# iMuon: Intrinsic Spectral Optimization on Riemannian Matrix Manifolds

**Yibang Li**[1*] **Bihari Lal Pandey**[2*] **Ravi Sah**[2] **Andi Han**[3] **Cyrus Mostajeran**[1]
**Pratik Jawanpuria**[2] **Bamdev Mishra**[4]
[1]Nanyang Technological University, Singapore [2]Indian Institute of Technology Bombay, India
[3]The University of Sydney, Australia [4]Microsoft India
yibang001@e.ntu.edu.sg, bihari@iitb.ac.in, ravi.sah@iitb.ac.in,
andi.han@sydney.edu.au, cyrussam.mostajeran@ntu.edu.sg,
pratik.jawanpuria@iitb.ac.in, bamdevm@microsoft.com

## Abstract

Muon and related norm-constrained optimizers, now central to large-scale learning, compute updates via a linear maximization oracle (LMO) over a matrix-norm ball in Euclidean space. However, they do not generalize cleanly to manifold-valued parameters such as low-rank factorizations, orthogonality constraints, or symmetric positive-definite (SPD) matrices. Naively restricting the Muon LMO to the tangent space (i) breaks quotient symmetries and (ii) couples the tangent-space constraint with an ambient norm bound, which obstructs closed-form solutions. We resolve both issues with a key observation: every Riemannian metric canonically lifts a unitarily invariant Euclidean norm to an intrinsic norm on each tangent space. The resulting intrinsic LMO is symmetry preserving and, under a structural condition on the scaled tangent space, reduces to a standard Euclidean LMO. Building on this, we introduce intrinsic Muon (iMuon), a unified framework with closed-form updates on the fixed-rank, SPD, Stiefel, and Grassmann manifolds for various unitarily invariant norms, including the spectral, Frobenius, and nuclear norms. We establish deterministic and stochastic convergence rates of iMuon with constants that depend only on the manifold dimension. Notably, on the fixed-rank manifold, the constant depends only on the rank, which makes the rate independent of factor conditioning and removes the runtime factor rescaling required by prior work. Experiments on LoRA fine-tuning of LLMs, image classification, and subspace learning illustrate the efficacy of the proposed approach.

## 1 Introduction

A growing family of matrix-variate optimizers is unified by a single design principle: at each step, select the update direction that maximizes the inner product with the (stochastic) gradient over a matrix-norm ball, a linear maximization oracle (LMO) [56, 8, 77]. The choice of norm determines the optimizer: the Frobenius norm recovers SGD, the spectral norm yields Muon [33, 8], and a combined spectral-nuclear constraint gives NuMuon [50]. This perspective has driven recent advances in LLM pretraining [84, 49, 73, 32] and low-rank adaptation (LoRA) fine-tuning [29, 61, 54, 21]. Despite their empirical success, the design of such LMO-based optimizers has been mostly confined to unconstrained Euclidean spaces.

However, many modern representation-learning problems are intrinsically manifold-valued: LoRA fine-tuning operates on the fixed-rank manifold [48, 81], orthogonality constraints define the Stiefel manifold, and robust covariance estimation lives on the SPD manifold. Extending norm-constrained

*Equal contribution.

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

matrix optimizers to such Riemannian settings remains relatively unexplored [21, 12]. The seemingly natural fix of restricting the LMO to the tangent space runs into two fundamental obstacles:

- **(O1) Symmetry breaking on quotient manifolds.** The Euclidean spectral norm is not invariant under the symmetry group of a quotient manifold (i.e., manifold arising out of symmetries). For example, on the fixed-rank manifold, equivalent factorizations $X = BA$ that share the same product induce different factor-wise Muon updates, so both the direction and the magnitude of the update depend on an arbitrary representative [32].
- **(O2) Coupling of the tangent space and norm constraints obstructs a closed-form solution.** Restricting the spectral-norm LMO to the tangent space couples two constraints, tangent space membership and an ambient norm bound, whose joint enforcement generally precludes the Euclidean polar-factor solution and admits no closed form on most manifolds of interest.

Existing literature addresses these issues only partially. Spectron [32] bounds the ambient update by an iterate-dependent factor rescaling computed dynamically. Riemannion [12] relies on an approximate tangent space orthogonalization. Manifold constrained steepest descent (MCSD) [80] drops the tangent constraint and projects back to the manifold instead. However, none of the existing approaches resolves (O1) and (O2) jointly.

In this paper, we propose intrinsic Muon (**iMuon**), a principled and unified framework for spectral, and, more generally, unitarily invariant norm-constrained optimization on Riemannian matrix manifolds (whose points can be represented as matrices) that resolves both (O1) and (O2) simultaneously. The key observation is that the Riemannian metric canonically induces an *intrinsic* norm on each tangent space, with the Frobenius case recovering the standard Riemannian norm and the spectral case giving a manifold-aware analog of Muon. Our main contributions are:

- **Intrinsic LMO with closed-form solutions on four manifolds (Sections 3 and 4).** We identify two structural properties that together resolve (O1) and (O2) and reduce the intrinsic LMO to a standard Euclidean problem: a left-right product form of the metric (Proposition 3.1) together with either singular value invariance (Definition 3.2) or $\varphi$-contractivity (Definition 3.3) of the scaled tangent space. This yields exact closed-form solutions (Proposition 3.4) on the fixed-rank, SPD, Stiefel, and Grassmann manifolds, for various norms including the spectral, Frobenius, and nuclear norms.
- **Unifying viewpoint.** Beyond enabling efficient computation, instantiating these intrinsic LMOs across different norms and geometries provides a generic approach that unifies established algorithms and yields new ones. In the Euclidean special case [33, 8], our proposed LMO recovers standard SGD, Muon, and NuMuon [50] under different norm ball constraints. On the fixed-rank manifold equipped with quotient geometry, the Frobenius instance recovers Scaled GD [47, 67, 81], while the spectral instance is, to our knowledge, the first quotient-Muon optimizer with an exact closed-form tangent space update.
- **Convergence rate (Section 5).** For both deterministic and stochastic iMuon, we prove an $O(1/\sqrt{T})$ rate in the dual intrinsic norm (up to a logarithmic factor and an additive noise term in the stochastic case) whose constant depends solely on the manifold dimensions (and not on the iterates). On the fixed-rank manifold with the spectral norm, this constant depends only on the rank, yielding a rate that is independent of factor conditioning, and, for example, obviates the runtime factor rescaling of Janson et al. [32].
- **Empirical validation (Section 6).** Experiments across LoRA fine-tuning, image classification, and subspace learning confirm the merits of the intrinsic approach in various settings.

## 2 Background and related work

We briefly review concepts of Riemannian geometry for matrix manifolds (Section 2.1), Muon and norm-constrained Euclidean LMO (Section 2.2), and recent attempts at generalizing the latter to low-rank and matrix manifold settings (Section 2.3). Additional background and related works are discussed in Sections I and J and detailed comparisons are provided in Section H.

### 2.1 Riemannian optimization on matrix manifolds

We consider the problem $\min_{x\in\mathcal{M}} f(x)$ on a Riemannian matrix manifold $\mathcal{M}$ with smooth $f : \mathcal{M} \to \mathbb{R}$. The metric $g_x : T_x\mathcal{M} \times T_x\mathcal{M} \to \mathbb{R}$ on the tangent space $T_x\mathcal{M}$ at $x \in \mathcal{M}$ admits a self-adjoint, positive-definite operator form $g_x(\xi, \zeta) = \langle G_x\xi, \zeta\rangle$, with induced Riemannian norm $\|\xi\|_x = \|G_x^{1/2}\xi\|_F$, where $\|\cdot\|_F$ is the Frobenius norm. The Riemannian gradient $\operatorname{grad} f(x) \in T_x\mathcal{M}$ satisfies $g_x(\operatorname{grad} f, \zeta) = \langle \nabla f(x), \zeta\rangle$ for all $\zeta \in T_x\mathcal{M}$. To update parameters, a retraction operator $R_x : T_x\mathcal{M} \to \mathcal{M}$ generalizes the Euclidean step $x + \eta\xi$ with step size $\eta$. On an *embedded submanifold* $\mathcal{M} \subset \mathbb{R}^{p\times q}$, $T_x\mathcal{M}$ is a linear subspace of the ambient space. On a *quotient manifold* $\mathcal{M} = \overline{\mathcal{M}}/\mathcal{G}$, obtained by symmetry invariance $\mathcal{G}$ in the total space $\overline{\mathcal{M}}$, the metric must be $\mathcal{G}$-invariant for the Riemannian gradient to be well-defined independently of the representative. Please refer to Appendix J and standard references [1, 13] for more details.

### 2.2 Muon and norm-constrained Euclidean optimizers

For an unconstrained matrix parameter $X \in \mathbb{R}^{p\times q}$ and gradient (or momentum) $M \in \mathbb{R}^{p\times q}$, Muon [33, 8, 15] has the update $X_+ = X - \eta\, O$, where $O$ is the solution to:

$$\text{Spectral-norm LMO} \qquad O \;=\; \operatorname*{arg\,max}_{\|O\|_2\le\tau} \langle O, M\rangle \;=\; \tau\, \operatorname{Ortho}(M), \tag{1}$$

with $\operatorname{Ortho}(M) = UV^\top$ being the polar factor of $M$, where $M = U\Sigma V^\top$ is the compact singular value decomposition with $\Sigma \succ 0$. Bernstein and Newhouse [8] formalize (1) as steepest descent under the spectral norm. Subsequent works generalize the LMO viewpoint to broader operator norms [56], modular norms across the network [9, 41], additional nuclear-norm budget [50], per-layer and product norms [18], and analyze convergence, preconditioning, and Newton–Schulz approximations of $\operatorname{Ortho}(\cdot)$ [62, 44, 60, 38, 3, 58, 75]. All of these methods reside in unconstrained Euclidean space and our intrinsic LMO (4) reduces to them in that special case (Section G).

### 2.3 Muon variants for structured and manifold settings

A natural attempt to use (1) on a Riemannian manifold $\mathcal{M}$ is to additionally restrict the search to the tangent space $T_x\mathcal{M}$ at $x \in \mathcal{M}$:

$$\xi^*_{\text{naive}} \;=\; \operatorname*{arg\,max}_{\xi\in T_x\mathcal{M},\ \|\xi\|_2\le\tau} \langle \xi,\, \nabla f(x)\rangle. \tag{2}$$

This naive lift is the implicit template behind several recent manifold/structured Muon proposals but it inherits the two challenges (O1) and (O2) introduced in Section 1.

**LoRA and the fixed-rank manifold.** For LoRA-style fine-tuning $X = BA$ with $B \in \mathbb{R}^{m\times r}, A \in \mathbb{R}^{r\times n}$, $X$ lies on the fixed-rank manifold $\mathcal{M}_r$, equipped with the $\mathrm{GL}(r)$ symmetry $(B, A) \sim (BN^{-1}, NA)$, where $N \in \mathrm{GL}(r)$ the set of $r \times r$ non-singular matrices. Existing works [47, 67, 81] propose Riemannian preconditioned gradient descent for such settings, which corresponds to the Frobenius instance of our framework on $\mathcal{M}_r$. Recently, [32, 34] study the following factor-wise Muon problem, which independently solves a Euclidean spectral-norm LMO on each factor, i.e.,

$$\dot{B}^* = \operatorname*{arg\,max}_{\|\dot{B}\|_2\le\tau}\langle \dot{B},\, \nabla_X f\, A^\top\rangle, \quad \dot{A}^* = \operatorname*{arg\,max}_{\|\dot{A}\|_2\le\tau}\langle \dot{A},\, B^\top\nabla_X f\rangle. \tag{3}$$

Solving (3) requires applying $\operatorname{Ortho}(\cdot)$ to each factor's gradient (or momentum). However, the resulting ambient update is not invariant under the $\mathrm{GL}(r)$ symmetry, exhibiting (O1). Kang et al. [34] prove a continuous-time convergence for the momentum-free limit, conditional on factor-norm boundedness. On the other hand, [32] bounds the ambient spectral norm of the update by an iterate-dependent rescaling $\rho = \eta/(\|B\|_2 + \|A\|_2 + 1)$ computed at runtime. While this circumvents (O1) numerically, our approach (Section 3) handles (O1) algebraically. Please also refer to Remark 5.4 and Section H.1 for more details.

**Embedded manifolds and the coupled tangent space LMO.** Restricting (1) to $T_x\mathcal{M}$ couples the spectral-norm constraint to a tangent-space constraint. The resulting LMO (2) generally has no closed form, which is (O2) described earlier. Riemannion [12] approximately solves the tangent-constrained spectral LMO on the embedded fixed-rank manifold, while MCSD/SPEL [80] drops the tangent-space constraint on the Stiefel manifold and projects the ambient solution onto the Stiefel

manifold. On the Stiefel manifold, the spectral-norm variant of (2) has also been studied in the context of manifold Muon [7, 16, 65], and in a concurrent work [64], which we discuss in Section H.3.

In the next section, we propose an intrinsic LMO that preserves quotient symmetries and identify conditions under which it admits closed-form solutions. In Section 4, we derive these solutions on the fixed-rank, SPD, Stiefel, and Grassmann manifolds.

# 3 The intrinsic LMO

We now generalize norm-constrained Euclidean LMOs, e.g., (1), to the geometry-aware manifold setting. We begin by discussing the key idea that the Riemannian metric itself canonically induces the appropriate notion of matrix norm on the tangent space. Later, we propose a novel family of intrinsic norm-constrained LMOs on the tangent space of manifolds and analyze it with respect to the challenges (O1) and (O2).

**Intrinsic norms on the tangent space.** The metric operator $G_x \succ 0$ induces the Riemannian norm $\|\xi\|_x := \|G_x^{1/2}\xi\|_F$, where $\xi \in T_x\mathcal{M}$ and $\|\cdot\|_F$ is the Frobenius norm. This suggests a canonical extension from the Frobenius norm to any unitarily invariant norm $\varphi$: for $\xi \in T_x\mathcal{M}$, we define the intrinsic $\varphi$-norm by $\|\xi\|_{\varphi,x} := \varphi(G_x^{1/2}\xi)$. In particular, choosing $\varphi = \|\cdot\|_2$, $\|\cdot\|_F$, or $\|\cdot\|_*$ yields the intrinsic spectral, Frobenius, and nuclear norms, respectively. We note that $\|\cdot\|_{\varphi,x}$ is intrinsic in the sense that it is determined entirely by the Riemannian metric. It reduces to the standard matrix norm in the Euclidean case $G_x = I$, while for $\varphi = \|\cdot\|_F$ it recovers the Riemannian norm itself.

## 3.1 Proposed LMO formulation

Given the intrinsic norm $\|\cdot\|_{\varphi,x}$ introduced above, we propose the update direction at $x \in \mathcal{M}$ for optimizing $\min_{x\in\mathcal{M}} f(x)$ through the following LMO:

$$\xi^* = \underset{\xi\in T_x\mathcal{M},\ \varphi(G_x^{1/2}\xi)\le\tau}{\arg\max} g_x(\xi,\ \mathrm{grad}\, f), \tag{4}$$

where $\varphi$ is any unitarily invariant norm and $g_x$ is the metric defined on the tangent space $T_x\mathcal{M}$ at $x \in \mathcal{M}$. The iterate is then updated by retracting along the $-\xi^*$ direction: $x_+ = R_x(-\eta\,\xi^*)$ with step size $\eta > 0$. Compared with the naive tangent space LMO (2), both the objective $g_x(\xi, \mathrm{grad}\, f)$ and the feasible set $\{\xi \in T_x\mathcal{M} : \|\xi\|_{\varphi,x} \le \tau\}$ in the proposed LMO (4) are now intrinsic to the manifold geometry.

Our formulation recovers several familiar special cases. If $\mathcal{M}$ is a Euclidean space with the standard metric, then (4) reduces exactly to the usual Euclidean norm-constrained LMO. If $\varphi = \|\cdot\|_F$ and $\tau = 1$, then (4) returns the normalized Riemannian gradient direction [1, 13]. In the next sections, we show that our LMO formulation simultaneously preserves the quotient symmetries (resolving (O1)) and admits efficient closed-form solutions (resolving (O2)).

## 3.2 Symmetry preservation

We next analyze the proposed intrinsic constraint in (4) under a group $\mathcal{G}$ of symmetries of the metric. This covers quotient manifolds $\mathcal{M} = \bar{\mathcal{M}}/\mathcal{G}$ ($\bar{\mathcal{M}}$ being the total space) with a $\mathcal{G}$-invariant Riemannian metric $g_x$, where $\mathcal{G}$ acts on representatives, as well as groups acting on $\mathcal{M}$ by isometries, such as affine-invariance in the context of symmetric positive-definite matrices. In such settings, we show that our intrinsic constraint respects the underlying symmetry if the metric admits a left-right form: $g_x(\xi,\zeta) = \mathrm{tr}(\xi^\top P_x \zeta Q_x)$, or equivalently, the metric operator factors as $G_x\xi = P_x\xi Q_x$, where $P_x$ and $Q_x$ are symmetric positive-definite matrices depending smoothly on $x$. Then, $G_x^{1/2}\xi = P_x^{1/2}\xi Q_x^{1/2}$ and the intrinsic norm reads $\|\xi\|_{\varphi,x} = \varphi(P_x^{1/2}\xi Q_x^{1/2})$.

In Section 4, we discuss popular manifolds where the metric has the left-right form. It should be noted that the intrinsic norm itself is well-defined for any positive self-adjoint $G_x$ and does not require the left-right factorization. However, under such symmetries, this factorization makes the metric scaling explicit and the symmetry behavior of the intrinsic norm transparent. Our next result shows that, under such metrics and a compatible transformation of the factors $P_x, Q_x$ under $\mathcal{G}$, the intrinsic norm constraint is $\mathcal{G}$-invariant.

**Proposition 3.1** (Symmetry invariance of the intrinsic constraint)**.** *Let a group $\mathcal{G}$ act on $\mathcal{M}$ (on representatives in $\overline{\mathcal{M}}$ when $\mathcal{M} = \overline{\mathcal{M}}/\mathcal{G}$), with $\tilde{x} = h \cdot x$ and induced tangent action $\xi \mapsto \tilde{\xi} = L_h \, \xi \, R_h^\top$. Suppose the metric operator on $T_x\mathcal{M}$ is $G_x\xi = P_x \, \xi \, Q_x$ with $P_x, Q_x \succ 0$, and $P_{\tilde{x}} = L_h^{-\top} P_x L_h^{-1}$, $Q_{\tilde{x}} = R_h^{-\top} Q_x R_h^{-1}$. Then $g$ is $\mathcal{G}$-invariant and $\varphi(G_{\tilde{x}}^{1/2}\tilde{\xi}) = \varphi(G_x^{1/2}\xi)$ for all $\xi \in T_x\mathcal{M}$ and every unitarily invariant norm $\varphi$. This holds on the manifolds considered in Section 4: the fixed-rank manifold $\mathcal{M}_r$ with $\mathrm{GL}(r)$ (block-wise), the Grassmann manifold $\mathrm{Gr}(m,r) = \mathrm{St}(m,r)/O(r)$, and the manifold $S_{++}^n$ of symmetric positive-definite matrices with $\mathrm{GL}(n)$ acting by congruence (an isometry group).*

The proof is given in Section D.1. With respect to the proposed LMO problem (4), the above result shows that the feasible set $\{\xi : \varphi(G_x^{1/2}\xi) \leq \tau\}$ is invariant under the symmetry group $\mathcal{G}$. When $f$ is $\mathcal{G}$-invariant, as on quotient manifolds, the objective $g_x(\xi, \operatorname{grad} f)$ is preserved as well, so the optimal update is equivariant, i.e., $\tilde{\xi}^* = L_h \, \xi^* R_h^\top$, and the resulting update does not depend on the chosen representative. Thus, our result shows that our intrinsic norm generalizes the Frobenius-norm invariance (that makes the Riemannian gradient well-defined on quotients [1, 46, 13]) to all unitarily invariant norms.

### 3.3 Efficient and optimal solution to the coupled norm and tangent-space constraints

A key computational challenge in solving (4) efficiently is the simultaneous presence of two constraint sets: $\{\xi : \xi \in T_x\mathcal{M}\}$ and $\{\xi : \varphi(G_x^{1/2}\xi) \leq \tau\}$. We now show that, under certain conditions, these two constraints can be simplified, which allows for obtaining closed-form solutions of (4).

In this regard, we first define the scaled tangent space $\mathcal{Z}_x \coloneqq \{G_x^{1/2}\eta : \eta \in T_x\mathcal{M}\}$. Since $G_x^{1/2}$ is invertible on $T_x\mathcal{M}$, the map $\xi \mapsto Z \coloneqq G_x^{1/2}\xi$ is an isomorphism between $T_x\mathcal{M}$ and $\mathcal{Z}_x$. The intrinsic norm may now be rewritten as $\varphi(G_x^{1/2}\xi) = \varphi(Z)$. Moreover, if we define the scaled gradient $H \coloneqq G_x^{1/2}\operatorname{grad} f$, then the objective of (4) satisfies $g_x(\xi, \operatorname{grad} f) = \langle Z, H\rangle$. To formalize our next result, we also require the following notions.

**Definition 3.2.** A linear subspace $\mathcal{S} \subseteq \mathbb{R}^{p\times q}$ is *singular value (SV) invariant* if, for every $Z \in \mathcal{S}$ with compact SVD $Z = U\Sigma V^\top$ ($\Sigma \succ 0$ diagonal), we have $UDV^\top \in \mathcal{S}$ for every nonnegative diagonal $D$ satisfying $D_{ii} = D_{jj}$ whenever $\Sigma_{ii} = \Sigma_{jj}$.

**Definition 3.3.** Let $\varphi$ be a unitarily invariant norm. A linear subspace $\mathcal{S} \subseteq \mathbb{R}^{p\times q}$ is *$\varphi$-contractive* if the orthogonal projection $P_{\mathcal{S}}$ onto $\mathcal{S}$ satisfies $\varphi(P_{\mathcal{S}}(W)) \leq \varphi(W)$ for all $W \in \mathbb{R}^{p\times q}$.

For example, every linear subspace is $\|\cdot\|_F$-contractive, and $\mathrm{Skew}(r)$ is $\varphi$-contractive for every unitarily invariant $\varphi$ (Lemma D.1).

**Proposition 3.4** (Closed-form solution)**.** *Let $\mathcal{Z}_x \coloneqq \{G_x^{1/2}\eta : \eta \in T_x\mathcal{M}\}$, let $\varphi$ be a unitarily invariant norm, and suppose $\mathcal{Z}_x$ is SV-invariant or $\varphi$-contractive. Then the solution of (4) is given by*

$$\xi^* = G_x^{-1/2} P_{\mathcal{Z}_x}(Z^*), \text{ where } Z^* = \operatorname*{arg\,max}_{\varphi(Z)\leq\tau} \langle Z, H\rangle, \tag{5}$$

*$H \coloneqq G_x^{1/2}(\operatorname{grad} f)$, and $P_{\mathcal{Z}_x}$ is the orthogonal projection onto $\mathcal{Z}_x$. Moreover, $Z^*$ can be computed as $Z^* = U \operatorname{diag}(z^*)V^\top$, where $H = U \operatorname{diag}(\sigma)V^\top$ is the compact SVD of $H$ and $z^*$ solves the vector problem $\max_{z:\varphi(z)\leq\tau}\langle z, \sigma\rangle = \tau\varphi^\circ(\sigma)$ with $z_i^* = z_j^*$ whenever $\sigma_i = \sigma_j$. With this choice, $P_{\mathcal{Z}_x}(Z^*) = Z^*$ when $\mathcal{Z}_x$ is SV-invariant.*

The proof is given in Section D.2. We note that the vector problem ($\arg\max_{z:\varphi(z)\leq\tau}\langle z, \sigma\rangle$) admits a closed-form solution for various unitarily invariant norms, including the spectral, Frobenius, and nuclear norms (Section C). In Section 4, we discuss several manifolds of interest where Proposition 3.4 is useful. In Section P, we also present a broader discussion including cases where SV invariance of $\mathcal{Z}_x$ or the left-right form of the metric fail.

## 4 Application of intrinsic LMO (4) on matrix manifolds

In this section, we discuss the closed-form solutions of the proposed LMO (4) for various manifolds and intrinsic spectral/Frobenius/nuclear norms.

**Product of tangent spaces.** When the tangent space decomposes as a product $T_x\mathcal{M} = \mathbb{R}^{p_1\times q_1} \times \cdots \times \mathbb{R}^{p_k\times q_k}$ and $G_x$ has the corresponding block structure, then $G_x^{1/2}$ acts component-wise. In such settings, we may choose $\varphi(Z_1,\dots,Z_k)$ as the $\ell_\infty$-product norm, i.e., $\varphi(Z_1,\dots,Z_k) \coloneqq \max_i \varphi(Z_i)$, as it decouples the intrinsic LMO into $k$ independent per-block subproblems:

$$\{Z_i^*\}_{i=1}^k = \operatorname*{arg\,max}_{\varphi(Z_1,\dots,Z_k)\le\tau} \sum\nolimits_i \langle Z_i,\, H_i\rangle \;\equiv\; Z_i^* = \operatorname*{arg\,max}_{\varphi(Z_i)\le\tau} \langle Z_i,\, H_i\rangle, \qquad i=1,\dots,k, \quad (6)$$

each solvable in closed form via Proposition 3.4, since the constraint $\max_i \varphi(Z_i) \le \tau$ separates into $\varphi(Z_i) \le \tau$ for each block independently. Interestingly, [18] has also studied $\ell_\infty$-based product norms in the context of combining norms across neural network layers.

**Fixed-rank manifold.** We now consider the set $\mathcal{M}_r$ of rank-$r$ matrices parameterized as $X = BA$ with $B \in \mathbb{R}_*^{m\times r}$, $A \in \mathbb{R}_*^{r\times n}$ (where $\mathbb{R}_*^{\cdot}$ denotes full-rank matrices). We note that the factorization $X = BA$ has the quotient symmetry: $(B, A) \sim (BN^{-1}, NA)$ for $N \in \mathrm{GL}(r)$, where $\mathrm{GL}(r)$ denotes the group of $r \times r$ invertible matrices. This results in $\mathcal{M}_r = (\mathbb{R}_*^{m\times r} \times \mathbb{R}_*^{r\times n})/\mathrm{GL}(r)$ and $T_x\mathcal{M}_r = \mathbb{R}^{m\times r} \times \mathbb{R}^{r\times n}$ [1, 48]. If $x = (B, A) \in \mathcal{M}_r$ and $\xi_i = (\dot B_i, \dot A_i) \in T_x\mathcal{M}_r$ for $i = 1, 2$, then the coupled metric is $g_x(\xi_1, \xi_2) = \mathrm{tr}(\dot B_1^\top \dot B_2\, AA^\top) + \mathrm{tr}(B^\top B\, \dot A_1 \dot A_2^\top)$ [47, 46, 81]. We observe that the metric is $\mathrm{GL}(r)$-invariant with the left-right form (Proposition 3.1) and has a block structure, with $G_x^{1/2}\xi_i = (\dot B_i(AA^\top)^{1/2},\ (B^\top B)^{1/2}\dot A_i)$. Thus, we decouple the LMO into $\dot B$ and $\dot A$ subproblems (6).

**Corollary 4.1** (Intrinsic LMO (4) for fixed-rank manifold)**.** *Let $\mathcal{M} = \mathcal{M}_r$ and at $X = BA \in \mathcal{M}_r$, let $\nabla_X f$ denote the Euclidean gradient w.r.t. $X$. Then, with the above discussed $\mathrm{GL}(r)$-invariant metric and $\varphi = \|\cdot\|_2$, an optimal solution of the proposed LMO (4) under the product constraint $\max\{\|\dot B(AA^\top)^{1/2}\|_2, \|(B^\top B)^{1/2}\dot A\|_2\} \le \tau$ is $\xi^* = (\dot B^*, \dot A^*)$, where $\dot B^* = \tau\,\mathrm{Ortho}((\nabla_B f)(AA^\top)^{-1/2})(AA^\top)^{-1/2}$, $\dot A^* = \tau(B^\top B)^{-1/2}\,\mathrm{Ortho}((B^\top B)^{-1/2}\nabla_A f)$, $\nabla_B f = \nabla_X f A^\top$, and $\nabla_A f = B^\top \nabla_X f$.*

The proof is given in Section D.3. We note that the intrinsic spectral LMO in Corollary 4.1 differs from prior $\mathcal{M}_r$-Muon constructions in two ways: (i) Factor-wise Muon [32, 34] solves (3) by applying $\mathrm{Ortho}(\cdot)$ to each factor gradient. The resulting ambient update is not $\mathrm{GL}(r)$-invariant and its spectral norm scales with $\|A\|_2 + \|B\|_2$. Hence, Spectron [32] rescales the solution of (3) during runtime to bound the ambient spectral norm. (ii) Riemannion [12] works in the embedded view with $G_X = I$ and approximates the tangent-constrained LMO by a rank-$2r$ truncated polar followed by tangent space re-projection. Therefore, the spectral-norm bound is satisfied only approximately. Corollary 4.1, in contrast, is the exact maximizer of (4) on each factor block, and its Frobenius instance recovers Riemannian preconditioned gradient for fixed-rank matrices [46, 67, 81]. Section H.1 provides detailed update-rule and norm-bound comparisons.

**SPD manifold.** The SPD manifold $S_{++}^n$ is the space of all $n \times n$ symmetric positive-definite matrices. At any $X \in S_{++}^n$, $T_X S_{++}^n = \mathrm{Sym}(n)$, the space of $n \times n$ symmetric matrices. Since $f$ is defined on symmetric matrices, the Euclidean gradient $\nabla_X f \in \mathrm{Sym}(n)$ is itself symmetric. Consider $S_{++}^n$ with the affine-invariant metric: $g_X(\xi, \zeta) = \mathrm{tr}(X^{-1}\xi X^{-1}\zeta)$ [10]. Both Propositions 3.1 and 3.4 are satisfied with $G_X^{1/2}\xi = X^{-1/2}\xi X^{-1/2}$ and the Riemannian gradient is $\mathrm{grad}\, f(X) = X(\nabla_X f)X$. The solution of the proposed LMO (4) with $\varphi = \|\cdot\|_2$ is $\xi^* = \tau\, X^{1/2}\,\mathrm{sign}(H)\, X^{1/2}$, where $H = G_X^{1/2}(\mathrm{grad}\, f) = X^{1/2}(\nabla_X f)X^{1/2}$ and $\mathrm{sign}(H) = Q\,\mathrm{diag}(\mathrm{sign}(\lambda_i))Q^\top$ [27] with $Q$ and $\{\lambda_i\}_i$ being eigenvectors and eigenvalues of $H$, respectively. We also note that the intrinsic Frobenius norm constraint in this setting recovers the natural gradient [2]. Alternative SPD metrics are discussed in Section P.5.

**Stiefel manifold.** The Stiefel manifold $\mathrm{St}(m, r) = \{X \in \mathbb{R}^{m\times r} : X^\top X = I_r\}$ is the set of $m \times r$ matrices with orthonormal columns, and its tangent space at $X$ is $T_X\mathrm{St} = \{\xi \in \mathbb{R}^{m\times r} : X^\top\xi + \xi^\top X = 0\}$. We equip $\mathrm{St}(m, r)$ with the Euclidean metric, i.e., $G_X = I$. Any tangent vector $\xi \in T_X\mathrm{St}$ can be uniquely parametrized as $\xi = [X, X_\perp][A; B]$, where $A \in \mathrm{Skew}(r)$ (set of $r \times r$ skew-symmetric matrices), $B \in \mathbb{R}^{(m-r)\times r}$, $[A; B] \coloneqq [A^\top, B^\top]^\top$, and $X_\perp \in \mathbb{R}^{m\times(m-r)}$ is an orthonormal basis for the orthogonal complement of $\mathrm{col}(X)$ [20, 1, 13]. Since $G_X = I$, the space $\mathcal{Z}_X = \{G_X^{1/2}\eta : \eta \in T_X\mathcal{M}\}$ is the same as $T_X\mathcal{M}$. As $\mathcal{Z}_X$ is not SV-invariant, Proposition 3.4 cannot be directly applied to $\mathrm{St}(m, r)$.

**Algorithm 1** Proposed Intrinsic Muon (iMuon) algorithm

**Require:** Initial point $x_0 \in \mathcal{M}$, step size $\eta > 0$, norm bound $\tau > 0$, norm $\varphi$, iterations $T$
1: **for** $t = 0, 1, \ldots, T-1$ **do**
2: Compute Euclidean gradient $\nabla f(x_t)$
3: Obtain $\xi_t^*$ by solving intrinsic LMO (4) using $\nabla f(x_t)$ (refer to Section 4 for specific cases)
4: Retract: $x_{t+1} = R_{x_t}(-\eta\, \xi_t^*)$
5: **end for**
6: **return** $x_T$

However, we note that $\varphi(\xi) = \varphi([A; B])$ and $g_X(\xi, \operatorname{grad} f) = \langle A,\, X^\top \nabla_X f\rangle + \langle B,\, X_\perp^\top \nabla_X f\rangle$. Furthermore, we employ the $\ell_\infty$-product norm $\max\{\varphi(A), \varphi(B)\} \le \tau$ to decouple the LMO into a skew and a normal subproblem as in (6), where the skew subproblem retains the constraint $A \in \operatorname{Skew}(r)$. In the spectral case, this relaxes the joint norm, since $\max\{\|A\|_2, \|B\|_2\} \le \|[A; B]\|_2 \le \sqrt{2}\max\{\|A\|_2, \|B\|_2\}$. Both blocks are SV-invariant (Section E), so Proposition 3.4 gives the following exact solution for this product-norm formulation.

**Lemma 4.2** (Intrinsic LMO (4) for Stiefel manifold)**.** *Let $\mathcal{M} = \operatorname{St}(m, r)$ and at $X \in \mathcal{M}$, let $G_X = I$. Let $\nabla_X f$ denote the Euclidean gradient w.r.t. $X$. Then, with $\varphi = \|\cdot\|_2$, an optimal solution of the proposed LMO (4) under the product constraint $\max\{\|A\|_2, \|B\|_2\} \le \tau$ is $\xi^* = \tau(X \operatorname{skew}(\operatorname{Ortho}(\operatorname{skew}(X^\top \nabla_X f))) + X_\perp \operatorname{Ortho}(X_\perp^\top \nabla_X f))$.*

The proof is given in Section D.4. Recently, Yang and Lai [80] studied LMO-based updates in the context of optimization over the Stiefel manifold. However, they sidestep the coupled LMO problem by solving the LMO in the ambient space and projecting the result onto the Stiefel manifold. Thus, the update direction need not lie in $T_X\mathcal{M}$. Other recent approaches to the joint-norm formulation are discussed in Section H.3.

**Grassmann manifold.** The Grassmann manifold $\operatorname{Gr}(m, r)$ is the set of $r$-dimensional linear subspaces of $\mathbb{R}^m$, represented by an orthonormal basis $X \in \mathbb{R}^{m\times r}$ with $X^\top X = I_r$ up to right multiplication by $O(r)$, i.e., $\operatorname{Gr}(m, r) = \operatorname{St}(m, r)/O(r)$ [20, 13]. However, unlike the Stiefel setting, the tangent space (horizontal space) $T_X\operatorname{Gr} = \{\xi : X^\top \xi = 0\}$ is SV-invariant. Using Proposition 3.4, we get the optimal solution of our intrinsic LMO (4) for $\varphi = \|\cdot\|_2$ as $\xi^* = \tau\ \operatorname{Ortho}((I - XX^\top)\nabla_X f)$.

**Overall.** A salient feature of our above analysis across various manifolds is that computing the optimal solution of the proposed problem (4) does not require the computation of the Riemannian gradient. Our closed-form optimal expressions rely only on Euclidean gradients, which are available via back-propagation. Please refer to Section F for proofs and additional results. In particular, Table 4 in Section F.1 states, for each manifold, whether the update solves the joint LMO Equation (4) or the product LMO Equation (6).

## 5 Convergence

In this section, we discuss the convergence analysis of our algorithm intrinsic Muon (**iMuon**), Algorithm 1 in both deterministic and stochastic settings. Our convergence analysis is parametrized by a geometric quantity $C_\varphi$, which captures the *size* of the constraint set in (4), defined as follows.

**Definition 5.1.** The squared Riemannian radius of the $\varphi$-ball on $T_x\mathcal{M}$ is $C_\varphi(x) \coloneqq \max\{\|\xi\|_x^2 : \xi \in T_x\mathcal{M}, \varphi(G_x^{1/2}\xi) \le 1\}$. Also define $C_\varphi \coloneqq \sup_{x\in\mathcal{M}} C_\varphi(x)$.

For $\varphi = \|\cdot\|_2$: $C_\varphi = 2r$ on $\mathcal{M}_r$, $n$ on $S_{++}^n$, $2\lfloor r/2\rfloor + \min(m-r, r)$ on $\operatorname{St}(m, r)$, $\min(m-r, r)$ on $\operatorname{Gr}(m, r)$. For $\varphi \in \{\|\cdot\|_F, \|\cdot\|_*\}$, $C_\varphi = O(1)$ on all manifolds. On all four manifolds, $C_\varphi$ depends only on the manifold dimensions, not on the iterates.

**Assumption 5.2** (Retraction-$L$-smoothness [13, Assumption 4.3])**.** There exist constants $L > 0$ and $\overline{\eta} > 0$ such that for all $x \in \mathcal{M}$, all $\xi \in T_x\mathcal{M}$, and all $\eta \in (0, \overline{\eta}]$, $f(R_x(-\eta\xi)) \le f(x) - \eta\, g_x(\xi, \operatorname{grad} f) + \frac{L\eta^2}{2}\|\xi\|_x^2$.

The next theorem shows that our algorithm iMuon converges at rate $O(1/\sqrt{T})$ with the constant determined by $C_\varphi$.

**Theorem 5.3** (Deterministic iMuon)**.** *Under Assumption 5.2, with constant step size* $\eta = c/(\tau\sqrt{T})$ *where* $c = \sqrt{2\Delta_0/(LC_\varphi)}$ *(assuming* $\eta \le \overline{\eta}$*), our Algorithm 1 achieves* $\min_{t\le T} \varphi^\circ(H_t) \le \sqrt{2LC_\varphi\Delta_0/T}$*, where* $\Delta_0 = f(x_0) - f^*$*,* $\varphi^\circ$ *is the dual norm of* $\varphi$*, and* $H_t = G_{x_t}^{1/2}(\operatorname{grad} f(x_t))$*. With a horizon-free decaying schedule defined as* $\eta_t = c/(\tau\sqrt{t+1})$*, Algorithm 1 achieves:* $\min_{t\le T} \varphi^\circ(H_t) \le \frac{\Delta_0(2+\ln T)}{2c(\sqrt{T}-1)}$.

The proof is provided in Section D.5. We note that the convergence criterion $\varphi^\circ(H_t) \to 0$ implies $\|\operatorname{grad} f(x_t)\|_{x_t} \to 0$ (first-order stationarity) as $H_t = G_{x_t}^{1/2}(\operatorname{grad} f(x_t))$ and $G_{x_t}^{1/2} \succ 0$. For $\varphi = \|\cdot\|_F$, the rate recovers standard Riemannian gradient descent with $C_\varphi = 1$. Our key technical ingredient is the uniform bound $\|\xi_t^*\|_{x_t}^2 \le C_\varphi\tau^2$ (Lemma D.3). We also note manifold specific convergence rates below for constant step size setting.

*Remark* 5.4 (Fixed-rank). On $\mathcal{M}_r$ with $\varphi = \|\cdot\|_2$, our rate is $O(\sqrt{Lr/T})$, depending only on the rank $r$ and independent of the factor norms $\|B\|, \|A\|$. In contrast, Kang et al. [34] show in continuous time that, from almost every initialization, the trajectory either converges to a global minimum or its factor norms diverge, and they obtain unconditional global convergence only with $\ell_2$ regularization. Janson et al. [32] instead apply an iterate-dependent spectral renormalization at runtime.

*Remark* 5.5 (SPD, Stiefel, Grassmann). Our convergence rates are $O(\sqrt{Ln/T})$ on $S_{++}^n$, $O(\sqrt{L(r+\min(m-r,r))/T})$ on $\mathrm{St}(m,r)$, and $O(\sqrt{L\min(m-r,r)/T})$ on $\mathrm{Gr}(m,r)$.

In the stochastic setting, stochastic gradients ($\widetilde{\nabla f}$) are available and we appropriately employ the Riemannian stochastic gradient ($\widetilde{\operatorname{grad} f}$) in the proposed LMO (4) to obtain the update direction $\tilde{\xi}_t^*$.

**Assumption 5.6.** There exists $\sigma_\varphi > 0$ such that the stochastic scaled gradient $\tilde{H}_t := G_{x_t}^{1/2}(\widetilde{\operatorname{grad} f_t})$ satisfies $\mathbb{E}[\varphi^\circ(\tilde{H}_t - H_t)^2 \mid x_0, \dots, x_t] \le \sigma_\varphi^2$, where $H_t := G_{x_t}^{1/2}(\operatorname{grad} f_t)$ is the true scaled gradient.

For $\varphi = \|\cdot\|_F$, this reduces to the standard Riemannian bounded-variance condition [13, 51].

**Theorem 5.7** (Stochastic iMuon)**.** *Under Assumptions 5.2 & 5.6, with step size* $\eta_t = c/(\tau\sqrt{t+1})$ *where* $c = \sqrt{2\Delta_0/(LC_\varphi)}$ *(assuming* $\eta_t \le \overline{\eta}$*), our Algorithm 1 achieves* $\min_{t\le T} \mathbb{E}[\varphi^\circ(H_t)] \le \sqrt{LC_\varphi\Delta_0/2}(2+\ln T)/\sqrt{T} + 2\sigma_\varphi$*, recovering Theorem 5.3 (up to the* $\ln T$ *factor from the decaying schedule) when* $\sigma_\varphi = 0$.

The proof is provided in Section D.8. The additive $2\sigma_\varphi$ arises from the lack of variance reduction [83, 36, 59] in the gradient estimator and can be removed with gradient averaging or momentum techniques [51, 63], which we leave to future work.

Several prior results arise as instances of our framework (see Section G for details): Euclidean spectral steepest descent [8] ($G_x = I$, $C_\varphi = \min(p,q)$), standard Riemannian GD [13] ($\varphi = \|\cdot\|_F$, $C_\varphi = 1$), factor-wise Muon [34] (the coupled metric removes its factor-norm dependence, $C_\varphi = 2r$), the rank-1 special case of NuMuon [50] ($\varphi = \|\cdot\|_*$, $C_\varphi = 1$), and general Euclidean LMOs [56] ($G_x = I$).

## 6 Experiments

We evaluate the proposed intrinsic LMO on two manifolds: (i) the fixed-rank manifold $\mathcal{M}_r$ via LoRA fine-tuning of GPT-2 Medium on E2E NLG and Mistral-7B on GLUE, and (ii) the SPD manifold $S_{++}^n$ via SPD covariance classification on CIFAR-100. In LoRA fine-tuning, we compare iMuon with factor-wise Euclidean Muon [33, 8] (as in (3)), SGD (factor-wise with Frobenius norm), Scaled GD [81] (preconditioned), Spectron [32], and Riemannion [12]. With momentum, SGD and Scaled GD become AdamW and Scaled AdamW. In the SPD setting, we include the Frobenius and nuclear instances of our framework (RGD, iMuon-Nu) and their Euclidean counterparts (EGD, NuMuon). Full details and additional experiments including those on Stiefel and Grassmann manifolds are deferred to Sections K to O. The code to reproduce our experiments is available at `https://github.com/1bang118/manifold-intrinsic-muon`.

**LoRA fine-tuning ($\mathcal{M}_r$).** We fine-tune GPT-2 Medium on E2E NLG [53] with LoRA rank $r = 4$, and 4-bit quantized Mistral-7B on GLUE [71] with LoRA rank $r = 16$. We follow the LoRA setup

Table 1: E2E NLG results with GPT-2 Medium and LoRA rank $r = 4$, reported separately for methods without and with optimizer momentum. For each method, we report the run with the best test-set BLEU. The remaining metrics are computed from the same run. Within each block, the best value in each column is in bold and the second best is underlined.

| Method | BLEU | NIST | METEOR | ROUGE-L | CIDEr |
|---|---|---|---|---|---|
| *Without momentum* | | | | | |
| SGD | 66.60 | 8.54 | 44.20 | 68.20 | 2.32 |
| Scaled GD | 69.20 | 8.71 | 46.30 | 70.90 | 2.48 |
| Factor-wise Muon | 70.02 | <u>8.81</u> | 46.77 | 71.68 | **2.53** |
| Riemannion (SGD) | 70.02 | 8.78 | <u>46.79</u> | <u>71.99</u> | <u>2.52</u> |
| Spectron | <u>70.25</u> | <u>8.81</u> | **46.85** | 71.81 | 2.51 |
| iMuon | **70.30** | **8.85** | 46.62 | **72.03** | **2.53** |
| *With momentum* | | | | | |
| AdamW | 68.90 | 8.69 | 46.50 | 71.30 | 2.51 |
| Scaled AdamW | 69.60 | <u>8.77</u> | 46.60 | 71.80 | <u>2.52</u> |
| Factor-wise Muon | <u>70.26</u> | **8.83** | 46.76 | 71.75 | <u>2.52</u> |
| Riemannion (Adam) | 70.21 | **8.83** | 46.55 | <u>71.87</u> | **2.53** |
| Spectron | 69.84 | <u>8.77</u> | **46.98** | 71.84 | 2.50 |
| iMuon | **70.41** | **8.83** | <u>46.84</u> | **71.96** | <u>2.52</u> |

of Hu et al. [29], tuning the learning rate per method via grid search. We also evaluate a momentum variant of iMuon that transports the momentum buffer by horizontal projection (Section K).

Tables 1 and 2 report the five standard E2E metrics and the GLUE per-task and average scores, separately for methods without and with momentum. The momentum-free block isolates the LMO direction itself, since momentum schemes differ across methods and accumulate gradients over iterations. On E2E NLG, iMuon without momentum is best on BLEU, NIST, and ROUGE-L and ties factor-wise Muon for the best CIDEr, and with momentum it is best or second best on all five metrics. Momentum improves its BLEU and METEOR and leaves the other metrics nearly unchanged. Across both blocks, iMuon attains the best BLEU, NIST, ROUGE-L, and (tied) CIDEr. On GLUE, the advantage without momentum is more pronounced: iMuon has the best average. With momentum, its average ($88.76$) is second only to Scaled AdamW ($89.01$). Across both blocks, iMuon attains the best SST-2 and STS-B scores and ties for the best WNLI. We attribute iMuon's edge over Spectron in both tables (e.g., GLUE average $88.15$ vs. $87.53$ without and $88.76$ vs. $87.58$ with momentum) to geometry: Spectron's scalar rescaling bounds the norm of the ambient update, but the ambient update still changes under the factor rescaling $(B, A) \mapsto (cB, A/c)$ that leaves $X = BA$ unchanged (Section H.1), whereas iMuon's $\mathrm{GL}(r)$-invariant metric makes its ambient update independent of the factorization (Proposition 3.1).

**SPD prototype classification ($S^n_{++}$).** Covariance matrices of local image features have been used as SPD descriptors since region covariances [68], e.g., for pedestrian detection [69] and in SPD neural networks [30]. We classify CIFAR-100 images into the $20$ coarse classes [40] using shrinkage covariance descriptors $C_i \in S^{32}_{++}$ [42] of frozen ResNet-18 `layer3` features [26]. As a Riemannian analog of nearest-class-mean classification, we learn one SPD prototype $X_c$ per class under the affine-invariant (AI) metric [55, 10] by minimizing the cross-entropy of the logits $-\beta\, d_{\mathrm{AI}}(C_i, X_c)^2$, regularized toward the log-Euclidean [4] class-mean initialization (Section M). Since the AI metric is not Euclidean ($G_X \neq I$), the intrinsic and Euclidean LMOs differ for every norm, so we compare all three pairs: Frobenius (RGD vs. EGD), spectral (iMuon vs. Muon), and nuclear (iMuon-Nu vs. NuMuon, here the pure nuclear-norm LMO [50], i.e., (1) with the nuclear norm in place of the spectral norm).

Figure 1 shows the test-accuracy trajectories. At the end of training, the intrinsic method improves the mean test accuracy in every pair: RGD $0.548 \pm 0.002$ vs. EGD $0.523 \pm 0.026$ (Frobenius), iMuon $0.475 \pm 0.009$ vs. Muon $0.407 \pm 0.016$ (spectral), and iMuon-Nu $0.546 \pm 0.000$ (std below $0.0005$) vs. NuMuon $0.422 \pm 0.032$ (nuclear), over three seeds at the validation-selected learning rate. For the Frobenius pair, this is consistent with the known advantages of the AI geometry [55, 10]. The spectral and nuclear pairs show that these advantages carry over to LMO directions. The gain grows

Table 2: GLUE results with 4-bit NF4-quantized Mistral-7B and LoRA rank $r = 16$, reported separately without and with optimizer momentum. Avg. is the unweighted average over the nine tasks. Within each block, the best value in each column is in bold and the second best is underlined.

| Method | MNLI | SST-2 | MRPC | CoLA | QNLI | QQP | RTE | STS-B | WNLI | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|
| *Without momentum* | | | | | | | | | | |
| SGD | 88.15 | 96.10 | 70.10 | 55.89 | 94.22 | 88.59 | 50.90 | 47.64 | 49.30 | 71.21 |
| Scaled GD | <u>90.21</u> | <u>96.90</u> | 81.62 | 68.17 | 94.40 | **91.15** | 54.15 | 90.31 | 56.34 | 80.36 |
| Factor-wise Muon | 89.77 | **97.25** | 82.35 | <u>69.32</u> | 94.47 | 88.42 | 84.12 | 84.39 | 45.07 | 81.68 |
| Riemannion (SGD) | **91.55** | <u>96.90</u> | 82.60 | 68.90 | **94.93** | 87.77 | 82.67 | 77.99 | 60.56 | 82.65 |
| Spectron | 89.97 | 96.79 | **90.69** | 66.62 | 94.34 | <u>88.84</u> | <u>87.73</u> | <u>91.12</u> | **81.69** | <u>87.53</u> |
| iMuon | 90.01 | <u>96.90</u> | <u>89.71</u> | **71.49** | <u>94.67</u> | 88.31 | **89.89** | **92.05** | <u>80.28</u> | **88.15** |
| *With momentum* | | | | | | | | | | |
| AdamW | 89.86 | 96.79 | 88.48 | 71.05 | 94.42 | 91.24 | <u>90.61</u> | 90.42 | <u>81.69</u> | 88.28 |
| Scaled AdamW | 90.68 | <u>97.25</u> | <u>89.46</u> | 71.30 | 94.67 | **92.22** | **91.34** | 91.10 | **83.10** | **89.01** |
| Factor-wise Muon | 91.29 | 96.90 | **89.71** | <u>71.45</u> | <u>94.97</u> | <u>91.74</u> | 89.89 | <u>92.18</u> | 78.87 | 88.56 |
| Riemannion (Adam) | 91.21 | 96.44 | 79.90 | 67.00 | 94.55 | 90.36 | 80.87 | 62.13 | **83.10** | 82.84 |
| Spectron | **91.60** | 97.13 | 87.25 | **72.30** | **95.28** | 91.43 | 88.81 | 91.22 | 73.24 | 87.58 |
| iMuon | <u>91.50</u> | **97.59** | 88.48 | 71.18 | 94.71 | 91.65 | 88.09 | **92.52** | **83.10** | <u>88.76</u> |

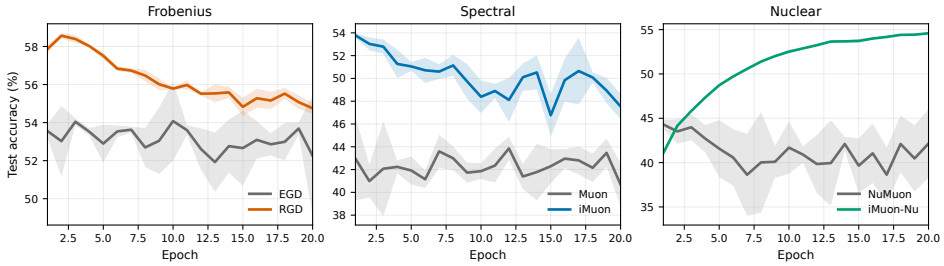



Figure 1: SPD classification on CIFAR-100 ($S^{32}_{++}$, 20 coarse classes). Each panel pairs a Euclidean LMO with its intrinsic counterpart under a common norm: Frobenius (EGD vs. RGD), spectral (Muon vs. iMuon), and nuclear (NuMuon vs. iMuon-Nu). Curves show mean test accuracy with $\pm 1$ std bands over three seeds at the validation-selected learning rate. From the second epoch on, the intrinsic method is above its Euclidean counterpart in every pair, and the final gap widens from Frobenius to spectral to nuclear.

from Frobenius ($+0.025$) to spectral ($+0.068$) to nuclear ($+0.124$). One possible reason for the large nuclear gain is that the rank-1 Euclidean step places the whole step on a single direction regardless of the scale of $X$ along it, whereas the metric scaling $X^{1/2}(\cdot)X^{1/2}$ adapts the step to $X$.

## 7 Conclusion

We introduced iMuon, a unified framework for unitarily invariant norm-constrained optimization on Riemannian matrix manifolds. By lifting the ambient norm through the Riemannian metric, iMuon resolves the symmetry breaking (O1) and tangent space and norm constraints coupling (O2) that obstruct prior manifold extensions of Muon, and admits closed-form LMO solutions on the fixed-rank, SPD, Stiefel, and Grassmann manifolds. The constant in the resulting convergence rate depends only on the manifold dimension. On the fixed-rank manifold this is independent of factor conditioning, removing the bounded-factor-norm assumption and runtime rescaling required by prior work. Empirically, iMuon performs favorably across LoRA fine-tuning and other learning problems on the fixed-rank, SPD, Stiefel, and Grassmann manifolds, with the benefit of the intrinsic geometry most pronounced when the underlying metric is non-trivial. We note that this advantage is established in the momentum-free setting, which isolates the update direction. With momentum, iMuon is competitive with existing momentum-based optimizers on LoRA fine-tuning but does not uniformly outperform adaptive baselines such as Scaled AdamW. Developing momentum tailored to the intrinsic LMO, along with variance-reduced variants and adaptive choice of $\varphi$, remains future work. Limitations are discussed in Section B.

## Acknowledgments and Disclosure of Funding

PJ acknowledges the SBI Foundation Hub for Data & Analytics grant and IIT Bombay seed grant. CM acknowledges the support of National Research Foundation of Singapore under its NRF Fellowship Programme Award No.: NRF-NRFF18-2026-0010.

## References

[1] P.-A. Absil, R. Mahony, and R. Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008.

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

[3] N. Amsel, D. Persson, C. Musco, and R. M. Gower. The Polar Express: Optimal matrix sign methods and their application to the Muon algorithm. In *International Conference on Learning Representations (ICLR)*, 2026.

[4] V. Arsigny, P. Fillard, X. Pennec, and N. Ayache. Geometric means in a novel vector space structure on symmetric positive-definite matrices. *SIAM Journal on Matrix Analysis and Applications*, 29(1):328–347, 2007.

[5] L. Balzano, R. Nowak, and B. Recht. Online identification and tracking of subspaces from highly incomplete information. In *Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, 2010.

[6] G. Bécigneul and O.-E. Ganea. Riemannian adaptive optimization methods. In *International Conference on Learning Representations (ICLR)*, 2019.

[7] J. Bernstein. Modular manifolds. Thinking Machines Lab: Connectionism, `https://thinkingmachines.ai/blog/modular-manifolds/`, 2025.

[8] J. Bernstein and L. Newhouse. Old optimizer, new norm: An anthology. Technical report, arXiv preprint arXiv:2409.20325, 2024.

[9] J. Bernstein and L. Newhouse. Modular duality in deep learning. In *International Conference on Machine Learning (ICML)*, 2025.

[10] R. Bhatia. *Positive Definite Matrices*. Princeton University Press, 2007.

[11] R. Bhatia, T. Jain, and Y. Lim. On the Bures–Wasserstein distance between positive definite matrices. *Expositiones Mathematicae*, 37(2):165–191, 2019.

[12] V. Bogachev, V. Aletov, A. Molozhavenko, D. Bobkov, V. Soboleva, A. Alanov, and M. Rakhuba. LoRA meets Riemannion: Muon optimizer for parametrization-independent low-rank adapters. In *International Conference on Learning Representations (ICLR)*, 2026.

[13] N. Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023.

[14] E. Candès and B. Recht. Exact matrix completion via convex optimization. *Communications of the ACM*, 55(6):111–119, 2012.

[15] D. Carlson, E. Collins, Y.-P. Hsieh, L. Carin, and V. Cevher. Preconditioned spectral descent for deep learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2015.

[16] F. L. Cesista. Heuristic solutions for steepest descent on the Stiefel manifold. `https://leloykun.github.io/ponder/steepest-descent-stiefel/`, 2025.

[17] L. Chen, J. Li, and Q. Liu. Muon optimizes under spectral norm constraints. *Transactions on Machine Learning Research*, 2026.

[18] M. Crawshaw, C. Modi, M. Liu, and R. M. Gower. An exploration of non-Euclidean gradient descent: Muon and its many variants. In *International Conference on Machine Learning (ICML)*, 2026.

[19] C. Davis, W. M. Kahan, and H. F. Weinberger. Norm-preserving dilations and their applications to optimal error bounds. *SIAM Journal on Numerical Analysis*, 19(3):445–469, 1982.

[20] A. Edelman, T. A. Arias, and S. T. Smith. The geometry of algorithms with orthogonality constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998.

[21] Y. Gu and Z. Xie. Mano: Restriking manifold optimization for LLM training. Technical report, arXiv preprint arXiv:2601.23000, 2026.

[22] J. Hamm and D. D. Lee. Grassmann discriminant analysis: A unifying view on subspace-based learning. In *International Conference on Machine Learning (ICML)*, 2008.

[23] A. Han, B. Mishra, P. Jawanpuria, and J. Gao. On Riemannian optimization over positive definite matrices with the Bures-Wasserstein geometry. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[24] A. Han, B. Mishra, P. Jawanpuria, and J. Gao. Riemannian accelerated gradient methods via extrapolation. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2023.

[25] F. M. Harper and J. A. Konstan. The MovieLens datasets: History and context. *ACM Transactions on Interactive Intelligent Systems*, 5(4):1–19, 2015.

[26] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2016.

[27] N. J. Higham. *Functions of Matrices: Theory and Computation*. SIAM, 2008.

[28] R. A. Horn and C. R. Johnson. *Matrix Analysis*. Cambridge University Press, 2nd edition, 2012.

[29] E. J. Hu, Y. 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.

[30] Z. Huang and L. Van Gool. A Riemannian network for SPD matrix learning. In *AAAI Conference on Artificial Intelligence (AAAI)*, 2017.

[31] Z. Huang, R. Wang, S. Shan, and X. Chen. Projection metric learning on Grassmann manifold with application to video based face recognition. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[32] P. Janson, E. Oyallon, and E. Belilovsky. Stabilizing native low-rank LLM pretraining. In *International Conference on Machine Learning (ICML)*, 2026.

[33] K. Jordan, Y. Jin, V. Boza, J. You, F. L. Cesista, L. Newhouse, and J. Bernstein. Muon: An optimizer for hidden layers in neural networks. `https://kellerjordan.github.io/posts/muon/`, 2024.

[34] C. Kang, J. Yun, B. Shin, Y. Cho, and C. Yun. Uniform spectral growth under factor-wise Muon orthogonalization in matrix factorization and LoRA. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2026. Cited theorem and proposition numbers refer to arXiv:2602.06385v1.

[35] H. Kasai and B. Mishra. Low-rank tensor completion: A Riemannian manifold preconditioning approach. In *International Conference on Machine Learning (ICML)*, 2016.

[36] H. Kasai, H. Sato, and B. Mishra. Riemannian stochastic recursive gradient algorithm. In *International Conference on Machine Learning (ICML)*, 2018.

[37] R. H. Keshavan, A. Montanari, and S. Oh. Matrix completion from a few entries. *IEEE Transactions on Information Theory*, 56(6):2980–2998, 2010.

[38] G. Y. Kim and M.-h. Oh. Convergence of Muon with Newton–Schulz. In *International Conference on Learning Representations (ICLR)*, 2026.

[39] Y. Koren, R. Bell, and C. Volinsky. Matrix factorization techniques for recommender systems. *Computer*, 42(8):30–37, 2009.

[40] A. Krizhevsky. Learning multiple layers of features from tiny images. Technical report, University of Toronto, 2009.

[41] T. Large, Y. Liu, M. Huh, H. Bahng, P. Isola, and J. Bernstein. Scalable optimization in the modular norm. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024.

[42] O. Ledoit and M. Wolf. A well-conditioned estimator for large-dimensional covariance matrices. *Journal of Multivariate Analysis*, 88(2):365–411, 2004.

[43] J. Li and M. Hong. A note on the convergence of Muon. Technical report, arXiv preprint arXiv:2502.02900, 2025.

[44] J. Ma, Y. Huang, Y. Chi, and Y. Chen. Preconditioning benefits of spectral orthogonalization in Muon. Technical report, arXiv preprint arXiv:2601.13474, 2026.

[45] L. Malagò, L. Montrucchio, and G. Pistone. Wasserstein Riemannian geometry of Gaussian densities. *Information Geometry*, 1(2):137–179, 2018.

[46] B. Mishra and R. Sepulchre. Riemannian preconditioning. *SIAM Journal on Optimization*, 26 (1):635–660, 2016.

[47] B. Mishra, K. A. Apuroop, and R. Sepulchre. A Riemannian geometry for low-rank matrix completion. Technical report, arXiv preprint arXiv:1211.1550, 2012.

[48] B. Mishra, G. Meyer, S. Bonnabel, and R. Sepulchre. Fixed-rank matrix factorizations and Riemannian low-rank optimization. *Computational Statistics*, 29(3–4):591–621, 2014.

[49] Z. Mo, L.-K. Huang, and S. J. Pan. Parameter and memory efficient pretraining via low-rank Riemannian optimization. In *International Conference on Learning Representations (ICLR)*, 2025.

[50] H. Mohaghegh Dolatabadi, T. Ajanthan, S. Ramasinghe, C. P. Hewa Koneputugodage, S. Siriwardhana, V. Shevchenko, K. Pajak, J. Snewin, G. Avraham, and A. Long. NuMuon: Nuclear-norm-constrained Muon for compressible LLM training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2026.

[51] A. Mokhtari, H. Hassani, and A. Karbasi. Stochastic conditional gradient methods: From convex minimization to submodular maximization. *Journal of Machine Learning Research*, 21(105):1–49, 2020.

[52] C. Mostajeran, N. Da Costa, G. Van Goffrier, and R. Sepulchre. Differential geometry with extreme eigenvalues in the positive semidefinite cone. *SIAM Journal on Matrix Analysis and Applications*, 45(2):1089–1113, 2024.

[53] J. Novikova, O. Dušek, and V. Rieser. The E2E dataset: New challenges for end-to-end generation. In *Annual SIGdial Meeting on Discourse and Dialogue (SIGDIAL)*, 2017.

[54] J. Park, M. Kang, S. Lee, H. Lee, S. Kim, and J. Lee. Riemannian optimization for LoRA on the Stiefel manifold. In *Findings of the Association for Computational Linguistics: EMNLP*, 2025.

[55] X. Pennec, P. Fillard, and N. Ayache. A Riemannian framework for tensor computing. *International Journal of Computer Vision*, 66(1):41–66, 2006.

[56] T. Pethick, W. Xie, K. Antonakopoulos, Z. Zhu, A. Silveti-Falls, and V. Cevher. Training deep learning models with norm-constrained LMOs. In *International Conference on Machine Learning (ICML)*, 2025.

[57] B. Recht. A simpler approach to matrix completion. *Journal of Machine Learning Research*, 12(104):3413–3430, 2011.

[58] A. Riabinin, E. Shulgin, K. Gruntkowska, and P. Richtárik. From Muon to Gluon: Bridging theory and practice of LMO-based optimizers for LLMs. In *International Conference on Machine Learning (ICML)*, 2026.

[59] H. Sato, H. Kasai, and B. Mishra. Riemannian stochastic variance reduced gradient algorithm with retraction and vector transport. *SIAM Journal on Optimization*, 29(2):1444–1472, 2019.

[60] N. Sato, H. Naganuma, and H. Iiduka. Convergence bound and critical batch size of Muon optimizer. Technical report, arXiv preprint arXiv:2507.01598, 2025.

[61] S. Schotthöfer, T. Klein, and J. Kusch. A geometric framework for momentum-based optimizers for low-rank training. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2025.

[62] W. Shen, R. Huang, M. Huang, C. Shen, and J. Zhang. On the convergence analysis of Muon. *Transactions on Machine Learning Research*, 2026.

[63] Z. Shen, C. Fang, P. Zhao, J. Huang, and H. Qian. Complexities in projection-free stochastic non-convex minimization. In *International Conference on Artificial Intelligence and Statistics (AISTATS)*, 2019.

[64] M. Solonko, A. Molozhavenko, and M. Rakhuba. Muon on the Stiefel manifold admits an exact closed-form update. Technical report, arXiv preprint arXiv:2608.06218, 2026.

[65] J. Su. Steepest descent on manifolds: 3. Muon + Stiefel. `https://www.kexue.fm/archives/11221`, 2025. Blog post (in Chinese).

[66] A. C. Thompson. On certain contraction mappings in a partially ordered vector space. *Proceedings of the American Mathematical Society*, 14(3):438–443, 1963.

[67] T. Tong, C. Ma, and Y. Chi. Accelerating ill-conditioned low-rank matrix estimation via scaled gradient descent. *Journal of Machine Learning Research*, 22(150):1–63, 2021.

[68] O. Tuzel, F. Porikli, and P. Meer. Region covariance: A fast descriptor for detection and classification. In *European Conference on Computer Vision (ECCV)*, 2006.

[69] O. Tuzel, F. Porikli, and P. Meer. Pedestrian detection via classification on Riemannian manifolds. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 30(10):1713–1727, 2008.

[70] B. Vandereycken. Low-rank matrix completion by Riemannian optimization. *SIAM Journal on Optimization*, 23(2):1214–1236, 2013.

[71] A. Wang, A. Singh, J. Michael, F. Hill, O. Levy, and S. R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *International Conference on Learning Representations (ICLR)*, 2019.

[72] F. Wang, J. Cheng, W. Liu, and H. Liu. Additive margin softmax for face verification. *IEEE Signal Processing Letters*, 25(7):926–930, 2018.

[73] Z. Wang, J. Liang, R. He, Z. Wang, and T. Tan. Taming momentum: Rethinking optimizer states through low-rank approximation. In *International Conference on Learning Representations (ICLR)*, 2026.

[74] M. Weber and S. Sra. Riemannian optimization via Frank–Wolfe methods. *Mathematical Programming*, 199(1–2):525–556, 2023.

[75] K. Wen, D. Hall, T. Ma, and P. Liang. Fantastic pretraining optimizers and where to find them. In *International Conference on Learning Representations (ICLR)*, 2026.

[76] L. Wolf, T. Hassner, and I. Maoz. Face recognition in unconstrained videos with matched background similarity. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011.

[77] T. Xie, H. Luo, H. Tang, Y. Hu, J. K. Liu, Q. Ren, Y. Wang, W. X. Zhao, R. Yan, B. Su, C. Luo, and B. Guo. Controlled LLM training on spectral sphere. In *International Conference on Machine Learning (ICML)*, 2026.

[78] G. Yang, E. J. Hu, I. Babuschkin, S. Sidor, X. Liu, D. Farhi, N. Ryder, J. Pachocki, W. Chen, and J. Gao. Tuning large neural networks via zero-shot hyperparameter transfer. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

[79] G. Yang, J. B. Simon, and J. Bernstein. A spectral condition for feature learning. Technical report, arXiv preprint arXiv:2310.17813, 2023.

[80] K. Yang and L. Lai. Manifold constrained steepest descent for smooth and closed-set optimization. Technical report, arXiv preprint arXiv:2601.21487, 2026.

[81] F. Zhang and M. Pilanci. Riemannian preconditioned LoRA for fine-tuning foundation models. In *International Conference on Machine Learning (ICML)*, 2024.

[82] H. Zhang and S. Sra. An estimate sequence for geodesically convex optimization. In *Conference on Learning Theory (COLT)*, 2018.

[83] H. Zhang, S. J. Reddi, and S. Sra. Riemannian SVRG: Fast stochastic optimization on Riemannian manifolds. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2016.

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

# Appendix


**A Broader Societal Impact** **17**

**B Limitations** **17**

**C Closed-Form LMO Solutions via Norm Duality** **18**

C.1 Reduction to a vector problem . . . . . 18

C.2 Spectral, Frobenius, and nuclear norms . . . . . 18

C.3 Other unitarily invariant norms . . . . . 19

**D Proofs of Main Results** **19**

D.1 Proof of Proposition 3.1 . . . . . 19

D.2 Proof of Proposition 3.4 . . . . . 20

D.3 Proof of Corollary 4.1 . . . . . 20

D.4 Proof of Lemma 4.2 . . . . . 21

D.5 Proof of Theorem 5.3 . . . . . 22

D.6 Computation of $C_\varphi$ for the fixed-rank manifold . . . . . 23

D.7 Computation of $C_\varphi$ for SPD, Stiefel, and Grassmann . . . . . 23

D.8 Proof of Theorem 5.7 . . . . . 23

**E Singular Value Invariance of the Scaled Tangent Spaces** **24**

**F Derivations and Implementation Details** **25**

F.1 Summary of closed-form solutions . . . . . 25

F.2 Computing the polar factor . . . . . 25

F.3 Fixed-rank manifold $\mathcal{M}_r$ . . . . . 26

F.4 SPD manifold $S^n_{++}$ . . . . . 27

F.5 Stiefel manifold $\mathrm{St}(m, r)$ . . . . . 27

F.6 Grassmann manifold $\mathrm{Gr}(m, r)$ . . . . . 28

**G Prior Convergence Results as Special Cases** **28**

**H Detailed Comparisons with Prior Methods** **29**

H.1 Fixed-rank manifold $\mathcal{M}_r$: iMuon vs. Spectron, Riemannion, and factor-wise Muon 29

H.2 SPD manifold $S^n_{++}$: spectral LMO vs. natural gradient . . . . . 32

H.3 Stiefel manifold $\mathrm{St}(m, r)$: product norm vs. joint norm . . . . . 33

H.4 Grassmann manifold $\mathrm{Gr}(m, r)$: connections to subspace tracking . . . . . 34

**I Background on Riemannian Optimization** **35**

I.1 Riemannian optimization: key concepts . . . . . 35

I.2 Quotient manifolds and scale invariance . . . . . 35

**J Extended Related Work** **35**

J.1 Riemannian Frank–Wolfe methods . . . . . 35

J.2 Euclidean spectral optimizers and LMO . . . . . 36

J.3 Muon extensions for low-rank/LoRA fine-tuning . . . . . 36

J.4 Convergence theory of Muon . . . . . 36

J.5 Riemannian and manifold-aware optimizers for LLMs . . . . . 37

**K Experimental Details for LoRA Fine-Tuning** **37**

K.1 E2E NLG . . . . . 38

K.2 GLUE . . . . . 39

**L Learning on Fixed-Rank Manifold** **40**

L.1 Matrix completion: synthetic experiments . . . . . 41

L.2 Fixed-rank CIFAR-100 classification heads . . . . . 42

L.3 MovieLens-1M . . . . . 44

**M SPD Prototype Learning for Covariance-Based Image Classification** **45**

**N Grassmann Fréchet Subspace Learning for Video Face Identification** **46**

**O Stiefel Sub-Center Prototype Learning for Frozen-Feature Image Classification** **47**

**P Additional Manifolds and Metrics** **49**

P.1 General applicability . . . . . 49

P.2 Fixed-rank manifold: embedded viewpoint . . . . . 49

P.3 Stiefel manifold: embedded viewpoint . . . . . 50

P.4 Spectral sphere . . . . . 51

P.5 Alternative SPD metrics: Log-Euclidean and Bures–Wasserstein . . . . . 51

# A Broader Societal Impact

This paper presents foundational research on optimization algorithms on Riemannian matrix manifolds. The contribution is methodological and theoretical. The framework is not tied to a particular application or deployment, and our empirical evaluation uses standard publicly available benchmarks. As with any optimizer, the downstream impact depends on the objective and the data it is applied to. We do not identify a direct path to negative societal impact beyond the general-purpose effect of more efficient training.

# B Limitations

Our closed-form solutions rely on two structural properties. The first is the left-right form of the metric, which gives symmetry preservation (Proposition 3.1). The second is singular value invariance or $\varphi$-contractivity of the scaled tangent space, which gives the closed-form reduction (Proposition 3.4). Both properties hold on the four manifolds we treat but not in general. Section P discusses cases where one of them fails. The LMO remains a convex problem in such cases but requires an iterative

solver. On $\mathrm{St}(m, r)$, we further use an $\ell_\infty$-product norm relaxation of the joint norm, which enlarges $C_\varphi$ (Section H.3). On $\mathcal{M}_r$, we use the $\ell_\infty$-product norm over the two factor blocks, for which no joint unitarily invariant norm is defined (Table 4). Extending the framework to tensor manifolds, such as the Tucker manifold [35], is left to future work. Our convergence analysis does not incorporate momentum or acceleration [82, 24]. The stochastic rate carries an additive $2\sigma_\varphi$ term, and the momentum variant used in our experiments (Section K) is a practical construction without a convergence guarantee.

Empirically, the advantage of iMuon is pronounced in the no-momentum setting, while with momentum, iMuon is competitive (Tables 1 and 2). Our experiments cover LoRA fine-tuning of GPT-2 Medium and 4-bit Mistral-7B, and learning problems on the fixed-rank, SPD, Stiefel, and Grassmann manifolds. Frontier-scale LLM pretraining, where Muon is most heavily used, is not explored in this work.

# C Closed-Form LMO Solutions via Norm Duality

We derive the closed-form solution of the intrinsic LMO for a general unitarily invariant norm $\varphi$.

## C.1 Reduction to a vector problem

The key tool is von Neumann's trace inequality, which reduces the matrix LMO to a vector problem over singular values.

**Lemma C.1.** *Let $H = U\,\mathrm{diag}(\sigma_1, \dots, \sigma_{\min(p,q)})V^\top$ with $\sigma_1 \geq \cdots \geq \sigma_{\min(p,q)} \geq 0$, and for a vector $z$ write $\varphi(z) := \varphi(\mathrm{diag}(z))$, with $\mathrm{diag}(z)$ the $p \times q$ rectangular diagonal matrix. Then $\max_{\varphi(Z)\leq\tau}\langle Z,\, H\rangle = \tau\,\varphi^\circ(\sigma)$, where $\varphi^\circ$ is the dual norm and $\sigma = (\sigma_1, \dots, \sigma_{\min(p,q)})$. A maximizer has the form $Z^* = U\,\mathrm{diag}(z_1^*, \dots, z_{\min(p,q)}^*)V^\top$, where $z^*$ solves the vector problem $\max_{\varphi(z)\leq\tau}\langle z,\, \sigma\rangle$.*

*Proof.* By von Neumann's trace inequality [28, Thm. 7.4.1.1], $\langle Z,\, H\rangle \leq \sum_i \sigma_i(Z)\sigma_i(H)$ with equality when $Z$ shares the singular vectors of $H$. Hence $Z^* = U\,\mathrm{diag}(z^*)V^\top$ for some nonnegative $z^*$ with $\varphi(z^*) \leq \tau$, reducing the problem to $\max_{\varphi(z)\leq\tau}\langle z,\, \sigma\rangle = \tau\,\varphi^\circ(\sigma)$ by definition of the dual norm. □

## C.2 Spectral, Frobenius, and nuclear norms

The vector problem $\max_{\varphi(z)\leq\tau}\langle z,\, \sigma\rangle$ has closed-form solutions for the three standard norms:

*Spectral norm* ($\varphi(z) = \|z\|_\infty = \max_i |z_i|$). Each $z_i$ is independently bounded by $\tau$, and $\sigma_i \geq 0$, so $z_i^* = \tau$ for all $i$. Thus $Z^* = \tau UV^\top = \tau\ \mathrm{Ortho}(H)$ and the optimal value is $\tau\sum_i \sigma_i = \tau\|H\|_*$. Entries with $\sigma_i = 0$ do not affect the objective and may be set to zero, which gives the thin polar factor used in Proposition 3.4. The dual norm is $\varphi^\circ = \|\cdot\|_1$ on singular values, i.e., the nuclear norm.

*Frobenius norm* ($\varphi(z) = \|z\|_2$). By Cauchy–Schwarz, $z^* = \tau\,\sigma/\|\sigma\|_2$, giving $Z^* = \tau\,H/\|H\|_F$. The optimal value is $\tau\|H\|_F$. Its dual norm is $\varphi^\circ = \|\cdot\|_F$, i.e., it is self-dual.

*Nuclear norm* ($\varphi(z) = \|z\|_1 = \sum_i |z_i|$). Under $\sum_i z_i \leq \tau$ with $z_i \geq 0$, all of the norm bound goes to the largest $\sigma_i$: $z^* = \tau\, e_1$. Writing $u_i$ and $v_i$ for the $i$-th columns of $U$ and $V$, we get $Z^* = \tau\, u_1 v_1^\top$ and the optimal value is $\tau\sigma_1 = \tau\|H\|_2$. If exactly $k > 1$ singular values of $H$ equal $\sigma_1$, the choice required in Proposition 3.4 is $Z^* = \frac{\tau}{k}\sum_{i=1}^k u_i v_i^\top$, which does not depend on the choice of singular vectors for $\sigma_1$. The dual norm is $\varphi^\circ = \|\cdot\|_\infty$, i.e., the spectral norm.

For a general unitarily invariant norm $\varphi$, norm duality gives $\max_{\varphi(z)\leq\tau}\langle z,\, \sigma\rangle = \tau\,\varphi^\circ(\sigma)$, and the maximizers are the elements of $\tau\,\partial\varphi^\circ(\sigma)$. Computing a maximizer thus amounts to a convex problem in $\min(p, q)$ variables, and Section C.3 gives closed forms for several other norms.

### C.3 Other unitarily invariant norms

The spectral and nuclear norms are the two extremes of this family (for $\varphi$ normalized so that $\varphi(e_1) = 1$, $\|z\|_\infty \le \varphi(z) \le \|z\|_1$), with the Frobenius norm in between. Here we note two natural interpolations that also admit closed-form LMOs within our framework. Throughout, $m := \min(p, q)$ denotes the number of singular values of $H$.

**Partial polar $k$-norms.** Let $\|Z\|_{(k)} := \sum_{i=1}^{k} \sigma_i(Z)$ denote the Ky Fan $k$-norm [28, §3.4], the sum of the $k$ largest singular values. We take as the constraint norm its dual,

$$\varphi(Z) = \|Z\|_{(k)}^\circ = \max\bigl(\|Z\|_2,\ \|Z\|_*/k\bigr), \qquad k \in \{1, \dots, m\},$$

which is unitarily invariant and interpolates between the nuclear norm at $k = 1$, where $\|Z\|_* \ge \|Z\|_2$ gives $\varphi = \|\cdot\|_*$, and the spectral norm at $k = m$, where $\|Z\|_*/m \le \|Z\|_2$ gives $\varphi = \|\cdot\|_2$. The LMO has a clean closed form: the vector problem $\max_{\varphi(z)\le\tau}\langle z,\, \sigma\rangle$ is solved by $z_i^* = \tau$ for $i \le k$ and $z_i^* = 0$ for $i > k$, giving

$$Z^* = \tau \sum_{i=1}^{k} u_i v_i^\top,$$

the *rank-$k$ partial polar factor* of $H$. If a tied block $\sigma_j = \cdots = \sigma_{j+\ell-1}$ straddles the cutoff $k$, the symmetric choice required by Proposition 3.4 spreads the remaining budget $(k - j + 1)\tau$ equally over the block, $z_i^* = \tau(k - j + 1)/\ell$ for $j \le i \le j + \ell - 1$, which attains the same optimal value and is independent of the choice of singular vectors within the block. The dual norm is $\varphi^\circ = \|\cdot\|_{(k)}$, the Ky Fan $k$-norm itself, and the squared Riemannian radius is $C_\varphi = k$. Thus Theorem 5.3 yields rate $O(\sqrt{Lk\Delta_0/T})$ with convergence in $\|H_t\|_{(k)}$. Setting $k$ between 1 and $m$ provides a tunable trade-off between aggressiveness and rank: $k = 1$ concentrates on the leading singular direction (conservative, $C_\varphi = 1$), while $k = m$ equalizes all $m$ directions (aggressive, $C_\varphi = m$).

**Schatten $s$-norms.** The Schatten $s$-norm $\|Z\|_{\mathcal{S}_s} := (\sum_i \sigma_i(Z)^s)^{1/s}$ for $s \in [1, \infty]$ is unitarily invariant and continuously interpolates between the nuclear ($s = 1$), Frobenius ($s = 2$), and spectral ($s = \infty$) norms. The dual norm is the Schatten $s^*$-norm with $1/s + 1/s^* = 1$, so the LMO $\max_{\|z\|_s\le\tau}\langle z,\, \sigma\rangle$ is solved by Hölder's inequality:

$$z_i^* = \tau\, \frac{\sigma_i^{s^*-1}}{\|\sigma^{s^*-1}\|_s}, \qquad Z^* = U\, \mathrm{diag}(z^*) V^\top,$$

where $\sigma^{s^*-1}$ denotes the vector $(\sigma_1^{s^*-1}, \dots, \sigma_m^{s^*-1})$. This is a *power-law reweighting* of the singular values. For $s = 2$ ($s^* = 2$), $z_i^* \propto \sigma_i$ (proportional, recovering Frobenius). For $s \to \infty$ ($s^* \to 1$), $z_i^* \to \tau$ (all equal, recovering spectral). For $s = 1$ ($s^* = \infty$), all budget concentrates on the largest $\sigma_i$ (recovering nuclear). Intermediate values of $s$ give a "soft Muon" that partially equalizes singular values without the hard cutoff of the partial polar norm.

The squared Riemannian radius is

$$C_\varphi = \begin{cases} 1 & s \le 2, \\ m^{1-2/s} & s \ge 2. \end{cases}$$

For $s \le 2$, the maximizer of $\|z\|_2^2$ subject to $\|z\|_s \le 1$ is a unit vector ($C_\varphi = 1$). For $s > 2$, it is the uniform vector $z = (m^{-1/s}, \dots, m^{-1/s})$ giving $C_\varphi = m^{1-2/s}$. Thus Theorem 5.3 yields rate $O(\sqrt{L \cdot \max(1, m^{1-2/s}) \cdot \Delta_0/T})$, smoothly interpolating between the $O(\sqrt{L/T})$ rate of Frobenius/nuclear norms and the $O(\sqrt{Lm/T})$ rate of the spectral norm. In practice, choosing $s$ slightly above 2 gives a mild singular value equalization ("softer" than full Muon) at a reduced effective-dimension cost.

## D Proofs of Main Results

### D.1 Proof of Proposition 3.1

*Proof of Proposition 3.1.* The map $\xi \mapsto P_x\, \xi\, Q_x$ is self-adjoint and positive definite on the ambient matrix space, and it maps $T_x\mathcal{M}$ into itself (it equals $G_x$ there). Hence, it also maps the orthogonal

complement of $T_x\mathcal{M}$ into itself. Consequently, its square root $\xi \mapsto P_x^{1/2}\xi\, Q_x^{1/2}$ maps $T_x\mathcal{M}$ into itself and coincides there with $G_x^{1/2}$, i.e., $G_x^{1/2}\xi = P_x^{1/2}\xi\, Q_x^{1/2}$; the same holds at $\tilde{x}$. Let $Z := G_x^{1/2}\xi$. Then

$$G_{\tilde{x}}^{1/2}\tilde{\xi} = P_{\tilde{x}}^{1/2}L_h\,\xi\,R_h^\top Q_{\tilde{x}}^{1/2} = U_h\,Z\,V_h^\top, \qquad U_h := P_{\tilde{x}}^{1/2}L_h P_x^{-1/2}, \quad V_h := Q_{\tilde{x}}^{1/2}R_h Q_x^{-1/2}.$$

By the transformation law, $U_h^\top U_h = P_x^{-1/2}L_h^\top P_{\tilde{x}} L_h P_x^{-1/2} = P_x^{-1/2}P_x P_x^{-1/2} = I$, and similarly $V_h^\top V_h = I$. Since $U_h$ and $V_h$ are square, they are orthogonal, so $\sigma_i(U_h Z V_h^\top) = \sigma_i(Z)$ for all $i$. Hence $\varphi(G_{\tilde{x}}^{1/2}\tilde{\xi}) = \varphi(G_x^{1/2}\xi)$ for every unitarily invariant $\varphi$. With $\varphi = \|\cdot\|_F$, this gives $g_{\tilde{x}}(\tilde{\xi},\tilde{\xi}) = \|G_{\tilde{x}}^{1/2}\tilde{\xi}\|_F^2 = \|G_x^{1/2}\xi\|_F^2 = g_x(\xi,\xi)$, so $g$ is $\mathcal{G}$-invariant by polarization. It remains to verify the hypotheses in the three cases.

*Fixed-rank manifold.* The action $(B, A) \mapsto (BN^{-1}, NA)$ acts separately on the two factor blocks of $T_x\mathcal{M}_r = \mathbb{R}^{m\times r} \times \mathbb{R}^{r\times n}$, which are full matrix spaces. For the $\dot{B}$-block, $\dot{B} \mapsto \dot{B}N^{-1}$ gives $L_h = I_m$ and $R_h = N^{-\top}$, with $P = I_m$ and $Q = AA^\top \mapsto NAA^\top N^\top = R_h^{-\top} Q\, R_h^{-1}$. For the $\dot{A}$-block, $\dot{A} \mapsto N\dot{A}$ gives $L_h = N$ and $R_h = I_n$, with $Q = I_n$ and $P = B^\top B \mapsto N^{-\top}B^\top B\, N^{-1} = L_h^{-\top} P\, L_h^{-1}$. The conclusion therefore holds for each block, and hence for the product norm $\max\{\varphi(\dot{B}(AA^\top)^{1/2}), \varphi((B^\top B)^{1/2}\dot{A})\}$.

*SPD manifold.* $T_X S_{++}^n = \mathrm{Sym}(n)$ is mapped into itself by $\xi \mapsto X^{-1}\xi X^{-1}$. The congruence $X \mapsto NXN^\top$ gives $L_h = R_h = N$, and $P = Q = X^{-1} \mapsto (NXN^\top)^{-1} = N^{-\top}X^{-1}N^{-1}$.

*Grassmann manifold.* On the horizontal space $\{\xi : X^\top\xi = 0\}$, the metric is Euclidean, $P = I_m$ and $Q = I_r$, and $O \in O(r)$ acts by $\xi \mapsto \xi O$, so $L_h = I_m$ and $R_h = O^\top$. The law holds because $R_h^{-\top}R_h^{-1} = O^\top O = I_r$. □

### D.2 Proof of Proposition 3.4

*Proof of Proposition 3.4.* Substitute $Z = G_x^{1/2}\xi$ in (4). Since $g_x(\xi, \operatorname{grad} f) = \langle G_x^{1/2}\xi,\, G_x^{1/2}(\operatorname{grad} f)\rangle = \langle Z,\, H\rangle$ (by self-adjointness of $G_x^{1/2}$) and $\varphi(G_x^{1/2}\xi) = \varphi(Z)$, the LMO becomes $\max_{Z\in\mathcal{Z}_x,\,\varphi(Z)\le\tau}\langle Z,\, H\rangle$, and its solution $\hat{Z}$ gives $\xi^* = G_x^{-1/2}\hat{Z}$. Note that $H \in \mathcal{Z}_x$ since $\operatorname{grad} f \in T_x\mathcal{M}$. Write $H = U\Sigma V^\top$ (compact SVD with $\Sigma = \operatorname{diag}(\sigma)$ and $\sigma_1 \ge \cdots \ge \sigma_r > 0$). By von Neumann's trace inequality [28, Thm. 7.4.1.1], $\langle Z,\, H\rangle \le \sum_i \sigma_i(Z)\sigma_i(H)$ with equality when $Z$ shares singular vectors with $H$. Thus the unconstrained problem $\max_{\varphi(Z)\le\tau}\langle Z,\, H\rangle$ has a maximizer of the form $Z^* = U\operatorname{diag}(z^*)V^\top$ (Lemma C.1), where $z^*$ solves $\max_{\varphi(z)\le\tau}\langle z,\, \sigma\rangle$, with optimal value $\tau\varphi^\circ(\sigma)$ by norm duality. We choose $z^*$ with $z_i^* = z_j^*$ whenever $\sigma_i = \sigma_j$. Such a maximizer exists: averaging any maximizer over permutations within each group of equal $\sigma_i$ preserves the objective and, by convexity and permutation invariance of $\varphi$, does not increase $\varphi$.

If $\mathcal{Z}_x$ is SV-invariant, then since $H \in \mathcal{Z}_x$, replacing $\Sigma$ with $\operatorname{diag}(z^*)$ preserves membership by Definition 3.2: $Z^* \in \mathcal{Z}_x$. Hence the unconstrained maximizer satisfies the $\mathcal{Z}_x$-constraint, so $\hat{Z} = Z^* = P_{\mathcal{Z}_x}(Z^*)$ solves the constrained problem.

If $\mathcal{Z}_x$ is $\varphi$-contractive, then $P_{\mathcal{Z}_x}(Z^*) \in \mathcal{Z}_x$ and $\varphi(P_{\mathcal{Z}_x}(Z^*)) \le \varphi(Z^*) \le \tau$, so it is feasible. Since $P_{\mathcal{Z}_x}$ is self-adjoint and $H \in \mathcal{Z}_x$, $\langle P_{\mathcal{Z}_x}(Z^*),\, H\rangle = \langle Z^*,\, P_{\mathcal{Z}_x}(H)\rangle = \langle Z^*,\, H\rangle$. As the constrained maximum cannot exceed the unconstrained one, $\hat{Z} = P_{\mathcal{Z}_x}(Z^*)$ solves the constrained problem. □

### D.3 Proof of Corollary 4.1

*Proof of Corollary 4.1.* On $\mathcal{M}_r$ with the coupled metric, $G_x^{1/2}(\dot{B}, \dot{A}) = (\dot{B}(AA^\top)^{1/2},\, (B^\top B)^{1/2}\dot{A})$ and $T_x\mathcal{M}_r = \mathbb{R}^{m\times r} \times \mathbb{R}^{r\times n}$.

*Step 1: Left-right form and symmetry (Proposition 3.1).* The $\dot{B}$-block metric is $g^B(\dot{B}_1, \dot{B}_2) = \operatorname{tr}(\dot{B}_1^\top \dot{B}_2\, AA^\top)$, which has the left-right form with $P = I_m$ and $Q = AA^\top$. Under $(B, A) \mapsto (BN^{-1}, NA)$, $Q = AA^\top \mapsto NAA^\top N^\top$ and $\dot{B} \mapsto \dot{B}N^{-1}$; the metric is invariant since $\operatorname{tr}((\dot{B}N^{-1})^\top(\dot{B}N^{-1})NAA^\top N^\top) = \operatorname{tr}(\dot{B}^\top\dot{B}\, AA^\top)$. By Proposition 3.1, the intrinsic constraint $\varphi(\dot{B}(AA^\top)^{1/2}) \le \tau$ is $\mathrm{GL}(r)$-invariant. The $\dot{A}$-block is symmetric (with $P = B^\top B$, $Q = I_n$).

*Step 2: SV-invariance and reduction (Proposition 3.4).* The scaled tangent space decomposes as $\mathcal{Z}_x = \mathbb{R}^{m\times r} \times \mathbb{R}^{r\times n}$ (full ambient space on each block), which is trivially singular value invariant (Section E). With the $\ell_\infty$-product norm $\max(\varphi(Z_B), \varphi(Z_A)) \le \tau$, the LMO (4) decouples into independent $\dot{B}$ and $\dot{A}$ subproblems via (6).

*Step 3: Closed form for the $\dot{B}$-block.* The objective is $g_x(\xi, \operatorname{grad} f) = \langle \dot{B}A, \nabla_X f\rangle = \langle \dot{B}, \nabla_B f\rangle$ where $\nabla_B f = \nabla_X f A^\top$. Substituting $Z_B = \dot{B}(AA^\top)^{1/2}$ gives objective $\langle Z_B, H_B\rangle$ with $H_B = \nabla_B f\,(AA^\top)^{-1/2}$. By Proposition 3.4 (with SV-invariance from Step 2), the tangent-space constraint is redundant and the spectral-norm LMO reduces to the unconstrained problem $\max_{\|Z_B\|_2\le\tau}\langle Z_B, H_B\rangle$. By Lemma C.1, $Z_B^* = \tau\,\operatorname{Ortho}(H_B)$, so $\dot{B}^* = \tau\,\operatorname{Ortho}(\nabla_B f\,(AA^\top)^{-1/2})\,(AA^\top)^{-1/2}$.

The $\dot{A}$-block follows by symmetry: $H_A = (B^\top B)^{-1/2}\nabla_A f$ with $\nabla_A f = B^\top \nabla_X f$, giving $\dot{A}^* = \tau\,(B^\top B)^{-1/2}\,\operatorname{Ortho}((B^\top B)^{-1/2}\nabla_A f)$. □

### D.4 Proof of Lemma 4.2

The proof requires showing that on each block of the product-norm decoupling, the tangent-space constraint is redundant. Both blocks are SV-invariant (Section E). For the skew block, the following result additionally shows that $\operatorname{Skew}(r)$ is $\varphi$-contractive, so that any maximizer can be projected onto $\operatorname{Skew}(r)$.

**Lemma D.1** (Skew-symmetric LMO compatibility)**.** *Let $S \in \operatorname{Skew}(r)$ and let $\varphi$ be any unitarily invariant norm. Then $\max_{A\in\operatorname{Skew}(r),\ \varphi(A)\le\tau}\langle A, S\rangle = \max_{\varphi(Z)\le\tau}\langle Z, S\rangle$. Thus, the $\operatorname{Skew}(r)$ constraint is redundant and the unconstrained maximizer can always be chosen skew-symmetric. In addition, $\operatorname{Skew}(r)$ is $\varphi$-contractive (Definition 3.3).*

*Proof.* The inequality "$\le$" is trivial (smaller feasible set). For "$\ge$": let $Z^*$ be any maximizer of the right-hand side. Define $\hat{Z} := \frac{1}{2}(Z^* - (Z^*)^\top) \in \operatorname{Skew}(r)$.

Since $S \in \operatorname{Skew}(r)$, we note that $\langle \hat{Z}, S\rangle = \frac{1}{2}\langle Z^*, S\rangle - \frac{1}{2}\langle (Z^*)^\top, S\rangle = \frac{1}{2}\langle Z^*, S\rangle - \frac{1}{2}\langle Z^*, S^\top\rangle = \frac{1}{2}\langle Z^*, S\rangle + \frac{1}{2}\langle Z^*, S\rangle = \langle Z^*, S\rangle$. Thus, the objective is preserved when $Z^*$ is replaced by $\hat{Z}$.

Since $\varphi$ is unitarily invariant, $\varphi(M^\top) = \varphi(M)$ and $\varphi(-M) = \varphi(M)$. By the triangle inequality: $\varphi(\hat{Z}) = \varphi\big(\frac{Z^* - (Z^*)^\top}{2}\big) \le \frac{1}{2}\varphi(Z^*) + \frac{1}{2}\varphi((Z^*)^\top) = \varphi(Z^*) \le \tau$. Hence, $\hat{Z}$ satisfies the norm constraint, since its $\varphi$-norm does not increase.

Overall, we showed that $\hat{Z} \in \operatorname{Skew}(r)$, $\varphi(\hat{Z}) \le \tau$, and $\langle \hat{Z}, S\rangle = \langle Z^*, S\rangle$. Therefore $\hat{Z}$ is also an optimizer of the left-hand side.

The norm bound above holds with $Z^*$ replaced by any $W \in \mathbb{R}^{r\times r}$, which shows that $\operatorname{Skew}(r)$ is $\varphi$-contractive. □

*Proof of Lemma 4.2.* On $\operatorname{St}(m, r)$ with $G_X = I$, a tangent vector $\xi \in T_X\operatorname{St}$ decomposes as $\xi = XA + X_\perp B$ with $A \in \operatorname{Skew}(r)$ and $B \in \mathbb{R}^{(m-r)\times r}$. Since $[X \mid X_\perp] \in O(m)$ and $\varphi$ is unitarily invariant, $\varphi(\xi) = \varphi\big(\begin{smallmatrix} A \\ B \end{smallmatrix}\big)$.

*Step 1: Product-norm decoupling.* Under the $\ell_\infty$-product norm $\max(\varphi(A), \varphi(B)) \le \tau$, the objective separates as $g_X(\xi, \operatorname{grad} f) = \langle A, S\rangle + \langle B, N\rangle$ where $S = \operatorname{skew}(X^\top \nabla_X f) \in \operatorname{Skew}(r)$ and $N = X_\perp^\top \nabla_X f \in \mathbb{R}^{(m-r)\times r}$. By (6), the LMO decouples into independent subproblems:

$$A^* = \operatorname*{arg\,max}_{A\in\operatorname{Skew}(r),\ \varphi(A)\le\tau} \langle A, S\rangle, \qquad B^* = \operatorname*{arg\,max}_{B\in\mathbb{R}^{(m-r)\times r},\ \varphi(B)\le\tau} \langle B, N\rangle.$$

*Step 2: Normal block ($B$).* The feasible set $\mathbb{R}^{(m-r)\times r}$ is the full ambient space, hence trivially SV-invariant. By Proposition 3.4 and Lemma C.1, $B^* = \tau\,\operatorname{Ortho}(N) = \tau\,\operatorname{Ortho}(X_\perp^\top \nabla_X f)$ for $\varphi = \|\cdot\|_2$.

*Step 3: Skew block ($A$).* The feasible set $\operatorname{Skew}(r)$ is SV-invariant (Section E). By Proposition 3.4 and Lemma C.1, $A^* = \tau\,\operatorname{Ortho}(S)$ with the thin polar factor, which is skew-symmetric. Since $\operatorname{Skew}(r)$

is also $\varphi$-contractive (Lemma D.1), $\tau\ \mathrm{skew}(\mathrm{Ortho}(S))$ is a solution for any choice of polar factor, and it coincides with $\tau\ \mathrm{Ortho}(S)$ for the thin polar factor.

Thus, $\xi^* = XA^* + X_\perp B^* = \tau\big(X\ \mathrm{skew}(\mathrm{Ortho}(\mathrm{skew}(X^\top \nabla_X f))) + X_\perp\ \mathrm{Ortho}(X_\perp^\top \nabla_X f)\big)$. □

*Remark* D.2. With the thin polar factor, $\mathrm{Ortho}(S)$ is skew-symmetric for every $r$. If the polar factor is computed from a full SVD and $S$ is singular, which always happens for odd $r$ since $\det S = 0$, it fills the null directions and need not be skew-symmetric. In this case, the skew projection $\hat{Z} = \frac{\tau}{2}(\mathrm{Ortho}(S) - \mathrm{Ortho}(S)^\top)$ achieves the same objective as $\tau\ \mathrm{Ortho}(S)$, since the null directions contribute zero to $\langle \cdot, S\rangle$, and satisfies $\|\hat{Z}\|_2 \le \tau \|\ \mathrm{Ortho}(S)\|_2 = \tau$. Either way, $A^* \in \mathrm{Skew}(r)$ with optimal value $\tau\, \varphi^\circ(S)$.

### D.5 Proof of Theorem 5.3

**Lemma D.3** (LMO objective value and update norm bound)**.** *Let $\xi_t^*$ be the solution of the intrinsic LMO* (4) *at $x_t$ with norm bound $\tau$, and let $Z_t^* = G_{x_t}^{1/2}\xi_t^*$ and $H_t = G_{x_t}^{1/2}(\mathrm{grad}\, f(x_t))$. Then:*

*(i)* (Objective value) $g_{x_t}(\xi_t^*, \mathrm{grad}\, f_t) = \tau\, \varphi^\circ(H_t)$.

*(ii)* (Update norm bound) $\|\xi_t^*\|_{x_t}^2 \le C_\varphi\, \tau^2$.

*Proof. Part (i).* By the change of variables $Z = G_{x_t}^{1/2}\xi$ (Proposition 3.4), the LMO objective equals $\langle Z_t^*, H_t\rangle$. By Proposition 3.4, $Z_t^*$ attains the maximum of $\langle Z, H_t\rangle$ over all $Z$ with $\varphi(Z) \le \tau$, without the constraint $Z \in \mathcal{Z}_{x_t}$. By Lemma C.1, this maximum equals $\tau\, \varphi^\circ(H_t)$.

*Part (ii).* The Riemannian norm of the update is $\|\xi_t^*\|_{x_t}^2 = \|G_{x_t}^{1/2}\xi_t^*\|_F^2 = \|Z_t^*\|_F^2$. By Definition 5.1, $C_\varphi(x_t) = \max\{\|Z\|_F^2 : \varphi(Z) \le 1,\ Z \in \mathcal{Z}_{x_t}\}$, so for any $Z \in \mathcal{Z}_{x_t}$ with $\varphi(Z) \le \tau$, we have $\|Z\|_F^2 = \tau^2\|Z/\tau\|_F^2 \le \tau^2 C_\varphi(x_t) \le C_\varphi\, \tau^2$. Applying this to $Z_t^*$ gives $\|Z_t^*\|_F^2 \le C_\varphi \tau^2$. □

*Proof of Theorem 5.3.* By retraction-$L$-smoothness (Assumption 5.2), with $x_{t+1} = R_{x_t}(-\eta\xi_t^*)$:

$$\begin{aligned} f(x_{t+1}) &\le f(x_t) - \eta\, g_{x_t}(\xi_t^*, \mathrm{grad}\, f_t) + \frac{L\eta^2}{2}\|\xi_t^*\|_{x_t}^2 \\ &= f(x_t) - \eta\, D_t + \frac{L\eta^2}{2}\|Z_t^*\|_F^2 \\ &\le f(x_t) - \eta\tau\, \varphi^\circ(H_t) + \frac{L\eta^2}{2}\, C_\varphi\, \tau^2. \end{aligned} \tag{7}$$

Here we used $\langle \xi_t^*, \nabla f_t\rangle = g_{x_t}(\xi_t^*, \mathrm{grad}\, f_t) = \langle Z_t^*, H_t\rangle = D_t$ and $\|Z_t^*\|_F^2 \le C_\varphi\tau^2$ (Lemma D.3). Summing (7) over $t = 0, \dots, T-1$:

$$f(x_T) - f(x_0) \le -\eta\tau \sum_{t=0}^{T-1} \varphi^\circ(H_t) + \frac{L\eta^2 C_\varphi \tau^2 T}{2}.$$

Rearranging and using $f(x_T) \ge f^*$:

$$\frac{1}{T}\sum_{t=0}^{T-1} \varphi^\circ(H_t) \le \frac{\Delta_0}{\eta\tau T} + \frac{L\eta C_\varphi \tau}{2},$$

where $\Delta_0 = f(x_0) - f^*$. Setting $\eta = \sqrt{2\Delta_0/(LC_\varphi\tau^2 T)}$:

$$\min_{t \le T} \varphi^\circ(H_t) \le \frac{1}{T}\sum_{t} \varphi^\circ(H_t) \le \sqrt{\frac{2LC_\varphi\Delta_0}{T}}.$$

□

**Proof for the decaying schedule.** With $\eta_t = \sqrt{2\Delta_0/(LC_\varphi)}/(\tau\sqrt{t+1})$, write $c = \sqrt{2\Delta_0/(LC_\varphi)}$ so that $\eta_t = c/(\tau\sqrt{t+1})$ and $Lc^2C_\varphi/2 = \Delta_0$. The descent step (7) with time-varying $\eta_t$ gives

$$f(x_{t+1}) \le f(x_t) - \frac{c\,\varphi^\circ(H_t)}{\sqrt{t+1}} + \frac{\Delta_0}{t+1}.$$

Summing over $t = 0, \dots, T-1$ and using $f(x_T) \ge f^*$:

$$c\sum_{t=0}^{T-1} \frac{\varphi^\circ(H_t)}{\sqrt{t+1}} \le \Delta_0 + \Delta_0 \sum_{t=0}^{T-1} \frac{1}{t+1} \le \Delta_0(2 + \ln T).$$

Since $\min_t \varphi^\circ(H_t) \cdot \sum_t 1/\sqrt{t+1} \le \sum_t \varphi^\circ(H_t)/\sqrt{t+1}$ and $\sum_{t=0}^{T-1} 1/\sqrt{t+1} \ge 2(\sqrt{T}-1)$:

$$\min_{t\le T} \varphi^\circ(H_t) \le \frac{\Delta_0(2+\ln T)}{2c(\sqrt{T}-1)} = O\left(\sqrt{\frac{LC_\varphi\Delta_0}{T}} \cdot \ln T\right),$$

matching the constant-step-size rate up to a logarithmic factor, without requiring the horizon $T$ in advance.

### D.6 Computation of $C_\varphi$ for the fixed-rank manifold

We compute $C_\varphi = 2r$ for $\mathcal{M}_r$ with $\varphi = \|\cdot\|_2$ (Remark 5.4). The decoupled LMO produces $Z_B^* = \mathrm{Ortho}(H_B) \in \mathbb{R}^{m\times r}$ and $Z_A^* = \mathrm{Ortho}(H_A) \in \mathbb{R}^{r\times n}$, where $H_A := (B^\top B)^{-1/2}\nabla_A f$. Each has rank $\le r$, so $\|Z_B^*\|_F^2 \le r$ and $\|Z_A^*\|_F^2 \le r$. The total Riemannian norm is $\|\xi^*\|_x^2 = \|Z_B^*\|_F^2 + \|Z_A^*\|_F^2 \le 2r$. Hence $C_\varphi = 2r$, and Theorem 5.3 gives rate $O(\sqrt{Lr/T})$.

### D.7 Computation of $C_\varphi$ for SPD, Stiefel, and Grassmann

**SPD ($C_\varphi = n$).** $Z^* = \mathrm{sign}(H) \in \mathrm{Sym}(n)$ has eigenvalues in $\{-1, 0, 1\}$, so $\|Z^*\|_F^2 \le n$.

**Stiefel ($C_\varphi = 2\lfloor r/2\rfloor + \min(m-r, r)$).** With the product-norm decoupling, the skew block gives $Z_S^* = \mathrm{skew}(\mathrm{Ortho}(S)) \in \mathrm{Skew}(r)$ with $\|Z_S^*\|_2 \le 1$. Since the singular values of a skew-symmetric matrix come in equal pairs, $\mathrm{rank}(Z_S^*) \le 2\lfloor r/2\rfloor$, so $\|Z_S^*\|_F^2 \le 2\lfloor r/2\rfloor$. The normal block gives $Z_N^* = \mathrm{Ortho}(N) \in \mathbb{R}^{(m-r)\times r}$ with rank $\le \min(m-r, r)$, so $\|Z_N^*\|_F^2 \le \min(m-r, r)$. The total is $C_\varphi = 2\lfloor r/2\rfloor + \min(m-r, r)$.

**Grassmann ($C_\varphi = \min(m-r, r)$).** $Z^* = \mathrm{Ortho}(\hat{H}) \in \mathbb{R}^{m\times r}$ with $\hat{H} = (I - XX^\top)\nabla_X f$ having column space in $\mathrm{col}(X)^\perp \cong \mathbb{R}^{m-r}$. Thus $\mathrm{Ortho}(\hat{H})$ has rank $\le \min(m-r, r)$, giving $\|Z^*\|_F^2 \le \min(m-r, r)$.

### D.8 Proof of Theorem 5.7

*Proof of Theorem 5.7.* The LMO applied to the stochastic gradient $\widetilde{\nabla} f_t$ produces $\tilde{\xi}_t^*$ with scaled variable $\tilde{Z}_t^* = G_{x_t}^{1/2}\tilde{\xi}_t^*$ satisfying $\varphi(\tilde{Z}_t^*) \le \tau$. The per-step descent (Assumption 5.2) gives

$$f(x_{t+1}) \le f(x_t) - \eta_t\, g_{x_t}(\tilde{\xi}_t^*, \mathrm{grad}\, f_t) + \frac{L\eta_t^2}{2}\, C_\varphi\, \tau^2.$$

The inner product decomposes as

$$g_{x_t}(\tilde{\xi}_t^*, \mathrm{grad}\, f_t) = \langle \tilde{Z}_t^*,\, H_t\rangle = \langle \tilde{Z}_t^*,\, \tilde{H}_t\rangle + \langle \tilde{Z}_t^*,\, H_t - \tilde{H}_t\rangle.$$

The first term equals $\tau\,\varphi^\circ(\tilde{H}_t)$ by LMO optimality (Lemma D.3). For the second, Hölder's inequality gives $|\langle \tilde{Z}_t^*,\, H_t - \tilde{H}_t\rangle| \le \varphi(\tilde{Z}_t^*)\,\varphi^\circ(H_t - \tilde{H}_t) \le \tau\,\varphi^\circ(\tilde{H}_t - H_t)$. By the reverse triangle inequality, $\varphi^\circ(\tilde{H}_t) \ge \varphi^\circ(H_t) - \varphi^\circ(\tilde{H}_t - H_t)$. Combining:

$$g_{x_t}(\tilde{\xi}_t^*, \mathrm{grad}\, f_t) \ge \tau\big[\varphi^\circ(H_t) - 2\,\varphi^\circ(\tilde{H}_t - H_t)\big]. \tag{8}$$

Substituting into the descent with $\eta_t = c/(\tau\sqrt{t+1})$:

$$f(x_{t+1}) \le f(x_t) - \frac{c\,\varphi^\circ(H_t)}{\sqrt{t+1}} + \frac{2c\,\varphi^\circ(\tilde{H}_t - H_t)}{\sqrt{t+1}} + \frac{Lc^2 C_\varphi}{2(t+1)}.$$

Taking expectations conditional on $x_0, \dots, x_t$ and using $\mathbb{E}[\varphi^\circ(\tilde{H}_t - H_t) \,|\, x_0, \dots, x_t] \le \sigma_\varphi$ (by Jensen and Assumption 5.6), then summing over $t = 0, \dots, T-1$:

$$c\sum_{t=0}^{T-1} \frac{\mathbb{E}[\varphi^\circ(H_t)]}{\sqrt{t+1}} \;\le\; \Delta_0 + 2c\sigma_\varphi \sum_{t=0}^{T-1} \frac{1}{\sqrt{t+1}} + \frac{Lc^2 C_\varphi}{2} \sum_{t=0}^{T-1} \frac{1}{t+1}.$$

Using $\sum 1/\sqrt{t+1} \le 2\sqrt{T}$ and $\sum 1/(t+1) \le 1 + \ln T$, dividing both sides by $c\sum_{t=0}^{T-1} 1/\sqrt{t+1}$, and noting that the $\sigma_\varphi$ term reduces exactly to $2\sigma_\varphi$:

$$\min_{t \le T} \mathbb{E}[\varphi^\circ(H_t)] \;\le\; \frac{\Delta_0}{c\sum_{t=0}^{T-1} 1/\sqrt{t+1}} + 2\sigma_\varphi + \frac{Lc\,C_\varphi(1+\ln T)}{2\sum_{t=0}^{T-1} 1/\sqrt{t+1}}.$$

Using $\sum_{t=0}^{T-1} 1/\sqrt{t+1} \ge \sqrt{T}$ in the denominator (each term is at least $1/\sqrt{T}$), which holds for all $T \ge 1$:

$$\min_{t \le T} \mathbb{E}[\varphi^\circ(H_t)] \;\le\; \frac{1}{\sqrt{T}}\Big(\frac{\Delta_0}{c} + \frac{Lc\,C_\varphi(1+\ln T)}{2}\Big) + 2\sigma_\varphi. \tag{9}$$

Setting $c = \sqrt{2\Delta_0/(LC_\varphi)}$ yields $O\big(\sqrt{LC_\varphi\Delta_0}\,\ln T/\sqrt{T}\big) + 2\sigma_\varphi$, recovering Theorem 5.3 (up to log factors) when $\sigma_\varphi = 0$. □

# E Singular Value Invariance of the Scaled Tangent Spaces

The closed-form solutions derived in Section 4 all rely on Proposition 3.4, through its condition that the scaled tangent space $\mathcal{Z}_x = \{G_x^{1/2}\xi : \xi \in T_x\mathcal{M}\}$ be singular value (SV) invariant (Definition 3.2). Recall that SV-invariance means: if $U\Sigma V^\top \in \mathcal{Z}_x$ is a compact SVD, then $UDV^\top \in \mathcal{Z}_x$ for every nonnegative diagonal $D$ that takes equal values on equal singular values. For such $D$, the matrix $UDV^\top$ does not depend on the choice of the compact SVD, so it suffices to check one compact SVD of each element. When this holds, the tangent-space membership constraint $Z \in \mathcal{Z}_x$ in the LMO becomes redundant: the unconstrained maximizer $Z^*$ of $\langle Z, H\rangle$ subject to $\varphi(Z) \le \tau$ automatically lands in $\mathcal{Z}_x$ (because $Z^*$ shares singular vectors with $H \in \mathcal{Z}_x$, and SV-invariance allows replacing singular values freely). This is what enables the clean closed forms $\mathrm{Ortho}(H)$, $H/\|H\|_F$, and $u_1 v_1^\top$ without any projection step.

Below we verify SV-invariance for each manifold treated in Section 4. For the fixed-rank and Grassmann manifolds, $\mathcal{Z}_x$ is a standard matrix space and the verification is immediate. For the SPD manifold, SV-invariance follows from the structure of symmetric matrices. For the Stiefel manifold, the full tangent space is *not* SV-invariant, which is why Section 4 uses the product-norm decoupling into skew and normal blocks. We verify SV-invariance of each block separately.

**Fixed-rank manifold $\mathcal{M}_r$.** With the coupled metric, $G_x^{1/2}(\dot{B}, \dot{A}) = (\dot{B}(AA^\top)^{1/2}, (B^\top B)^{1/2}\dot{A})$ and the tangent space is $T_x\mathcal{M}_r = \mathbb{R}^{m\times r} \times \mathbb{R}^{r\times n}$. The scaled tangent space decomposes blockwise as $\mathcal{Z}_x = \mathbb{R}^{m\times r} \times \mathbb{R}^{r\times n}$, where each block is the full ambient matrix space of that dimension. Full matrix spaces are trivially SV-invariant (every $m \times r$ matrix lies in $\mathbb{R}^{m\times r}$, regardless of its singular values). This is why the LMO on $\mathcal{M}_r$ (Corollary 4.1) reduces to independent unconstrained problems on each factor block.

**SPD manifold $S_{++}^n$.** With the affine-invariant metric, $G_X^{1/2}\xi = X^{-1/2}\xi X^{-1/2}$ maps $T_X S_{++}^n = \mathrm{Sym}(n)$ to $\mathcal{Z}_X = \mathrm{Sym}(n)$. To check SV-invariance: a symmetric matrix $M$ has a compact SVD $M = U\Sigma V^\top$ where $V = U\,\mathrm{diag}(\pm 1)$ (the sign pattern encodes the signs of the eigenvalues). Replacing $\Sigma$ with any admissible non-negative diagonal $D$ gives $UDV^\top = UD\,\mathrm{diag}(\pm 1)U^\top$, which is symmetric since $D\,\mathrm{diag}(\pm 1)$ is diagonal. Hence $\mathrm{Sym}(n)$ is SV-invariant, and the LMO on $S_{++}^n$ reduces to the unconstrained problem yielding $\mathrm{sign}(H)$, $H/\|H\|_F$, or rank-1 truncation.

**Stiefel manifold $\mathbf{St}(m, r)$ (block-wise).** With $G_X = I$, the scaled tangent space coincides with the tangent space itself: $\mathcal{Z}_X = T_X\mathrm{St} = \{XA + X_\perp B : A \in \mathrm{Skew}(r),\ B \in \mathbb{R}^{(m-r)\times r}\}$. Since $[X \mid X_\perp] \in O(m)$, any $\xi \in T_X\mathrm{St}$ can be written as $[X, X_\perp]\left(\begin{smallmatrix} A \\ B \end{smallmatrix}\right)$, so $\mathcal{Z}_X$ is isometric to $\{\left(\begin{smallmatrix} A \\ B \end{smallmatrix}\right) : A \in \mathrm{Skew}(r),\ B \in \mathbb{R}^{(m-r)\times r}\}$. This space is *not* jointly SV-invariant: if $\left(\begin{smallmatrix} A \\ B \end{smallmatrix}\right) = U\Sigma V^\top$, replacing $\Sigma$ with a different $D$ can produce a top block that is no longer skew-symmetric. This is precisely why Section 4 uses the product-norm $\max(\varphi(A), \varphi(B)) \le \tau$ to decouple the two blocks, and SV-invariance is checked for each block independently:

- *Normal block* $\mathbb{R}^{(m-r)\times r}$: the full ambient space, trivially SV-invariant.
- *Skew block* $\mathrm{Skew}(r)$: a skew-symmetric matrix has real Schur form $S = Q\,\mathrm{blkdiag}(\sigma_1 J, \sigma_2 J, \ldots)\,Q^\top$ with $J = \left(\begin{smallmatrix} 0 & 1 \\ -1 & 0 \end{smallmatrix}\right)$, so its singular values come in equal pairs $(\sigma_1, \sigma_1, \sigma_2, \sigma_2, \ldots)$. Replacing $\sigma_i$ with any $d_i \ge 0$ gives $Q\,\mathrm{blkdiag}(d_1 J, d_2 J, \ldots)\,Q^\top$, which remains skew-symmetric. Since an admissible $D$ takes equal values on each pair of equal singular values, $\mathrm{Skew}(r)$ is SV-invariant.

For $\varphi = \|\cdot\|_2$, the skew-block LMO sets all nonzero pairs to $(1, 1)$, producing the thin polar factor $\mathrm{Ortho}(S)$. For $\varphi = \|\cdot\|_*$, it concentrates on the leading pair, giving a rank-2 skew-symmetric matrix.

**Grassmann manifold $\mathbf{Gr}(m, r)$.** With $G_X = I$ and $X^\top X = I_r$, the horizontal space is $\mathcal{Z}_X = T_X\mathrm{Gr} = \{Z \in \mathbb{R}^{m\times r} : X^\top Z = 0\}$. If $M = U\Sigma V^\top \in \mathcal{Z}_X$, then $X^\top M = 0$ implies $\mathrm{col}(M) \subseteq \mathrm{col}(X)^\perp$. In a compact SVD, the columns of $U$ span $\mathrm{col}(M) \subseteq \mathrm{col}(X)^\perp$. Replacing $\Sigma$ with any non-negative diagonal $D$ gives $UDV^\top$, whose column space is still in $\mathrm{col}(X)^\perp$, so $X^\top(UDV^\top) = 0$. Hence $T_X\mathrm{Gr}$ is SV-invariant. No product-norm relaxation is needed on Grassmann, and the LMO reduces directly to the unconstrained problem.

# F Derivations and Implementation Details

This appendix provides complete derivations of the intrinsic LMO solutions for all three standard norms (spectral, Frobenius, nuclear) on each manifold, along with step-by-step implementation procedures for the spectral instance. The general pipeline from Proposition 3.4 is: (i) scale $H = G_x^{1/2}(\mathrm{grad}\, f)$, (ii) solve $Z^* = \arg\max_{\varphi(Z)\le\tau}\langle Z,\ H\rangle$ in closed form via Section C, (iii) invert $\xi^* = G_x^{-1/2} Z^*$. The three standard norms give $Z^* = \tau\ \mathrm{Ortho}(H)$ (spectral), $Z^* = \tau\, H/\|H\|_F$ (Frobenius), and $Z^* = \tau\, u_1 v_1^\top$ (nuclear). Throughout, $\nabla_X f$ denotes the Euclidean gradient of $f$ at $X \in \mathcal{M}$.

## F.1 Summary of closed-form solutions

Table 3 collects the closed-form intrinsic LMO solutions on four manifolds under three norms.

Table 4 complements Table 3 by stating, for each manifold, whether the update solves the joint LMO (4) or the $\ell_\infty$-product LMO (6), and the resulting relaxation factor where a joint norm exists.

Table 3: Closed-form intrinsic LMO solutions $\xi^*$ (up to the norm bound $\tau$) on four manifolds. $H_B = \nabla_B f\,(AA^\top)^{-1/2}$ and $H_A = (B^\top B)^{-1/2}\nabla_A f$ are the scaled factor gradients; $H = X^{1/2}(\nabla_X f)X^{1/2}$ is the scaled SPD gradient; $S = \mathrm{skew}(X^\top\nabla_X f)$ and $N = X_\perp^\top\nabla_X f$ are the Stiefel skew and normal blocks; $\hat{H} = (I - XX^\top)\nabla_X f$ is the Grassmann horizontal gradient.

| **Manifold** | $G_x^{1/2}(\xi)$ | $\varphi = \|\cdot\|_2$ (spectral) | $\varphi = \|\cdot\|_F$ (Frobenius) | $\varphi = \|\cdot\|_*$ (nuclear) |
|---|---|---|---|---|
| $\mathcal{M}_r$ ($\dot{B}$-block) | $\dot{B}(AA^\top)^{1/2}$ | $\mathrm{Ortho}(H_B)\,(AA^\top)^{-1/2}$ | $\nabla_B f\,(AA^\top)^{-1}/\|H_B\|_F$ | $u_1 v_1^\top (AA^\top)^{-1/2}$ |
| $S_{++}^n$ | $X^{-1/2}\xi\, X^{-1/2}$ | $X^{1/2}\,\mathrm{sign}(H)X^{1/2}$ | $X(\nabla_X f)X/\|\mathrm{grad}\, f\|_X$ | $\mathrm{sign}(\lambda_1)\, X^{1/2} q_1 q_1^\top X^{1/2}$ |
| $\mathrm{St}(m, r)$ | $\xi$ | $X\ \mathrm{skew}(\mathrm{Ortho}(S)) + X_\perp\ \mathrm{Ortho}(N)$ | $X(S/\|S\|_F) + X_\perp(N/\|N\|_F)$ | rank-2 skew + rank-1 |
| $\mathrm{Gr}(m, r)$ | $\xi$ | $\mathrm{Ortho}(\hat{H})$ | $\hat{H}/\|\hat{H}\|_F$ | $u_1 v_1^\top$ of $\hat{H}$ |

## F.2 Computing the polar factor

Table 4: Intrinsic LMO formulation solved on each manifold. For every row the solution is in closed form for the spectral, Frobenius, nuclear, partial polar, and Schatten norms (Section C). Any other unitarily invariant norm reduces to a $\min(p, q)$-dimensional convex program over singular values (Proposition 3.4). The last column gives the relaxation factor of the product constraint relative to the joint norm when a joint norm exists.

| Manifold | LMO solved | Relaxation of the joint norm |
|---|---|---|
| $\mathcal{M}_r$ | $\ell_\infty$-product LMO (6) over the factor blocks $(\dot{B}, \dot{A})$ | not applicable, since the two blocks have different shapes and no joint unitarily invariant norm is defined on the pair |
| $S^n_{++}$ | joint LMO (4) | none, the joint LMO is solved exactly |
| $\mathrm{St}(m, r)$ | $\ell_\infty$-product LMO (6) over the skew and normal blocks $(A, B)$ | factor $\sqrt{2}$ for the spectral norm, since $\max\{\|A\|_2, \|B\|_2\} \le \|[A; B]\|_2 \le \sqrt{2}\max\{\|A\|_2, \|B\|_2\}$ |
| $\mathrm{Gr}(m, r)$ | joint LMO (4) | none, the joint LMO is solved exactly |

**Newton–Schulz iterations.** All spectral-norm instances require computing the polar factor $\mathrm{Ortho}(M) = UV^\top$ from $M = U\Sigma V^\top$. Following Jordan et al. [33], this can be approximated by Newton–Schulz iterations: $X_0 = M/\|M\|_F$, $X_{k+1} = \frac{3}{2}X_k - \frac{1}{2}X_kX_k^\top X_k$. Each iteration costs $O(pr^2)$ for $M \in \mathbb{R}^{p\times r}$ and converges quadratically when $\|X_0\|_2 < \sqrt{3}$, which is ensured by the normalization since $\|X_0\|_2 = \|M\|_2/\|M\|_F \le 1$. In practice, 5–10 iterations suffice.

The choice of algorithm depends on the shape and size of $M$:

*Tall-skinny or short-wide matrices* ($M \in \mathbb{R}^{p\times r}$ with $r \ll p$): dominant case on $\mathcal{M}_r$, $\mathrm{Gr}(m, r)$, and the normal block of $\mathrm{St}(m, r)$. Thin SVD costs $O(pr^2)$; Newton–Schulz avoids the SVD and is GPU-friendly.

*Small square matrices* ($M \in \mathbb{R}^{r\times r}$): arises for the skew block on $\mathrm{St}(m, r)$. For skew-symmetric $S$, the real Schur decomposition gives $\mathrm{Ortho}(S)$ directly.

*Large square matrices* ($M \in \mathbb{R}^{n\times n}$): arises on $S^n_{++}$ where the "polar factor" is the matrix sign $\mathrm{sign}(H)$. Newton iterations $X_{k+1} = \frac{1}{2}(X_k + X_k^{-1})$ converge quadratically; eigendecomposition is also standard.

### F.3 Fixed-rank manifold $\mathcal{M}_r$

**Derivation (all three norms).** On $\mathcal{M}_r$ with the coupled metric, $G_x^{1/2}(\dot{B}, \dot{A}) = (\dot{B}(AA^\top)^{1/2}, (B^\top B)^{1/2}\dot{A})$. The LMO decouples into independent $\dot{B}$ and $\dot{A}$ subproblems via (6). For the $\dot{B}$-block, the scaled gradient is $H_B = \nabla_B f\,(AA^\top)^{-1/2}$ where $\nabla_B f = \nabla_X f A^\top$. By Proposition 3.4, $Z_B^*$ is the unconstrained maximizer of $\langle Z_B, H_B\rangle$ subject to $\varphi(Z_B) \le \tau$, and $\dot{B}^* = Z_B^*(AA^\top)^{-1/2}$. The three norms give:

- *Spectral* ($\varphi = \|\cdot\|_2$): $Z_B^* = \tau\,\mathrm{Ortho}(H_B)$, so $\dot{B}^* = \tau\,\mathrm{Ortho}(\nabla_B f\,(AA^\top)^{-1/2})\,(AA^\top)^{-1/2}$.
- *Frobenius* ($\varphi = \|\cdot\|_F$): $Z_B^* = \tau\,H_B/\|H_B\|_F$, so $\dot{B}^* = \tau\,\nabla_B f\,(AA^\top)^{-1}/(\|H_B\|_F)$. This is the Riemannian preconditioned gradient of Mishra and Sepulchre [46], Zhang and Pilanci [81], normalized to unit intrinsic Frobenius norm.
- *Nuclear* ($\varphi = \|\cdot\|_*$): $Z_B^* = \tau\,u_1v_1^\top$ from the SVD of $H_B$, so $\dot{B}^* = \tau\,u_1v_1^\top(AA^\top)^{-1/2}$ (rank-1 update per block).

The $\dot{A}$-block follows by symmetry with $H_A = (B^\top B)^{-1/2}\nabla_A f$.

**QR-based implementation (spectral norm).** We show that $\dot{B} = \tau\,\mathrm{Ortho}(H_B)\,(AA^\top)^{-1/2}$ can be computed using only QR factors. Let $A^\top = \hat{Q}_A R_A$ be a thin QR factorization ($\hat{Q}_A \in \mathbb{R}^{n\times r}$ with orthonormal columns, $R_A \in \mathbb{R}^{r\times r}$ upper triangular with positive diagonal); similarly $B = \hat{Q}_B R_B$.

The scaled factor gradient is the $m \times r$ matrix

$$H_B = \nabla_B f \, (AA^\top)^{-1/2} = \nabla_X f A^\top (AA^\top)^{-1/2} = \nabla_X f \hat{Q}_A R_A (AA^\top)^{-1/2}.$$

Writing $Q_A := (AA^\top)^{-1/2} A \in \mathbb{R}^{r \times n}$ (orthonormal rows), we have $H_B = \nabla_X f Q_A^\top$. Since $Q_A^\top$ and $\hat{Q}_A$ span the same column space, $Q_A^\top = \hat{Q}_A \Omega$ for some $\Omega \in O(r)$. Thus $\mathrm{Ortho}(H_B) = \mathrm{Ortho}(\nabla_X f \hat{Q}_A)\, \Omega$. Moreover, $AA^\top = R_A^\top R_A$, so $R_A = \Omega \, (AA^\top)^{1/2}$ is the polar decomposition of $R_A$ and $\Omega (AA^\top)^{-1/2} = R_A^{-\top}$ exactly.

In practice we compute $\dot{B} = \tau \ \mathrm{Ortho}(\nabla_X f \hat{Q}_A)\, R_A^{-\top}$ and $\dot{A} = \tau \, R_B^{-1} \ \mathrm{Ortho}(\hat{Q}_B^\top \nabla_X f)$.

**Step-by-step procedure** ($\dot{B}$ computation; $\dot{A}$ is symmetric):

1. **QR factorization**: $A^\top = \hat{Q}_A R_A$; similarly $B = \hat{Q}_B R_B$. $O((m+n)r^2)$.
2. **Projected gradient**: $\nabla_X f \, \hat{Q}_A = (\nabla_B f)\, R_A^{-1} \in \mathbb{R}^{m \times r}$. $O(mr^2)$.
3. **Polar factor**: $\mathrm{Ortho}(\nabla_X f \, \hat{Q}_A)$ via Newton–Schulz. $O(mr^2)$ per iteration.
4. **Triangular solve**: $\dot{B} = \tau \ \mathrm{Ortho}(\nabla_X f \, \hat{Q}_A)\, R_A^{-\top}$. $O(mr^2)$.
5. **Retract**: $B_+ = B - \eta \, \dot{B}$, $A_+ = A - \eta \, \dot{A}$. $O((m+n)r)$.

**Total**: $O((m+n)r^2 + r^3)$ per step. No eigendecomposition or matrix square root is needed.

### F.4 SPD manifold $S_{++}^n$

**Derivation (all three norms).** With the affine-invariant metric, $G_X^{1/2} \xi = X^{-1/2} \xi X^{-1/2}$ and $\mathrm{grad}\, f(X) = X(\nabla_X f)X$. The scaled gradient is $H = G_X^{1/2}(\mathrm{grad}\, f) = X^{1/2}(\nabla_X f)X^{1/2} \in \mathrm{Sym}(n)$. Since $\mathcal{Z}_X = \mathrm{Sym}(n)$ is SV-invariant, Proposition 3.4 applies. The three norms give $Z^*$ and $\xi^* = X^{1/2} Z^* X^{1/2}$:

- *Spectral*: $Z^* = \tau \ \mathrm{sign}(H) = \tau \, Q \, \mathrm{diag}(\mathrm{sign}(\lambda_i)) Q^\top$, so $\xi^* = \tau \, X^{1/2} \, \mathrm{sign}(H) X^{1/2}$.
- *Frobenius*: $Z^* = \tau \, H / \|H\|_F$, so $\xi^* = \tau \, \mathrm{grad}\, f / \|\mathrm{grad}\, f\|_X$ (the normalized natural gradient).
- *Nuclear*: $Z^* = \tau \ \mathrm{sign}(\lambda_1)\, q_1 q_1^\top$, so $\xi^* = \tau \ \mathrm{sign}(\lambda_1)\, X^{1/2} q_1 q_1^\top X^{1/2}$ (rank-1 symmetric).

All three are $\mathrm{GL}(n)$-equivariant by Proposition 3.1: under $X \mapsto NXN^\top$, the update transforms as $\xi^* \mapsto N \xi^* N^\top$.

**Implementation (spectral norm).**

1. **Matrix square root**: Compute $X^{1/2}$ and $X^{-1/2}$ via eigendecomposition of $X$. $O(n^3)$.
2. **Scaled gradient**: $H = X^{1/2} \, \nabla_X f \, X^{1/2}$. $O(n^3)$.
3. **Matrix sign**: $\mathrm{sign}(H) = Q \, \mathrm{diag}(\mathrm{sign}(\lambda_i)) Q^\top$ via eigendecomposition, or Newton iterations $Z_{k+1} = \frac{1}{2}(Z_k + Z_k^{-1})$. $O(n^3)$.
4. **Invert**: $\xi^* = \tau \, X^{1/2} \ \mathrm{sign}(H) \, X^{1/2}$. $O(n^3)$.
5. **Retract**: $X_+ = X^{1/2} \exp(-\eta \, X^{-1/2} \xi^* X^{-1/2}) \, X^{1/2}$. $O(n^3)$.

**Total**: $O(n^3)$ per step, dominated by eigendecompositions.

### F.5 Stiefel manifold $\mathrm{St}(m, r)$

**Derivation (all three norms).** With $G_X = I$, the tangent space decomposes as $\xi = XA + X_\perp B$ with $A \in \mathrm{Skew}(r)$, $B \in \mathbb{R}^{(m-r) \times r}$. Under the product norm $\max(\varphi(A), \varphi(B)) \le \tau$, the LMO decouples into skew and normal blocks with $S = \mathrm{skew}(X^\top \nabla_X f)$ and $N = X_\perp^\top \nabla_X f$.

- *Spectral*: $A^* = \tau \ \mathrm{skew}(\mathrm{Ortho}(S))$, $B^* = \tau \ \mathrm{Ortho}(N)$. Result: $\xi^* = \tau(X \ \mathrm{skew}(\mathrm{Ortho}(S)) + X_\perp \ \mathrm{Ortho}(N))$ (Lemma 4.2).

- *Frobenius*: $A^* = \tau\, S/\|S\|_F$, $B^* = \tau\, N/\|N\|_F$ (block-wise normalized).
- *Nuclear*: $A^* = \frac{\tau}{2}(u_1 v_1^\top + u_2 v_2^\top)$ from the leading pair of singular vectors of $S$, a rank-2 skew-symmetric matrix with $\|A^*\|_* = \tau$, and $B^* = \tau\, p_1 q_1^\top$ from the leading singular vectors of $N$ (rank-1).

**Implementation (spectral norm).** $X_\perp \in \mathbb{R}^{m\times(m-r)}$ is never formed explicitly; instead $X_\perp\ \mathrm{Ortho}(X_\perp^\top \nabla_X f) = \mathrm{Ortho}((I - XX^\top)\nabla_X f)$, where $\mathrm{Ortho}(M) = U_1 V_1^\top$ is formed from the nonzero singular values of $M$. The Newton–Schulz iteration of Section F.2 returns this factor directly. If a thin SVD is used instead, the singular vectors associated with zero singular values are discarded, which is equivalent to forming $X_\perp$ explicitly and using $X_\perp\ \mathrm{Ortho}(X_\perp^\top \nabla_X f)$.

1. **Skew gradient**: $P = X^\top \nabla_X f$, then $S = \frac{1}{2}(P - P^\top)$. $O(mr^2)$.
2. **Skew polar factor**: $\mathrm{skew}(\mathrm{Ortho}(S))$ via real Schur decomposition. $O(r^3)$.
3. **Normal gradient**: $\hat{N} = \nabla_X f - X\, P$. $O(mr^2)$.
4. **Normal polar factor**: $\mathrm{Ortho}(\hat{N})$ via Newton–Schulz. $O(mr^2)$.
5. **Assemble**: $\xi^* = \tau[X\ \mathrm{skew}(\mathrm{Ortho}(S)) + \mathrm{Ortho}(\hat{N})]$. $O(mr^2)$.
6. **Retract**: $X_+ = \mathrm{qf}(X - \eta\, \xi^*)$. $O(mr^2)$.

**Total**: $O(mr^2 + r^3)$ per step. Steps 2 and 4 can be parallelized.

### F.6 Grassmann manifold $\mathbf{Gr}(m, r)$

**Derivation (all three norms).** With $G_X = I$ and $X^\top X = I_r$, the horizontal space $T_X\mathrm{Gr} = \{\xi : X^\top \xi = 0\}$ is SV-invariant. The horizontal gradient is $\hat{H} = (I - XX^\top)\nabla_X f$.

- *Spectral*: $\xi^* = \tau\ \mathrm{Ortho}(\hat{H})$, orthonormal columns in $\mathrm{col}(X)^\perp$.
- *Frobenius*: $\xi^* = \tau\, \hat{H}/\|\hat{H}\|_F$, normalized Riemannian gradient on $\mathrm{Gr}(m, r)$ [20].
- *Nuclear*: $\xi^* = \tau\, u_1 v_1^\top$ from the SVD of $\hat{H}$, rank-1 horizontal update.

**Implementation (spectral norm).**

1. **Horizontal projection**: $\hat{H} = \nabla_X f - X\,(X^\top \nabla_X f)$. $O(mr^2)$.
2. **Polar factor**: $\mathrm{Ortho}(\hat{H})$ via Newton–Schulz. $O(mr^2)$.
3. **Assemble**: $\xi^* = \tau\ \mathrm{Ortho}(\hat{H})$. $O(mr)$.
4. **Retract**: $X_+ = \mathrm{qf}(X - \eta\, \xi^*)$. $O(mr^2)$.

**Total**: $O(mr^2)$ per step. The simplest instantiation: $G_X = I$ eliminates scaling and inversion.

## G Prior Convergence Results as Special Cases

We show how several known rates arise as special cases of Theorems 5.3 and 5.7.

**Euclidean spectral steepest descent [15, 8].** Setting $\mathcal{M} = \mathbb{R}^{p\times q}$ (unconstrained), $G_x = I$, and $\varphi = \|\cdot\|_2$ recovers the Euclidean Muon update $\xi^* = \tau\ \mathrm{Ortho}(\nabla f)$. The use of this spectral-norm steepest-descent operator for deep learning was first proposed by Carlson et al. [15], who also combined it with element-wise adaptive preconditioning; Bernstein and Newhouse [8, Prop. 5] later formalize the connection to steepest descent. The squared Riemannian radius is $C_\varphi = \min(p, q)$ (rank of $\mathrm{Ortho}(\nabla f)$), and Theorem 5.3 gives the nonasymptotic rate $O(\sqrt{L\min(p, q)/T})$.

**Standard Riemannian gradient descent [13].** Setting $\varphi = \|\cdot\|_F$ with the joint LMO (4) gives $\xi^* \propto \operatorname{grad} f/\|\operatorname{grad} f\|_x$ and $C_\varphi = 1$ (under the $\ell_\infty$-product norm on $\mathcal{M}_r$ and $\mathrm{St}(m,r)$, the direction is normalized block-wise and $C_\varphi = 2$). Theorem 5.3 yields $\min_t \|\operatorname{grad} f(x_t)\|_{x_t} \le \sqrt{2L\Delta_0/T}$, which is the standard retraction-smooth Riemannian GD rate [13, Cor. 4.9] (with the normalized-gradient step replacing the unnormalized one; the rates coincide up to constants when $\eta$ is optimized).

**Factor-wise Muon on $\mathcal{M}_r$ [34].** Kang et al. [34] analyze factor-wise Muon (SpecGF) on $\mathcal{M}_r$ with $\varphi = \|\cdot\|_2$. Since their orthogonalization operator satisfies $\|T(\cdot)\|_2 = 1$, the ambient update norm satisfies $\|\dot{X}_t\|_F^2 \le 2\tau^2(\|A_t\|_F^2 + \|B_t\|_F^2)$, which depends on factor conditioning. Accordingly, their main convergence result (Theorem 6.2) is conditional: SpecGF either converges to a global minimum or $\|A_t\|_F^2 + \|B_t\|_F^2 \to \infty$; unconditional global convergence is established only with $\ell_2$ regularization. Our quotient formulation sidesteps this issue entirely: the Gram-root scaling yields a rank-only bound $\|\xi^*\|_x^2 \le 2r\tau^2$ with $C_\varphi = 2r$ (Remark 5.4, computed in Section D.6), giving a nonasymptotic rate $O(\sqrt{Lr/T})$ without any bounded-factor-norm condition.

**NuMuon [50].** Setting $G_x = I$ and $\varphi = \|\cdot\|_*$ gives $\xi^* = \tau\, u_1 v_1^\top$ (rank-1 update) with $C_\varphi = 1$. Theorem 5.3 gives rate $O(\sqrt{L/T})$ with convergence in $\varphi^\circ = \|\cdot\|_2$ (spectral norm of the gradient). NuMuon [50] uses a joint spectral-and-nuclear-norm constraint ($\|O\|_2 \le \rho$, $\|O\|_* \le \tau$), yielding a rank-$k$ update ($k = \lfloor \tau/\rho \rfloor$); their Theorem 3.6 gives a rate with an $Lk$-dependent term and convergence in $\|\nabla f\|_*$ (nuclear norm of the gradient). Our pure nuclear instance ($k = 1$) recovers the rank-1 special case. On manifolds with $G_x \neq I$, our quotient extensions add symmetry invariance while preserving $C_\varphi = O(1)$.

**General Euclidean LMOs [56].** Pethick et al. [56] study LMOs constrained by general norms (including spectral, RMS, and sign norms) in Euclidean space ($G_x = I$), and propose Scion, a unified algorithmic family. Their spectral-norm instance corresponds to our framework with $G_x = I$ on $\mathbb{R}^{p\times q}$. Our contribution beyond their setting is the non-trivial metric $G_x \neq I$ (which introduces the metric-scaling and inversion steps and the symmetry proposition) and the per-manifold closed forms.

**The role of $C_\varphi$.** Across all these instances, the squared Riemannian radius $C_\varphi$ is the unifying quantity: it is $\min(p,q)$ for Euclidean Muon, $O(1)$ for Frobenius/nuclear norms, $2r$ for quotient-Muon on $\mathcal{M}_r$, and also $2r$ for factor-wise Muon (Euclidean metric on the factors), whose dependence on $\|A_t\|_F^2 + \|B_t\|_F^2$ enters instead through the ambient update norm and hence the smoothness constant. The spectral norm always gives the largest $C_\varphi$ but the strongest convergence criterion ($\varphi^\circ = \|\cdot\|_*$). The nuclear norm gives the smallest $C_\varphi$ but the weakest criterion ($\varphi^\circ = \|\cdot\|_2$).

# H Detailed Comparisons with Prior Methods

In the main text (Section 4), we briefly compare our framework to prior methods on the fixed-rank manifold. Here we provide detailed calculations and side-by-side comparisons on all four manifolds. Table 5 summarizes convergence guarantees for selected prior methods and iMuon. The entries use the convergence criteria and assumptions of the cited results.

## H.1 Fixed-rank manifold $\mathcal{M}_r$: iMuon vs. Spectron, Riemannion, and factor-wise Muon

We compare three approaches to spectral-norm optimization on the rank-$r$ manifold $\mathcal{M}_r$ with factorization $X = BA$. Throughout, $\nabla_X f$ denotes the Euclidean gradient.

**Prior methods.** *Factor-wise Muon* applies the standard Euclidean Muon [33, 9, 8] polar factor $\mathrm{Ortho}(\cdot)$ to each factor's gradient independently: $\dot{B} = \tau\,\mathrm{Ortho}(\nabla_X f A^\top)$ and $\dot{A} = \tau\,\mathrm{Ortho}(B^\top \nabla_X f)$. This is simple and requires no Riemannian structure, but the resulting update $\dot{X} = \dot{B}A + B\dot{A}$ has $\|\dot{X}\|_F^2 \le 2\tau^2(\|A\|_F^2 + \|B\|_F^2)$, which depends on factor conditioning. The convergence analysis of Kang et al. [34] for this update is discussed in Section G and Table 5.

Table 5: Convergence guarantees and key assumptions of various prior works and iMuon.

| Method | Convergence guarantee | Key assumptions |
|---|---|---|
| Riemannian AMSGrad / AdamNC [6, Thms. 1–2] | Average regret: $\widetilde{O}(T^{-1/2})$ | Geodesic convexity, compact domain, bounded gradients |
| Scaled GD [81, Thm. 7.4] | Local linear, condition-number independent | Least squares, RIP, spectral initialization |
| Spectron [32] | $\times$ | N/A |
| Riemannion [12] | $\times$ | N/A |
| SpecGF [34, Thm. 6.2] | Global minimum or factor divergence | Least-squares factorization, almost every initialization |
| SpecGF [34, Prop. 6.7] | Conditional exponential loss decay | Convergence to a global minimum, distinct target singular values |
| iMuon, deterministic, Theorem 5.3 | Stationarity: $O(T^{-1/2})$ | Retraction smoothness |
| iMuon, stochastic, Theorem 5.7 | Expected stationarity: $\widetilde{O}(T^{-1/2}) + 2\sigma_\varphi$ | Retraction smoothness; bounded scaled noise |

$\times$: no optimization convergence rate established in the cited work. $\widetilde{O}$ suppresses logarithmic factors. The Riemannian adaptive bounds depend on curvature and domain diameter. SpecGF is a continuous-time flow (with a smoothed orthogonalization map) for least-squares factorization. For iMuon, rates concern best-iterate stationarity in $\varphi^\circ(H_t)$, with constants scaling as $\sqrt{LC_\varphi\Delta_0}$; $C_\varphi = 2r$ for the fixed-rank spectral product formulation. Full assumptions and step-size conditions are given in the cited works.

*Spectron* [32] addresses the same factor-growth pathology from a practical angle. Like factor-wise Muon, Spectron orthogonalizes each factor's momentum via Newton–Schulz iterations, producing $\mathrm{Ortho}(M_A)$ and $\mathrm{Ortho}(M_B)$. The key innovation is a *spectral renormalization* step that adaptively scales the updates to control the composite update norm. Since the ambient update satisfies $\|\dot{X}\|_2 \le \rho(\|A\|_2+\|B\|_2+1)$ (via submultiplicativity and the triangle inequality), Spectron sets the per-factor constraint radius to $\rho = \eta/(\|A\|_2 + \|B\|_2 + 1)$, yielding factor updates $\dot{A} = \rho\ \mathrm{Ortho}(M_A)$ and $\dot{B} = \rho\ \mathrm{Ortho}(M_B)$, which guarantee $\|\dot{X}\|_2 \le \eta$. The spectral norms $\|A\|_2, \|B\|_2$ are estimated at each iteration via a single power iteration step. This makes training stable and enables native low-rank pretraining that matches dense-model performance.

Spectron and our method address the same instability but through fundamentally different mechanisms. Spectron is an *extrinsic* fix: it applies standard (Euclidean) orthogonalization to each factor and then rescales the result by an iterate-dependent quantity $1/(\|A\|_2 + \|B\|_2 + 1)$ to bound the ambient update. The rescaling shrinks as the factors grow, which is effective in practice but means the effective step size is not a fixed hyperparameter. It depends on the current factor norms. Our method is *intrinsic*: the Gram-root preconditioning $(AA^\top)^{-1/2}$ from the coupled metric absorbs the factor scaling *before* orthogonalization, so the resulting update is $\mathrm{GL}(r)$-equivariant by construction (Proposition 3.1), its induced update of $X = BA$ does not depend on the factorization, and its squared Riemannian norm is bounded by $2r\tau^2$ regardless of factor magnitudes (Section D.6), with no power iteration or adaptive rescaling needed.

*Riemannion* [12] takes an *embedded-manifold* approach. It optimizes $X$ directly on $\mathcal{M}_r = \{X \in \mathbb{R}^{m\times n} : \mathrm{rank}(X) = r\}$, parameterized via a compact SVD $X = A_L\Sigma_r B_R^\top$ with $A_L \in \mathrm{St}(m,r)$, $B_R \in \mathrm{St}(n,r)$, and $\Sigma_r \in \mathbb{R}^{r\times r}$ diagonal positive. At a point $X \in \mathcal{M}_r$, the tangent space $T_X\mathcal{M}_r$ consists of rank-$\le 2r$ matrices; in SVD coordinates, $\dot{X} = A_L M B_R^\top + U_\perp B_R^\top + A_L V_\perp^\top$ for $M \in \mathbb{R}^{r\times r}$, $U_\perp \perp A_L$, and $V_\perp \perp B_R$. Riemannion maintains a momentum $M_t \in T_X\mathcal{M}_r$ (rank $\le 2r$) via Riemannian heavy-ball updates with vector transport. To generalize Muon, it applies a rank-$2r$ orthogonalization $\mathrm{Ortho}_r(M_t)$, replacing only the first $2r$ singular values of $M_t$ with 1 while setting the rest to 0, and then projects back onto the tangent space: $\tilde{M}_t = P_{T_X\mathcal{M}_r}(\mathrm{Ortho}_r(M_t))$. Because $\mathrm{Ortho}_r$ preserves the column and row spaces of $M_t \in T_X\mathcal{M}_r$, the result is a good approximation (the authors report singular values in $(0.9, 1.1)$ in practice), but not exact. The iterate is then updated via retraction: $X_{t+1} = R_X(-\eta\,\tilde{M}_t)$.

We can equivalently describe the entire procedure in the $X = BA$ notation used throughout this section. The embedded tangent space is $T_X\mathcal{M}_r = \{\dot{B}A + B\dot{A} : \dot{B} \in \mathbb{R}^{m\times r},\ \dot{A} \in \mathbb{R}^{r\times n}\}$, a subspace of rank-$\leq 2r$ matrices in $\mathbb{R}^{m\times n}$. The orthogonal projection of any $Z \in \mathbb{R}^{m\times n}$ onto $T_X\mathcal{M}_r$ (under the Euclidean metric $G_X = I$) is $P_{T_X\mathcal{M}_r}(Z) = P_B\, Z + Z\, P_A - P_B\, Z\, P_A$, where $P_B = B(B^\top B)^{-1}B^\top$ projects onto $\mathrm{col}(B)$ and $P_A = A^\top(AA^\top)^{-1}A$ projects onto $\mathrm{row}(A)$. Riemannion's update in $BA$ language is then: (i) maintain momentum $M_t \in T_X\mathcal{M}_r$ (rank $\leq 2r$); (ii) orthogonalize $\mathrm{Ortho}_r(M_t)$ (set all $2r$ nonzero singular values to 1); (iii) project back via $\tilde{M}_t = P_B\ \mathrm{Ortho}_r(M_t) + \mathrm{Ortho}_r(M_t)\, P_A - P_B\ \mathrm{Ortho}_r(M_t)\, P_A$; (iv) retract $X_{t+1} = R_X(-\eta\, \tilde{M}_t)$. The projection in step (iii) couples $B$ and $A$ through $P_B$ and $P_A$: the result depends on both factors simultaneously, and there is no block-diagonal decoupling into independent $\dot{B}$ and $\dot{A}$ subproblems. Moreover, the projection perturbs the singular values away from 1, so the spectral-norm constraint $\|\tilde{M}_t\|_2 \leq 1$ is only approximately satisfied.

Because Riemannion works on $\mathcal{M}_r$ directly (not on factors $(A, B)$), it is reparameterization-invariant: different factorizations of the same $X$ yield the same update. However, it uses the Euclidean metric ($G_X = I$) on the embedded manifold, which provides no preconditioning. To obtain an exact spectral-norm solution, Riemannion would need to solve the tangent-constrained LMO $\max_{S\in T_X\mathcal{M}_r,\, \|S\|_2\leq 1}\langle M_t,\, S\rangle$, for which Proposition 3.4 does not apply, since $T_X\mathcal{M}_r$ in the embedded viewpoint is neither SV-invariant nor spectral-norm contractive (Section P.2).

*Our method (iMuon)* resolves both issues by working on the *quotient* $\mathcal{M}_r = (\mathbb{R}^{m\times r}_* \times \mathbb{R}^{r\times n}_*)/\mathrm{GL}(r)$ with the coupled metric. The intrinsic scaling $G_x^{1/2}$ converts the factor gradients into Euclidean coordinates via the Gram-root preconditioning $(AA^\top)^{-1/2}$, and the product-norm decoupling yields independent closed-form LMOs on each factor block. Unlike Riemannion, the spectral-norm constraint is *exactly* satisfied on each block (no approximation), and the coupled metric provides automatic preconditioning that makes the convergence rate independent of factor conditioning ($C_\varphi = 2r$). Unlike Spectron, no power iteration or adaptive rescaling is needed: the $\mathrm{GL}(r)$-invariance of the intrinsic norm handles factor-scale dependence algebraically rather than numerically.

**Side-by-side update rules.**

| **Method** | $\dot{B}$ | $\dot{A}$ |
|---|---|---|
| Factor-wise Muon | $\tau\ \mathrm{Ortho}(\nabla_X f A^\top)$ | $\tau\ \mathrm{Ortho}(B^\top \nabla_X f)$ |
| Spectron | $\rho\ \mathrm{Ortho}(M_B)$ | $\rho\ \mathrm{Ortho}(M_A)$ |
| Riemannion | $\dot{X} = P_{T_X\mathcal{M}_r}(\mathrm{Ortho}_r(M_t))$, embedded manifold with $G_X = I$ | |
| Ours (iMuon) | $\tau\ \mathrm{Ortho}(\nabla_X f \hat{Q}_A)\, R_A^{-\top}$ | $\tau\, R_B^{-1}\ \mathrm{Ortho}(\hat{Q}_B^\top \nabla_X f)$ |

where $A^\top = \hat{Q}_A R_A$ and $B = \hat{Q}_B R_B$ are thin QR factorizations, $M_A, M_B$ are per-factor momentum buffers, and $\rho = \eta/(\|A\|_2 + \|B\|_2 + 1)$ is Spectron's adaptive constraint radius. The key differences are:

- *Ortho input:* factor-wise Muon and Spectron both apply $\mathrm{Ortho}$ to raw factor gradients (or momenta). Our method applies $\mathrm{Ortho}$ to the *scaled* gradient $\nabla_X f \hat{Q}_A$, which projects $\nabla_X f$ onto the row space of $A$ with orthonormalized rows, a different polar direction in general.
- *Norm control mechanism:* factor-wise Muon uses a fixed $\tau$; Spectron uses the adaptive radius $\rho$ that shrinks as factors grow; Riemannion uses tangent-space projection after rank-$2r$ orthogonalization (approximate). Our method uses the Gram-root preconditioning $R_A^{-\top}$ (equivalently $(AA^\top)^{-1/2}$), which absorbs the factor scale *before* orthogonalization.
- *Reparameterization invariance:* our update is $\mathrm{GL}(r)$-equivariant by Proposition 3.1, i.e., $(\dot{B}^*, \dot{A}^*) \mapsto (\dot{B}^* N^{-1}, N\dot{A}^*)$, so the induced update of $X = BA$ does not depend on the chosen factorization. Riemannion is also reparameterization-invariant (it works on $\mathcal{M}_r$ directly), but violates the spectral-norm constraint after projection. Factor-wise Muon and Spectron are not reparameterization-invariant: different factorizations of the same $X$ yield different updates. Spectron compensates via the adaptive $\rho$, which bounds

the *ambient* update norm but does not make the *direction* invariant. For example, without momentum, the rescaling $(B, A) \mapsto (cB, A/c)$ with $c > 0$ turns Spectron's ambient update $\rho\,[\mathrm{Ortho}(\nabla_X f A^\top)A + B\ \mathrm{Ortho}(B^\top \nabla_X f)]$ into $\rho'\,[c^{-1}\,\mathrm{Ortho}(\nabla_X f A^\top)A + c\,B\ \mathrm{Ortho}(B^\top \nabla_X f)]$, so the two terms are reweighted by $1/c$ and $c$, which the common scalar $\rho'$ cannot undo.

**Ambient update norm comparison.** The ambient update $\dot{X} = \dot{B}A + B\dot{A}$ enters the convergence rate through the descent lemma. We compute $\|\dot{X}\|_F^2$ for each method:

*Our method.* Define $Q_A = (AA^\top)^{-1/2}A \in \mathbb{R}^{r\times n}$, which has orthonormal rows ($Q_A Q_A^\top = I_r$). Then $\dot{B}A = \tau\ \mathrm{Ortho}(H_B)\,(AA^\top)^{-1/2}A = \tau\ \mathrm{Ortho}(H_B)\,Q_A$. Since $\mathrm{Ortho}(H_B) \in \mathbb{R}^{m\times r}$ has orthonormal columns:

$$\|\dot{B}A\|_F^2 = \tau^2\ \mathrm{tr}(Q_A^\top\, I_r\, Q_A) = \tau^2\ \mathrm{tr}(I_r) = \tau^2 r.$$

Similarly $\|B\dot{A}\|_F^2 = \tau^2 r$. By the triangle inequality: $\|\dot{X}\|_F \le 2\tau\sqrt{r}$, so $\|\dot{X}\|_F^2 \le 4r\tau^2$. This bound depends only on the rank $r$, not on the factor norms. The Gram root converts $A$ into the partial isometry $Q_A$, exactly canceling the factor norm dependence.

*Factor-wise Muon.* $\dot{B}A = \tau\ \mathrm{Ortho}(\nabla_X f A^\top)A$. Since $\mathrm{Ortho}(\nabla_X f A^\top)$ has orthonormal columns, $\|\dot{B}A\|_F^2 = \tau^2\|A\|_F^2$. Similarly $\|B\dot{A}\|_F^2 = \tau^2\|B\|_F^2$. Thus $\|\dot{X}\|_F^2 \le 2\tau^2(\|A\|_F^2 + \|B\|_F^2)$, *grows with factor norms*.

*Riemannion.* $\dot{X} = P_{T_X\mathcal{M}_r}(\mathrm{Ortho}_r(M_t))$, where $\mathrm{Ortho}_r$ replaces the first $2r$ singular values of $M_t$ with 1. The tangent-space projection after rank-$2r$ orthogonalization perturbs singular values (they lie in $(0.9, 1.1)$ rather than exactly 1), so the spectral-norm constraint $\|\dot{X}\|_2 \le \tau$ is only approximately satisfied.

**Spectral-norm control.** Our method also controls the spectral norm of the ambient update: $\|\dot{X}\|_2 \le \|\dot{B}A\|_2 + \|B\dot{A}\|_2 \le 2\tau$, since $\|\dot{B}A\|_2 = \tau\|\,\mathrm{Ortho}(H_B)\,Q_A\|_2 \le \tau \cdot 1 \cdot 1 = \tau$ (orthonormal columns times orthonormal rows). Factor-wise Muon gives $\|\dot{X}\|_2 \le \tau(\|A\|_2 + \|B\|_2)$, which can be arbitrarily large.

**Nuclear norm ($\varphi = \|\cdot\|_*$).** The nuclear instance replaces $\mathrm{Ortho}(H_B)$ with the leading left/right singular vectors $u_1 v_1^\top$ of $H_B$, giving a rank-1 update per factor. The ambient update $\dot{X} = \dot{B}A + B\dot{A}$ has rank $\le 2$ and $\|\dot{X}\|_F^2 \le 4\tau^2$ (each rank-1 block contributes $\tau^2$, independent of factor norms). This is the intrinsic analog of NuMuon [50]: NuMuon constrains the update with both a spectral-norm cap $\rho$ and a nuclear-norm budget $\tau$, yielding a rank-$k$ update ($k = \lfloor \tau/\rho \rfloor$) from the top-$k$ singular vectors of the momentum. In the rank-1 special case ($k = 1$), i.e., $\tau = \rho$, NuMuon produces $u_1 v_1^\top$, while our nuclear instance applies the rank-1 truncation to the *scaled* gradient $H_B = \nabla_B f\,(AA^\top)^{-1/2}$, inheriting the same $\mathrm{GL}(r)$-invariance and condition-number-free bound as the spectral instance.

### H.2 SPD manifold $S_{++}^n$: spectral LMO vs. natural gradient

On the SPD manifold with the affine-invariant metric, the standard Riemannian optimization method is the *natural gradient* [2], which corresponds to the Frobenius instance ($\varphi = \|\cdot\|_F$) of our framework: $\xi^* \propto \mathrm{grad}\, f/\|\mathrm{grad}\, f\|_X$ where $\mathrm{grad}\, f(X) = X(\nabla_X f)X$. The intrinsic spectral norm $\|X^{-1/2}\xi X^{-1/2}\|_2$ is the Finsler norm of the Thompson metric on $S_{++}^n$ [66, 52], so the spectral instance is the steepest-descent direction of this Finsler geometry. The natural gradient preserves the relative magnitudes of the scaled gradient's eigenvalues, directing more step budget toward larger-eigenvalue directions.

The spectral instance ($\varphi = \|\cdot\|_2$) of our framework produces a qualitatively different update: the matrix-sign $\xi^* = \tau\, X^{1/2}\,\mathrm{sign}(H) X^{1/2}$, where $H = X^{1/2}(\nabla_X f)X^{1/2}$. This equalizes all eigenvalues to $\pm 1$, redistributing the step uniformly across eigendirections, the SPD analog of how Euclidean Muon replaces the gradient with its polar factor.

The difference is analogous to Muon vs. SGD in Euclidean space:

- The natural gradient preserves the magnitude profile of $H$'s eigenvalues, so large eigenvalues of $H$ receive proportionally larger step components.
- The matrix-sign update equalizes all eigenvalues to $\pm 1$, redistributing the step uniformly across eigendirections. This is more aggressive (higher squared Riemannian radius $C_\varphi = n$ vs. 1) but drives all eigenvalues of the scaled gradient to zero simultaneously.

The naive alternative, applying $\mathrm{Ortho}$ or $\mathrm{sign}$ to the Euclidean gradient $\nabla_X f$ directly (without the scaling $X^{1/2}(\cdot)X^{1/2}$), would not be $\mathrm{GL}(n)$-equivariant. For example, under $X \mapsto NXN^\top$ the Euclidean gradient becomes $N^{-\top}(\nabla_X f)N^{-1}$, and $\mathrm{sign}(N^{-\top}(\nabla_X f)N^{-1}) \neq N\ \mathrm{sign}(\nabla_X f)\, N^\top$ in general, whereas $\mathrm{sign}(H) = \mathrm{sign}(X^{1/2}(\nabla_X f)X^{1/2})$ transforms covariantly because the scaling absorbs the change of basis.

**Nuclear norm ($\varphi = \|\cdot\|_*$).** The nuclear instance selects the leading eigenpair $(q_1, \lambda_1)$ of $H = X^{1/2}(\nabla_X f)X^{1/2}$ and produces a rank-1 symmetric update $\xi^* = \tau\ \mathrm{sign}(\lambda_1)\, X^{1/2} q_1 q_1^\top X^{1/2}$. This concentrates the entire norm bound on the eigendirection of $H$ with the largest eigenvalue magnitude. Compared to the spectral instance ($C_\varphi = n$, all eigendirections active), the nuclear instance has $C_\varphi = 1$ and converges in the spectral norm of the gradient ($\varphi^\circ = \|\cdot\|_2$), which is weaker but cheaper.

### H.3 Stiefel manifold $\mathrm{St}(m, r)$: product norm vs. joint norm

On the Stiefel manifold with $G_X = I$, the natural formulation is the *joint-norm* LMO: constrain $\varphi(\xi) \leq \tau$ over the full tangent space $T_X\mathrm{St} \subset \mathbb{R}^{m\times r}$. As discussed in Section P.3, this tangent space is neither SV-invariant nor spectral-norm contractive, so Proposition 3.4 does not apply directly. For $\varphi = \|\cdot\|_2$, this problem has been studied in the context of manifold Muon. Bernstein [7] solves it via dual ascent on a symmetric Lagrange multiplier, and Cesista [16], Su [65] propose other iterative procedures. A concurrent work [64] shows that the joint-norm problem admits an exact closed-form solution, which we discuss at the end of this section.

Our framework instead uses the tangent space decomposition $\xi = XA + X_\perp B$ with the *product norm* $\max\{\varphi(A), \varphi(B)\} \leq \tau$. This relaxation decouples the LMO into closed-form subproblems on the two blocks, and the skew block $A$, which rotates within the current column space, and the normal block $B$, which moves orthogonally to it, each receive their own budget. For $\varphi = \|\cdot\|_2$, since $\max\{\|A\|_2, \|B\|_2\} \leq \|[A; B]\|_2 \leq \sqrt{2}\max\{\|A\|_2, \|B\|_2\}$, the product constraint permits tangent vectors with $\|\xi\|_2$ up to $\sqrt{2}\tau$. For $\varphi = \|\cdot\|_F$, let $S = \mathrm{skew}(X^\top \nabla_X f)$ and $N = X_\perp^\top \nabla_X f$ be the gradient blocks paired with $A$ and $B$. The product-norm solution gives $\|A^*\|_F = \|B^*\|_F = \tau$, whereas the joint-norm solution, which is the normalized Riemannian gradient, keeps the ratio $\|A^*\|_F/\|B^*\|_F = \|S\|_F/\|N\|_F$. The two solutions are parallel only when $\|S\|_F = \|N\|_F$, and even then the product-norm solution is $\sqrt{2}$ times the joint-norm solution.

The relaxation also enlarges the squared Riemannian radius $C_\varphi$ (Definition 5.1), which enters the rate constant of Theorem 5.3. Since the product-norm ball contains the joint-norm ball, $C_\varphi$ under the product norm is at least that under the joint norm. For $\varphi = \|\cdot\|_2$, the product norm gives $C_\varphi = 2\lfloor r/2 \rfloor + \min(m - r, r)$, whereas the joint norm gives at most $r$, since every $\xi \in \mathbb{R}^{m\times r}$ with $\|\xi\|_2 \leq 1$ satisfies $\|\xi\|_F^2 \leq r$. For example, when $r$ is even and $m \geq 2r$, the two radii are $2r$ and $r$, respectively. For $\varphi = \|\cdot\|_F$ and $\varphi = \|\cdot\|_*$, the joint norm gives $C_\varphi = 1$, whereas the product norm gives $C_\varphi = 2$ and $C_\varphi = 3/2$, respectively, when $r \geq 2$ and $m > r$.

**Nuclear norm ($\varphi = \|\cdot\|_*$).** The nuclear instance concentrates on the leading singular direction of each block. Since the singular values of $S$ come in equal pairs, the skew update is the rank-2 skew-symmetric matrix $A^* = \frac{\tau}{2}(u_1 v_1^\top + u_2 v_2^\top)$, where $u_1, u_2$ and $v_1, v_2$ are the left and right singular vectors of $S$ for its leading pair, so that $\|A^*\|_* = \tau$ and $\|A^*\|_F^2 = \tau^2/2$. The normal update is the rank-1 matrix $B^* = \tau\, p_1 q_1^\top$, where $p_1$ and $q_1$ are the leading left and right singular vectors of $N$. We note that the joint-norm construction of Solonko et al. [64] does not extend to the nuclear norm (see the end of this section), so the product norm provides the closed-form nuclear instance on $\mathrm{St}(m, r)$.

**Comparison with MCSD/SPEL [80].** MCSD (Manifold Constrained Steepest Descent) is a recent single-loop framework for norm-constrained optimization on embedded submanifolds. Its Stiefel specialization, SPEL, computes the spectral-norm LMO of the Riemannian gradient $\nabla_{\mathcal{M}} f(X_t) = P_{T_{X_t}\mathrm{St}}(\nabla f(X_t))$ in the *ambient* Euclidean space, $d_t = -\operatorname{Ortho}(\nabla_{\mathcal{M}} f(X_t))$, and then projects back to $\mathrm{St}(m,r)$ via the polar map: $X_{t+1} = \operatorname{Ortho}(X_t + \eta_t\, d_t) = \operatorname{Ortho}\big(X_t - \eta_t\, \operatorname{Ortho}(\nabla_{\mathcal{M}} f(X_t))\big)$. A defining design choice of MCSD is that the update direction $d_t$ *need not* lie in $T_X\,\mathrm{St}$; manifold feasibility is restored solely by the polar projection step, which is what makes MCSD single-loop.

Both SPEL and our Stiefel LMO (Lemma 4.2) avoid the joint-norm tangent space problem $\max_{\xi \in T_X\,\mathrm{St},\, \|\xi\|_2 \le \tau} \langle \xi,\, \nabla f \rangle$, but they use different relaxation strategies:

- *Our approach* decomposes the tangent space into skew and normal blocks with a product norm $\max(\|A\|_2, \|B\|_2) \le \tau$, solving each block's LMO in closed form. The update remains in $T_X\mathrm{St}$ by construction, and each block receives independent spectral control.
- *SPEL* drops the tangent-space membership constraint, solving the LMO over all of $\mathbb{R}^{m\times r}$, and compensates with the polar manifold projection. This yields a simpler per-iteration formula (one $\operatorname{Ortho}$ of an $m \times r$ matrix to obtain the LMO direction, plus one $\operatorname{Ortho}$ for the polar projection) but couples the skew (rotation within $\operatorname{col}(X)$) and normal (extension into $\operatorname{col}(X)^{\perp}$) contributions into a single ambient direction. The MCSD framework also extends beyond Stiefel to other embedded submanifolds (Grassmann, sphere) but does not natively handle quotient geometries such as $\mathcal{M}_r$ with the coupled metric.

**Joint-norm closed form [64].** For $\varphi = \|\cdot\|_2$, a concurrent work [64] solves the joint-norm LMO in closed form by lifting the tangent space to $m \times m$ skew-symmetric matrices. The map $Y \mapsto YX$ sends skew-symmetric matrices onto $T_X\mathrm{St}$, and a norm-preserving dilation argument [19] shows that it maps the spectral-norm ball of skew-symmetric matrices onto the spectral-norm ball of $T_X\mathrm{St}$. Since $\langle YX,\, \nabla_X f\rangle = \langle Y,\, \operatorname{skew}(\nabla_X f\, X^\top)\rangle$ for skew-symmetric $Y$, the joint-norm LMO reduces to a spectral-norm LMO over skew-symmetric matrices, which is solved as in Lemma D.1. This gives

$$\xi^* = \tau\, \operatorname{skew}\big(\operatorname{Ortho}(\operatorname{skew}(\nabla_X f\, X^\top))\big)\, X.$$

This construction relies on the norm-preserving dilation property of the spectral norm, and for other unitarily invariant norms the lifted problem differs from the intrinsic LMO (4). For example, for $\varphi = \|\cdot\|_F$ the lift yields a ball of the canonical metric, and for $\varphi = \|\cdot\|_*$ it doubles the norm of $\xi = X_\perp B$.

### H.4 Grassmann manifold $\mathrm{Gr}(m,r)$: connections to subspace tracking

The Grassmann manifold $\mathrm{Gr}(m,r)$ represents $r$-dimensional subspaces of $\mathbb{R}^m$. With $G_X = I$ and $X^\top X = I_r$, the horizontal space $\{\xi : X^\top \xi = 0\}$ is SV-invariant, so Proposition 3.4 applies directly and no product-norm relaxation is needed. The three norm instances of our framework produce geometrically distinct updates that connect to known methods in subspace estimation:

- *Spectral* ($\varphi = \|\cdot\|_2$): $\xi^* = \tau\, \operatorname{Ortho}((I - XX^\top)\nabla_X f)$, an $m \times r$ matrix with orthonormal columns in $\operatorname{colspan}(X)^{\perp}$, updating all $\min(m-r, r)$ subspace directions simultaneously.
- *Frobenius* ($\varphi = \|\cdot\|_F$): $\xi^* = \tau\,(I - XX^\top)\nabla_X f / \|(I - XX^\top)\nabla_X f\|_F$, the normalized Riemannian gradient on $\mathrm{Gr}(m,r)$ [20], whose discrete updates are closely related to Oja's subspace rule.
- *Nuclear* ($\varphi = \|\cdot\|_*$): $\xi^* = \tau\, u_1 v_1^\top$ from the SVD of $(I - XX^\top)\nabla_X f$, a rank-1 horizontal update ($C_\varphi = 1$).

The nuclear instance is closely related to GROUSE (Grassmannian Rank-One Update Subspace Estimation) [5], which performs online subspace tracking via rank-1 updates on $\mathrm{Gr}(m,r)$. Both produce a rank-1 tangent vector in the horizontal space; the difference is that GROUSE derives its direction from a *single observation's residual* (streaming setting with one sample at a time), whereas our LMO uses the *leading singular vectors of the full horizontal gradient* $(I - XX^\top)\nabla_X f$ (batch or mini-batch). Geometrically, each rank-1 update adds one new direction $u_1 \perp \operatorname{colspan}(X)$ while

downweighting the direction $Xv_1$ of the current subspace. Since $G_X = I$ on $\mathrm{Gr}(m, r)$, no scaling is needed, the Euclidean and intrinsic nuclear norms coincide, and the nuclear iMuon on $\mathrm{Gr}(m, r)$ is simply NuMuon restricted to the horizontal space.

# I Background on Riemannian Optimization

This appendix provides a more thorough discussion of the background material summarized in Section 2, aimed at readers less familiar with Riemannian optimization on matrix manifolds.

## I.1 Riemannian optimization: key concepts

Riemannian optimization [1, 13] generalizes gradient-based methods to curved spaces (manifolds) by replacing Euclidean gradients with Riemannian gradients and straight-line updates with retractions. The key ingredients are:

- A *tangent space* $T_x\mathcal{M}$: the set of valid update directions at $x \in \mathcal{M}$.
- A *Riemannian metric* $g_x$: a smoothly varying inner product on $T_x\mathcal{M}$ that measures lengths and angles. The Riemannian gradient $\operatorname{grad} f(x)$ is the unique tangent vector satisfying $g_x(\operatorname{grad} f, \zeta) = Df(x)[\zeta]$ for all $\zeta \in T_x\mathcal{M}$.
- A *retraction* $R_x : T_x\mathcal{M} \to \mathcal{M}$: a map that moves from $x$ along a tangent direction back onto $\mathcal{M}$, generalizing the Euclidean update $x + \eta\,\xi$.

Standard convergence analyses [13] assume *retraction-$L$-smoothness*: $f(R_x(\xi)) \le f(x) + Df(x)[\xi] + \frac{L}{2}\|\xi\|_x^2$. Under this condition, Riemannian gradient descent achieves an $O(\sqrt{L/T})$ rate to an approximate first-order stationary point, matching the Euclidean rate.

## I.2 Quotient manifolds and scale invariance

Many structured parameterizations involve redundancies. For example, a rank-$r$ matrix $X = BA$ can be factored as $(B, A)$ or $(BN^{-1}, NA)$ for any invertible $N \in \mathrm{GL}(r)$. These are different points in parameter space representing the same $X$. The mathematically clean way to handle this is the *quotient manifold*: define an equivalence relation $(B, A) \sim (BN^{-1}, NA)$ and work on the quotient space $\mathcal{M}_r = (\mathbb{R}_*^{m\times r} \times \mathbb{R}_*^{r\times n})/\mathrm{GL}(r)$, whose points are equivalence classes $[(B, A)]$. The tangent space of the total space $\overline{\mathcal{M}}$ at a representative $x$ splits into a *vertical space* $\mathcal{V}_x$ (directions along the symmetry orbit, which do not change the equivalence class) and a complementary *horizontal space* $\mathcal{H}_x$ (directions that represent genuine motion on the quotient); only horizontal tangent vectors carry information about the quotient geometry.

For optimization on a quotient manifold to be well-defined, the Riemannian metric must be *invariant* under the equivalence group: $g_{(BN^{-1},NA)}(\cdot,\cdot) = g_{(B,A)}(\cdot,\cdot)$ (after appropriate pushforward of tangent vectors). If the metric is not invariant, the Riemannian gradient depends on which representative $(B, A)$ of the equivalence class is used, a mathematical inconsistency that manifests as training instability in practice [32].

The *coupled metric* of Mishra et al. [48], Mishra and Sepulchre [46] achieves invariance by weighting each factor by the *other* factor's Gram matrix: $g(\dot{B}, \dot{A}) = \operatorname{tr}(\dot{B}^\top \dot{B}\, AA^\top) + \operatorname{tr}(B^\top B\, \dot{A}\dot{A}^\top)$. Under $(B, A) \mapsto (BN^{-1}, NA)$, the cross-Gram weighting transforms as $AA^\top \mapsto NAA^\top N^\top$, which exactly compensates the change $\dot{B} \mapsto \dot{B}N^{-1}$, preserving lengths. The resulting Riemannian gradient has the form $\nabla_B f\,(AA^\top)^{-1}$ (for the $B$-factor), which can be recognized as a preconditioned gradient that automatically equalizes the contribution of large and small singular directions of $A$.

# J Extended Related Work

## J.1 Riemannian Frank–Wolfe methods

Weber and Sra [74] study Riemannian Frank–Wolfe methods for optimization over geodesically convex constraint sets on manifolds. Their LMO operates over a *feasible set on the manifold* to keep

iterates within a constraint (e.g., a geodesic ball or a convex subset of the manifold). Our LMO is structurally different: it constrains a norm in the *tangent space* to select the step direction, then retracts onto the manifold. This is a trust-region-like norm bound on the update magnitude rather than feasibility of the iterates. The two approaches are complementary; ours targets unconstrained manifold optimization with spectral control of the step.

### J.2 Euclidean spectral optimizers and LMO

The idea of using spectral-norm (Schatten-$\infty$) steepest descent for matrix-valued parameters in deep learning predates the recent Muon line of work. Carlson et al. [15] observe that the objective function in feedforward neural networks admits a tighter majorization bound under the Schatten-$\infty$ norm than under the Frobenius norm, and derive the corresponding steepest-descent operator, which is precisely the orthogonal polar factor $\mathrm{Ortho}(\cdot)$ used by Muon. They further combine this non-Euclidean gradient with element-wise adaptive learning rates (RMSprop, ADAgrad), yielding preconditioned spectral descent algorithms (RMSspectral, ADAspectral) that demonstrate speedups on RBMs, feedforward nets, and CNNs. Our framework can be seen as replacing their element-wise diagonal preconditioning with a principled Riemannian metric $G_x$ that is determined by the manifold geometry: where Carlson et al. scale each matrix entry independently, the intrinsic norm $\varphi(G_x^{1/2}\xi)$ scales the entire tangent vector via the metric operator, simultaneously achieving preconditioning, symmetry invariance on quotient manifolds, and closed-form solutions.

More recently, Bernstein and Newhouse [8] formalize the connection between the polar factor and steepest descent under the spectral norm, and Pethick et al. [56] (Scion) generalize the LMO viewpoint to a broad family of induced *operator* norms in Euclidean space, including the spectral norm, sign / $\ell_\infty$, ColNorm, and RMS-to-RMS norms, which goes beyond the unitarily invariant family used in our framework. Our intrinsic LMO (4) recovers all of these unitarily invariant instances as the special case $G_x = I$ on $\mathbb{R}^{p\times q}$, and extends them to non-trivial metrics and manifold-structured parameters.

### J.3 Muon extensions for low-rank/LoRA fine-tuning

A first wave of works ports Muon-like spectral control to LoRA-style and natively low-rank training. Riemannion [12] works on the embedded fixed-rank manifold $\mathcal{M}_r \subset \mathbb{R}^{m\times n}$ with the ambient Frobenius metric, projects the Euclidean gradient to $T_X\mathcal{M}_r$, and approximates a tangent space spectral LMO by replacing all $\leq 2r$ singular values of the tangent space momentum with 1 via a tangent space '$\mathrm{Ortho}_r$' operator and re-projecting; the spectral-norm constraint is satisfied only approximately (singular values of the resulting tangent space step lie in $(0.9, 1.1)$ in their experiments). Kang et al. [34] prove a continuous-time convergence dichotomy for SpecGF on $\|BA - Y\|_F^2$ (see Section G). Janson et al. [32] (Spectron) renormalizes the orthogonalized step of each factor by $\rho = \eta/(\|A\|_2 + \|B\|_2 + 1)$, with $\|A\|_2, \|B\|_2$ estimated by a single power iteration per step. Park et al. [54] (Stiefel-LoRA) restrict LoRA's $B$ factor to the Stiefel manifold with Riemannian SGD/AdamW. NuMuon [50] adds a nuclear-norm budget to Muon's spectral LMO, yielding rank-$\lfloor\tau/\rho\rfloor$ updates (nuclear budget $\tau$, spectral cap $\rho$) suitable for compressible LLMs. Wang et al. [73] factorize the EMA momentum buffer to compress optimizer-state memory in both Adam-style and Muon-style optimizers. Schotthöfer et al. [61] provide a geometric framework for momentum-based low-rank training. None of these works simultaneously (i) preserves quotient symmetries on $\mathcal{M}_r$, (ii) gives an exact closed-form tangent space spectral LMO, and (iii) attains a condition-number-free rate, the three properties of our framework specialized to $\mathcal{M}_r$.

### J.4 Convergence theory of Muon

Discrete-time convergence guarantees for Muon and its variants have appeared in a sequence of works: Li and Hong [43] provide a first convergence note; Sato et al. [60] establish convergence and a critical batch-size analysis under both Nesterov and weight-decay variants; Shen et al. [62], Kim and Oh [38], Chen et al. [17] treat Muon convergence under various smoothness assumptions, and Ma et al. [44] analyze preconditioning benefits of spectral orthogonalization. Riabinin et al. [58] (Gluon) bridge theory and practice in LMO-based optimizers. Amsel et al. [3] study optimal polynomial approximations of the matrix sign function used inside Muon. Our Theorems 5.3 and 5.7 cover both deterministic and stochastic settings on Riemannian matrix manifolds, with the squared

Riemannian radius $C_\varphi$ as the unifying complexity parameter, and recover the Euclidean Muon rate of Bernstein and Newhouse [8], Shen et al. [62] as the $G_x = I$ special case (Section G).

### J.5 Riemannian and manifold-aware optimizers for LLMs

A complementary line of work injects manifold-derived primitives into LLM-scale pretraining without strictly constraining iterates to a manifold. Mano [21] projects momentum onto the tangent space of the Oblique manifold and applies alternating column-wise and row-wise normalization across odd/even iterations, which imposes only a soft manifold constraint on the update direction (no retraction is applied to the iterate), so the parameter trajectory itself remains Euclidean. SSO [77] constrains weights to the spectral sphere $\mathcal{S}_R = \{X : \|X\|_2 = R\}$ and runs steepest descent under the spectral norm via bisection on a single Lagrange multiplier; we discuss this case (where SV-invariance fails) in Section P.4. Yang et al. [79, 78] provide the spectral / $\mu$P framework that motivates much of this line. These works are largely orthogonal to our setting, which targets problems where the parameters genuinely live on a Riemannian manifold (LoRA-style fixed-rank factorizations, SPD covariance descriptors, subspaces). Within that setting, Tong et al. [67], Zhang and Pilanci [81], Mishra et al. [48], Mishra and Sepulchre [46] are the closest direct precursors of our Frobenius-norm instance on $\mathcal{M}_r$.

## K Experimental Details for LoRA Fine-Tuning

**Compared methods.** Each LoRA adapter parameterizes the trainable update as $X = BA$, where $B \in \mathbb{R}^{m\times r}$ and $A \in \mathbb{R}^{r\times n}$, while the pretrained weight is kept fixed. This representation exposes the fixed-rank structure used by the geometry-aware methods. We compare Euclidean, Riemannian-preconditioned, and spectral-update methods under this common LoRA parameterization.

SGD and AdamW [81] update the LoRA factors directly as ordinary Euclidean parameters. Scaled GD and Scaled AdamW [47, 67, 81] use the $\mathrm{GL}(r)$-invariant preconditioning induced by the quotient geometry of $\mathcal{M}_r$: Scaled GD corresponds to the Frobenius-norm instance of the intrinsic LMO, which recovers the Riemannian gradient direction, while Scaled AdamW uses the same geometry-aware preconditioning inside an AdamW-style update. Factor-wise Muon [33, 8, 34, 32] applies the Euclidean spectral LMO factor-wise as in (3). Spectron [32] applies the same factor-wise spectral LMO and rescales the step by $\rho = \eta/(\|A\|_2 + \|B\|_2 + 1)$. In contrast, iMuon is the spectral-norm instance of our intrinsic LMO on $\mathcal{M}_r$ (Corollary 4.1). We also include Riemannion [12], reproduced from the authors' released code, as an embedded fixed-rank Muon baseline with approximate tangent-space orthogonalization.

**Momentum implementation.** For the Adam family methods, we use standard Adam first- and second-moment updates. For iMuon and factor-wise Muon, we use the *Nesterov* SGD style momentum update. For Spectron, we follow the momentum recurrence of Janson et al. [32] with $\beta = 0.95$. For a LoRA pair $(B_t, A_t)$ with $X_t = B_t A_t$ at iteration $t$, let

$$G_{B,t} = \nabla_X f(X_t) A_t^\top, \qquad G_{A,t} = B_t^\top \nabla_X f(X_t)$$

denote the Euclidean gradients of the LoRA factors. For factor-wise Muon, the momentum buffers are

$$M_{B,t} = \beta M_{B,t-1} + G_{B,t}, \qquad M_{A,t} = \beta M_{A,t-1} + G_{A,t},$$

with $\beta = 0.95$ and the quantities passed to its spectral step are

$$\widetilde{G}_{B,t} = G_{B,t} + \beta M_{B,t}, \qquad \widetilde{G}_{A,t} = G_{A,t} + \beta M_{A,t}.$$

Factor-wise Muon applies the spectral LMO separately to these two factor directions. iMuon instead uses projection-transported momentum. To describe its construction, consider full-rank factors and define

$$P_t = B_t^\top B_t, \qquad Q_t = A_t A_t^\top.$$

We transport the previous momentum pair $Z = M_{t-1}$ by projecting it into the horizontal space at the current point $(B_t, A_t)$:

$$\mathcal{T}_t(Z) = \big(Z_B - B_t\Omega,\ Z_A + \Omega A_t\big),$$

Table 6: Per-batch runtime profile on E2E NLG for one training epoch with batch size 8. Timings are reported in milliseconds as mean ± standard deviation over training batches.

| Stage | Factor-wise Muon | iMuon (QR) | iMuon (CholeskyQR) | Riemannion (Adam) |
|---|---|---|---|---|
| Forward pass | $135.97 \pm 1.24$ | $133.45 \pm 1.30$ | $135.48 \pm 1.42$ | $127.43 \pm 0.32$ |
| Backward pass | $158.00 \pm 0.34$ | $157.41 \pm 0.31$ | $157.45 \pm 0.28$ | $155.72 \pm 0.68$ |
| Optimizer step | $3.22 \pm 0.09$ | $13.08 \pm 0.14$ | $5.49 \pm 0.13$ | $54.16 \pm 0.30$ |
| Total per batch | $297.20 \pm 1.21$ | $303.95 \pm 1.26$ | $298.41 \pm 1.38$ | $337.31 \pm 1.00$ |

where

$$\Omega = \frac{1}{2}\left(P_t^{-1} B_t^\top Z_B - Z_A A_t^\top Q_t^{-1}\right).$$

With the metric-preconditioned gradient pair

$$R_t = \left(G_{B,t} Q_t^{-1},\ P_t^{-1} G_{A,t}\right),$$

we update the momentum buffer and form the Nesterov direction as

$$M_t = \beta \mathcal{T}_t(M_{t-1}) + (1-\beta) R_t, \qquad D_t = \beta M_t + (1-\beta) R_t,$$

with $M_0 = (0, 0)$ and $\beta = 0.95$. Here $P_t$ and $Q_t$ are invertible since the LoRA factors have full rank $r$. The map $\mathcal{T}_t$ is the metric-orthogonal projection onto the horizontal space at $(B_t, A_t)$ and $R_t$ is the Riemannian gradient $G^{-1}\nabla f$, so $D_t$ transforms as a tangent vector under $(B, A) \mapsto (BN^{-1}, NA)$ and the resulting update is representative-invariant. The intrinsic LMO of Corollary 4.1 is then applied to $D_t$ in place of the gradient pair.

**Compute Resource.** All the experiments are run on Nvidia GPU clusters using L40, A40, A100, RTX 6000 GPUs depending on availability of the resources. The E2E NLG experiments use single-GPU training. For GLUE, all tasks except MNLI and QQP use single-GPU training, while MNLI and QQP use two-GPU DDP. No hardware specific tuning is performed and the same training code path and optimizer settings are used across GPU types.

We additionally compare the per-batch runtime of factor-wise Muon, iMuon, and Riemannion (Adam) over one E2E NLG epoch at batch size 8. We profile each batch into three stages: forward pass, backward pass, and optimizer step. The optimizer-step overhead motivates a CholeskyQR variant of iMuon, which replaces the thin QR by a Cholesky-based factorization of $AA^\top$. Table 6 reports per-stage means and standard deviations for factor-wise Muon, iMuon (PyTorch thin-QR), iMuon (CholeskyQR), and Riemannion (Adam).

### K.1 E2E NLG

**Setup** We fine-tune GPT-2 Medium (355M parameters) on the E2E NLG Challenge [53], a table-to-text generation task with approximately 42K training and 4.7K test examples in the restaurant domain. LoRA adapters are applied to the query and value projections within the merged `c_attn` linear layer, following LoRA [29] and Riemannian LoRA [81]. The LoRA configuration uses rank $r = 4$, scaling $\alpha = 32$, and dropout $0.1$. Training runs for 5 epochs with batch size 8, a linear learning-rate schedule with 500 warmup steps, weight decay $0.01$, label smoothing $0.1$, and no gradient clipping. Generation uses beam search with beam size 10, length penalty $0.8$, and no-repeat $4$-gram decoding. All single-seed comparisons use seed 110, which is also the seed used by Zhang and Pilanci [81]. We compute BLEU, NIST, METEOR, ROUGE-L, and CIDEr using the standard `e2e-metrics` package.

Learning rates are tuned separately for each method. Following the E2E reporting protocol, we report the best test-set BLEU run from the sweep, with the remaining metrics computed from the same run. The full sweeps are reported in Tables 8 and 9. The SGD, Scaled GD, AdamW, and Scaled AdamW entries in Table 1 are taken from Zhang and Pilanci [81].

**Results.** Table 1 reports, for each method, the run with the best test-set BLEU, with the remaining metrics computed from the same run. Tables 8 and 9 report the full learning-rate sweeps from which these runs are selected. Table 7 reports the seed sensitivity of iMuon with and without momentum. For each configuration, the learning rate selected in Table 1 is fixed and the run is repeated with

Table 7: Seed sensitivity of iMuon on E2E NLG with LoRA rank $r = 4$. Entries report mean $\pm$ sample standard deviation across seeds $0, 42, 110$.

| Configuration | BLEU | NIST | METEOR | ROUGE-L | CIDEr |
|---|---|---|---|---|---|
| Without momentum | $70.06 \pm 0.22$ | $8.83 \pm 0.02$ | $46.61 \pm 0.05$ | $71.54 \pm 0.46$ | $2.524 \pm 0.007$ |
| With momentum | $69.95 \pm 0.49$ | $8.79 \pm 0.04$ | $46.80 \pm 0.10$ | $71.62 \pm 0.31$ | $2.517 \pm 0.012$ |

Table 8: E2E NLG learning-rate sweep without momentum. For Riemannion (SGD), low-LR points $\{5\times10^{-5}, 10^{-4}\}$ are omitted because pilot runs showed the useful range was at substantially larger learning rates.

| Method | LR | BLEU | NIST | METEOR | ROUGE-L | CIDEr |
|---|---|---|---|---|---|---|
| Factor-wise Muon | $5\times10^{-5}$ | 64.22 | 8.33 | 42.46 | 67.49 | 2.21 |
| | $10^{-4}$ | 65.31 | 8.42 | 43.40 | 67.85 | 2.30 |
| | $5\times10^{-4}$ | 69.09 | 8.73 | 46.20 | 71.17 | 2.48 |
| | $10^{-3}$ | 70.02 | 8.81 | 46.77 | 71.68 | 2.53 |
| Riemannion (SGD) | $5\times10^{-4}$ | 65.05 | 8.40 | 43.30 | 68.05 | 2.29 |
| | $10^{-3}$ | 66.08 | 8.48 | 43.88 | 68.07 | 2.36 |
| | $5\times10^{-3}$ | 68.60 | 8.70 | 45.84 | 70.49 | 2.45 |
| | $10^{-2}$ | 69.40 | 8.75 | 46.28 | 71.18 | 2.47 |
| | $5\times10^{-2}$ | 70.02 | 8.78 | 46.79 | 71.99 | 2.52 |
| Spectron | $5 \times 10^{-5}$ | 53.00 | 3.67 | 35.04 | 64.71 | 1.34 |
| | $10^{-4}$ | 57.63 | 5.95 | 37.58 | 65.52 | 1.61 |
| | $5 \times 10^{-4}$ | 64.91 | 8.31 | 41.86 | 67.39 | 2.13 |
| | $10^{-3}$ | 65.49 | 8.40 | 43.44 | 68.33 | 2.30 |
| | $5 \times 10^{-3}$ | 68.17 | 8.62 | 45.78 | 70.13 | 2.43 |
| | $10^{-2}$ | 68.45 | 8.66 | 46.03 | 70.79 | 2.43 |
| | $5 \times 10^{-2}$ | 70.25 | 8.81 | 46.85 | 71.81 | 2.51 |
| iMuon | $5 \times 10^{-5}$ | 66.00 | 8.40 | 44.67 | 69.09 | 2.35 |
| | $10^{-4}$ | 66.00 | 8.41 | 45.22 | 69.32 | 2.40 |
| | $5 \times 10^{-4}$ | 69.85 | 8.80 | 46.49 | 71.41 | 2.52 |
| | $10^{-3}$ | 70.25 | 8.86 | 46.55 | 71.42 | 2.53 |
| | $5 \times 10^{-3}$ | 69.64 | 8.78 | 46.43 | 71.38 | 2.50 |
| | $10^{-2}$ | 69.98 | 8.79 | 46.70 | 71.48 | 2.51 |
| | $5 \times 10^{-2}$ | 70.30 | 8.85 | 46.62 | 72.03 | 2.53 |

seeds 0, 42, and 110. Entries are the mean and sample standard deviation across the three seeds. Table 10 reports results without momentum at LoRA rank 32 for factor-wise Muon, Riemannion, Spectron, and iMuon, with the learning rate selected by a sweep at this rank.

### K.2 GLUE

**Setup** We evaluate LoRA fine-tuning on the GLUE benchmark, following the Mistral-7B setup of Zhang and Pilanci [81]. We fine-tune a 4-bit quantized Mistral-7B-v0.1 model whose frozen base weights are stored in NF4 and use bfloat16 compute. LoRA adapters use rank $r = 16$, scaling $\alpha = 16$, dropout $0.05$, and no bias, and are applied to five target modules: `q_proj`, `k_proj`, `v_proj`, `o_proj`, and `gate_proj`. Riemannion uses its doubled-rank PEFT parameterization with rank $r = 32$ and $\alpha = 32$, corresponding to effective rank 16. The doubling is used for its tangent-space implementation rather than as a rank-32 forward adapter.

Training uses per-device batch size $8$ for up to $5$ epochs, with two-GPU DDP for MNLI and QQP and single-GPU training for the remaining seven tasks. We use a linear learning-rate schedule with no warmup, weight decay $0.01$, and maximum gradient norm 1. AdamW-family methods use $\beta_1 = 0.9$, $\beta_2 = 0.999$, and $\epsilon = 10^{-6}$. Early stopping is enabled with patience 10, evaluation is performed every $500$ steps, and the best checkpoint is selected by the validation metric. Single-seed results use seed $42$. The seed-sensitivity experiments use seeds $0$, $42$, and $110$.

Table 9: E2E NLG learning-rate sweep with momentum. Results of the no-momentum LR sweep were used to narrow the LR sweep of the momentum-based runs.

| Method | LR | BLEU | NIST | METEOR | ROUGE-L | CIDEr |
|---|---|---|---|---|---|---|
| Factor-wise Muon | $5\times10^{-5}$ | 68.53 | 8.67 | 45.78 | 70.80 | 2.43 |
| | $10^{-4}$ | 69.57 | 8.77 | 46.40 | 71.53 | 2.50 |
| | $5\times10^{-4}$ | 69.97 | 8.81 | 46.70 | 71.44 | 2.53 |
| | $7\times10^{-4}$ | 70.26 | 8.83 | 46.76 | 71.75 | 2.52 |
| | $10^{-3}$ | 70.22 | 8.82 | 46.61 | 71.93 | 2.53 |
| Riemannion (Adam) | $5\times10^{-5}$ | 61.80 | 7.57 | 39.83 | 65.24 | 1.95 |
| | $10^{-4}$ | 64.59 | 8.16 | 41.17 | 67.11 | 2.10 |
| | $5\times10^{-4}$ | 65.23 | 8.38 | 43.87 | 67.82 | 2.31 |
| | $10^{-3}$ | 66.46 | 8.54 | 44.50 | 67.90 | 2.34 |
| | $5\times10^{-3}$ | 69.27 | 8.79 | 46.03 | 70.12 | 2.47 |
| | $10^{-2}$ | 69.58 | 8.79 | 46.57 | 70.85 | 2.52 |
| | $5\times10^{-2}$ | 70.21 | 8.83 | 46.55 | 71.87 | 2.53 |
| Spectron | $5\times10^{-3}$ | 69.28 | 8.74 | 46.31 | 71.10 | 2.51 |
| | $10^{-2}$ | 69.63 | 8.76 | 46.67 | 71.40 | 2.51 |
| | $5\times10^{-2}$ | 69.84 | 8.77 | 46.98 | 71.84 | 2.50 |
| iMuon | $5\times10^{-4}$ | 69.66 | 8.78 | 46.41 | 71.41 | 2.51 |
| | $10^{-3}$ | 70.16 | 8.80 | 46.57 | 71.87 | 2.53 |
| | $5\times10^{-3}$ | 70.19 | 8.83 | 46.83 | 71.84 | 2.54 |
| | $10^{-2}$ | 70.41 | 8.83 | 46.84 | 71.96 | 2.52 |

Table 10: Higher-rank E2E NLG results without momentum at LoRA rank 32.

| Method | BLEU | NIST | METEOR | ROUGE-L | CIDEr |
|---|---|---|---|---|---|
| | | *Rank* 32 | | | |
| Factor-wise Muon | 70.05 | 8.84 | 46.53 | 71.76 | 2.52 |
| Riemannion (SGD) | 70.27 | 8.83 | 46.60 | 71.91 | 2.52 |
| Spectron | 69.95 | 8.81 | 46.70 | 71.83 | 2.52 |
| iMuon | 70.05 | 8.83 | 46.41 | 71.56 | 2.51 |

We evaluate on the nine GLUE tasks: MNLI, SST-2, MRPC, CoLA, QNLI, QQP, RTE, STS-B, and WNLI. Following the GLUE convention, the metric is accuracy for classification tasks, Matthews correlation for CoLA, and Pearson correlation for STS-B, and this metric is used both for checkpoint selection and for reporting. The reported average is the unweighted average across the nine tasks. The SGD, Scaled GD, AdamW, and Scaled AdamW entries in Table 2 are taken from Zhang and Pilanci [81].

**Results.** Table 2 reports the GLUE results for 4-bit Mistral-7B fine-tuning with LoRA rank $r = 16$. The learning rate is selected separately for each method and task by validation performance, since this setting is sensitive to the learning rate across both tasks and optimizers. This follows the task-specific selection of Zhang and Pilanci [81] for the same model. Table 11 lists the candidate set searched for each method and the learning rate selected for each task.

Table 12 reports the seed sensitivity of iMuon with and without momentum. For each task, the learning rate selected with seed 42 is fixed and the run is repeated with seeds 0 and 110. Entries are the mean and sample standard deviation across the three seeds.

## L Learning on Fixed-Rank Manifold

**Common LMO step-size convention.** All LMO-based methods use constraint radius $\tau = 1$. The reported learning rate scales the resulting unit-radius direction. When Spectron is included, we use its adaptive extrinsic normalization separately.

Table 11: Learning-rate candidates and selected learning rates for the GLUE entries evaluated in this work. Candidate sets summarize method-specific searches and can differ across tasks; the task columns report the learning rate selected by validation performance. The SGD, Scaled GD, AdamW, and Scaled AdamW rows in Table 2 are taken from Zhang and Pilanci [81] and are therefore not included here.

| Method | LR candidates | MNLI | SST-2 | MRPC | CoLA | QNLI | QQP | RTE | STS-B | WNLI |
|---|---|---|---|---|---|---|---|---|---|---|
| *Without momentum* | | | | | | | | | | |
| Factor-wise Muon | $\{5\times10^{-5}, 10^{-4}\}$ | $5\times10^{-5}$ | $5\times10^{-5}$ | $10^{-4}$ | $10^{-4}$ | $5\times10^{-5}$ | $5\times10^{-5}$ | $10^{-4}$ | $10^{-4}$ | $10^{-4}$ |
| Riemannion (SGD) | $\{5\times10^{-5}, 10^{-4}, 5\times10^{-4}, 10^{-3}, 2\times10^{-3}, 5\times10^{-3}, 10^{-2}, 5\times10^{-2}\}$ | $5\times10^{-3}$ | $5\times10^{-3}$ | $5\times10^{-3}$ | $10^{-2}$ | $5\times10^{-3}$ | $10^{-2}$ | $5\times10^{-3}$ | $2\times10^{-3}$ | $10^{-2}$ |
| Spectron | $\{10^{-4}, 10^{-3}, 10^{-2}, 5\times10^{-2}\}$ | $10^{-3}$ | $10^{-3}$ | $10^{-2}$ | $10^{-3}$ | $10^{-3}$ | $10^{-4}$ | $10^{-2}$ | $10^{-2}$ | $10^{-2}$ |
| iMuon | $\{1, 2, 3, 5, 10\}\times10^{-5}$ | $2\times10^{-5}$ | $3\times10^{-5}$ | $10^{-4}$ | $5\times10^{-5}$ | $2\times10^{-5}$ | $10^{-5}$ | $10^{-4}$ | $10^{-4}$ | $10^{-4}$ |
| *With momentum* | | | | | | | | | | |
| Factor-wise Muon | $\{5\times10^{-5}, 10^{-4}\}$ | $5\times10^{-5}$ | $5\times10^{-5}$ | $10^{-4}$ | $10^{-4}$ | $5\times10^{-5}$ | $5\times10^{-5}$ | $10^{-4}$ | $10^{-4}$ | $10^{-4}$ |
| Riemannion (Adam) | $\{10^{-5}, 2\times10^{-5}, 3\times10^{-5}, 4\times10^{-5}, 5\times10^{-5}, 10^{-4}, 5\times10^{-4}\}$ | $10^{-4}$ | $4\times10^{-5}$ | $10^{-4}$ | $10^{-4}$ | $5\times10^{-5}$ | $10^{-4}$ | $10^{-4}$ | $5\times10^{-5}$ | $5\times10^{-4}$ |
| Spectron | $\{10^{-4}, 10^{-3}, 10^{-2}, 5\times10^{-2}\}$ | $10^{-4}$ | $10^{-3}$ | $10^{-3}$ | $10^{-3}$ | $10^{-4}$ | $10^{-3}$ | $10^{-3}$ | $10^{-3}$ | $10^{-3}$ |
| iMuon | $\{5\times10^{-5}, 10^{-4}\}$ | $5\times10^{-5}$ | $10^{-4}$ | $10^{-4}$ | $5\times10^{-5}$ | $5\times10^{-5}$ | $10^{-4}$ | $5\times10^{-5}$ | $10^{-4}$ | $10^{-4}$ |

Table 12: Seed sensitivity of iMuon on GLUE with 4-bit Mistral-7B and LoRA rank $r = 16$. Entries report mean $\pm$ sample standard deviation across seeds $0, 42, 110$. Avg. is the unweighted nine-task average; its standard deviation is computed across the three per-seed averages.

| Configuration | MNLI | SST-2 | MRPC | CoLA | QNLI | QQP | RTE | STS-B | WNLI | Avg. |
|---|---|---|---|---|---|---|---|---|---|---|
| Without momentum | $89.46 \pm 0.51$ | $96.67 \pm 0.23$ | $90.28 \pm 0.62$ | $71.20 \pm 1.29$ | $94.40 \pm 0.25$ | $88.46 \pm 0.13$ | $90.13 \pm 1.10$ | $92.40 \pm 0.35$ | $81.22 \pm 2.93$ | $88.25 \pm 0.64$ |
| With momentum | $91.65 \pm 0.14$ | $97.52 \pm 0.24$ | $88.97 \pm 0.42$ | $69.86 \pm 1.74$ | $94.79 \pm 0.12$ | $91.95 \pm 0.29$ | $87.97 \pm 0.21$ | $92.34 \pm 0.36$ | $83.10 \pm 0.00$ | $88.68 \pm 0.18$ |

### L.1 Matrix completion: synthetic experiments

**Setup.** We study synthetic fixed-rank matrix completion at large scale, following the standard sampled-entry low-rank recovery setup [14, 37, 57] and the quotient fixed-rank factorization used in Riemannian matrix completion [70, 47, 48, 46]. For each trial, we generate a rank-$r$ ground-truth matrix $X^* \in \mathbb{R}^{m\times n}$, with $m = n = 5000$ and $r = 10$. The nonzero singular values are ordered as $\sigma_1 \geq \cdots \geq \sigma_r > 0$, and we control the target conditioning by setting $\kappa = \sigma_1(X^*)/\sigma_r(X^*)$. Given an observed index set $\Omega$, we optimize

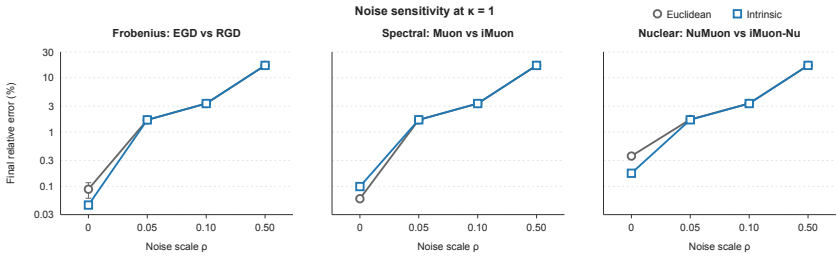
$$\min_{X=BA} f(X) = \frac{1}{2|\Omega|} \left\|\mathcal{P}_\Omega(X - Y)\right\|_F^2,$$

where $B \in \mathbb{R}^{m\times r}$, $A \in \mathbb{R}^{r\times n}$, and $Y$ contains the observed entries. In the noiseless case, $Y_{ij} = X^*_{ij}$ for $(i, j) \in \Omega$. In the noisy case, $Y_{ij} = X^*_{ij} + \varepsilon_{ij}$, with Gaussian observation noise scaled as described below. We use an oversampling ratio $s$, meaning that $|\Omega| = sr(m + n)$, and all reported runs use $s = 10$. All methods are initialized from the rank-$r$ SVD of the zero-filled observed matrix $\mathcal{P}_\Omega(Y)$. The step size follows $\eta_t = \eta_0/\sqrt{t}$, where $\eta_0$ is selected separately for each method and each $(\kappa, \rho)$ setting.

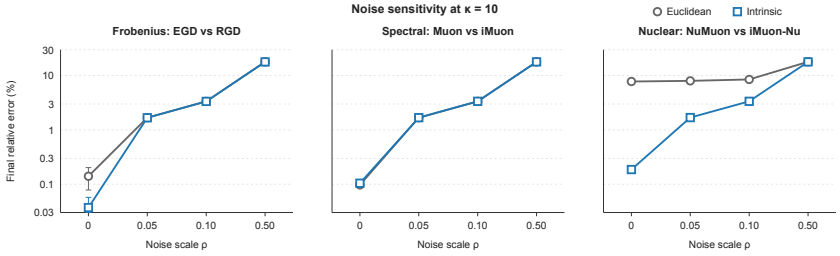
**Methods.** We compare norm-matched Euclidean and intrinsic updates:

- **Frobenius norm:** The intrinsic LMO recovers standard (normalized) Riemannian gradient descent, denoted by **RGD**, compared with (normalized) Euclidean gradient descent on the chosen representative, denoted by **EGD**.
- **Spectral norm: iMuon** denotes the intrinsic spectral norm LMO induced by the Riemannian metric, while factor-wise **Muon** denotes the corresponding Euclidean spectral norm LMO applied to the same representative as in (3).
- **Nuclear norm: iMuon-Nu** denotes the intrinsic nuclear norm LMO. **NuMuon** is its Euclidean nuclear norm counterpart, following the nuclear norm constrained Muon terminology of Mohaghegh Dolatabadi et al. [50]. In this section, we use only the nuclear norm constraint so that the comparison is norm matched to iMuon-Nu.

Each comparison is made within the same norm family, isolating the effect of using the intrinsic metric induced geometry rather than the ambient Euclidean geometry.




(a) $\kappa = 1$




(b) $\kappa = 10$

Figure 2: Noise sensitivity in large-scale synthetic fixed-rank matrix completion. Each row fixes the condition number and plots final relative recovery error as the relative observed-entry noise scale $\rho$ varies. The three panels in each row compare the Frobenius, spectral, and nuclear norm pairs. The y-axis is logarithmic, and lower is better.

**Dataset.** We vary the target condition number over $\kappa \in \{1, 10\}$. To match the matrix-completion convention of reporting noise by its variance or standard deviation, we parameterize the observation noise by the relative entrywise scale

$$\rho = \sigma_\varepsilon / \mathrm{rms}_\Omega(X^*),$$

where $\varepsilon_{ij} \sim \mathcal{N}(0, \sigma_\varepsilon^2)$ is added only on observed entries and $\mathrm{rms}_\Omega(X^*)$ is the root mean square magnitude of the sampled clean entries. We report the noiseless case $\rho = 0$, a mild noise setting $\rho = 0.05$, and two additional noise-stress settings $\rho = 0.10$ and $\rho = 0.50$. For each $(\kappa, \rho)$ setting and each seed, we generate a fresh rank-10 target matrix $X^*$, sample the observation set $\Omega$, and add observation noise when applicable. Results are averaged over seeds $\{0, 1, 2\}$.

**Results.** Figure 2 plots final recovery error against the relative noise level for each condition number. For both condition numbers, the Frobenius and nuclear intrinsic updates improve over their Euclidean counterparts for low and moderate noise. The spectral pair is close at all noise levels, with factor-wise Muon slightly ahead in the noiseless case at $\kappa = 1$. At $\rho = 0.50$, observation noise dominates and the ordering becomes mixed.

### L.2 Fixed-rank CIFAR-100 classification heads

The task is to assign each CIFAR-100 image to one of 100 semantic categories using frozen ResNet-18 features [40, 26]. A low-rank linear head maps these features to class scores while making the class weight vectors share a small set of latent visual directions, so the learned classifier is the fixed-rank parameter $X = BA$. We consider two regimes: a representative-sensitivity diagnostic and a balanced spectral comparison with Spectron [32].

For both regimes, given frozen features $h_i \in \mathbb{R}^{512}$ and labels $y_i$, we train a rank-$r$ classifier $X = BA$ and bias $b \in \mathbb{R}^{100}$ by minimizing

$$\min_{B,A,b} \frac{1}{N} \sum_{i=1}^{N} \ell_{\mathrm{CE}}\big(h_i^\top BA + b, y_i\big) + \frac{\lambda}{2}\|BA\|_F^2 + \frac{\lambda_b}{2}\|b\|_2^2.$$

We use $\lambda = \lambda_b = 10^{-4}$. The rank $r$, initialization, and learning-rate grids are specified below.

Table 13: Representative-sensitivity diagnostic on a frozen-feature CIFAR-100 rank-10 classification head with $\alpha = 10^3$. We report mean test accuracy $\pm$ standard deviation over three seeds. Boldface marks the higher-accuracy method within each Euclidean/intrinsic pair.

| Norm | Euclidean | Intrinsic | Euclidean acc. | Intrinsic acc. |
|---|---|---|---|---|
| Frobenius | EGD ($10^{-5}$) | RGD (0.1) | $0.075 \pm 0.007$ | $\mathbf{0.342 \pm 0.004}$ |
| Spectral | Muon ($10^{-5}$) | iMuon (0.03) | $0.068 \pm 0.005$ | $\mathbf{0.339 \pm 0.003}$ |
| Nuclear | NuMuon ($10^{-4}$) | iMuon-Nu (0.3) | $0.100 \pm 0.003$ | $\mathbf{0.343 \pm 0.004}$ |

### L.2.1 Representative-sensitivity diagnostic

**Setup.** The representative-sensitivity diagnostic uses a rank-10 head,

$$X = BA, \qquad B \in \mathbb{R}^{512\times 10}, \quad A \in \mathbb{R}^{10\times 100}.$$

To test representative sensitivity, we keep the initial product fixed and rescale only the factors,

$$(B_0, A_0) = (\alpha \widetilde{B}_0, \alpha^{-1}\widetilde{A}_0), \qquad \alpha = 10^3.$$

Thus $B_0A_0$ is unchanged up to numerical precision, while the two factors are imbalanced. This is a real supervised task with a deliberately chosen representative of the same initial classifier.

**Methods.** We compare the three norm-matched Euclidean/intrinsic pairs used above. The Euclidean methods use the grid

$$\{10^{-5}, 10^{-4}, 10^{-3}, 10^{-2}, 10^{-1}\}.$$

For the intrinsic methods, we use local grids,

$$\text{RGD} : \{0.1, 0.3, 1.0\}, \qquad \text{iMuon} : \{0.03, 0.1, 0.3\}, \qquad \text{iMuon-Nu} : \{0.3, 1.0, 3.0\}.$$

The learning rate is selected by validation accuracy for each method.

**Dataset.** We use the standard CIFAR-100 train/test split, reserve $10\%$ of the training set for validation, extract frozen ResNet-18 features, and apply feature $\ell_2$-normalization. We train for $50$ epochs with batch size 32 and average over seeds $\{0, 1, 2\}$.

**Results.** Table 13 reports validation-selected test accuracy. Under the same rescaled initial product, the intrinsic methods reach $33.9\%$ to $34.3\%$, while their Euclidean counterparts range from $6.8\%$ to $10.0\%$.

### L.2.2 Balanced spectral comparison with Spectron

**Setup.** The balanced spectral comparison uses a rank-$40$ head:

$$X = BA, \qquad B \in \mathbb{R}^{512\times 40}, \quad A \in \mathbb{R}^{40\times 100},$$

with $\alpha = 1$. For each seed, all methods use the same train/validation split and the same rank-$40$ class-mean-SVD initialization of $B_0A_0$. This regime compares spectral fixed-rank updates without artificial factor rescaling.

**Methods.** We compare Muon, Spectron [32], and iMuon. Muon applies factor-wise Euclidean spectral LMO updates to $B$ and $A$ [33, 8]. Spectron orthogonalizes the factor gradients and rescales the update by $1/(\|B\|_2 + \|A\|_2 + 1)$. We run Spectron without momentum to match the optimizer protocol used here. iMuon uses the coupled fixed-rank metric and the intrinsic spectral LMO from Corollary 4.1. We train for $50$ epochs with batch size $2048$, use feature $\ell_2$-normalization, and select the learning rate by validation accuracy. The learning-rate grids are

$$\begin{aligned} \text{Muon} &: \{0.03, 0.05, 0.07, 0.10, 0.15\}, \\ \text{Spectron} &: \{0.5, 0.7, 1.0, 1.5, 2.0\}, \\ \text{iMuon} &: \{0.05, 0.07, 0.10, 0.15, 0.18, 0.20, 0.25\}. \end{aligned}$$

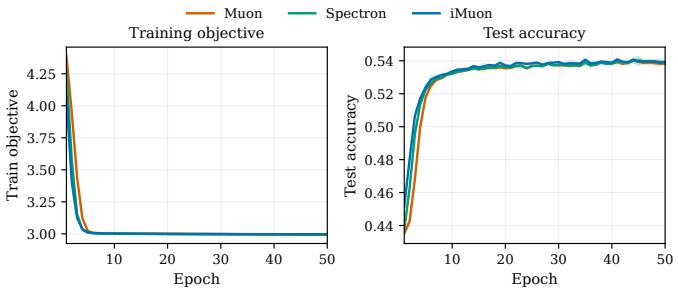



Figure 3: Training trajectories for the balanced fixed-rank CIFAR-100 rank-head comparison. Curves show means over three seeds with standard-deviation bands. The three methods reach similar training objectives, while iMuon remains competitive in test accuracy.

**Dataset.** We use the standard CIFAR-100 train/test split and reserve $10\%$ of the training split for validation. Results are averaged over seeds $\{0, 1, 2\}$.

**Results.** Table 14 reports validation-selected test accuracy. iMuon obtains the highest mean accuracy, $53.93\%$, compared with $53.82\%$ for factor-wise Muon and $53.92\%$ for Spectron. The gaps are small, so we view this as a balanced comparison with fixed-rank spectral baselines rather than a large separation. Figure 3 shows the corresponding training trajectories. The three methods reach similar training objectives, while iMuon remains competitive in test accuracy throughout training.

Table 14: Frozen-feature CIFAR-100 classification with a rank-40 fixed-rank head. We report validation-selected test accuracy over three seeds. Higher is better.

| Method | Selected LR | Test acc. (%) |
|---|---|---|
| Factor-wise Muon | 0.05 | $53.82 \pm 0.02$ |
| Spectron | 0.5 | $53.92 \pm 0.09$ |
| iMuon | 0.18 | $\mathbf{53.93 \pm 0.09}$ |

### L.3 MovieLens-1M

**Setup.** We also test MovieLens-1M [25], a standard real-data matrix completion benchmark. The data contain 6040 users, 3952 movies, and 1,000,209 observed ratings. Following standard recommender preprocessing [39], we subtract a global mean, user biases, and movie biases. We then fit the residual ratings with a rank-5 fixed-rank factorization $X = BA$. On the training entries $\Omega_{\text{train}}$, the objective is

$$f(B, A) = \frac{1}{2|\Omega_{\text{train}}|} \sum_{(u,i)\in\Omega_{\text{train}}} \big((BA)_{ui} - R_{ui}\big)^2 + \frac{\lambda}{2} \left(\|B\|_F^2 + \|A\|_F^2\right),$$

where $R$ is the residual rating matrix after bias correction. All methods start from the rank-5 SVD of the zero filled observed residual matrix. We use $\lambda = 1$.

**Methods.** We compare the same norm-matched Euclidean/intrinsic pairs as above and add Spectron [32] to the spectral group. Spectron is run without momentum, matching the optimizer protocol used in this section. Each method uses the decaying schedule $\eta_t = \eta_0/\sqrt{t}$, with $\eta_0$ selected by validation RMSE from the same grid

$$\{0.3, 0.5, 1, 1.5, 2, 3, 5, 7, 10, 15, 20, 30, 50, 75, 100, 150, 200, 300, 500, 700, 1000\}.$$

The values of $\eta_0$ are not directly comparable across methods because the update directions have different normalizations.

**Dataset.** We use a random $20\%$ test split and hold out $10\%$ of the remaining observations for validation. Validation RMSE is used for model selection. Results are averaged over seeds $\{0, 1, 2\}$.

Table 15: MovieLens-1M fixed-rank matrix completion with observed SVD initialization, rank 5, factor $\ell_2$ regularization $\lambda = 1$, and $\eta_t = \eta_0/\sqrt{t}$. We report mean test RMSE $\pm$ standard deviation over three seeds. Lower is better. Boldface marks the lowest RMSE within each norm group.

| Norm | Method | Selected $\eta_0$ | Test RMSE |
|---|---|---|---|
| Frobenius | EGD | 5 | $0.8694 \pm 0.0023$ |
| | RGD | 500 | $\mathbf{0.8691 \pm 0.0021}$ |
| Spectral | Muon | 20 | $0.8693 \pm 0.0019$ |
| | Spectron | 700 | $\mathbf{0.8691 \pm 0.0022}$ |
| | iMuon | 300 | $0.8692 \pm 0.0018$ |
| Nuclear | NuMuon | 7 | $0.8729 \pm 0.0026$ |
| | iMuon-Nu | 100 | $\mathbf{0.8715 \pm 0.0023}$ |

Table 16: Euclidean and intrinsic LMO directions for the SPD prototype experiment. Here $(\mu_1, u_1)$ and $(\lambda_1, q_1)$ are eigenpairs of $G$ and $H$, respectively, associated with the largest absolute eigenvalue, and $\mathrm{sign}(\cdot)$ denotes the matrix sign function.

| Norm | Euclidean $\xi^*$ | Intrinsic $\xi^*$ |
|---|---|---|
| Frobenius | $G/\|G\|_F$ | $XGX/\|H\|_F$ |
| Spectral | $\mathrm{sign}(G)$ | $X^{1/2}\,\mathrm{sign}(H)X^{1/2}$ |
| Nuclear | $\mathrm{sign}(\mu_1)u_1u_1^\top$ | $\mathrm{sign}(\lambda_1)X^{1/2}q_1q_1^\top X^{1/2}$ |

**Results.** Table 15 reports validation-selected test RMSE. RGD and iMuon-Nu improve over their Euclidean counterparts. In the spectral group, Spectron gives the lowest mean RMSE, while iMuon remains close and improves over factor-wise Muon. But these margins are small.

## M SPD Prototype Learning for Covariance-Based Image Classification

Table 16 gives the SPD update directions for the six methods: EGD, Muon, and NuMuon use the Euclidean column, while RGD, iMuon, and iMuon-Nu use the intrinsic column. Let $G = \nabla_X f$ and $H = X^{1/2}GX^{1/2}$. The table reports the LMO maximizer $\xi^*$. The optimizer retracts along $-\eta\xi^*$.

**Setup.** We consider supervised prototype classification from SPD covariance descriptors. Each example is a covariance matrix $C_i \in S^{32}_{++}$ with coarse label $y_i \in \{1, \dots, K\}$, where $K = 20$ for CIFAR-100 coarse labels. We learn one SPD prototype $X_c \in S^{32}_{++}$ per class, so the optimization variable is $X = (X_1, \dots, X_K) \in (S^{32}_{++})^K$.

Given prototypes $X_1, \dots, X_K$, we form logits from negative squared AI distance

$$z_{ic} = -\beta\, d_{\mathrm{AI}}(C_i, X_c)^2, \qquad d_{\mathrm{AI}}(C, X)^2 = \left\|\log\left(X^{-1/2}CX^{-1/2}\right)\right\|_F^2.$$

We minimize the regularized cross-entropy objective

$$f(X) = \frac{1}{N}\sum_{i=1}^{N} \ell_{\mathrm{CE}}(z_i, y_i) + \lambda \sum_{c=1}^{K} d_{\mathrm{AI}}(X_c, \bar{X}_c)^2,$$

where $\bar{X}_c$ is the log-Euclidean class-mean initialization for class $c$. We use the AI metric from the main text, initialize prototypes by log-Euclidean class means, set the logit scale to $\beta = 8.0$, and use $\lambda = 10^{-3}$. We train for 20 epochs with training batch size 64 and evaluation batch size 512, and report results over seeds $\{0, 1, 2\}$.

**Methods.** We use the same norm-matched comparison structure: RGD vs. EGD for the Frobenius norm, iMuon vs. Muon for the spectral norm, and iMuon-Nu vs. NuMuon for the nuclear norm. All methods are tuned over the shared learning-rate grid

$$\{10^{-3}, 3{\cdot}10^{-3}, 10^{-2}, 3{\cdot}10^{-2}, 10^{-1}\}.$$

Table 17: SPD covariance prototype classification on CIFAR-100 coarse labels. We report mean test accuracy $\pm$ standard deviation over three seeds. Boldface marks the higher accuracy method within each pair.

| Method | Selected LR | Mean test acc. |
|---|---|---|
| RGD | 0.01 | **0.548 $\pm$ 0.002** |
| EGD | 0.001 | 0.523 $\pm$ 0.026 |
| iMuon | 0.01 | **0.475 $\pm$ 0.009** |
| Muon | 0.001 | 0.407 $\pm$ 0.016 |
| iMuon-Nu | 0.001 | **0.546 $\pm$ 0.000** |
| NuMuon | 0.001 | 0.422 $\pm$ 0.032 |

The selected values are $0.01$ for RGD and iMuon, and $0.001$ for iMuon-Nu, EGD, Muon, and NuMuon. Model selection uses validation accuracy.

**Dataset.** We use CIFAR-100 coarse labels (20 classes) [40] together with frozen ResNet-18 visual features [26]. We extract the `layer3` feature map for each image, keep the backbone fixed, and project each spatial token from the backbone channel dimension to a 32-dimensional space using a fixed random projector with seed 17. We then build a shrinkage regularized covariance descriptor $C_i \in S^{32}_{++}$ [42, 69, 30] for each image using covariance shrinkage $0.1$ and diagonal stabilization $\varepsilon = 10^{-4}$. We use the standard CIFAR-100 train/test split and reserve $10\%$ of the training set as a validation split for learning-rate selection.

**Results.** Table 17 reports validation-selected test accuracy for the three norm-matched Euclidean/intrinsic pairs. Each intrinsic method improves over its Euclidean counterpart. The Frobenius gap is positive (RGD $0.548 \pm 0.002$ vs. EGD $0.523 \pm 0.026$), the spectral gap is positive (iMuon $0.475 \pm 0.009$ vs. Muon $0.407 \pm 0.016$), and the largest endpoint gap is for the nuclear pair (iMuon-Nu $0.546 \pm 0.000$ vs. NuMuon $0.422 \pm 0.032$, where the reported zero is a standard deviation below $0.0005$).

Figure 4 reports a companion convergence plot for the classification term. It uses the same covariance features and norm-matched pairs, removes the prototype anchoring term, and uses a smaller CPU split. In the Frobenius and spectral pairs, the Euclidean method reaches a lower training cross-entropy, but its test accuracy declines after the first few epochs while that of the intrinsic method continues to rise. In the nuclear pair, iMuon-Nu is ahead on both. Endpoint comparisons use Table 17.

## N Grassmann Fréchet Subspace Learning for Video Face Identification

**Setup.** We consider video-based face identification on YouTube Faces [76]. Each video clip is encoded as a $k$-dimensional subspace $S_i \in \mathrm{Gr}(64, k)$, and each identity class has a learned Grassmann prototype $X_c \in \mathrm{Gr}(64, k)$. For each class $c$, we optimize the Fréchet prototype objective

$$\min_{X_c \in \mathrm{Gr}(64,k)} \frac{1}{|\mathcal{I}_c|} \sum_{i \in \mathcal{I}_c} d_{\mathrm{Gr}}(X_c, S_i)^2,$$

where $\mathcal{I}_c$ indexes the training clips of class $c$, and $d_{\mathrm{Gr}}$ is the geodesic distance induced by principal angles. A test clip is classified by the nearest learned prototype under $d_{\mathrm{Gr}}$.

We use the canonical Grassmann metric. With a basis $X \in \mathbb{R}^{64 \times k}$ for a subspace, tangent vectors lie in the horizontal space $\{\xi : X^\top \xi = 0\}$, where $G_X = I$. Hence $\varphi(G_X^{1/2}\xi) \le \tau$ reduces to the Euclidean constraint $\varphi(\xi) \le \tau$. This boundary case checks that the intrinsic LMO recovers the projected Euclidean update when the metric scaling is identity.

**Methods.** We optimize the three norm-matched pairs for 100 iterations, using the squared-geodesic loss with scale $c_{\mathrm{Gr}} = 0.25$ and $\varepsilon = 10^{-6}$. We also report GDA [22] and PML [31] as reference methods. GDA learns a discriminative Grassmann projection, while PML learns a projection metric for subspaces.

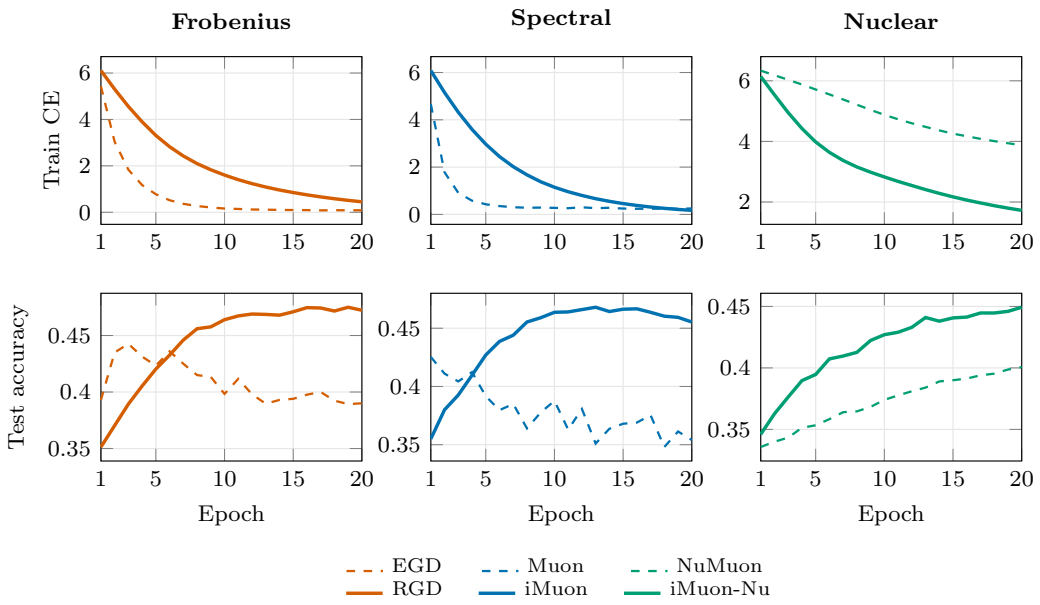



Figure 4: SPD convergence plot on frozen covariance features. The top row reports training cross-entropy and the bottom row reports test accuracy. Columns compare the Frobenius, spectral, and nuclear norm pairs. This variant omits the prototype anchoring term, so the accuracy values are not directly comparable with Table 17.

Table 18: Grassmann Fréchet prototype learning on YouTube Faces. We report mean classification accuracy $\pm$ standard deviation over 5 splits. Boldface marks the highest accuracy in each row. Since $G_X = I$ on the horizontal space, each intrinsic method coincides with its Euclidean projected counterpart.

| $k$ | RGD | iMuon | iMuon-Nu | EGD | Muon | NuMuon | GDA | PML |
|---|---|---|---|---|---|---|---|---|
| 3 | $0.136 \pm 0.017$ | $0.147 \pm 0.012$ | $0.138 \pm 0.018$ | $0.136 \pm 0.017$ | $0.147 \pm 0.012$ | $0.138 \pm 0.018$ | $0.087 \pm 0.012$ | $\mathbf{0.160 \pm 0.046}$ |
| 5 | $\mathbf{0.219 \pm 0.037}$ | $0.206 \pm 0.050$ | $0.211 \pm 0.038$ | $\mathbf{0.219 \pm 0.037}$ | $0.206 \pm 0.050$ | $0.211 \pm 0.038$ | $0.094 \pm 0.028$ | $0.179 \pm 0.031$ |
| 8 | $0.232 \pm 0.016$ | $\mathbf{0.238 \pm 0.042}$ | $\mathbf{0.238 \pm 0.034}$ | $0.232 \pm 0.016$ | $\mathbf{0.238 \pm 0.042}$ | $\mathbf{0.238 \pm 0.034}$ | $0.166 \pm 0.022$ | $0.194 \pm 0.031$ |
| 10 | $0.262 \pm 0.016$ | $0.264 \pm 0.038$ | $\mathbf{0.266 \pm 0.034}$ | $0.262 \pm 0.016$ | $0.264 \pm 0.038$ | $\mathbf{0.266 \pm 0.034}$ | $0.164 \pm 0.032$ | $0.213 \pm 0.011$ |

**Dataset.** We use YouTube Faces clips [76] represented as $k$-dimensional subspaces of a 64-dimensional feature space, with $k \in \{3, 5, 8, 10\}$. We use the standard Grassmann geometry [20, 1] and evaluate over 5 random train/test splits with training fraction 0.7.

**Results.** Table 18 reports classification accuracy on YouTube Faces. Across all $k$, each intrinsic method matches its Euclidean counterpart within the displayed precision. This is expected: with $G_X = I$, the intrinsic LMO and the Euclidean LMO applied to the horizontal gradient $(I - XX^\top)\nabla_X f$ are the same problem, so the paired columns are the same algorithm. The informative comparison in the table is therefore across norms, and against GDA and PML.

## O Stiefel Sub-Center Prototype Learning for Frozen-Feature Image Classification

**Setup.** We consider supervised classification with class prototypes constrained to the Stiefel manifold [20, 1]. Given frozen image features $h_i \in \mathbb{R}^{512}$ and labels $y_i \in \{1, \dots, C\}$, with $C = 100$ for CIFAR-100, we learn $q = 4$ prototypes per class, giving $Q = Cq$ prototypes in total. The prototypes are the columns of

$$X = [x_{1,1}, \dots, x_{1,q}, \dots, x_{C,1}, \dots, x_{C,q}] \in \mathrm{St}(512, Q), \qquad X^\top X = I_Q.$$

Table 19: Stiefel sub-center prototype classification on frozen CIFAR-100 features. We train an additive-margin classifier with $q = 4$ prototypes per class on $\mathrm{St}(512, 400)$, select the learning rate by validation accuracy for each seed, and report mean test accuracy and test cross entropy. CE denotes average negative log likelihood. Higher accuracy and lower CE are better.

| Method | Selected LRs | Test acc. | Test CE |
|---|---|---|---|
| RGD | $0.1, 0.1, 0.1$ | $0.542 \pm 0.013$ | $1.853 \pm 0.026$ |
| iMuon | $0.03, 0.02, 0.02$ | $\mathbf{0.553 \pm 0.006}$ | $\mathbf{1.839 \pm 0.024}$ |
| SPEL | $0.03, 0.03, 0.05$ | $0.545 \pm 0.012$ | $1.869 \pm 0.066$ |

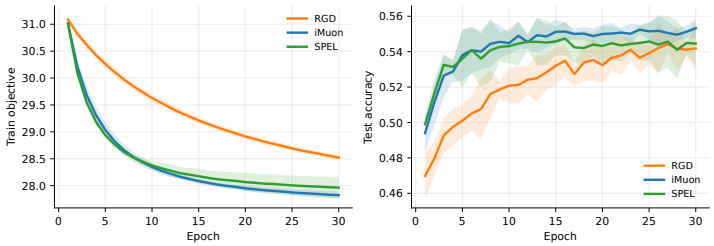



Figure 5: Stiefel sub-center prototype classification trajectories. The left panel reports the training objective and the right panel reports test accuracy over epochs. Curves show means over three seeds with standard deviation bands.

For class $c$, the classifier score is the best matching prototype score

$$s_c(h_i, X) = \max_{1 \le j \le q} h_i^\top x_{c,j}.$$

We use an additive margin softmax objective [72]. For the target class $y_i$, the score is shifted to $s_{y_i}(h_i, X) - m$ before multiplying all class scores by a logit scale $\gamma$. The optimization problem is

$$\min_{X \in \mathrm{St}(512,Q)} \frac{1}{N} \sum_{i=1}^{N} \ell_{\mathrm{CE}}\big(\gamma \left(s_1(h_i, X), \ldots, s_{y_i}(h_i, X) - m, \ldots, s_C(h_i, X)\right), y_i\big) .$$

We use margin $m = 0.5$ and logit scale $\gamma = 64$. The Stiefel constraint enforces global orthogonality among all class prototypes. This gives a real supervised task for comparing intrinsic and extrinsic spectral updates on $\mathrm{St}(512, Q)$, $Q = 400$.

**Methods.** We compare RGD, iMuon, and SPEL. RGD is normalized Riemannian gradient descent on $\mathrm{St}(512, Q)$. iMuon is the spectral norm instance of our intrinsic LMO, using the closed form tangent decomposition from the main text. SPEL is the Stiefel specialization of manifold constrained steepest descent [80]. It forms a spectral steepest descent direction from the Riemannian gradient using extrinsic normalization, then retracts to the manifold. The updates are discussed in Section H.3. All methods use polar retraction. Learning rates are selected by validation accuracy from $\{0.02, 0.03, 0.05, 0.1\}$.

**Dataset.** We use CIFAR-100 [40] with frozen ResNet-18 features [26]. Features are $\ell_2$-normalized before training the Stiefel classifier. Each seed uses 50 training examples, 10 validation examples, and 30 test examples per class. We train for 30 epochs with batch size 64, evaluation batch size 4096, and select the learning rate by validation accuracy.

**Results.** Table 19 reports test accuracy and test cross entropy (CE), the average negative log likelihood under the additive margin softmax classifier. iMuon has the best mean accuracy and the lowest CE. It improves over RGD by 1.1 percentage points and is slightly above SPEL. Figure 5 shows that iMuon and SPEL reduce the training objective faster than RGD, with iMuon slightly higher in mean test accuracy near the end of training.

# P Additional Manifolds and Metrics

The LMO framework of Section 3 is stated for general Riemannian matrix manifolds. In the main text we instantiate it on four manifolds: two quotient manifolds ($\mathcal{M}_r$, $\mathrm{Gr}(m, r)$), the SPD manifold $S^n_{++}$ (with $\mathrm{GL}(n)$ acting by isometries), and the Stiefel manifold $\mathrm{St}(m, r)$. Here we discuss broader applicability and several instructive boundary cases: (i) general conditions for closed-form solutions, (ii) the fixed-rank manifold from the *embedded* viewpoint, why the joint spectral-norm constraint on $\dot{X}$ does not decouple, and how the quotient formulation resolves this, (iii) the Stiefel manifold from the *embedded submanifold* viewpoint, why the naive approach fails and how the block decomposition of Section 4 resolves it, (iv) the spectral sphere $\{X : \|X\|_2 = R\}$, another non-SV-invariant example, recently used in the SSO optimizer [77], and (v) alternative SPD metrics (Log-Euclidean and Bures–Wasserstein).

## P.1 General applicability

The LMO (4) is well-defined on *any* Riemannian manifold $\mathcal{M}$ with a metric operator $G_x$: it is simply a convex optimization over the tangent space. The two structural properties that enable closed-form solutions and symmetry preservation are:

1. **Left-right structure** (Proposition 3.1): the scaled map $\rho_h = G_{\tilde{x}}^{1/2} \circ D_x h \circ G_x^{-1/2}$ acts as $Z \mapsto U_h Z V_h^\top$ for orthogonal $U_h, V_h$, which holds when the factors of a left-right metric transform as in Proposition 3.1. This ensures $\varphi(\rho_h Z) = \varphi(Z)$ for unitarily invariant $\varphi$.
2. **Singular value invariance or $\varphi$-contractivity** (Proposition 3.4): the scaled tangent space $\mathcal{Z}_x$ is SV-invariant (Definition 3.2) or $\varphi$-contractive (Definition 3.3), so the tangent-space constraint is redundant and the LMO reduces to a standard Euclidean problem with closed-form solutions.

For *embedded submanifolds* (no quotient symmetry), the symmetry condition (1) is not required. Condition (2) alone suffices for closed-form solutions. If $\mathcal{Z}_x$ satisfies neither condition, the LMO remains a convex problem, but its solution is in general no longer given by the closed forms of Proposition 3.4, such as $\mathrm{Ortho}(H)$, and may require an iterative solver (see, e.g., Section P.4).

## P.2 Fixed-rank manifold: embedded viewpoint

In Section 4 we work on the quotient $\mathcal{M}_r = (\mathbb{R}_*^{m\times r} \times \mathbb{R}_*^{r\times n})/\mathrm{GL}(r)$ with the coupled metric, which gives $G_x^{1/2} \neq I$ and decouples the LMO into independent $\dot{B}$ and $\dot{A}$ subproblems. Here we discuss the alternative *embedded* viewpoint, working directly on $\{X \in \mathbb{R}^{m\times n} : \mathrm{rank}(X) = r\}$ with $G_x = I$, and explain why it does not yield the same clean decoupled solutions.

**The embedded LMO.** On the embedded rank-$r$ manifold, the tangent space at $X = BA$ is $T_X\mathcal{M}_r = \{\dot{B}A + B\dot{A} : \dot{B} \in \mathbb{R}^{m\times r},\ \dot{A} \in \mathbb{R}^{r\times n}\}$. With $G_x = I$, the LMO becomes

$$\dot{X}^* = \underset{\dot{X}\in T_X\mathcal{M}_r,\ \varphi(\dot{X})\le\tau}{\arg\max} \langle \dot{X},\, \nabla_X f\rangle.$$

For $\varphi = \|\cdot\|_2$, the unconstrained maximizer over all of $\mathbb{R}^{m\times n}$ is $\mathrm{Ortho}(\nabla_X f)$, which has rank $\min(m, n)$ and generally does not lie in $T_X\mathcal{M}_r$ (a subspace of rank $\le 2r$). Riemannion [12] addresses this by first computing a rank-$2r$ orthogonalization $\mathrm{Ortho}_r(M_t)$ of the momentum (which lies in $T_X\mathcal{M}_r$), and then projecting the result back onto $T_X\mathcal{M}_r$. This tangent-space projection changes the singular values, so the spectral-norm constraint is only approximately satisfied.

**Failure of SV-invariance.** The scaled tangent space $\mathcal{Z}_X = T_X\mathcal{M}_r = \{\dot{B}A + B\dot{A}\} \subset \mathbb{R}^{m\times n}$ is *not* singular value invariant: given $\dot{X} = U\Sigma V^\top \in T_X\mathcal{M}_r$, replacing $\Sigma$ with a different non-negative diagonal $D$ yields $UDV^\top$ which generally does not decompose as $\dot{B}A + B\dot{A}$ for any $(\dot{B}, \dot{A})$. The tangent space is not spectral-norm contractive either (simple counter-examples exist). Thus Proposition 3.4 does not apply, and the constrained LMO requires solving

$\max_{\dot{X} \in T_X \mathcal{M}_r,\, \|\dot{X}\|_2 \le \tau} \langle \dot{X},\, \nabla_X f \rangle$, a constrained problem for which we are not aware of a closed form.

**Why the quotient formulation resolves this.** On the quotient, each factor block lives in a full ambient space ($\mathbb{R}^{m\times r}$ for $\dot{B}$, $\mathbb{R}^{r\times n}$ for $\dot{A}$), which is trivially SV-invariant. The product-norm decoupling $\max(\varphi(Z_B), \varphi(Z_A)) \le \tau$ then gives two independent LMOs, each with a clean closed form. The price is twofold: (a) the product norm is a relaxation of the joint norm $\varphi(\dot{X}) \le \tau$ (specifically, $\|\dot{X}\|_2 \le 2\tau$ rather than $\tau$), and (b) the non-trivial metric $G_x^{1/2} \neq I$ must be inverted (the Gram-root inversion). But this price buys symmetry invariance (Proposition 3.1) and a condition-number-free bound $\|\xi^*\|_x^2 \le 2r\tau^2$ (Section D.6).

**Frobenius norm.** For $\varphi = \|\cdot\|_F$, the embedded LMO does have a closed form: $\dot{X}^* \propto P_{T_X}(\nabla_X f)$, the tangent-space projection of the gradient, which is the Riemannian gradient (up to normalization).

### P.3 Stiefel manifold: embedded viewpoint

In Section 4, we derived closed-form LMO solutions on $\mathrm{St}(m, r)$ using the decomposition $T_X\mathrm{St} = X \cdot \mathrm{Skew}(r) \oplus X_\perp \cdot \mathbb{R}^{(m-r)\times r}$ and the product norm. Here, we discuss the alternative *embedded* viewpoint, which treats $T_X\mathrm{St}$ as a single subspace of $\mathbb{R}^{m\times r}$, and explain why Proposition 3.4 does not apply to it.

**The embedded LMO.** Since $G_X = I$, the LMO (4) on $\mathrm{St}(m, r)$ is

$$\xi^* = \operatorname*{arg\,max}_{\xi \in T_X\mathrm{St},\ \varphi(\xi) \le \tau} \langle \xi,\, \nabla_X f \rangle.$$

We write $S = \mathrm{skew}(X^\top \nabla_X f)$ and $N = X_\perp^\top \nabla_X f$ for the gradient blocks paired with $A$ and $B$ in $\xi = XA + X_\perp B$, so that $\langle \xi,\, \nabla_X f\rangle = \langle A,\, S\rangle + \langle B,\, N\rangle$. For $\varphi = \|\cdot\|_F$, the solution is the normalized projected gradient $\xi^* = \tau\,(XS + X_\perp N)/(\|S\|_F^2 + \|N\|_F^2)^{1/2}$. For $\varphi = \|\cdot\|_2$, the unconstrained maximizer $\tau\ \mathrm{Ortho}(\nabla_X f)$ generally does not lie in $T_X\mathrm{St}$, and projecting it onto $T_X\mathrm{St}$ changes its singular values, so the spectral-norm constraint may be violated.

**Failure of the conditions of Proposition 3.4.** The tangent space $T_X\mathrm{St}$ is not SV-invariant (Definition 3.2), since replacing the singular values of $\xi \in T_X\mathrm{St}$ generally violates $X^\top\xi + \xi^\top X = 0$. It is not spectral-norm contractive either. For $X = [e_1, e_2] \in \mathrm{St}(3, 2)$ and $W = [e_3, e_1]$, we have $\|W\|_2 = 1$ but $\|P_{T_X\mathrm{St}}(W)\|_2 = \sqrt{5}/2 > 1$. Hence, for $\varphi = \|\cdot\|_2$, the LMO above does not reduce to the unconstrained Euclidean LMO. In block form, the joint constraint $\|[A; B]\|_2 \le \tau$ couples $A$ and $B$, and the unconstrained maximizer $\mathrm{Ortho}([S; N])$ generally has a top block that is not skew-symmetric.

**Resolution via the product norm.** The product norm $\max\{\varphi(A), \varphi(B)\} \le \tau$ of Section 4 decouples the LMO into a skew and a normal subproblem. The normal block $\mathbb{R}^{(m-r)\times r}$ is trivially SV-invariant, and the skew block $\mathrm{Skew}(r)$ is SV-invariant because the singular values of a skew-symmetric matrix come in equal pairs (Section E). Hence, Proposition 3.4 applies to each block and yields the closed forms of Lemma 4.2. For $\varphi = \|\cdot\|_F$, the product-norm solution normalizes the two blocks separately and is therefore parallel to the normalized projected gradient above only when $\|S\|_F = \|N\|_F$ (in which case it is $\sqrt{2}$ times larger) (Section H.3).

**Conditioning of the product norm when the skew block vanishes.** Since the product norm constrains the two blocks independently, the spectral-norm skew update satisfies $\|A^*\|_2 = \tau$ whenever $S \neq 0$, irrespective of the magnitude of $\|S\|_F$. As $S \to 0$, the polar factor is computed from a matrix whose entries lie at the level of rounding error, its singular vectors become arbitrary, and the update retains a rotation of full magnitude within $\mathrm{col}(X)$ that carries no gradient information. This is the Grassmann limit. Objectives invariant under $X \mapsto XO$ for $O \in O(r)$, such as $f(X) = -\frac{1}{2}\,\mathrm{tr}(X^\top C X)$ with $C$ symmetric, belong to this case, since $X^\top \nabla_X f$ is then symmetric,

so $S$ vanishes in exact arithmetic and lies at the level of rounding error in floating point. Such objectives depend only on $\mathrm{col}(X)$, so the rotation leaves the function value unchanged. When the $O(r)$ symmetry is broken, however, the rotation contributes no descent and the step is used inefficiently. This behavior is a consequence of orthogonalizing the two blocks separately. A simple safeguard is to set $A^* = 0$ once $\|S\|_F$ falls below a prescribed tolerance relative to $\|N\|_F$.

**Relation to the joint-norm formulation.** For $\varphi = \|\cdot\|_2$, the joint-norm LMO above has been solved via dual ascent [7] and other iterative procedures [16, 65], and in closed form in a concurrent work [64]. We discuss these approaches in Section H.3.

### P.4 Spectral sphere

The spectral sphere $\mathcal{S}_R := \{X \in \mathbb{R}^{m\times n} : \|X\|_2 = R\}$ constrains only the *largest* singular value of $X$ to equal $R$. The remaining singular values are free to take any values in $[0, R]$. When the top singular value is simple (a generic condition), the manifold is smooth with tangent space

$$T_X\mathcal{S}_R = \{\xi \in \mathbb{R}^{m\times n} : \langle u_1 v_1^\top, \xi\rangle = 0\},$$

where $(u_1, v_1)$ are the leading left and right singular vectors of $X$, and the gradient of the spectral norm is $\nabla_X \|X\|_2 = u_1 v_1^\top$.

**Failure of SV-invariance.** The tangent space is the hyperplane $\{\xi : \langle u_1 v_1^\top, \xi\rangle = 0\}$, defined by an inner product with a *data-dependent* rank-1 matrix. This is not SV-invariant: given $\xi = U\Sigma V^\top \in T_X\mathcal{S}_R$, replacing $\Sigma$ with a different non-negative diagonal $D$ yields $UDV^\top$ whose inner product $\langle u_1 v_1^\top, UDV^\top\rangle = \sum_i d_i (u_1^\top a_i)(v_1^\top b_i)$ (with $a_i, b_i$ the columns of $U, V$) generally differs from zero. The tangent space is not spectral-norm contractive either. Thus Proposition 3.4 does not apply, and the spectral-norm LMO $\max\{\langle \nabla_X f, \xi\rangle : \xi \in T_X\mathcal{S}_R,\ \|\xi\|_2 \le \tau\}$ is not covered by the closed forms of Proposition 3.4.

**Connection to SSO.** Although SV-invariance fails, the LMO on $\mathcal{S}_R$ remains a well-posed convex problem, namely the maximization of a linear functional over the intersection of a spectral-norm ball and a hyperplane. This is precisely the problem solved by the Spectral Sphere Optimizer (SSO) [77], which performs steepest descent on $\mathcal{S}_R$ under the spectral norm for $\mu$P-aligned [79] LLM training. In other words, SSO is an instance of our framework (4) with $G_X = I$ and $\varphi = \|\cdot\|_2$ on the manifold $\mathcal{S}_R$. It solves exactly the same intrinsic LMO, but without the benefit of a closed-form solution. The single hyperplane constraint requires one Lagrange multiplier $\lambda$, and SSO solves for it via bisection on the monotone function

$$h(\lambda) := \langle u_1 v_1^\top,\, \mathrm{Ortho}(\nabla_X f + \lambda\, u_1 v_1^\top)\rangle = 0, \qquad \xi^*(\lambda) = \mathrm{Ortho}(\nabla_X f + \lambda\, u_1 v_1^\top).$$

Each bisection step requires one $\mathrm{Ortho}$ evaluation and approximately 5–7 bisection steps are needed in practice. If the tangent constraint $\langle u_1 v_1^\top, \xi\rangle = 0$ were dropped (i.e., no weight-norm preservation), the solution would reduce to $\xi^* = \mathrm{Ortho}(\nabla_X f)$, standard Euclidean Muon. Thus the bisection cost is the computational price of the non-SV-invariant tangent constraint, which arises from simultaneously constraining both the update norm and the weight norm.

**Takeaway.** The spectral sphere illustrates that our framework applies beyond manifolds with SV-invariant tangent spaces: the LMO (4) is always well-posed, and SV-invariance determines only whether the solution is available in closed form or requires a solver. On the four manifolds of Section 4, SV-invariance holds (blockwise on $\mathrm{St}(m, r)$ under the product norm) and the LMO reduces to $\mathrm{Ortho}$, $\mathrm{sign}$, or rank-1 truncation. On the spectral sphere, a single linear constraint breaks SV-invariance, but a one-dimensional bisection restores tractability at the cost of a few additional $\mathrm{Ortho}$ evaluations per step. The Frobenius instance ($\varphi = \|\cdot\|_F$) remains closed-form in all cases, as the LMO reduces to projecting the gradient onto the tangent space and normalizing.

### P.5 Alternative SPD metrics: Log-Euclidean and Bures–Wasserstein

In Section 4 we use the affine-invariant (AI) metric on $S^n_{++}$. Here we analyze how two other common SPD metrics interact with the two structural properties of the framework: left-right structure (Proposition 3.1, enabling symmetry invariance) and SV-invariance of the scaled tangent space

(Proposition 3.4, enabling closed-form reduction). All three metrics share the same tangent space $T_X S^n_{++} = \mathrm{Sym}(n)$, which is SV-invariant (Section E). Thus Proposition 3.4 applies to all three, and the LMO always reduces to an unconstrained Euclidean problem with the standard closed forms ($\mathrm{sign}(H)$, $H/\|H\|_F$, or rank-1 truncation). The metrics differ only in how the two propositions' *other* aspects, symmetry group size and the explicitness of the scale/invert steps, play out.

**Affine-invariant (AI) metric (recap).** $G^{AI}_X(\xi) = X^{-1}\xi X^{-1}$, so $G^{1/2}_X(\xi) = X^{-1/2}\xi X^{-1/2}$ (left-right form with $P = Q = X^{-1}$).

- *Proposition 3.1 (left-right structure):* holds with $\mathrm{GL}(n)$-invariance.
- *Proposition 3.4 (SV-invariance):* $\mathcal{Z}_X = \mathrm{Sym}(n)$, which is SV-invariant.
- *Scale/invert:* coordinate-free sandwich formulas $H = X^{1/2}(\nabla_X f)X^{1/2}$ and $\xi^* = X^{1/2}Z^*X^{1/2}$.

This is the ideal case: both propositions apply with the maximal symmetry group, and all three pipeline steps (scale, solve, invert) have explicit matrix-coordinate formulas.

**Log-Euclidean metric.** The Log-Euclidean metric [4] uses $\Phi : X \mapsto \log X$ as an isometry from $(S^n_{++}, g^{LE})$ to $(\mathrm{Sym}(n), \|\cdot\|_F)$, with $g^{LE}_X(\xi,\eta) = \mathrm{tr}\big(D_X \log(\xi)^\top D_X \log(\eta)\big)$. In the log-coordinates $Y = \log X$, the metric is flat Euclidean ($G_Y = I$).

- *Proposition 3.1:* trivially satisfied ($G_Y = I$), but invariant only under similarities $X \mapsto cQXQ^\top$ with $c > 0$ and $Q \in O(n)$, not $\mathrm{GL}(n)$-invariant.
- *Proposition 3.4:* $\mathcal{Z}_Y = \mathrm{Sym}(n)$, SV-invariant. Since $G_Y = I$, the reduction is trivial: no scaling or inversion needed.
- *Scale/invert:* identity (nothing to compute).

The spectral-norm LMO gives $\Delta Y^* = \tau\ \mathrm{Ortho}(\nabla_Y f)$ in log-space, followed by $X_+ = \exp(Y - \eta\,\Delta Y^*)$. This is simply standard Euclidean Muon applied to $\log X$; the framework adds nothing beyond reparameterization. Both propositions hold, but only vacuously. No non-trivial preconditioning or symmetry preservation emerges.

**Bures–Wasserstein (BW) metric.** The BW metric [11, 45, 23] is $g^{BW}_X(\xi,\zeta) = \frac{1}{2}\,\mathrm{tr}(L_\xi \zeta)$, where $L_\xi$ is the unique symmetric solution of the Lyapunov equation $XL_\xi + L_\xi X = \xi$. The metric operator is $G^{BW}_X = \frac{1}{2}\mathcal{L}_X$, where $\mathcal{L}_X(\xi) := L_\xi$. In the eigenbasis $X = U\Lambda U^\top$ with eigenvalues $\lambda_1, \ldots, \lambda_n$, entry $(i,j)$ of $G^{BW}_X(\xi)$ scales as $\xi_{ij}/(\lambda_i + \lambda_j)$.

- *Proposition 3.1 (left-right structure):* the left-right assumption **fails** because the BW scaling $1/(\lambda_i + \lambda_j)$ does not factor as $f(\lambda_i)\cdot g(\lambda_j)$, so $G^{1/2}_X$ cannot be written as $P\xi Q^\top$ and the proof of Proposition 3.1 does not apply directly. However, the *conclusion* of Proposition 3.1 still holds for the symmetry group $O(n)$ via a direct argument: under $X \mapsto QXQ^\top$ with $Q \in O(n)$, the eigenbasis rotates $U \mapsto QU$ while eigenvalues $\lambda_i$ are unchanged, so $G^{1/2}_{\tilde{X}}(Q\xi Q^\top) = Q\,G^{1/2}_X(\xi)\,Q^\top$. Since left- and right-multiplication by the same orthogonal $Q$ preserves all unitarily invariant norms, $\varphi(G^{1/2}_{\tilde{X}}\tilde{\xi}) = \varphi(G^{1/2}_X \xi)$. Thus the intrinsic constraint is $O(n)$-invariant, but not $\mathrm{GL}(n)$-invariant (for example, $N = cI$ with $|c| \neq 1$ gives $\varphi(G^{1/2}_{\tilde{X}}(N\xi N^\top)) = |c|\,\varphi(G^{1/2}_X\xi)$, because $N\xi N^\top = c^2\xi$ while the Hadamard weights $1/\sqrt{\lambda_i + \lambda_j}$ scale by $1/|c|$).
- *Proposition 3.4 (SV-invariance):* **holds.** The scaled tangent space $\mathcal{Z}_X = \{(G^{BW}_X)^{1/2}\xi : \xi \in \mathrm{Sym}(n)\} = \mathrm{Sym}(n)$ (since $(G^{BW}_X)^{1/2}$ is an invertible self-adjoint operator on $\mathrm{Sym}(n)$), which is SV-invariant. Thus the LMO reduces to the unconstrained problem $Z^* = \tau\ \mathrm{sign}(H)$ for the spectral norm, etc.

- *Scale/invert:* closed-form but requires the eigenbasis. In the eigenbasis $X = U\Lambda U^\top$, the operators $(G_X^{BW})^{1/2}$ and $(G_X^{BW})^{-1/2}$ act entrywise on $\hat{\xi} := U^\top \xi U$ as

$$\big[U^\top (G_X^{BW})^{1/2}(\xi)\, U\big]_{ij} = \frac{\hat{\xi}_{ij}}{\sqrt{2(\lambda_i + \lambda_j)}}, \qquad \big[U^\top (G_X^{BW})^{-1/2}(Z)\, U\big]_{ij} = \sqrt{2(\lambda_i + \lambda_j)}\, \hat{Z}_{ij},$$

where $\hat{Z} = U^\top Z U$. Thus the full pipeline is: (i) diagonalize $X = U\Lambda U^\top$ once ($O(n^3)$); (ii) scale: $H_{ij} = \hat{G}_{ij}/\sqrt{2(\lambda_i + \lambda_j)}$ where $\hat{G} = U^\top(\operatorname{grad} f)U$; (iii) solve: $Z^* = \tau\, \operatorname{sign}(H)$ via eigendecomposition of $H$; (iv) invert: $\hat{\xi}^*_{ij} = \sqrt{2(\lambda_i + \lambda_j)}\, Z^*_{ij}$, then $\xi^* = U\hat{\xi}^* U^\top$. All steps are $O(n^3)$, the same asymptotic cost as the AI case.

The key difference from AI is therefore not computational cost but geometric: (a) the inversion formula is a Hadamard (entrywise) rescaling in the eigenbasis rather than a coordinate-free matrix sandwich $X^{1/2}(\cdot)X^{1/2}$, and (b) the symmetry group is $O(n)$ rather than $\mathrm{GL}(n)$, so the intrinsic constraint is invariant under a smaller group.

**Summary.** We summarize the status of the two key structural properties across the three metrics:

| | **Affine-invariant** | **Log-Euclidean** | **Bures–Wasserstein** |
|---|---|---|---|
| Symmetry group | $\mathrm{GL}(n)$ | $O(n)$ and scalings | $O(n)$ |
| Proposition 3.1 (left-right) | ✓ | trivial ($G = I$) | × (but $O(n)$-inv. holds) |
| Proposition 3.4 (SV-inv.) | ✓ | ✓ | ✓ |
| Scale/invert formula | $X^{1/2}(\cdot)X^{1/2}$ | identity | eigenbasis Hadamard |
| Cost | $O(n^3)$ | $O(n^3)$ | $O(n^3)$ |

## 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 abstract and introduction state four contributions: a unified intrinsic LMO framework, closed-form solutions on four manifolds, a convergence analysis parametrized by $C_\varphi$, and experiments. These match the corresponding sections of the paper.

   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: The conclusion points to a dedicated Limitations appendix (Section B) and additional details are in Section P.

   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: All theorems, propositions, lemmas, and corollaries are numbered and cross-referenced. The retraction-smoothness and bounded-noise assumptions are stated explicitly. All proofs (of the propositions, lemmas, corollaries, and convergence theorems) are given in the appendix. Singular-value invariance is verified per-manifold in the appendix.

Guidelines:

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

4. **Experimental result reproducibility**

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

Answer: [Yes]

Justification: The algorithm proposed in the paper specifies the full update rule. The appendix provides step-by-step implementation details (QR-based computation, Newton–Schulz iterations) for every manifold. Experimental hyperparameters, seeds, and hardware are reported in the appendix.

Guidelines:

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

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

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

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

   Answer: [Yes]

   Justification: Code to reproduce all experiments is publicly available at `https://github.com/1bang118/manifold-intrinsic-muon`. All datasets used 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: The experiments section summarizes the setup for each experiment. Learning-rate grids and selected values, batch sizes, seeds, data splits, and epoch counts are given in the appendix (Sections K to O).

   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: [Yes]

   Justification: Mean $\pm$ standard deviation (one sigma, sample standard deviation across random seeds) is reported for all manifold experiments (Tables 13 to 15 and 17 to 19 and Figures 1 to 5) and for the LoRA seed-sensitivity experiments (Tables 7 and 12). The

seeds and the source of randomness (initialization and data order, and train/test splits where applicable) are stated in the corresponding appendix sections.

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: Section K reports the GPU types used (L40, A40, A100, RTX 6000), the single-GPU versus two-GPU DDP setup for E2E NLG and GLUE, and a per-batch runtime profile (Table 6); it also states that no hardware-specific tuning was used. The non-LLM experiments (SPD, Grassmann, Stiefel, fixed-rank synthetic and MovieLens-1M) are run on CPU/GPU with the seeds, batch sizes, and epoch counts reported in the corresponding appendix sections.

   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: The research is foundational work on optimization algorithms. It does not involve human subjects, sensitive data, or dual-use concerns. We have reviewed and conform to the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.

- If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
- The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

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

    Answer: [Yes]

    Justification: Section A discusses broader societal impact: the contribution is foundational and methodological, the framework is not tied to a particular application or deployment, and the empirical evaluation uses standard publicly available benchmarks. We do not identify a direct path to negative societal impact beyond the general-purpose effect of more efficient ML training.

    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: The paper does not release pre-trained models, large language models, image generators, or scraped datasets. The released code implements optimization algorithms only.

    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 datasets (E2E NLG, GLUE, CIFAR-100, MovieLens-1M, YouTube Faces), pretrained models (GPT-2 Medium, Mistral-7B-v0.1, ResNet-18), evaluation code (`e2e-metrics`), and baseline implementations (Riemannion) are cited with their original papers and are used under their respective licenses and terms of use.

    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 datasets or pre-trained models. Code is released for reproducibility but does not constitute a new benchmark or asset.

    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: The paper does not involve crowdsourcing or research with human subjects.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.

- 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: The paper does not involve human subjects research.

    Guidelines:

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

16. **Declaration of LLM usage**

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

    Answer: [N/A]

    Justification: LLMs are not a component of the proposed method. Generative AI tools were used to assist with drafting and editing text, formatting references, writing and debugging experiment code, and formulating and checking mathematical claims and proofs. They were not used to design the method or the experiments, to produce or alter experimental results, or to interpret results.

    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.