Generalizing the Geometry of Model Merging Through Fréchet Averages

Marvin F. da Silva[1,2*] **Mohammed Adnan**[2,3] **Felix Dangel**[4,5,6,†] **Sageev Oore**[1,2,†]
[1]Faculty of Computer Science, Dalhousie University
[2]Vector Institute for Artificial Intelligence
[3] Schulich School of Engineering, University of Calgary
[4] Department of Computer Science and Software Engineering, Concordia University
[5] Mila – Quebec AI Institute [6] Canada CIFAR AI Chair
[†]Equal-contribution senior authors.

Abstract

Model merging aims to combine multiple models into one without additional training. Naïve parameter-space averaging can be fragile under architectural symmetries, as their geometry does not take them into account. In this work, we argue that not only the geometry but also the averaging procedure itself must be symmetry-invariant to achieve symmetry-aware merges. Consequently, we propose a general solution: merging as Fréchet averaging, i.e. selecting parameters that minimize a sum of geodesic distances on an appropriate manifold. In this view, the key design choice is the overall geometry, i.e., the choice of metric, manifold, and distance approximation, that determines what it means for two models to be "close." We show that Fréchet averaging, combined with simplifying assumptions, contains Fisher merging. Building on this, we examine the particular case of low-rank adapters (LoRA), whose symmetries induce a distinct geometry: that of a quotient manifold. We outline limitations of current LoRA merging methods, propose a practical algorithm for this setting, and support the effectiveness of our method with empirical results.

1 Introduction

Model merging aims to combine multiple trained networks into a single model that preserves the union of their capabilities while avoiding additional end-to-end training. This objective is increasingly salient because modern workflows routinely produce *families* of related models: different random seeds and hyperparameters, domain- or task-specialists, safety patches, preference updates, and personalized variants. Existing approaches broadly fall into two classes: *(1) Data-dependent* methods require data and/or gradients at merge time; examples include Fisher-weighted merging [33], RegMean [24], and MaTS [41], which are strong when data is available but less applicable when it is proprietary or costly to process. *(2) Data-free* (weight-only) methods operate purely in parameter space and remain deployable with only model checkpoints; this includes model soups [45], Task Arithmetic [23], and TIES [47]. We focus on this class throughout this work.

Misalignment complicates adapter merging. When models are full-rank fine-tuned from the same pretrained initialization, these operations can yield strong merged performance without further training [45, 23, 33, 47]. *However, this success does not reliably transfer to models produced via parameter-efficient fine-tuning (PEFT).* Low-Rank Adaptation (LoRA) [21] constructs specialists by learning low-rank updates to selected matrices, yet weight-only merge rules that behave benignly for full-rank fine-tunes can degrade sharply when applied to LoRA specialists, even when specialists share the same base model [40]. Stoica et al. [40] propose that this gap is fundamentally about *alignment of task updates*. Comparing pairwise centered kernel alignment (CKA) of representations

*Corresponding Author. marvinf.silva@dal.ca

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

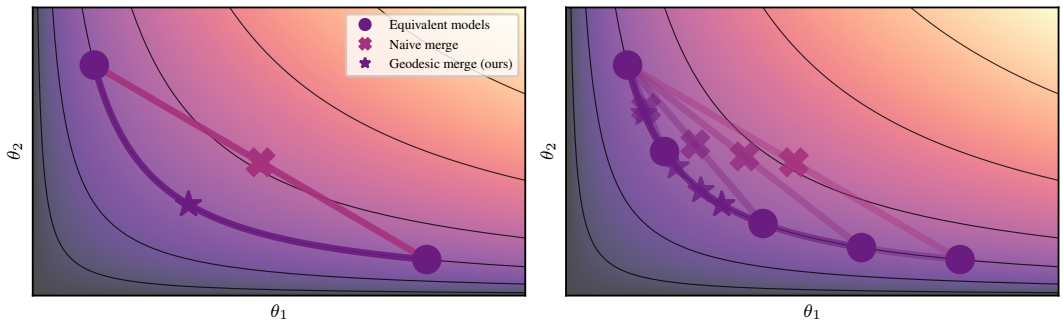


Figure 1: Visual comparison of different model merging approaches, highlighting failure scenarios due to symmetry unawareness. *Left:* Naive averaging steps off the orbit because it uses the wrong geometry. *Right:* Naive merging is ambiguous and lands on different orbits depending on parameterization. Geodesic merging always stays on the same orbit as it uses symmetry-invariant averaging.

attributable to fine-tuning updates alone, they find that full-rank fine-tuned models exhibit high CKA, while LoRA counterparts show substantially lower CKA, suggesting different LoRA task-updates process inputs through *misaligned* subspaces. This misalignment correlates with destructive interference under naive addition or averaging and motivates an explicit alignment stage: express task-updates in a shared basis (e.g., via an SVD construction), then apply merge operators (e.g., Task Arithmetic or TIES) in the aligned coordinate system. Alignment, however, raises a deeper question:

What exactly are we trying to align? Neural network parameters are not canonical coordinates for functions, and modern architectures have large symmetry groups under which distinct parameter settings implement exactly the same input-output map. Classic examples include neuron permutations [18] and rescaling symmetries in positively homogeneous networks [9]. Transformers exhibit richer continuous symmetries: attention blocks admit high-dimensional $\mathrm{GL}(h)$ symmetries that generate directions along which the network function (and loss) is unchanged [8, 48]. These symmetries can obscure both alignment analyses and procedures. For low-rank updates, the usual factorization introduces a $\mathrm{GL}(r)$ symmetry $(G, H) \sim (GA^{-1}, HA^{\top})$ that preserves the product $GH^{\top}$, so apparent misalignment can be partly a coordinate artifact. Wang et al. [43] found that applying one of these $\mathrm{GL}(r)$ transformations, the performance of merged models collapses. This suggests that robust alignment requires distances and averages intrinsic to the space of functions (or updates), not to an arbitrary parameterization.

In summary. Figure 1 illustrates the core geometric failure mode behind many weight-only merges under symmetry: when multiple parameter settings represent the same underlying model, Euclidean averaging can (i) step off the orbit because it measures "closeness" in the wrong geometry, and (ii) become ambiguous because different but equivalent parameterizations yield different merges.

Our solution: averaging on Riemannian manifolds. A principled way to do this averaging is to treat the parameter space as a *quotient* by the relevant symmetry group and define geometric notions intrinsically on that quotient. da Silva et al. [8] develop this viewpoint for transformers, arguing that symmetry can invalidate Euclidean geometric measures and proposing symmetry-corrected objects on the quotient manifold, mapped back to parameter space via horizontal lifts.

We ask whether the geometric methodology can be elevated to a *constructive* principle for merging:

> Can we formulate model merging as averaging on a Riemannian manifold whose geometry respects parameter symmetries to eliminate symmetry-related failures?

We develop **GeoMerge**, a geometric framework that treats merging as computing a *Fréchet mean* in a chosen Riemannian representation space. In this view, a merge method amounts to geometric choices: (a) an appropriate parameter manifold; (b) an appropriate metric; and (c) any equivalence on the parameters we wish to choose. Weight-only merging uses representations and metrics computable without data, while data-dependent merging uses metrics estimated from activations, gradients, or curvature, linking merging to information geometry. Our contributions are:

- **Sections 2 and 3:** We cast model merging as Fréchet averaging on a Riemannian manifold and argue that, under architectural symmetries, the correct objects live on a quotient manifold, yielding invariance with respect to gauge choices [37, 4].
- **Section 4**: We connect importance-weighted merging, specifically Fisher merging [33], to Fréchet objectives under information-geometric metrics, highlighting their symmetry non-invariance despite employing a refined geometry, due to their symmetry-unaware average.
- **Sections 5 and 6:** Motivated by GL symmetries in low-rank updates [LoRA, 21] and attention [8], we develop quotient-compatible primitives that yield symmetry-corrected merging algorithms for LoRA factors. We provide a quotient-aware lifting scheme for embedding low rank adapters into a higher dimensional space to assist in merging. We illustrate geodesic merging analytically for a small toy model (Section A) and on two real adapters (Section F), and as proof of concept, we scale the computational algorithm to LoRA adapters for larger models (ViT-B/32 and Llama 3 8B in Section 6).

1.1 Other Related Work

Weight averaging and basin structure. Model soups [45] show that averaging fine-tuned checkpoints can improve accuracy without inference-time overhead. A common explanation invokes flatness and (approximate) linear mode connectivity [15, 11], which makes Euclidean averaging benign when solutions live in one convex-like region. GeoMerge does not contradict this; rather, it asks what happens when Euclidean geometry is *not* the right notion of closeness—a question that becomes especially pressing in the presence of symmetries.

Symmetries, quotient geometry, and deep learning. Weight-space symmetries are known to invalidate naive Euclidean notions of sharpness and flatness [9]. da Silva et al. [8] study the high-dimensional $\mathrm{GL}(h)$ symmetries in transformers' attention heads, and develop symmetry-corrected sharpness measures on quotient manifolds [4]. We extend their recipe by studying Riemannian averaging operators and target *model merging* rather than sharpness.

KnOTS and LoRA. KnOTS [40] improves LoRA merging via joint SVD-based transformations to better align low-rank updates before applying existing merge rules. GeoMerge complements this; it lets us see SVD-based alignment as a particular gauge fixing for $\mathrm{GL}(r)$ symmetries of low-rank factorizations, selecting comparable representatives of the same underlying updates before averaging.

2 Preliminary Definitions, Notation & Math

We consider T task fine-tuned adapters $\{\theta_t\}_{t=1}^{T}$ derived from a common base model. We will define a geometric merge on a (possibly quotient) parameter manifold $\mathcal{M}$.

2.1 General Remarks

Riemannian manifolds, distance, and Exp/Log. A *Riemannian manifold* is a pair $(\mathcal{M}, g)$, where $\mathcal{M}$ is a smooth manifold and the metric g assigns to each $p \in \mathcal{M}$ an inner product $g_p : T_p\mathcal{M} \times T_p\mathcal{M} \to \mathbb{R}$ on the tangent space $T_p\mathcal{M}$ at p. For the tangent vectors ξ, ζ we write $\langle \xi, \zeta \rangle_p := g_p(\xi, \zeta)$ and $\|\xi\|_p := \sqrt{\langle \xi, \xi \rangle_p}$. A *geodesic* is a curve $\gamma : [0, 1] \to \mathcal{M}$ that locally minimizes length. The *geodesic distance* between two points $p, q \in \mathcal{M}$ is

$$d_{\mathcal{M}}(p, q) \;:=\; \inf_{\gamma(0)=p,\, \gamma(1)=q} \int_0^1 \|\dot\gamma(t)\|_{\gamma(t)}\, dt.$$

The *exponential map* $\mathrm{Exp}_p : T_p\mathcal{M} \to \mathcal{M}$ maps a tangent vector ξ to the endpoint of the geodesic starting at p with initial velocity ξ: $\mathrm{Exp}_p(\xi) := \gamma_\xi(1)$ where $\gamma_\xi(0) = p, \dot\gamma_\xi(0) = \xi$. Its (local) inverse is the *logarithm map* $\mathrm{Log}_p : \mathcal{M} \to T_p\mathcal{M}$ satisfying $\mathrm{Exp}_p(\mathrm{Log}_p(q)) = q$. These mappings are later used to obtain the Fréchet mean.

Quotients induced by symmetries. Neural network parameterizations often admit symmetries: many distinct parameter settings encode the same function. A principled way to enforce symmetry-invariance is to work on a *quotient manifold*. Let $\overline{\mathcal{M}}$ be a (total) manifold with a smooth right action of a Lie group $\mathcal{G}$ so that $\Psi : \overline{\mathcal{M}} \times \mathcal{G} \to \overline{\mathcal{M}}$ and $(\bar p, G) \mapsto \bar p \cdot G$. This action induces an equivalence

relation $\bar{p} \sim \bar{q}$ if $\bar{q} = \bar{p} \cdot G$ for some $G \in \mathcal{G}$. The quotient space is $\mathcal{M} := \overline{\mathcal{M}}/\mathcal{G}$; we write $[\bar{p}] \in \mathcal{M}$ for the orbit and $\pi : \overline{\mathcal{M}} \to \mathcal{M}$ for the projection $\pi(\bar{p}) = [\bar{p}]$.

Vertical and horizontal spaces. At a point $\bar{p} \in \overline{\mathcal{M}}$, the *vertical space* $V_{\bar{p}}\overline{\mathcal{M}} \subset T_{\bar{p}}\overline{\mathcal{M}}$ is the tangent space to the orbit through $\bar{p}$, i.e. the set of infinitesimal symmetry directions. If $\overline{\mathcal{M}}$ has a Riemannian metric $\bar{g}$ under which the action is by isometries (metric is symmetry invariant), then we define the *horizontal space* as the orthogonal complement w.r.t the Riemannian metric $\bar{g}$

$$H_{\bar{p}}\overline{\mathcal{M}} \;:=\; \big(V_{\bar{p}}\overline{\mathcal{M}}\big)^{\perp} \subset T_{\bar{p}}\overline{\mathcal{M}}.$$

Tangent vectors in the quotient space are set-valued and therefore inconvenient for numerical manipulation. Conveniently, horizontal vectors provide representatives of tangent vectors on the quotient space: each $\xi \in T_{[\bar{p}]}\mathcal{M}$ has a unique *horizontal lift* $\bar{\xi} \in H_{\bar{p}}\overline{\mathcal{M}}$ with $d\pi_{\bar{p}}(\bar{\xi}) = \xi$ [1, 4].

Quotient distance as orbit alignment. If $\bar{g}$ is $\mathcal{G}$-invariant, it induces a well-defined metric g on the quotient. The resulting geodesic distance on $\mathcal{M}$ can be written as an *orbit-minimum*:

$$d_{\mathcal{M}}\big([\bar{p}], [\bar{q}]\big) \;=\; \min_{G \in \mathcal{G}} d_{\bar{\mathcal{M}}}\big(\bar{p}, \bar{q} \cdot G\big). \tag{1}$$

We will refer to a minimizer (a symmetry transformation that leads to the smallest geodesic distance)

$$G^{\star}(\bar{p}; \bar{q}) \in \arg\min_{G \in \mathcal{G}} d_{\bar{\mathcal{M}}}\big(\bar{p}, \bar{q} \cdot G\big)^2 \tag{2}$$

as an *alignment* (or *gauge choice*) of $\bar{q}$ to $\bar{p}$, and write the aligned representative as $\bar{q}^{\star} := \bar{q} \cdot G^{\star}(\bar{p}; \bar{q})$. Even when $\bar{q}$ and $\bar{q}^{\star}$ are different points in $\overline{\mathcal{M}}$, they represent the same quotient point: $[\bar{q}] = [\bar{q}^{\star}]$. Alignment implies horizontal geodesics: whenever $\bar{p}$ and $\bar{q}$ are aligned, $\mathrm{Log}_{\bar{p}}(\bar{q})$ is horizontal (e.g., Huckemann et al. [22]), that is, the points are connected by a horizontal geodesic that is always perpendicular to orbits, i.e. $\dot{\gamma}(t) \in H_{\gamma(t)}\mathcal{M}$. This will be a representation of the geodesic on the quotient space (which is an abstract space) that we can work with numerically.

2.2 Low-rank updates and their symmetries.

A common PEFT primitive is a rank-r update to a weight matrix. We consider updates of the form

$$\Delta \boldsymbol{W} \in \mathbb{R}^{d_1 \times d_2}, \quad \mathrm{rank}(\Delta \boldsymbol{W}) = r\,,$$

that admit several factorizations. We consider the following [21, 34]:

(A) Standard factorization ($\mathrm{GL}(r)$ symmetry). $\Delta \boldsymbol{W}$ is usually trained and stored as a factorization $\Delta \boldsymbol{W} = \boldsymbol{G}\boldsymbol{H}^{\top}$ with $\boldsymbol{G} \in \mathbb{R}^{d_1 \times r}_{\star}$ and $\boldsymbol{H} \in \mathbb{R}^{d_2 \times r}_{\star}$ ($_{\star}$ indicating full column rank). Then, for any $\boldsymbol{A} \in \mathrm{GL}(r)$, $(\boldsymbol{G}\boldsymbol{A}^{-1})(\boldsymbol{H}\boldsymbol{A}^{\top})^{\top} = \boldsymbol{G}\boldsymbol{H}^{\top}$ and we define the equivalence relationship

$$(\boldsymbol{G}, \boldsymbol{H}) \sim (\boldsymbol{G}\boldsymbol{A}^{-1},\, \boldsymbol{H}\boldsymbol{A}^{\top}).$$

(B) Polar factorization ($\mathrm{O}(r)$ symmetry). Following Mishra et al. [34], a rank-r matrix also admits the three-factor factorization

$$\Delta \boldsymbol{W} = \boldsymbol{U}\boldsymbol{B}\boldsymbol{V}^{\top}, \tag{3}$$

where $\boldsymbol{U} \in \mathrm{St}(d_1, r)$, $\boldsymbol{V} \in \mathrm{St}(d_2, r)$ and $\boldsymbol{B} \in \mathbb{S}_{++}(r)$. Here $\mathrm{St}(d, r) = \{\boldsymbol{U} \in \mathbb{R}^{d \times r} : \boldsymbol{U}^{\top}\boldsymbol{U} = \boldsymbol{I}_r\}$ is the Stiefel manifold and $\mathbb{S}_{++}(r)$ is the manifold of $r \times r$ symmetric positive-definite (SPD) matrices. We follow Mishra et al. [34] in calling (3) the *polar factorization*; it differs from the classical polar decomposition, which has two factors, one orthogonal and one positive semidefinite [20], and from the SVD, whose middle factor is diagonal rather than SPD. This representation has the (smaller) orthogonal symmetry group: for any $\boldsymbol{O} \in \mathrm{O}(r)$,

$$(\boldsymbol{U}, \boldsymbol{B}, \boldsymbol{V}) \sim (\boldsymbol{U}\boldsymbol{O},\, \boldsymbol{O}^{\top}\boldsymbol{B}\boldsymbol{O},\, \boldsymbol{V}\boldsymbol{O}),$$

since $(\boldsymbol{U}\boldsymbol{O})(\boldsymbol{O}^{\top}\boldsymbol{B}\boldsymbol{O})(\boldsymbol{V}\boldsymbol{O})^{\top} = \boldsymbol{U}\boldsymbol{B}\boldsymbol{V}^{\top}$. This $\mathrm{O}(r)$-quotient viewpoint is particularly convenient computationally because geometric primitives like Exp, and Log either admit exact expressions or accurate numerical routines [12, 32]. Section C shows how we compute $(\boldsymbol{U}, \boldsymbol{B}, \boldsymbol{V})$ from the trained factors $(\boldsymbol{G}, \boldsymbol{H})$.

The quotient geometry of the polar factorization. We follow the construction from Mishra et al. [34]. On $\mathbb{S}_{++}$ we will use the affine-invariant metric:

$$g_{\boldsymbol{B}}(\boldsymbol{\zeta}_{\boldsymbol{B}}, \boldsymbol{\eta}_{\boldsymbol{B}}) = \langle \boldsymbol{\zeta}_{\boldsymbol{B}}, \boldsymbol{\eta}_{\boldsymbol{B}} \rangle_{\boldsymbol{B}} \;:=\; \mathrm{Tr}\big(\boldsymbol{B}^{-1}\boldsymbol{\zeta}_{\boldsymbol{B}}\boldsymbol{B}^{-1}\boldsymbol{\eta}_{\boldsymbol{B}}\big)$$

where $\boldsymbol{\eta}, \boldsymbol{\zeta} \in T_{\boldsymbol{B}}\mathbb{S}_{++} = \{\boldsymbol{\eta} = \boldsymbol{\eta}^\top\}$ and with closed-form Riemannian exponential and logarithm:

$$\mathrm{Exp}_{\boldsymbol{B}}(\boldsymbol{\eta}) = \boldsymbol{B}^{1/2} \exp\big(\boldsymbol{B}^{-1/2}\boldsymbol{\eta}\boldsymbol{B}^{-1/2}\big)\boldsymbol{B}^{1/2}\,, \tag{4a}$$

$$\mathrm{Log}_{\boldsymbol{B}}(C) = \boldsymbol{B}^{1/2} \log\big(\boldsymbol{B}^{-1/2}C\boldsymbol{B}^{-1/2}\big)\boldsymbol{B}^{1/2}\,, \tag{4b}$$

where $\exp_m(\cdot)$ denotes the matrix exponential. For the Stiefel factors, we deviate from Mishra et al. [34] and use the canonical Stiefel metric

$$\langle \boldsymbol{\zeta}_{\boldsymbol{U}}, \boldsymbol{\eta}_{\boldsymbol{U}} \rangle_{\boldsymbol{U}} = \mathrm{Tr}\left(\boldsymbol{\zeta}_{\boldsymbol{U}}^\top (\boldsymbol{I} - \frac{1}{2}\boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{\eta}_{\boldsymbol{U}}\right)$$

which has a better-behaved numerical routine for calculating its Log (using the algorithm from Mataigne et al. [32]). Edelman et al. [12] derived the exponential map: Given a point $\boldsymbol{U} \in \mathrm{St}(n, r)$ and a tangent vector $\boldsymbol{\eta}_{\boldsymbol{U}} \in T_{\boldsymbol{U}}\mathrm{St}(n, r)$, first create the compact QR factorization $(\boldsymbol{I} - \boldsymbol{U}\boldsymbol{U}^\top)\boldsymbol{\eta} = \boldsymbol{Q}\boldsymbol{R}$, with $\boldsymbol{Q} \in \mathrm{St}(n, r)$, $R \in \mathbb{R}^{r\times r}$. Then, $\boldsymbol{A} := \boldsymbol{U}^\top\boldsymbol{\eta}$ is skew-symmetric, i.e., $\boldsymbol{A}^\top = -\boldsymbol{A}$. With these definitions, the exponential mapping is

$$\mathrm{Exp}_{\boldsymbol{U}}(\boldsymbol{\eta}) = \begin{pmatrix}\boldsymbol{U} & \boldsymbol{Q}\end{pmatrix} \exp\left(\begin{pmatrix}\boldsymbol{A} & -\boldsymbol{R}^\top \\ \boldsymbol{R} & 0\end{pmatrix}\right)\begin{pmatrix}\boldsymbol{I}_p \\ 0\end{pmatrix} \tag{5}$$

and becomes particularly efficient when $r < n/2$ which is always the case for the LoRA adapters we will study—typically $(n = 4096) \times (r = 16)$ matrices.

Because we use the canonical rather than the Euclidean Stiefel metric, the horizontal space and the horizontal projection differ from those of Mishra et al. [34]; Section C.1 states which of their results carry over and derives the projection

$$\eta^{\mathrm{hor}} = (\eta_{\boldsymbol{U}} - \boldsymbol{U}\boldsymbol{\Omega},\ \eta_{\boldsymbol{B}} - (\boldsymbol{B}\boldsymbol{\Omega} - \boldsymbol{\Omega}\boldsymbol{B}),\ \eta_{\boldsymbol{V}} - \boldsymbol{V}\boldsymbol{\Omega}),$$

where the skew-symmetric $\boldsymbol{\Omega}$ is the unique solution of

$$\boldsymbol{B}^{-1}\boldsymbol{\Omega}\boldsymbol{B} + \boldsymbol{B}\boldsymbol{\Omega}\boldsymbol{B}^{-1} - \boldsymbol{\Omega} = \frac{1}{2}(\boldsymbol{V}^\top\eta_{\boldsymbol{V}} + \boldsymbol{U}^\top\eta_{\boldsymbol{U}}) - (\boldsymbol{B}^{-1}\eta_{\boldsymbol{B}} - \eta_{\boldsymbol{B}}\boldsymbol{B}^{-1})\,,$$

which has a closed form in the eigenbasis of $\boldsymbol{B}$ (Section C.1).

3 GeoMerge

We formulate model merging as a (possibly quotient) Fréchet mean on a Riemannian manifold and approximate it in practice by alternating orbit alignment with retraction-based geodesic updates. We work through an analytically solvable toy model of this section in Section A.

3.1 Fréchet averaging

The central object in our framework is the *Fréchet mean*: given points $x_1, \ldots, x_T$ on a metric space (or Riemannian manifold) $(\mathcal{M}, d)$ and weights $w_i \geq 0$ with $\sum_i w_i = 1$, a Fréchet mean is any minimizer of the *Fréchet functional*

$$\mu^\star \in \arg\min_{\mu\in\mathcal{M}} F_\mu = \arg\min_{\mu\in\mathcal{M}} \frac{1}{2}\sum_{i=1}^T w_i d(\mu, x_i)^2. \tag{6}$$

This definition is coordinate-free: it depends only on a notion of distance that reflects what it means for two objects (models, distributions, adapters) to be "close". In Euclidean space $\mathcal{M} = \mathbb{R}^D$ and $d(\mu, x) = \|\mu - x\|_2$, and (6) reduces to weighted averaging; on curved spaces it produces an *intrinsic* average that respects the underlying geometry.

3.2 Why Fréchet means help in model merging

Casting merging as (6) yields three benefits. First, it cleanly separates *what* we want (a geometry-respecting average) from *how* we compute it (distance approximations and optimization algorithms). Second, it makes invariances explicit: if a symmetry acts by isometries, the induced distance is symmetry-invariant and the merge is invariant *by construction*. Third, it unifies many existing merges as special cases obtained by different choices of $\mathcal{M}$ and d (or approximations thereof), and provides a principled way to derive new merges by swapping in more appropriate geometry. Furthermore, we are not restricted to d^2 averages: one could take the Riemannian median (d^1) [13], or even the Huber mean [29], which mixes elements of both.

3.3 Quotient manifolds imply symmetry invariance

Many model parameterizations are *not identifiable*: multiple parameter settings represent the same intrinsic object due to architectural symmetries (permutations, scalings, low-rank gauge freedoms). Let a Lie group $\mathcal{G}$ act on a total space $(\overline{\mathcal{M}}, \bar{g})$ by *isometries*. The intrinsic space is the quotient $\mathcal{M} := \overline{\mathcal{M}}/\mathcal{G}$ whose points are equivalence classes $[\bar{p}] = \{\bar{p} \cdot G \mid G \in \mathcal{G}\}$. The quotient distance can be written as an orbit-minimum:

$$d_{\mathcal{M}}\big([\bar{\mu}], [\bar{p}]\big) = \min_{G \in \mathcal{G}} \; d_{\overline{\mathcal{M}}}\big(\bar{\mu}, \bar{p} \cdot G\big). \tag{7}$$

Because the action is isometric, $d_{\mathcal{M}}$ is well-defined and independent of representatives. Consequently, the Fréchet objective on the quotient, $\sum_i d_{\mathcal{M}}([\bar{\mu}], [\bar{p}_i])^2$, is invariant to reparameterizations within each orbit: the merge depends only on intrinsic content, not on a particular gauge choice. In this sense the GeoMerge objective is *symmetry-invariant by construction*; Algorithm 1 inherits this invariance, up to numerical tolerance, whenever its initialization depends only on the orbits $[\bar{p}_i]$.

3.4 Generalization via geometry: a menu of aligned distances

A key advantage of the geometric formulation is that we can choose distances that are (i) appropriate for the object being merged and (ii) consistent with desired invariances.

Distribution-space geometries. When models are viewed as distributions (or predictive conditionals), the Fisher metric induces the Fisher-Rao geodesic distance; Fisher merging can be interpreted as a tractable approximation to Fréchet averaging under this geometry (via a Gaussian/Laplace approximation and a quadratic localization). Related divergences can serve as substitutes or bounds: for instance, symmetrized Jensen divergences provide upper bounds on Fisher-Rao distance, offering alternative (though not always tractable) objectives.

Adapter-space geometries. For low-rank adapters, the intrinsic object is naturally a quotient, and we can equip the total space with a product metric that respects the $O(r)$ gauge symmetry in the polar factorization. We use the canonical Stiefel metric for the Stiefel factors, and the affine-invariant metric for the SPD factor (see Section 2). *Once a metric is chosen*, its induced distance is automatically "aligned" with the constraints and symmetries encoded by the manifold/quotient structure.

3.5 Computing Fréchet means: the algorithmic recipe

Computing (6) generally requires three ingredients: (i) a way to evaluate or approximate $d_{\mathcal{M}}$, and/or (ii) access to the exponential and logarithm maps $\mathrm{Exp}_\mu : T_\mu\mathcal{M} \to \mathcal{M}$ and $\mathrm{Log}_\mu : \mathcal{M} \to T_\mu\mathcal{M}$, and (iii) an optimization scheme. A standard approach is Riemannian gradient descent on the Fréchet functional. When $\mathrm{Log}_\mu(x_i)$ is well-defined (e.g. within a geodesically convex neighborhood), the Riemannian gradient takes the simple form

$$\mathrm{grad}\, F(\mu) = -\textstyle\sum_{i=1}^{T} w_i \,\mathrm{Log}_\mu(\theta_i), \tag{8}$$

leading to the update

$$\mu_{t+1} = \mathrm{Exp}_{\mu_t}\Big(\alpha_t \textstyle\sum_{i=1}^{T} w_i \,\mathrm{Log}_{\mu_t}(\theta_i)\Big), \tag{9}$$

with step size $\alpha_t > 0$. If an exact closed-form for $d_{\mathcal{M}}$ is available and differentiable, one can also differentiate (6) directly; otherwise (9) provides a practical route whenever $\mathrm{Exp}/\mathrm{Log}$ (exact or approximate) are available.

3.6 Quotient GeoMerge: alignment & horizontal descent

Intuition. Figure 2 gives a schematic for quotient GeoMerge: the same intrinsic update can have many equivalent representatives, and naïvely merging unaligned representatives can produce a result that is not a sensible intrinsic average. Quotient GeoMerge therefore first aligns each input by orbit minimization, choosing the best symmetry transformation, and then averages only horizontal directions. A tangent vector between two representatives can have a component along the symmetry orbit, its vertical component, which changes the factors but not the update $\Delta \boldsymbol{W}$. Keeping it would make the merge depend on which factorization happened to be stored. After exact alignment the logarithm has no vertical component, so restricting to horizontal directions adds no constraint.

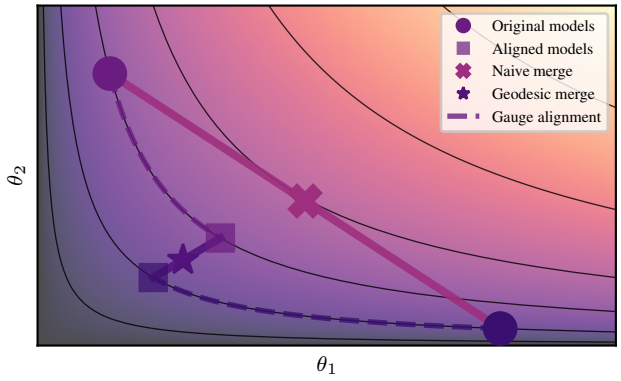


Figure 2: *Geodesic merging in a toy two-parameter setting with a scaling symmetry.* Two checkpoints (•) represent distinct models, each with infinitely many gauge-equivalent representatives along its symmetry orbit (dashed). Naïve ambient averaging yields a merge (×) that depends on the chosen representatives and can drift to a different orbit. GeoMerge instead aligns orbits (■), then averages intrinsically along a horizontal geodesic, yielding a symmetry-consistent merge (⋆).

Algorithm. Practically, as in generalized Procrustes analysis [16], we implement quotient Fréchet descent by alternating:

1. **Alignment (orbit minimization).** For a current representative $\bar{\mu}_t \in \bar{\mathcal{M}}$, choose alignments
$$G_i^\star(\bar{\mu}_t) \in \arg\min_{G\in\mathcal{G}} \tfrac{1}{2}\, d_{\bar{\mathcal{M}}}\big(\bar{\mu}_t, \bar{\theta}_i \cdot G\big)^2, \quad \bar{\theta}_i^\star(\bar{\mu}_t) := \bar{\theta}_i \cdot G_i^\star(\bar{\mu}_t). \tag{10}$$
2. **Intrinsic averaging in the total space.** Compute the horizontal logarithms $\eta_i := \Pi^H_{\bar{\mu}_t}\big(\mathrm{Log}_{\bar{\mu}_t}(\bar{\theta}_i^\star(\bar{\mu}_t))\big)$ and update
$$\bar{\mu}_{t+1} = \mathrm{Exp}_{\bar{\mu}_t}\Big(\alpha_t \textstyle\sum_{i=1}^T w_i\, \eta_i\Big). \tag{11}$$

With exact alignment and the exact logarithm, $\mathrm{Log}_{\bar{\mu}_t}(\bar{\theta}_i^\star(\bar{\mu}_t))$ is already horizontal at $\bar{\mu}_t$ [22], so Π^H has no effect; where the optimal alignments are unique, $\sum_i w_i\eta_i$ is then the horizontal lift of $-\operatorname{grad} F_{\mathcal{M}}([\bar{\mu}_t])$, where $F_{\mathcal{M}}([\bar{\mu}]) = \frac{1}{2}\sum_{i=1}^T w_i\, d_{\mathcal{M}}([\bar{\mu}],[\bar{\theta}_i])^2$ is the quotient Fréchet functional, and the update is Riemannian gradient descent on $F_{\mathcal{M}}$ [37, 2]. Our implementation solves the alignment only approximately and replaces the Stiefel logarithm and exponential by Cayley pseudo-logarithms and retractions (Section C.2). Its raw logarithm is then only approximately horizontal, which is why we apply Π^H explicitly; Section C gives the full procedure. GeoMerge is thus defined by the objective $F_{\mathcal{M}}$, and Algorithm 1 is an first-order Riemannian method for it, not a global solver. Replacing the exponential map by a retraction changes the update but not the objective, since a retraction agrees with the exponential map to first order [1].

4 Relation of GeoMerge to Fisher Merging

We connect *Fisher merging* [33] to GeoMerge by showing how it arises as tractable approximations to the Fréchet objective with an information-geometric distance. Given inputs $\{x_j\}_{j=1}^N$ and a conditional model $p_\theta(y \mid x)$, the estimated diagonal of the Fisher information matrix (FIM) at parameters θ is
$$\widehat{F}_\theta \;:=\; \frac{1}{N}\textstyle\sum_{j=1}^N \mathbb{E}_{y\sim p_\theta(y|x_j)}\Big[\big(\nabla_\theta \log p_\theta(y \mid x_j)\big)^{\odot 2}\Big]\,,$$
where $(\cdot)^{\odot 2}$ denotes elementwise squaring. Given models $\theta_1,\dots,\theta_T$ with corresponding diagonal FIM estimates $\widehat{F}_{\theta_i} \in \mathbb{R}^D$, Fisher merging computes a per-coordinate precision-weighted average:
$$\theta^\star_{\text{Fisher}} \;=\; \Big(\textstyle\sum_{i=1}^T\big(\widehat{F}_{\theta_i}\odot\theta_i\big)\Big) \oslash \Big(\textstyle\sum_{i=1}^T \widehat{F}_{\theta_i}\Big)\,,$$
with $\odot$ ($\oslash$) denoting elementwise multiplication (division). This weighting is motivated probabilistically, by a Gaussian approximation to each model's posterior [33]. The geometric lens of GeoMerge provides us a complementary view: *Fisher merging arises by replacing an analytically intractable Fréchet objective under the Fisher-Rao geometry with a tractable local quadratic surrogate.*

Concretely, view each checkpoint θ_i as specifying an (approximate) Gaussian distribution over parameters $q_i(\vartheta) \;\approx\; \mathcal{N}(\theta_i,\; F_i^{-1})$, where F_i denotes (an estimate of) the FIM at θ_i. Consider

the statistical manifold $\mathcal{Q}$ of Gaussians equipped with the Fisher information metric (the canonical metric in information geometry [3]), whose induced geodesic distance we denote by d_{FR} (Fisher-Rao distance). A more principled merge is then the Fréchet mean of $\{q_i\}_{i=1}^T$:

$$q^\star \in \arg\min_{q\in\mathcal{Q}} \ \frac{1}{2}\sum_{i=1}^T d_{\mathrm{FR}}(q, q_i)^2. \tag{12}$$

While d_{FR} between general Gaussians is not analytically available, it admits simple closed forms on two important submanifolds: (i) for fixed covariance, it reduces to a Mahalanobis distance in the mean; and (ii) for fixed mean, it reduces to a standard SPD distance in the covariance/precision (equivalently, an affine-invariant/log-Euclidean form).

An upper bound that yields Fisher merging. Restrict the candidate to the Laplace family $q(\vartheta) = \mathcal{N}(\theta, F^{-1})$ with free mean θ and (possibly) free precision F. For each i, introduce the intermediate Gaussian $\tilde{q}_i(\vartheta) := \mathcal{N}(\theta,\ F_i^{-1})$, which shares covariance with q_i and shares mean with q. By the triangle inequality for d_{FR} and the elementary bound $(a+b)^2 \le 2(a^2+b^2)$, we obtain (see, e.g. Pinele et al. [38] for derivations of these distances)

$$d_{\mathrm{FR}}\big(q_i, q\big)^2 \le 2\, d_{\mathrm{FR}}\big(q_i, \tilde{q}_i\big)^2 + 2\, d_{\mathrm{FR}}\big(\tilde{q}_i, q\big)^2 = 2\,(\theta-\theta_i)^\top F_i(\theta-\theta_i) + 2\, d_{\mathrm{SPD}}\big(F_i, F\big)^2,$$

where d_{SPD} denotes the canonical geodesic distance on the SPD manifold, induced by the affine-invariant metric. Crucially, the first term in the equation above is a *quadratic* (Mahalanobis) distance in parameter space, and the second term depends only on the choice of F (not on θ).

If we are just concerned with a pointwise estimate and simply drop the covariance SPD term, then minimizing (12) can be approximated by the quadratic surrogate

$$\theta^\star \ \in \ \arg\min_{\theta} \ \sum_{i=1}^T (\theta-\theta_i)^\top F_i(\theta-\theta_i). \tag{13}$$

The objective (13) is strictly convex when $\sum_i F_i \succ 0$ and has the closed-form minimizer

$$\theta^\star = \Big(\sum_{i=1}^T F_i\Big)^{-1}\Big(\sum_{i=1}^T F_i\,\theta_i\Big), \tag{14}$$

which is exactly Fisher merging (with F_i replaced by the chosen FIM approximation, e.g., its diagonal).

5 GeoMerge: High-Rank

Some LoRA merging methods first embed rank-r adapters into a larger rank budget and then apply a merge rule in the higher-dimensional representation. E.g., KnOTS aligns task updates in an SVD-derived coordinate system, while Core Space builds shared left/right bases before merging the induced cores [40, 36]. From our perspective, these methods combine two choices: a rank-increasing lift and an averaging rule. GeoMerge works as an averaging rule: a Fréchet mean on a quotient manifold. We thus need a quotient-compatible lift from $\mathcal{M}_r$ to $\mathcal{M}_R$, where $R > r$ is the target rank budget. Let

$$\theta_t = [U_t, B_t, V_t] \in \mathcal{M}_r, \qquad \mathcal{M}_r = \big(\mathrm{St}(d_{\mathrm{out}}, r) \times \mathbb{S}_{++}(r) \times \mathrm{St}(d_{\mathrm{in}}, r)\big)/O(r),$$

and write the dense rank-r update as $\Delta_t = U_t B_t V_t^\top$. A high-rank GeoMerge lift is a map

$$\mathcal{L}_t : \{\theta_s\}_{s=1}^T \longmapsto L_t = [\widehat{U}_t, \widehat{B}_t, \widehat{V}_t] \in \mathcal{M}_R,$$

followed by the same quotient Fréchet objective as before: $\mu_R^\star \in \arg\min_{\mu\in\mathcal{M}_R} \frac{1}{2}\sum_{t=1}^T w_t\, d_{\mathcal{M}_R}(\mu, L_t)^2$. We argue that the lift should satisfy two main conditions. First, it must be defined on quotient points rather than on arbitrary LoRA coordinates. Second, it must use the other task adapters, since we want the added rank to utilize cross-task structure.

The range of possible lifts is rather broad and we make no claims as to optimality; we present the simplest lift we could think of that satisfies the two criteria above. We leave a more thorough investigation of better choices for the lift to future work.

We choose orthonormal complements $U_t^\perp \in \mathbb{R}^{d_{\mathrm{out}}\times(R-r)}$ and $V_t^\perp \in \mathbb{R}^{d_{\mathrm{in}}\times(R-r)}$, and set $\widehat{U}_t = [U_t, U_t^\perp]$, $\widehat{V}_t = [V_t, V_t^\perp]$. We define the projectors $P_t^U = I - U_t U_t^\top$ and $P_t^V = I - V_t V_t^\top$. The task-conditioned residual is $R_t = \sum_{s\neq t} P_t^U \Delta_s P_t^V$. We then take the leading paired singular directions

Table 1: Merged-model performance on vision tasks for Vit-B/32 in normalized accuracies (%).

Method	Space	Cars	DTD	EuroSAT	GTSRB	MNIST	RESISC	SUN397	SVHN	Avg
TA	Full	81.97	73.72	48.97	42.24	53.12	71.50	97.46	41.25	63.78
Fisher	Full	84.5	72.4	44.4	55.8	47.8	70.9	96.1	39.2	63.9
TIES	Full	82.37	72.72	49.91	36.62	57.16	69.38	96.92	44.56	63.70
	KnOTS	83.75	74.45	50.36	47.31	67.01	71.79	96.51	50.64	67.73
	Core	84.74	76.46	52.19	50.41	67.36	71.21	96.45	50.18	68.63
Iso-C	Full	80.16	83.03	51.44	74.76	70.72	79.89	98.66	50.20	73.60
	KnOTS	80.33	79.29	57.50	67.60	65.63	79.54	99.26	46.62	71.97
	Core	83.35	84.30	50.13	81.97	71.07	83.46	99.17	53.90	75.92
TSV+Iso-C	Full	79.38	80.38	57.99	65.64	64.22	79.74	98.59	46.49	71.55
	KnOTS	80.81	83.03	58.25	74.34	67.66	79.69	98.54	49.86	74.02
	Core	82.98	85.12	50.95	84.25	71.14	84.39	99.06	53.53	76.43
Ours	Iso-C on our lift	82.64	85.31	51.44	83.29	72.86	82.96	98.97	54.85	76.54
	TSV+Iso-C on our lift	82.47	84.21	53.75	81.96	72.48	79.88	99.26	53.72	75.97
	GeoMerge+lift	83.48	84.30	53.27	82.14	75.45	82.22	98.76	57.18	**77.10**

$R_t = \widetilde{U}_t \Sigma_t \widetilde{V}_t^\top$ and use them as the columns of $U_t^\perp$ and $V_t^\perp$. We provide details on how we avoid instantiating the full dense matrix in Section E. The added rank is quotient-compatible: Δ_s, P_t^U, and P_t^V are invariant to the $O(r)$ gauge of the input factors, while any sign or rotation ambiguity in singular spaces is absorbed by the $O(R)$ quotient.

The lifted SPD factor can be picked in many different ways, but we keep it as simple as possible

$$\widehat{B}_t = \begin{bmatrix} B_t & 0 \\ 0 & cI_{R-r} \end{bmatrix}, \qquad c = \left(\prod_{s=1}^{T} \lambda_{\min}(B_s) \right)^{1/T}.$$

Scalar c is a conservative SPD filler: it keeps L_t in the fixed-rank manifold $\mathcal{M}_R$ without disturbing the lift in a large way.

6 Experiments

While we see our main contribution as conceptual, we empirically validate our approach for proof of concept. **Benchmarks.** Following experimental setup from prior work [40], we consider eight LoRA-adapted specialists derived from the same base ViT-B/32 model [10], where each was fine-tuned on a different vision dataset: Cars [27], DTD [7], EuroSAT [19], GTSRB [39], MNIST [28], RESISC [6], SUN397 [46], and SVHN [35]. These eight tasks are easy to fine-tune individually but hard to merge as LoRA adapters: Task Arithmetic reaches only 63.78% average normalized accuracy and the LoRA-specific KnOTS-TIES 67.73% (Table 1), which leaves a large gap between merging methods.

Baselines. We compare with Task Arithmetic [23] and with three merge rules, TIES, Iso-C and TSV+Iso-C [47, 30, 14], each applied in full weight space and in the spaces of KnOTS [40] and Core [36], with which they are traditionally used. We also report Fisher merging [33] as given by Stoica et al. [40]. Unlike the other methods, it needs gradients on data from each task, but we include it because Section 4 derives it from a Fréchet objective. **Metrics.** We report *normalized accuracy* on each dataset, defined per task as merged model accuracy divided by corresponding specialist accuracy, as in Stoica et al. [40] and Panariello et al. [36]; it removes differences in task difficulty between datasets. We summarize results using average normalized accuracy across tasks. **Results.** Table 1 summarizes the results on ViT-B/32. We outperform both Knots and Core, reaching new state-of-the-art performance on this benchmark, without the benefit of the vast literature around Euclidean weight averaging. As an ablation we also include the best performing Euclidean averaging method on our lift (after aligning the lifted adapters). The ablation confirms that the performance gains do not solely come from our lift, since we actually underperform the Core lift. Applying Iso-C instead of the Fréchet mean to the same lifted points, with the same coefficient, gives 76.54%. Fisher merging reaches 63.9%, close to Task Arithmetic and 13.2 points below GeoMerge; one possible reason is that diagonal Fisher merging averages coordinates and so inherits the symmetry failure of Euclidean averaging (Section A). Section F shows a merge of two of these adapters (Figure 3).

Language. Following Stoica et al. [40] and Panariello et al. [36], we also merge six LoRA-adapted specialists derived from the same Llama 3 8B base model [17], each fine-tuned on a different NLI

Table 2: Merged-model performance on NLI tasks for Llama-3 8B in normalized accuracies (%), with the evaluation protocol of Panariello et al. [36].

Method	Space	SNLI	MNLI	SICK	QNLI	RTE	SciTail	Avg
TA	Full	93.57	95.28	87.96	68.71	100.0	96.73	90.38
TIES	Full	95.17	96.19	84.18	74.18	100.0	96.78	91.08
	KnOTS	91.82	94.19	92.97	78.57	100.0	97.61	92.53
	Core	92.07	93.51	93.63	83.72	99.19	97.66	93.30
Iso-C	Full	55.00	39.04	76.54	55.90	46.77	69.25	57.08
	KnOTS	85.28	52.86	89.43	54.90	75.00	77.73	72.53
	Core	91.54	90.10	87.87	75.85	99.19	97.42	90.33
TSV	Full	95.38	95.12	88.83	76.80	101.61	97.56	92.55
	KnOTS	92.53	95.83	82.77	77.01	100.0	97.08	90.87
	Core	95.86	95.70	89.25	83.89	102.42	97.86	**94.16**
Ours	GeoMerge+lift	96.73	93.67	91.34	81.81	102.42	97.56	93.92

dataset: SNLI [5], MNLI [44], SICK [31], QNLI [42], RTE [42], and SciTail [26], and evaluate them with the protocol of Panariello et al. [36]. The isotropic filler of the vision lift flattens the spectrum, which suits the vision tasks but not NLI, where spectrum-flattening merges such as Iso-C perform poorly (Table 2). Since the lift is a separate module of GeoMerge, we use a spectrum-preserving lift here, which differs from Algorithm 2 in two places: the stacked peer frames are orthonormalized symmetrically, $W = F(F^\top F)^{-1/2}$, before the projection, and the lower-right block of $\widehat{B}_t$ holds the leading $R - r$ singular values of K_t instead of cI_{R-r}. The merge is otherwise unchanged ($R = 96$, uniform weights, Algorithm 1 with $N = 20$). GeoMerge reaches 93.92%, second only to TSV in Core Space (94.16%) among the methods in Table 2.

7 Remarks, Limitations & Future Work

While KnOTS operates on the full weight matrix space, i.e., directly on the $\Delta \boldsymbol{W}$s, and hence could reasonably be expected to be as fully symmetry invariant as it is possible to be, we posit that the difference in performance comes from the fact that the underlying geometry, once the symmetries are fully accounted for, is non-Euclidean and as such has curvature corrections not captured by KnOTS. While performance of GeoMerge is good, it is unclear what the best lifting procedure should be. This could potentially limit the widespread applicability of our approach and merits further study. Most merging methods in a PEFT setting use one of TIES or DARE-TIES as an add-on to a baseline merging/alignment procedure. This is not yet integrated into our framework. TIES and DARE-TIES, since they rely on parameter magnitudes and signs, are inherently coordinate-dependent procedures, whereas we take a completely coordinate-agnostic perspective. We leave this integration to future work. While we believe the main contribution of this paper is conceptual, computational efficiency could still be improved, and might limit widespread usability.

8 Conclusion

GeoMerge proposes an alternative, geometric view on model merging by treating it as computing a Fréchet mean under an explicit geometry, rather than an arbitrary average in parameter space. This perspective directly addresses architectural and parameterization symmetries: when models live on orbits of equivalent representations. By operating on the appropriate manifold via orbit alignment and geodesic updates, our merges are symmetry-consistent by construction and grant access to the deeper geometric structure of parameter space.

Our framework also yields practical algorithms and connections to prior work. We show how Fisher-weighted merging arises as a tractable approximation to an information-geometric Fréchet objective, and we instantiate geometric merging for LoRA adapters where symmetries are unavoidable. Empirical evidence supports the effectiveness of our proposed method.

Acknowledgments

Mohammed Adnan is supported by the NSERC Postgraduate Scholarship, RBC Borealis through the Borealis AI Global Fellowship Award, Killam Doctoral Fellowship, and the Digital Research Alliance of Canada EDIA Champions program. Felix Dangel is supported by a Canada CIFAR AI Chair, Mila. Resources used in preparing this research were provided, in part, by NSERC, the Province of Ontario, the Government of Canada through CI- FAR, and companies sponsoring the Vector Institute.

References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization algorithms on matrix manifolds*. 2008.

[2] B. Afsari, R. Tron, and R. Vidal. On the convergence of gradient descent for finding the riemannian center of mass. *SIAM Journal on Control and Optimization*, 51(3):2230–2260, 2013. doi: 10.1137/12086282x.

[3] S. Amari and H. Nagaoka. Methods of information geometry. 2000. URL `https://api.semanticscholar.org/CorpusID:116976027`.

[4] N. Boumal. *An introduction to optimization on smooth manifolds*. Cambridge University Press, 2023.

[5] S. R. Bowman, G. Angeli, C. Potts, and C. D. Manning. A large annotated corpus for learning natural language inference. In L. Màrquez, C. Callison-Burch, and J. Su, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 632–642, Lisbon, Portugal, Sept. 2015. Association for Computational Linguistics. doi: 10.18653/v1/D15-1075. URL `https://aclanthology.org/D15-1075/`.

[6] G. Cheng, J. Han, and X. Lu. Remote sensing image scene classification: Benchmark and state of the art. *Proceedings of the IEEE*, 105(10):1865–1883, 2017.

[7] M. Cimpoi, S. Maji, I. Kokkinos, S. Mohamed, and A. Vedaldi. Describing textures in the wild. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3606–3613, 2014.

[8] M. F. da Silva, F. Dangel, and S. Oore. Hide & seek: Transformer symmetries obscure sharpness & riemannian geometry finds it. In *ICML*, 2025. URL `https://arxiv.org/abs/2505.05409`.

[9] L. Dinh, R. Pascanu, S. Bengio, and Y. Bengio. Sharp minima can generalize for deep nets, 2017.

[10] A. Dosovitskiy, L. Beyer, A. Kolesnikov, D. Weissenborn, X. Zhai, T. Unterthiner, M. Dehghani, M. Minderer, G. Heigold, S. Gelly, J. Uszkoreit, and N. Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. *CoRR*, abs/2010.11929, 2020. URL `https://arxiv.org/abs/2010.11929`.

[11] F. Draxler, K. Veschgini, M. Salmhofer, and F. Hamprecht. Essentially no barriers in neural network energy landscape. In J. Dy and A. Krause, editors, *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 1309–1318. PMLR, 10–15 Jul 2018. URL `https://proceedings.mlr.press/v80/draxler18a.html`.

[12] A. Edelman, T. A. Arias, and S. T. Smith. The Geometry of Algorithms with Orthogonality Constraints. 1998.

[13] P. T. Fletcher, S. Venkatasubramanian, and S. Joshi. The geometric median on riemannian manifolds with application to robust atlas estimation. *NeuroImage*, 45(1):S143–S152, 2009. doi: 10.1016/j.neuroimage.2008.10.052.

[14] A. A. Gargiulo, D. Crisostomi, M. S. Bucarelli, S. Scardapane, F. Silvestri, and E. Rodolà. Task singular vectors: Reducing task interference in model merging. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 18695–18705, 2025. doi: 10.1109/CVPR52734.2025.01742.

[15] T. Garipov, P. Izmailov, D. Podoprikhin, D. P. Vetrov, and A. G. Wilson. Loss surfaces, mode connectivity, and fast ensembling of dnns. *Advances in neural information processing systems*, 31, 2018.

[16] J. C. Gower. Generalized procrustes analysis. *Psychometrika*, 40(1):33–51, 1975. doi: 10.1007/bf02291478.

[17] A. Grattafiori, A. Dubey, A. Jauhri, A. Pandey, A. Kadian, A. Al-Dahle, A. Letman, A. Mathur, A. Schelten, A. Vaughan, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

[18] R. Hecht-Nielsen. *On The Algebraic Structure of Feedforward Network Weight Spaces*. North-Holland, Amsterdam, 1990. ISBN 978-0-444-88400-8. doi: https://doi.org/10.1016/B978-0-444-88400-8.50019-4.

[19] P. Helber, B. Bischke, A. Dengel, and D. Borth. Eurosat: A novel dataset and deep learning benchmark for land use and land cover classification. *IEEE Journal of Selected Topics in Applied Earth Observations and Remote Sensing*, 12(7):2217–2226, 2019.

[20] N. J. Higham. Computing the polar decomposition—with applications. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1160–1174, 1986. doi: 10.1137/0907079.

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

[22] S. Huckemann, T. Hotz, and A. Munk. Intrinsic shape analysis: Geodesic pca for riemannian manifolds modulo isometric lie group actions. *Statistica Sinica*, 20(1):1–100, 2010.

[23] G. Ilharco, M. T. Ribeiro, M. Wortsman, S. Gururangan, L. Schmidt, H. Hajishirzi, and A. Farhadi. Editing models with task arithmetic. In *ICLR*, 2023. URL `https://arxiv.org/abs/2212.04089`.

[24] X. Jin, X. Ren, D. Preotiuc-Pietro, and P. Cheng. Dataless knowledge fusion by merging weights of language models, 2025. URL `https://arxiv.org/abs/2212.09849`.

[25] T. Kaneko, S. Fiori, and T. Tanaka. Empirical arithmetic averaging over the compact stiefel manifold. *IEEE Transactions on Signal Processing*, 61(4):883–894, 2013. doi: 10.1109/TSP.2012.2226167.

[26] T. Khot, A. Sabharwal, and P. Clark. SciTail: A textual entailment dataset from science question answering. In *Proceedings of the 32nd AAAI Conference on Artificial Intelligence*, 2018.

[27] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3d object representations for fine-grained categorization. In *Proceedings of the IEEE international conference on computer vision workshops*, pages 554–561, 2013.

[28] Y. LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[29] J. Lee and S. Jung. Huber means on riemannian manifolds, 2025. URL `https://arxiv.org/abs/2407.15764`.

[30] D. Marczak, S. Magistri, S. Cygert, B. Twardowski, A. D. Bagdanov, and J. van de Weijer. No Task Left Behind: Isotropic Model Merging with Common and Task-Specific Subspaces. In *International Conference on Machine Learning*, 2025.

[31] M. Marelli, S. Menini, M. Baroni, L. Bentivogli, R. Bernardi, and R. Zamparelli. A SICK cure for the evaluation of compositional distributional semantic models. In N. Calzolari, K. Choukri, T. Declerck, H. Loftsson, B. Maegaard, J. Mariani, A. Moreno, J. Odijk, and S. Piperidis, editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 216–223, Reykjavik, Iceland, May 2014. European Language Resources Association (ELRA). URL `https://aclanthology.org/L14-1314/`.

[32] S. Mataigne, R. Zimmermann, and N. Miolane. An efficient algorithm for the riemannian logarithm on the stiefel manifold for a family of riemannian metrics, 2024. URL `https://arxiv.org/abs/2403.11730`.

[33] M. Matena and C. Raffel. Merging models with fisher-weighted averaging. In *NeurIPS*, 2022. URL `https://arxiv.org/abs/2111.09832`.

[34] B. Mishra, G. Meyer, S. Bonnabel, and R. Sepulchre. Fixed-rank matrix factorizations and riemannian low-rank optimization. *Computational Statistics*, November 2013. URL `https://arxiv.org/abs/1209.0430`.

[35] Y. Netzer, T. Wang, A. Coates, A. Bissacco, B. Wu, A. Y. Ng, et al. Reading digits in natural images with unsupervised feature learning. In *NIPS workshop on deep learning and unsupervised feature learning*, volume 2011, page 4. Granada, 2011.

[36] A. Panariello, D. Marczak, S. Magistri, A. Porrello, B. Twardowski, A. D. Bagdanov, S. Calderara, and J. van de Weijer. Accurate and efficient low-rank model merging in core space, 2025. URL `https://arxiv.org/abs/2509.17786`.

[37] X. Pennec. Intrinsic statistics on riemannian manifolds: Basic tools for geometric measurements. *Journal of Mathematical Imaging and Vision*, 25:127–154, 2006. URL `https://api.semanticscholar.org/CorpusID:653972`.

[38] J. Pinele, J. E. Strapasson, and S. I. R. Costa. The fisher–rao distance between multivariate normal distributions: Special cases, bounds and applications. *Entropy*, 22(4), 2020. ISSN 1099-4300. doi: 10.3390/e22040404. URL https://www.mdpi.com/1099-4300/22/4/404.

[39] J. Stallkamp, M. Schlipsing, J. Salmen, and C. Igel. The german traffic sign recognition benchmark: a multi-class classification competition. In *The 2011 international joint conference on neural networks*, pages 1453–1460. IEEE, 2011.

[40] G. Stoica, P. Ramesh, B. Ecsedi, L. Choshen, and J. Hoffman. Model merging with svd to tie the knots. *ICLR*, 2025. URL https://arxiv.org/abs/2410.19735.

[41] D. Tam, M. Bansal, and C. Raffel. Merging by matching models in task parameter subspaces. *Transactions on Machine Learning Research*, 3 2024. URL https://arxiv.org/abs/2312.04339.

[42] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In T. Linzen, G. Chrupała, and A. Alishahi, editors, *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pages 353–355, Brussels, Belgium, Nov. 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446/.

[43] Z. Wang, E. Yang, L. Yin, S. Liu, and L. Shen. Model unmerging: Making your models unmergeable for secure model sharing, 2025. URL https://arxiv.org/abs/2509.01548.

[44] A. Williams, N. Nangia, and S. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122, 2018.

[45] M. Wortsman, G. Ilharco, S. Yitzhak Gadre, R. Roelofs, R. Gontijo-Lopes, A. S. Morcos, H. Namkoong, A. Farhadi, Y. Carmon, S. Kornblith, and L. Schmidt. Model soups: averaging weights of multiple fine-tuned models improves accuracy without increasing inference time. *ICML*, 2022.

[46] J. Xiao, K. A. Ehinger, J. Hays, A. Torralba, and A. Oliva. Sun database: Exploring a large collection of scene categories. *International Journal of Computer Vision*, 119(1):3–22, 2016.

[47] P. Yadav, D. Tam, L. Choshen, C. Raffel, and M. Bansal. Ties-merging: Resolving interference when merging models. In *NeurIPS*, 2023. URL https://arxiv.org/abs/2306.01708.

[48] B. Zhang, Z. Zheng, Z. Chen, and J. Li. Beyond the permutation symmetry of transformers: The role of rotation for model fusion. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=wBJIO15pBV.

Appendix

A GeoMerge: A Toy Problem

We analyze a minimal two-layer linear model with a scaling symmetry to make the symmetry-induced failure modes of naïve parameter averaging concrete and to illustrate GeoMerge as quotient/geodesic merging in closed form.

A two-layer linear network with a scaling gauge. We start with the smallest setting where a continuous architectural symmetry already makes "parameter-space averaging" ill-posed. Consider a scalar two-layer linear network

$$f_\theta(x) \;=\; \theta_2\theta_1 x, \quad \theta = (\theta_1, \theta_2) \in \overline{\mathcal{M}} := (\mathbb{R}^\star)^2 \tag{15}$$

where $\mathbb{R}^\star := \mathbb{R} \setminus \{0\}$. This parametrization has a multiplicative gauge symmetry: for any $a \in \mathbb{R}^\star$,

$$(\theta_1, \theta_2) \cdot a \;:=\; (a\,\theta_1,\ \theta_2/a),$$

and $f_{\theta\cdot a} \equiv f_\theta$ since the *predictor* $\beta(\theta) := \theta_1\theta_2$ is invariant under the action. Thus the intrinsic object is the orbit $[\theta] \in \mathcal{M} := \overline{\mathcal{M}}/\mathbb{R}^\star$.

A symmetry-invariant metric (scalar specialization of the $GH^\top$ metric). To respect the scaling symmetry, we equip $\overline{\mathcal{M}}$ with the scalar version of the symmetry-invariant metric proposed for $(\boldsymbol{G}, \boldsymbol{H})$-type parameters by da Silva et al. [8]:

$$g_{(\theta_1,\theta_2)}\big[(\eta_1,\eta_2),(\zeta_1,\zeta_2)\big] \;:=\; \theta_2^2\,\eta_1\zeta_1 \;+\; \theta_1^2\,\eta_2\zeta_2, \tag{16}$$

with $(\eta_1, \eta_2), (\zeta_1, \zeta_2) \in T_\theta\bar{\mathcal{M}} \cong \mathbb{R}^2$. A direct calculation shows the group action acts by isometries under g (so the quotient $\mathcal{M}$ inherits a well-defined metric).

Vertical vs. horizontal directions. Let $a = \exp(s)$ and differentiate the group action at $s = 0$ to obtain the vertical (orbit) direction $v(\theta) \;:=\; \frac{d}{ds}\big|_{s=0}(\theta \cdot e^s) = (\theta_1, -\theta_2)$. The horizontal space is the g-orthogonal complement of $\mathrm{span}\{v(\theta)\}$, i.e. $\eta = (\eta_1, \eta_2)$ is horizontal iff

$$g_\theta\big[\eta, v(\theta)\big] \;=\; 0 \iff \quad \frac{\eta_1}{\theta_1} \;=\; \frac{\eta_2}{\theta_2}. \tag{17}$$

Intuitively, horizontal motion changes the intrinsic predictor $w = \theta_1\theta_2$, while vertical motion changes only the representative (the gauge).

Horizontal geodesics make the predictor evolve linearly. For a horizontal initial velocity, geodesics admit a closed form. Writing $\gamma(t) = (\theta_1(t), \theta_2(t))$ with initial velocity $\dot\gamma(0) = (\eta_1(0), \eta_2(0))$ satisfying (17), that is, $\eta_1(0)/\theta_1(0) = \eta_2(0)/\theta_2(0) =: B$, substitution into the geodesic equations of (16) shows that

$$\theta_i(t) \;=\; \theta_i(0)\sqrt{1 + 2Bt}, \qquad i = 1, 2. \tag{18}$$

The ratio $\theta_1(t)/\theta_2(t) = \theta_1(0)/\theta_2(0)$ is therefore fixed, and

$$\beta(t) := \theta_1(t)\theta_2(t) \;=\; \theta_1(0)\theta_2(0)\big(1 + 2Bt\big). \tag{19}$$

Thus, once two points are *gauge-aligned* so that a horizontal geodesic connects them, the intrinsic interpolation is simply linear in predictor space.

Quotient distance reduces to predictor distance (after alignment). Because the action is isometric, the quotient distance can be written as an orbit-minimum:

$$d_{\mathcal{M}}\big([\theta],[\kappa]\big) \;=\; \min_{a\in\mathbb{R}^\star} d_{\overline{\mathcal{M}}}\big(\theta,\ \kappa\cdot a\big).$$

In this toy model, choosing a so that θ and $\kappa \cdot a$ lie on the same horizontal geodesic (i.e. they share the same ratio θ_1/θ_2 and lie in the same connected component/quadrant) yields a closed-form distance that depends only on the invariant predictors:

$$d_{\mathcal{M}}\big([\theta],[\kappa]\big) \;\propto\; \big|\, w(\theta) - w(\kappa) \,\big| \;=\; \big|\, \theta_1\theta_2 - \kappa_1\kappa_2 \,\big|. \tag{20}$$

GeoMerge becomes "average the predictors". Given checkpoints $\theta^{(1)}, \dots, \theta^{(N)}$, define $w_i := w(\theta^{(i)}) = \theta_1^{(i)}\theta_2^{(i)}$. With (20), the quotient Fréchet objective becomes

$$\begin{aligned}\mu^\star &= \arg\min_{[\mu]\in\mathcal{M}} \frac{1}{2}\textstyle\sum_{i=1}^T d_{\mathcal{M}}\big([\mu],[\theta^{(i)}]\big)^2 \\ &= \arg\min_{w\in\mathbb{R}} \frac{1}{2}\textstyle\sum_{i=1}^T (w-w_i)^2,\end{aligned}$$

so the merged intrinsic predictor is simply $w^\star = \frac{1}{T}\sum_{i=1}^T w_i$. Mapping back to parameters corresponds to choosing any representative $\mu = (\mu_1, \mu_2)$ with $\mu_1\mu_2 = w^\star$; this is exactly "pick a gauge" (e.g. enforce $\mu_1 = \mu_2 = \sqrt{|w^\star|}$ with a consistent sign choice).

Why this toy problem matters: a symmetry-induced pathology of Fisher/Euclidean averaging. Take two checkpoints that are *the same function* but different representatives, e.g. $\theta^{(2)} = \theta^{(1)}\cdot(-1) = (-\theta_1^{(1)}, -\theta_2^{(1)})$. Naïve Euclidean averaging gives $(\theta^{(1)} + \theta^{(2)})/2 = (0,0)$, which does not even lie in $\mathcal{M}$ and corresponds to the zero predictor. Moreover, in this toy setting, Fisher-weighted averaging exhibits the same failure mode: because the FIMs can coincide for $\theta^{(1)}$ and $\theta^{(2)}$ while the parameters cancel, the merged parameters collapse to $(0,0)$. In contrast, GeoMerge works on the quotient where $[\theta^{(1)}] = [\theta^{(2)}]$ and hence $d_{\mathcal{M}}([\theta^{(1)}],[\theta^{(2)}]) = 0$. We provide a quick derivation of the Fisher information matrix in Section B.

B Fisher Information Matrix for the Toy Model

We consider the regression model defined by the function $f(x;\boldsymbol{\theta}) = \theta_1\theta_2 x$. The observed data consists of n pairs $\{(x_i, y_i)\}_{i=1}^n$, and we assume additive Gaussian noise with variance σ^2. The model is given by:

$$y_i = \theta_1\theta_2 x_i + \epsilon_i, \quad \epsilon_i \sim \mathcal{N}(0,\sigma^2) \tag{21}$$

The parameter vector is $\boldsymbol{\theta} = [\theta_1, \theta_2]^\top$. The log-likelihood function $\ell(\boldsymbol{\theta})$ for the observations is:

$$\ell(\boldsymbol{\theta}) = -\frac{n}{2}\log\big(2\pi\sigma^2\big) - \frac{1}{2\sigma^2}\sum_{i=1}^n (y_i - \theta_1\theta_2 x_i)^2 \tag{22}$$

The score function is the gradient of the log-likelihood with respect to the parameters, $\nabla_{\boldsymbol{\theta}}\ell(\boldsymbol{\theta})$. The partial derivatives are:

$$\frac{\partial\ell}{\partial\theta_1} = \frac{1}{\sigma^2}\sum_{i=1}^n (y_i - \theta_1\theta_2 x_i)(\theta_2 x_i) \tag{23}$$

$$\frac{\partial\ell}{\partial\theta_2} = \frac{1}{\sigma^2}\sum_{i=1}^n (y_i - \theta_1\theta_2 x_i)(\theta_1 x_i) \tag{24}$$

The Hessian matrix $\mathbf{H}$ consists of the second-order partial derivatives:

$$\frac{\partial^2\ell}{\partial\theta_1^2} = -\frac{1}{\sigma^2}\sum_{i=1}^n \theta_2^2 x_i^2 \tag{25}$$

$$\frac{\partial^2\ell}{\partial\theta_2^2} = -\frac{1}{\sigma^2}\sum_{i=1}^n \theta_1^2 x_i^2 \tag{26}$$

$$\frac{\partial^2\ell}{\partial\theta_1\partial\theta_2} = \frac{1}{\sigma^2}\sum_{i=1}^n (y_i x_i - 2\theta_1\theta_2 x_i^2) \tag{27}$$

The Fisher Information Matrix (FIM), $\mathcal{I}(\boldsymbol{\theta})$, is defined as the negative expectation of the Hessian matrix:

$$\mathcal{I}(\boldsymbol{\theta}) = -E\left[\mathbf{H}\right] \tag{28}$$

We compute the expectation of the mixed partial derivative term using the relation $E[y_i] = \theta_1\theta_2 x_i$:

$$E\left[\frac{\partial^2 \ell}{\partial\theta_1\partial\theta_2}\right] = \frac{1}{\sigma^2}\sum_{i=1}^{n}(\theta_1\theta_2 x_i^2 - 2\theta_1\theta_2 x_i^2) = -\frac{1}{\sigma^2}\sum_{i=1}^{n}\theta_1\theta_2 x_i^2 \tag{29}$$

Let $S_{xx} = \sum_{i=1}^{n} x_i^2$. The Fisher Information Matrix is therefore:

$$\mathcal{I}(\boldsymbol{\theta}) = \frac{S_{xx}}{\sigma^2}\begin{bmatrix}\theta_2^2 & \theta_1\theta_2 \\ \theta_1\theta_2 & \theta_1^2\end{bmatrix} \tag{30}$$

Note that $\det(\mathcal{I}(\boldsymbol{\theta})) = 0$, indicating that the matrix is singular and the parameters θ_1 and θ_2 are unidentifiable as already pointed out.

C GeoMerge: Implementation Details

This appendix gives the implementation-level details omitted from the main text. The central routine approximates a quotient Fréchet mean: the same solver is used once at rank r to obtain an anchor for gauge alignment, and once at rank R after the endpoint lift. We write a point in either manifold as $p = [U, B, V] \in \mathcal{M}_k$, where $k = r$ or $k = R$.

Computing the polar factors. Adapters are trained and stored as $\Delta W = GH^\top$ with $G \in \mathbb{R}^{d_{\text{out}}\times r}$ and $H \in \mathbb{R}^{d_{\text{in}}\times r}$. We take thin QR factorizations $G = Q_G R_G$ and $H = Q_H R_H$, so that $\Delta W = Q_G M Q_H^\top$ with the small core $M := R_G R_H^\top \in \mathbb{R}^{r\times r}$. The polar decomposition of the core, $M = B\,O_M$ with $B := (MM^\top)^{1/2} \in \mathbb{S}_{++}(r)$ and $O_M := B^{-1}M \in O(r)$, gives $U := Q_G$ and $V := Q_H O_M^\top$, so that $UBV^\top = Q_G M Q_H^\top = \Delta W$. This costs $\mathcal{O}\big((d_{\text{out}} + d_{\text{in}})r^2 + r^3\big)$ and never forms a $d_{\text{out}} \times d_{\text{in}}$ matrix. B is positive definite exactly when $\operatorname{rank}(\Delta W) = r$; numerically, we floor the eigenvalues of $MM^\top$ at 10^{-6}. We convert each adapter once, before merging; the conversion leaves ΔW unchanged.

The only subtlety is that the logarithm $\operatorname{Log}_p(q)$ must compare representatives in a common gauge. Thus each Fréchet iteration first aligns every sample q_i to the current iterate p, then averages the resulting horizontal logarithms. For $q = [U_q, B_q, V_q]$, the $O(k)$ gauge action is

$$q \cdot O = [U_q O,\ O^\top B_q O,\ V_q O], \qquad O \in O(k).$$

The alignment step approximately solves

$$O^\star(p; q) \in \arg\min_{O\in O(k)} d_{\bar{\mathcal{M}}_k}\big(p, q\cdot O\big)^2,$$

where $\bar{\mathcal{M}}_k = \operatorname{St}(d_{\text{out}}, k) \times \mathbb{S}_{++}(k) \times \operatorname{St}(d_{\text{in}}, k)$ is the total space. In practice, we use a Procrustes initialization followed by a few horizontalization steps. The Procrustes initialization aligns the Stiefel factors: $O_0 = \operatorname{polar}\big(U_q^\top U + V_q^\top V\big)$. Given a current gauge O, we form the raw total-space logarithm from p to $q \cdot O$ and solve the horizontal projection equation for the skew-symmetric drift Ω. We then update

$$O \leftarrow O\exp(-\tau\Omega),$$

for a small fixed number of inner iterations (2 in every iteration for the vision experiments; 5 in the first iteration and 2 thereafter for NLI). The final logarithm returned to the Fréchet solver is the horizontal projection of the aligned raw logarithm.

Why the inner loop aligns. The update $O \leftarrow O\exp(-\tau\Omega)$ is a descent step on the alignment objective $f(O) := \frac{1}{2} d_{\bar{\mathcal{M}}_k}(p, q\cdot O)^2$. Because $\mathrm{O}(k)$ acts by isometries, X_Λ in (32) is a Killing field, so $\bar{g}(\dot\gamma, X_\Lambda)$ is constant along every geodesic γ. Applying this to the geodesic from p to $x := q \cdot O$, where the logarithm is defined, and using the first-variation formula gives

$$\frac{d}{ds}\Big|_{s=0} f\big(O\exp(s\Lambda)\big) = -\bar{g}_x\big(\operatorname{Log}_x(p), X_\Lambda(x)\big) = \bar{g}_p\big(\operatorname{Log}_p(x), X_\Lambda(p)\big) = \bar{g}_p\big(X_\Omega(p), X_\Lambda(p)\big),$$

where $X_\Omega(p)$ is the vertical part of $\operatorname{Log}_p(x)$, that is, Ω solves (34) with $\zeta = \operatorname{Log}_p(x)$. Taking $\Lambda = -\Omega$ makes the derivative $-\|X_\Omega(p)\|^2 \le 0$, and $\Omega = 0$, where the logarithm is horizontal, is the first-order optimality condition of the alignment. With the Cayley pseudo-logarithm these statements hold only approximately. We use $\tau = 1$ without tuning.

Algorithm 1 Quotient Fréchet Mean with Orbit Alignment

Require: Quotient points $\{q_i = [U_i, B_i, V_i]\}_{i=1}^T \subset \mathcal{M}_k$, weights $\{w_i\}_{i=1}^T$, initializer p_0, step size α, maximum iterations N, alignment iterations A, tolerance ε
Ensure: Fréchet mean estimate $\mu \in \mathcal{M}_k$

$p \leftarrow p_0$
for $n = 1$ to N **do**
  for $i = 1$ to T **do**
    $O_i \leftarrow \mathrm{polar}(U_i^\top U_p + V_i^\top V_p)$ ▷ Procrustes initialization
    for $a = 1$ to A **do**
      $\tilde{q}_i \leftarrow q_i \cdot O_i$
      $\zeta_i \leftarrow \mathrm{Log}_p^{\mathrm{total}}(\tilde{q}_i)$
      $\Omega_i \leftarrow \mathrm{GaugeDrift}(p, \zeta_i)$ ▷ skew drift from horizontal equation
      $O_i \leftarrow O_i \exp(-\tau\Omega_i)$
    $\tilde{q}_i \leftarrow q_i \cdot O_i$
    $\eta_i \leftarrow \Pi_p^H \mathrm{Log}_p^{\mathrm{total}}(\tilde{q}_i)$ ▷ horizontal quotient log
  $\bar{\eta} \leftarrow \sum_{i=1}^T w_i \eta_i$
  if $\|\bar{\eta}\|_p < \varepsilon$ **then**
    return p
  $p \leftarrow \mathrm{Exp}_p(\alpha\bar{\eta})$
return p

Here Π_p^H denotes horizontal projection. For a raw tangent $\zeta = (\zeta_U, \zeta_B, \zeta_V)$ at $p = [U, B, V]$, this projection has the form

$$\Pi_p^H \zeta = \big(\zeta_U - U\Omega,\ \zeta_B - (B\Omega - \Omega B),\ \zeta_V - V\Omega\big),$$

where $\Omega^\top = -\Omega$ is obtained from the horizontal gauge equation

$$B^{-1}\Omega B + B\Omega B^{-1} - \Omega = \frac{1}{2}\left(V^\top\zeta_V + U^\top\zeta_U\right) - \left(B^{-1}\zeta_B - \zeta_B B^{-1}\right).$$

Thus $\mathrm{GaugeDrift}(p, \zeta)$ in Algorithm 1 is the skew matrix Ω solving this equation. In the final line, Exp_p is implemented as the product update on the total-space factors: a Stiefel update for U and V, and the affine-invariant SPD exponential for B.

We use step size $\alpha = 1$, which is stable in our experiments, tolerance $\varepsilon = 10^{-8}$ and $N = 20$ Fréchet iterations, both for the rank-r anchor and for the rank-R merge.

We next describe the algorithm for the rank-lifted procedure. After constructing lifted points $L_t \in \mathcal{M}_R$, it computes the final rank-R quotient Fréchet mean.

C.1 Quotient Geometry under the Canonical Product Metric

Mishra et al. [34] construct the quotient $\mathcal{M}_k = \bar{\mathcal{M}}_k/\mathrm{O}(k)$, with $\bar{\mathcal{M}}_k = \mathrm{St}(d_{\mathrm{out}}, k) \times \mathbb{S}_{++}(k) \times \mathrm{St}(d_{\mathrm{in}}, k)$, using the Euclidean metric on the Stiefel factors. We use the canonical Stiefel metric instead, so the total space carries the product metric

$$\bar{g}_{(U,B,V)}(\xi, \eta) = \mathrm{tr}\big(\xi_U^\top(I - \tfrac{1}{2}UU^\top)\eta_U\big) + \mathrm{tr}\big(B^{-1}\xi_B B^{-1}\eta_B\big) + \mathrm{tr}\big(\xi_V^\top(I - \tfrac{1}{2}VV^\top)\eta_V\big), \quad (31)$$

and $\mathrm{O}(k)$ acts by $(U, B, V) \cdot Q = (UQ,\ Q^\top BQ,\ VQ)$. Most of their construction does not involve the metric and carries over unchanged. The action is free, since $UQ = U$ implies $Q = U^\top UQ = U^\top U = I$, and proper, since $\mathrm{O}(k)$ is compact. It is also isometric for (31): its differential maps (ξ_U, ξ_B, ξ_V) to $(\xi_U Q,\ Q^\top\xi_B Q,\ \xi_V Q)$, and $QQ^\top = I$ gives

$$\mathrm{tr}\big(Q^\top\xi_U^\top(I - \tfrac{1}{2}UQQ^\top U^\top)\eta_U Q\big) = \mathrm{tr}\big(\xi_U^\top(I - \tfrac{1}{2}UU^\top)\eta_U\big),$$
$$\mathrm{tr}\big((Q^\top BQ)^{-1}Q^\top\xi_B Q\,(Q^\top BQ)^{-1}Q^\top\eta_B Q\big) = \mathrm{tr}\big(B^{-1}\xi_B B^{-1}\eta_B\big),$$

and likewise for V. Hence (31) induces a Riemannian metric on $\mathcal{M}_k$ for which the projection $\pi : \bar{\mathcal{M}}_k \to \mathcal{M}_k$ is a Riemannian submersion [1]. The vertical space is the tangent space of the orbit, so it depends only on the action; differentiating $t \mapsto (U, B, V) \cdot \exp(t\Lambda)$ at $t = 0$ gives

$$\mathcal{V}_{(U,B,V)} = \big\{X_\Lambda := (U\Lambda,\ B\Lambda - \Lambda B,\ V\Lambda)\ :\ \Lambda^\top = -\Lambda\big\}. \quad (32)$$

Algorithm 2 Lifted GeoMerge

Require: Rank-r quotient points $\{\theta_t = [U_t, B_t, V_t]\}_{t=1}^T \subset \mathcal{M}_r$, target rank $R > r$, weights $\{w_t\}_{t=1}^T$
Ensure: Rank-R merged point $\mu_R \in \mathcal{M}_R$

$c \leftarrow \Big(\prod_{t=1}^T \lambda_{\min}(B_t)\Big)^{1/T}$
for $t = 1$ to T **do**
  $\mathcal{P}_t \leftarrow \{s : s \neq t,\ a_{ts} \neq 0\}$
  for $s \in \mathcal{P}_t$ **do**
    $A_{ts} \leftarrow U_s - U_t(U_t^\top U_s)$
    $C_{ts} \leftarrow V_s - V_t(V_t^\top V_s)$ ▷ applies P_t^U and P_t^V without forming them
  $A_t \leftarrow [A_{ts}]_{s\in\mathcal{P}_t}, \quad C_t \leftarrow [C_{ts}]_{s\in\mathcal{P}_t}$
  $M_t \leftarrow \mathrm{blkdiag}(a_{ts}B_s)_{s\in\mathcal{P}_t}$
  Thin span factorizations $A_t = E_t^U G_t, \qquad C_t = E_t^V H_t.$
  $K_t \leftarrow G_t M_t H_t^\top$ ▷ small core for the residual
  Compute $K_t = \bar{U}_t \Sigma_t \bar{V}_t^\top$
  $U_t^\perp \leftarrow E_t^U \bar{U}_t, \quad V_t^\perp \leftarrow E_t^V \bar{V}_t$
  Keep the leading $R - r$ columns of $U_t^\perp$ and $V_t^\perp$
  $\widehat{U}_t \leftarrow [U_t, U_t^\perp], \quad \widehat{V}_t \leftarrow [V_t, V_t^\perp]$
  $\widehat{B}_t \leftarrow \begin{bmatrix} B_t & 0 \\ 0 & cI_{R-r} \end{bmatrix}$
  $L_t \leftarrow [\widehat{U}_t, \widehat{B}_t, \widehat{V}_t] \in \mathcal{M}_R$
$\mu_R \leftarrow \mathrm{FrechetMean}_{\mathcal{M}_R}(\{L_t\}_{t=1}^T; \{w_t\}_{t=1}^T)$ ▷ Algorithm 1
return μ_R

The horizontal space is the $\bar{g}$-orthogonal complement of (32), and this is where the metric matters. For a tangent vector $\zeta = (\zeta_U, \zeta_B, \zeta_V)$ and a skew-symmetric Λ,

$$\begin{aligned} \mathrm{tr}\big(\zeta_U^\top (I - \tfrac{1}{2}UU^\top)U\Lambda\big) &= \tfrac{1}{2}\,\mathrm{tr}(\zeta_U^\top U\Lambda), \\ \mathrm{tr}\big(B^{-1}\zeta_B B^{-1}(B\Lambda - \Lambda B)\big) &= \mathrm{tr}\big((B^{-1}\zeta_B - \zeta_B B^{-1})\Lambda\big), \end{aligned}$$

and similarly for V. Because ζ is tangent, $\zeta_U^\top U = -U^\top \zeta_U$ and $\zeta_V^\top V = -V^\top \zeta_V$, so

$$\bar{g}(\zeta, X_\Lambda) = \mathrm{tr}(K\Lambda), \qquad K := (B^{-1}\zeta_B - \zeta_B B^{-1}) - \tfrac{1}{2}(U^\top \zeta_U + V^\top \zeta_V).$$

K is skew-symmetric, because ζ_B is symmetric and $U^\top \zeta_U$ and $V^\top \zeta_V$ are skew-symmetric. If $\mathrm{tr}(K\Lambda) = 0$ for every skew-symmetric Λ, then taking $\Lambda = K$ gives $\|K\|_F^2 = 0$, so $K = 0$. Hence

$$\mathcal{H}_{(U,B,V)} = \big\{\zeta \,:\, \tfrac{1}{2}(U^\top \zeta_U + V^\top \zeta_V) = B^{-1}\zeta_B - \zeta_B B^{-1}\big\}. \tag{33}$$

The horizontal projection subtracts a vertical vector, $\Pi^H \zeta = \zeta - X_\Omega$ with $\Omega^\top = -\Omega$. Substituting $\zeta - X_\Omega$ into (33) and using $U^\top U = V^\top V = I$ gives the gauge equation

$$B^{-1}\Omega B + B\Omega B^{-1} - \Omega = \tfrac{1}{2}(U^\top \zeta_U + V^\top \zeta_V) - (B^{-1}\zeta_B - \zeta_B B^{-1}) \tag{34}$$

used in Section 2 and Algorithm 1. With the Euclidean Stiefel metric of Mishra et al. [34], the weight $I - \frac{1}{2}UU^\top$ is replaced by I and the factor $\frac{1}{2}$ by 1, so their projection formula does not carry over unchanged.

Equation (34) has a closed-form solution. Let $B = PDP^\top$ with $D = \mathrm{diag}(d_1, \ldots, d_k)$ and all $d_i > 0$, and let C denote the right-hand side of (34). With $\tilde{\Omega} = P^\top \Omega P$ and $\tilde{C} = P^\top C P$, the equation decouples entrywise:

$$\Big(\frac{d_i}{d_j} + \frac{d_j}{d_i} - 1\Big)\tilde{\Omega}_{ij} = \tilde{C}_{ij}.$$

Since $d_i/d_j + d_j/d_i \geq 2$, every coefficient is at least 1. The solution therefore exists and is unique, it is skew-symmetric because $\tilde{C}$ is, and the division never amplifies $\tilde{C}$. It costs one $k \times k$ eigendecomposition, and our implementation computes Ω this way.

C.2 Cayley Stiefel Updates and Approximate Logs

The quotient Fréchet solver requires repeated Stiefel logarithm-like operations during orbit alignment and mean updates. Exact canonical Stiefel logarithms are expensive in this setting, since they are evaluated for every task, layer, and alignment iteration. We therefore use Cayley pseudo-lifts and Cayley retractions for the Stiefel factors, while keeping the SPD factor using the exact affine-invariant log/exp.

We use the notation of Kaneko et al. [25]. For a square matrix M, define

$$\mathrm{sk}(M) := \frac{1}{2}(M^\top - M). \tag{35}$$

Thus $\mathrm{sk}^{-1}(M)$ denotes $\big(\mathrm{sk}(M)\big)^{-1}$ when this inverse exists. Given $X, Q \in \mathrm{St}(n,k)$, the full-Cayley pseudo-lift of Kaneko et al. [25] is the ambient skew-symmetric matrix

$$\widehat{P}_X^{-1}(Q) = \frac{1}{2}(Q - X)\,\mathrm{sk}^{-1}(X^\top Q)(Q - X)^\top \in \mathfrak{so}(n), \tag{36}$$

provided that $\mathrm{sk}(X^\top Q)$ is nonsingular. This is the Cayley "log" used for the Stiefel factor in our quotient implementation.

The corresponding Cayley pseudo-retraction is

$$\widehat{P}_X(\Omega) = \mathrm{Cay}(\Omega)X, \qquad \mathrm{Cay}(\Omega) = (I_n + \Omega)(I_n - \Omega)^{-1}, \tag{37}$$

for $\Omega \in \mathfrak{so}(n)$.

Inside the quotient solver, after aligning the target representative to the current iterate, the raw total-space log is assembled as

$$\widetilde{\mathrm{Log}}^{\mathrm{total}}_{[U,B,V]}([U', B', V']) = \Big(\widehat{P}_U^{-1}(U'), \mathrm{Log}_B^{\mathrm{SPD}}(B'), \widehat{P}_V^{-1}(V')\Big), \tag{38}$$

where the first and third components are the full-Cayley pseudo-lifts in (36), and the middle component is the affine-invariant SPD logarithm.

For an horizontal tangent $(\Delta_U, \Delta_B, \Delta_V)$ at $[U, B, V]$, the update uses Cayley retractions on the Stiefel factors and the affine-invariant exponential on the SPD factor:

$$U^+ = \mathrm{Cay}(\alpha\Omega_U)U, \qquad B^+ = \mathrm{Exp}_B^{\mathrm{SPD}}(\alpha\Delta_B), \qquad V^+ = \mathrm{Cay}(\alpha\Omega_V)V. \tag{39}$$

Empirically, these approximations substantially reduce the cost of the repeated alignment loop, with only a modest degradation relative to the exact Stiefel-log reference path.

D Infrastructure Details

Compute time. All merges run on one NVIDIA Titan V GPU with an Intel(R) Xeon(R) W-2133 CPU (12 cores, 3.60 GHz). Table 3 lists the time of each merge, without evaluation.

Table 3: Merge time on one Titan V GPU. –: not run.

Method	ViT-B/32, 8 tasks	Llama-3 8B, 6 tasks
Task Arithmetic	3 s	9 s
KnOTS	16 min	70 min
GeoMerge, $R = r$, exact Stiefel Log/Exp	–	2 h
GeoMerge, $R = r$, Cayley	–	60 min
GeoMerge + lift, Cayley	20 min	55 min

Complexity. Let $D = d_{\mathrm{out}} + d_{\mathrm{in}}$, T the number of tasks, R the quotient rank, N the number of Fréchet iterations and A the number of alignment steps per iteration. For one task, one alignment step costs $\mathcal{O}(DR^2 + R^3)$: the Cayley pseudo-logarithm, the Procrustes initialization and the projection (34) need only products of $D \times R$ factors and $R \times R$ solves and eigendecompositions. One Fréchet iteration repeats this for T tasks and $A + 1$ logarithms, then retracts once. Our implementation forms

the Cayley retraction of Section C.2 with dense $n \times n$ matrices, where $n = d_{\text{out}}$ for U and $n = d_{\text{in}}$ for V, at cost $\mathcal{O}(n^3)$. One Fréchet solve therefore costs $\mathcal{O}\big(N\,[\,T(A+1)(DR^2+R^3)+d_{\text{out}}^3+d_{\text{in}}^3\,]\big)$ per layer. The skew matrix in the retraction has rank at most $2R$, so the Woodbury identity would reduce each retraction to $\mathcal{O}(nR^2+R^3)$ and the total to $\mathcal{O}\big(NT(A+1)(DR^2+R^3)\big)$; our implementation does not do this. The lift of Section 5 adds, per task, thin QR factorizations of $d \times (T-1)r$ matrices and one SVD of a core of size at most $(T-1)r \times (T-1)r$, that is, $\mathcal{O}\big(TD(Tr)^2+T(Tr)^3\big)$ per layer.

Dataset Licenses. Among the datasets we use, we were able to determine the following licenses and/or usage permissions:

- Cars [27] uses Creative Commons License.
- GTSRB [39] uses Creative Commons License.
- EuroSAT [19] uses MIT license.
- MNIST [28] uses Gnu General Public License and elsewhere is listed under Creative Commons Attribution-Share Alike 3.0 license.
- SUN397 [46] is listed as "for research purposes only".
- SVHN [35] is listed as being "for non-commercial use only".
- SNLI [5] uses CC BY-SA 4.0
- MNLI [44] appears to incorporate an aggregate of multiple licenses.
- SICK [31] uses a Creative Commons Attribution-NonCommercial-ShareAlike license.
- QNLI [42] is derived from SQuAD, which in turn uses a CC BY-SA 4.0 license.
- RTE [42] appears to incorporate an aggregate of multiple licenses.
- SciTail [26] uses an Apache 2.0 license.

E Efficient High-Rank Lift

We explain how we avoid instantiating the dense form of the completion basis used in Section 5. For a task t

$$R_t = \sum_{s \neq t} P_t^U \Delta_s P_t^V, \qquad P_t^U = I - U_t U_t^\top, \qquad P_t^V = I - V_t V_t^\top, \tag{40}$$

where $\Delta_s = U_s B_s V_s^\top$ is the rank-r polar representation of task s. The completion basis is then obtained from the leading paired singular directions of R_t. We now show how to obtain the same subspaces without forming any dense matrix.

For each task $s \neq t$, define

$$A_{ts} := P_t^U U_s = U_s - U_t(U_t^\top U_s), \qquad C_{ts} := P_t^V V_s = V_s - V_t(V_t^\top V_s). \tag{41}$$

Then

$$P_t^U \Delta_s P_t^V = (P_t^U U_s) B_s (P_t^V V_s)^\top = A_{ts} B_s C_{ts}^\top, \tag{42}$$

so

$$R_t = \sum_{s \neq t} A_{ts} B_s C_{ts}^\top. \tag{43}$$

Stack all $s_i \neq t, i \in \{1, \ldots, T-1\}$

$$A_t = \big[A_{s_1}, \ldots, A_{s_{T-1}}\big], \qquad C_t = \big[C_{s_1 Sur}, \ldots, C_{s_{T-1}}\big], \tag{44}$$

and let

$$M_t = \operatorname{blkdiag}\big(B_{s_1}, \ldots, B_{s_K}\big). \tag{45}$$

Then

$$R_t = A_t M_t C_t^\top. \tag{46}$$

Compute thin span factorizations

$$A_t = E_t^U G_t, \qquad C_t = E_t^V H_t, \qquad (E_t^U)^\top E_t^U = I, \qquad (E_t^V)^\top E_t^V = I, \tag{47}$$

Substituting into (46) gives

$$R_t = E_t^U K_t (E_t^V)^\top, \qquad K_t := G_t M_t H_t^\top. \tag{48}$$

The matrix K_t has size at most $(T-1)r \times (T-1)r$.

Now take the small SVD

$$K_t = \bar{U}_t \Sigma_t \bar{V}_t^\top. \tag{49}$$

Then

$$R_t = (E_t^U \bar{U}_t) \Sigma_t (E_t^V \bar{V}_t)^\top. \tag{50}$$

Therefore the dense residual's nonzero singular directions are

$$U_t^\perp = E_t^U \bar{U}_t, \qquad V_t^\perp = E_t^V \bar{V}_t. \tag{51}$$

This construction is exactly equivalent to forming R_t and taking its SVD: the nonzero singular values of R_t and K_t coincide, and their singular vectors are related by the isometries E_t^U and E_t^V. If singular values are repeated, equality is understood as equality of singular subspaces, up to signs and orthogonal rotations inside degenerate subspaces. This ambiguity is absorbed by the $O(R)$ quotient gauge.

Finally, the construction is quotient-defined. Under a peer gauge transformation

$$(U_s, B_s, V_s) \mapsto (U_s Q_s, Q_s^\top B_s Q_s, V_s Q_s), \qquad Q_s \in O(r), \tag{52}$$

one has

$$A_{ts} \mapsto A_{ts} Q_s, \qquad C_{ts} \mapsto C_{ts} Q_s, \tag{53}$$

and hence

$$(A_{ts} Q_s)(Q_s^\top B_s Q_s)(C_{ts} Q_s)^\top = A_{ts} B_s C_{ts}^\top. \tag{54}$$

The anchor projectors $U_t U_t^\top$ and $V_t V_t^\top$ are themselves invariant under the anchor gauge. Thus R_t and its completion subspaces depend only on the quotient points, not on the chosen representatives.

F Merging Two Real LoRA Adapters

Figure 3 illustrates GeoMerge on real adapters, for one pair of the tasks of Table 1: DTD and RESISC45, merged with the same settings. All scores in this section are average normalized accuracies on fixed 1,000-image subsets of the validation sets.

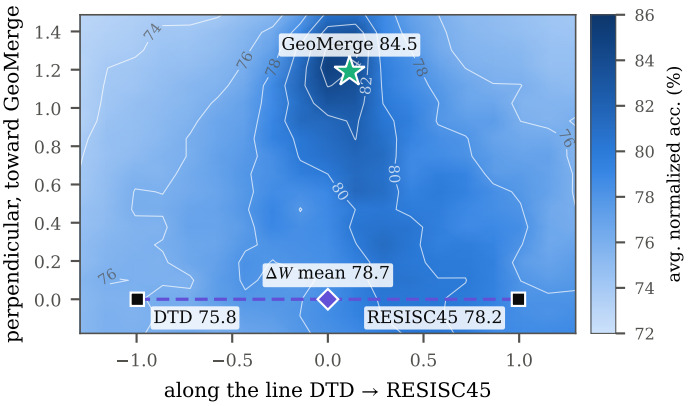


Figure 3: **Merging two real LoRA adapters** (ViT-B/32, KnOTS, rank 16) for DTD and RESISC45. The plane passes through both task updates and GeoMerge (settings of Table 1). Axes (Frobenius norm): along the line from ΔW_{DTD} to $\Delta W_{\mathrm{RESISC45}}$, which holds their Euclidean interpolations and has the ΔW mean at its midpoint, and perpendicular to it, toward GeoMerge. Color: normalized validation accuracy, averaged over the two tasks, of the base model plus each point of the plane. GeoMerge scores 84.5; the Euclidean interpolations we evaluated score at most 80.2, and the factor averages, which lie off the plane, 77.6, or 77.2–80.9 after random $\mathrm{GL}(16)$ reparametrizations that leave every ΔW unchanged.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The reason that the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope is because that is exactly one of the main purposes of the abstract and introduction. So, in turn, that was exactly one of our main objectives when writing the abstract and introduction.

   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 have included a separate section entitled "Remarks, Limitations & Future Work" in which we discuss the limitations of this work.

   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: We see this paper as making a primary theoretical contribution, supported by strong empirical evidence. We introduce and develop theory for Riemannian-based model merging, and we begin this by providing our definitions and assumptions. While we have not formulated the paper in terms of theorems and proofs (this would have been unnecessary), we have taken considerable efforts to explain our reasoning carefully, and to provide appropriate references when making use of a result that has been previously established. Where no reference is provided is because we assume the statement to be covered in standard mathematical references on differential and riemannian geometry.

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: Our intention in writing this paper is that if someone were to follow our recipe as we have described it—which we believe is entirely doable based on our provided description—then they will get essentially the same results.

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: We intend to make our code available upon acceptance. The datasets we use are publicly available.

   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 follow experimental setups previously reported in the literature and with publicly available results. Details about our optimization and hyperparameters are reported in Section C.

   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: Due to computational constraints we were not able to provide these.

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: [Yes]

   Justification: Provided in Section D

   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: This work conforms in every respect with 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: [N/A]

Justification: This paper presents work whose goal is to advance the field of machine learning, specifically through the use of Riemannian geometry to improve model merging. While the broad field of machine learning itself has many potential societal consequences, we do not feel that there are any societal impacts specific to this work that need to be highlighted.

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: This paper poses no 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: All of the datasets and models we use are publicly available and have been described in prior work; we have, to the best of our knowledge, properly credited that prior work. We have also provided a section in the appendix that lists the names of the licenses we were able to identify for each of the datasets we used.

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: The 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: This work included neither crowdsourcing nor 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: This work included neither crowdsourcing nor 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: The core methodology does not use LLMs.

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.