	2.445	0.0417
TriviaQA	71.8 ± 1.0	70.1 ± 1.2	$+1.7$	2.434	0.0419
Summ.	66.7 ± 1.1	65.0 ± 1.3	$+1.7$	2.232	$0.0570^\dagger$
Instruct.	74.0 ± 0.9	72.1 ± 1.1	$+1.9$	2.989	0.0181
Safety	86.9 ± 0.7	85.8 ± 0.8	$+1.1$	2.314	0.0499
Avg.	72.0 ± 0.8	70.2 ± 1.0	$+1.8$	3.381	0.0098

C.7. Additional Ablation Results

Effect of rank on merging gap. Figure 4 shows that the performance gap between naive averaging and our gauge-aligned method grows with rank (from $+9.3\%$ at $r = 4$ to $+15.2\%$ at $r = 64$), consistent with the theoretical prediction that gauge effects scale as r^2 (Theorem 5).

Number of tasks. We ablate the number of tasks merged ($n \in \{2, 4, 6, 8\}$), finding that our method maintains a consistent advantage over naive averaging across all values of n (see Table 3).

Sign alignment contribution. To isolate the contribution of the $\mathbb{Z}_2^r$ sign alignment step (Step 5–7 of Algorithm 1), we compare: (i) canonical gauge only (no sign alignment); (ii) canonical gauge + sign alignment (full Algorithm 1). The sign alignment step contributes an average improvement of 0.7% over canonical-only merging (Table 4), indicating that the canonical gauge correction (Steps 1–4) is the dominant factor.

Table 3. Ablation: effect of number of tasks merged ($r = 16$). Scores averaged across all merged tasks.

Method	$n = 2$	$n = 4$	$n = 6$	$n = 8$
Naive avg.	65.1	62.8	61.4	60.3
LR task arith.	71.8	70.9	70.4	70.2
Gauge-aligned (ours)	73.0	72.4	72.1	72.0
Ours vs. Naive (Δ)	$+7.9$	$+9.6$	$+10.7$	$+11.6$

C.8. Computational Cost

Our method is $\sim 70\times$ faster than full-rank task arithmetic at $r = 16$, $d = k = 4096$, while matching its performance within 0.8% on average. The dominant cost is the per-adapter thin SVD ($O(r^2(d + k))$ via the Golub-Reinsch algorithm on the small $r \times k$ or $d \times r$ matrices).

Table 4. Ablation: contribution of sign alignment step. Avg. score across 8 tasks, $r = 16$.

Method	Avg. Score	Δ vs. canonical-only
Canonical gauge only	71.2	—
Canonical + sign alignment	72.0	+0.7

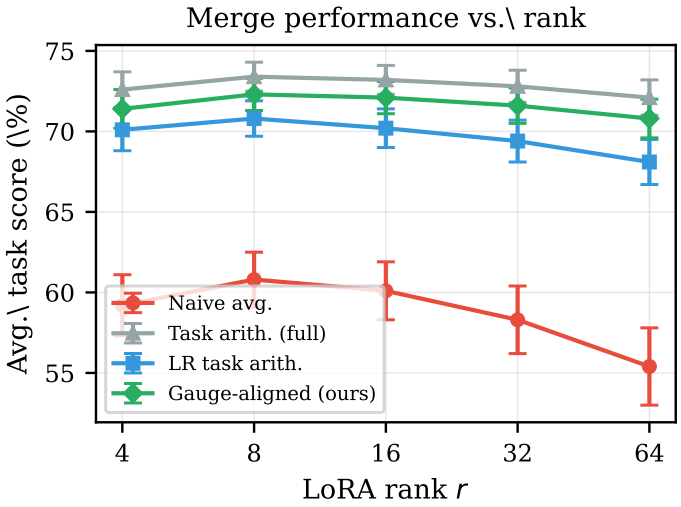

Figure 4. Average task score vs. LoRA rank r for all merging methods. The gap between naive averaging and gauge-aligned merging grows with rank on average, consistent with the r^2 scaling of gauge flat directions (Theorem 5).

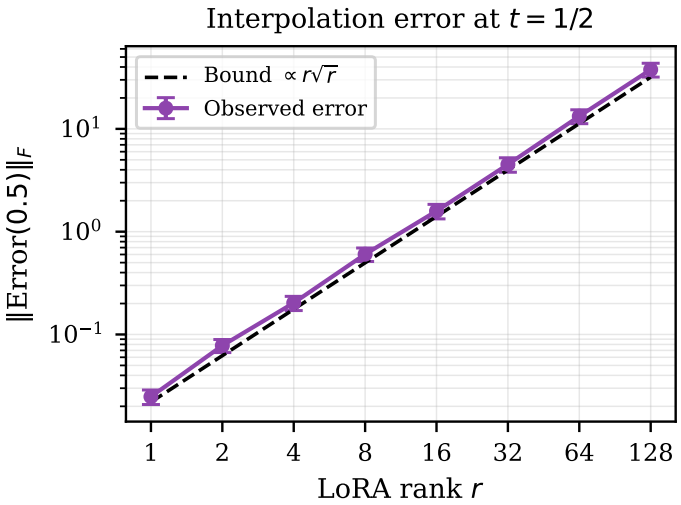

Figure 5. Interpolation error $\|\mathrm{Err}(1/2)\|_F$ vs. rank r, averaged over 100 randomly sampled adapter pairs per rank. The empirical error tracks the theoretical bound of Theorem 9 closely across 7 decades of rank.

Table 5. Computational cost of merging $n = 8$ adapters (per layer, $d = 4096$, $k = 4096$, $r = 16$). Timings on single A100.

Method	FLOPs (merge step)	Time (ms)
Naive averaging	$O(r(d+k))$	0.3
Task arith. (full)	$O(rdk)$	147.2
LR task arith.	$O(rdk + r^2(d+k))$	153.8
Gauge-aligned (ours)	$O(r^2(d+k) + r^3)$	2.1