PLUM: LIGHTWEIGHT LOW-RANK UPDATES FOR PARAMETER- AND MEMORY-EFFICIENT TRAINING

Anonymous authors
Paper under double-blind review

ABSTRACT

Training large models is constrained by memory for parameters, gradients, and optimizer states. Low-rank parameterization reduces these costs, but factor-wise optimization is not invariant to the non-unique factorization: equivalent representations of the same weight can induce different weight updates. We introduce PLUM (**P**arameter-efficient **L**ow-rank **U**pdates with **M**uon), a lightweight optimizer that represents each weight by an orthonormal subspace basis and a coefficient matrix, separating subspace motion from coefficient changes with a simple implementation. The unit nonzero singular values of Muon's msign update make the Grassmann geodesic and momentum transport particularly simple, reducing both to inexpensive matrix operations without matrix exponentials or repeated QR/polar retractions. A fixed Newton–Schulz approximation further yields a fully GEMM-based implementation with one momentum state per factor and no additional factor- or geometry-specific hyperparameters to tune. We establish an $O(T^{-1/4})$ expected first-order stationarity rate and an $O(T^{-1/2})$ bound on the expected maximum orthogonality error for fixed finite Newton–Schulz iterations under standard assumptions. Experiments on C4 pretraining and GLUE fine-tuning up to models with over 1B parameters show lower pretraining perplexity than prior low-rank methods together with a lightweight, GPU-friendly implementation.

1 INTRODUCTION

Training large models requires substantial memory for parameters, gradients, and optimizer states. Native low-rank pretraining directly reduces these persistent costs by keeping selected weight matrices in factorized form throughout optimization (Mo et al., 2025; Janson et al., 2026). Unlike low-rank adaptation, which trains compact corrections to a frozen dense backbone (Hu et al., 2022), or gradient-projection methods such as GaLore, which retain dense weights (Zhao et al., 2024), native low-rank pretraining also compresses the trainable weights themselves.

A selected weight $W \in \mathbb{R}^{m\times n}$ is stored as BA, with $B \in \mathbb{R}^{m\times r}$ and $A \in \mathbb{R}^{r\times n}$, replacing mn entries by $r(m+n)$. Retaining these factors throughout training reduces parameter, gradient, and optimizer-state storage (Mo et al., 2025; Janson et al., 2026). The resulting optimization, however, depends on the factor representation. Equivalent factorizations of the same weight can induce different weight updates under the same factor-wise optimizer.

This representation dependence follows directly from the product structure. Simultaneous updates satisfy

$$\Delta W = \Delta BA + B\Delta A + \Delta B\Delta A, \qquad BA = (BS)(S^{-1}A), \quad S \in \mathrm{GL}(r). \tag{1}$$

This symmetry is standard in low-rank matrix optimization (Mishra et al., 2014; Luo et al., 2024) and motivates factor-aware preconditioning in LoRA (Zhang & Pilanci, 2024; Yen et al., 2025). Its effect is particularly transparent for ideal factor-wise Muon. Positive rescaling leaves the matrix-sign updates Q_B, Q_A unchanged, but under $(B, A) \mapsto (cB, c^{-1}A)$ the first-order weight update becomes

$$-\eta(Q_B A + BQ_A) \quad \longmapsto \quad -\eta(c^{-1}Q_B A + cBQ_A).$$

The two contributions are reweighted even though the represented model is identical. LoRA-RITE connects this dependence to imbalanced factor updates (Yen et al., 2025). LORO instead optimizes

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the represented weight on the fixed-rank manifold rather than treating its factors as independent Euclidean variables (Mo et al., 2025). Spectron further links rapid growth of $\|\Delta W\|_2$ in native low-rank pretraining to large activation changes and constrains the factor updates to stabilize training through per-step power-iteration subroutine estimates of both factor spectral norms (Janson et al., 2026). Together, these results suggest that low-rank optimization should respect the represented weight rather than arbitrary factor coordinates. Existing approaches provide such control through Riemannian optimization or explicit constraints on factor-induced weight updates, often at the cost of additional geometric operations or additional hyperparameter tuning. A desirable alternative is therefore to retain product-level control within a simple, hardware-efficient optimizer.

To this end, we propose PLUM (**P**arameter-efficient **L**ow-rank **U**pdates with **M**uon), a lightweight optimizer that separates the column subspace from its coefficients:

$$W = UY^\top, \quad U \in \mathrm{St}(m, r), \quad Y \in \mathbb{R}^{n\times r}, \quad \|W\|_F = \|Y\|_F,$$

where the standard Stiefel manifold is defined as $\mathrm{St}(m, r) = \{U \in \mathbb{R}^{m\times r} : U^\top U = I_r\}$. Orthonormality removes the nonorthogonal rescaling in (1). At rank r, the remaining equivalent representations are $(U, Y) \sim (UO, YO)$ with $O^\top O = I_r$. Identifying these orthogonal rotations yields a smooth quotient parameter manifold, which is diffeomorphic to the fixed-rank manifold optimized by LORO (Mishra et al., 2014; Luo et al., 2024; Mo et al., 2025). The subspace $\mathrm{col}(U)$ belongs to the Grassmann manifold $\mathrm{Gr}(m, r) = \mathrm{St}(m, r)/\mathrm{O}(r)$, while Y determines the weight within it. The remaining equivalent representations differ only by basis rotations within the same column space, while PLUM restricts the basis update to $U^\top \Delta U = 0$, removing the in-subspace component that only rotates the basis without changing $\mathrm{col}(U)$.

Preserving this structure normally requires nontrivial geometric operations. A Grassmann geodesic with horizontal velocity $D = L\Sigma V^\top$ rotates its principal directions through angles $\theta_j = \eta\sigma_j(D)$ (Edelman et al., 1998; Bendokat et al., 2024). Both the geodesic and parallel transport depend on these singular directions. Since gradients and momenta at successive iterates lie in different tangent spaces, the previous momentum should be transported before being intrinsically combined with the new gradient (Absil et al., 2008; Bendokat et al., 2024). A simple projection is generally not isometric. QR and polar retractions also require matrix factorizations (Absil et al., 2008; Boumal, 2023). Across many layers, these operations can offset the savings of low-rank restriction, and LORO therefore amortizes exact Riemannian updates with cheaper Euclidean steps (Mo et al., 2025).

Our central observation is that Muon's update makes this geometry unusually simple. Muon (Jordan et al., 2024) replaces matrix momentum by its rectangular matrix sign. As a right-side spectral preconditioning, msign preserves the column space and therefore preserves horizontality. If $M_{U,t} = L\Sigma V^\top$ is horizontal, then

$$Q_{U,t} = \mathrm{msign}(M_{U,t}) = LV^\top, \qquad U_t^\top Q_{U,t} = 0, \qquad \sigma_j(Q_{U,t}) = 1 \Longrightarrow \theta_j = \eta_{U,t}$$

for every nonzero singular direction and $0 \le \eta_{U,t} < \pi/2$. All active principal directions therefore rotate through the same angle. When $M_{U,t}$ has full column rank, the basis step and parallel transport of its momentum can be reduced to

$$U_{t+1} = \cos\eta_{U,t} U_t - \sin\eta_{U,t} Q_{U,t}, \quad \mathcal{T}_{\eta_{U,t}}(M_{U,t}) = 1/\cos\eta_{U,t}[M_{U,t} - U_{t+1}(U_{t+1}^\top M_{U,t})].$$

Thus the equal-angle structure simplifies both subspace motion and momentum transport. Section 3 develops these identities and their spectral interpretation, while Appendix A provides the geometric background.

PLUM combines this basis step with a Muon update of Y. The relative factor learning rates maximize predicted first-order decrease per unit of modeled weight change without separately tuned factor multipliers. A fixed number of Newton–Schulz (NS) iterations approximates the matrix sign using general matrix–matrix multiplications (GEMMs). The resulting optimizer is fully GEMM-based, stores one momentum matrix per factor, and requires no additional factor- or geometry-specific hyperparameters to tune.

With finite NS, $Q_{U,t}$ is approximately column-orthonormal. To isolate this error, suppose $U_t^\top U_t = I_r$ and $U_t^\top Q_{U,t} = 0$. The basis update then gives $U_{t+1}^\top U_{t+1} - I_r = \sin^2\eta_{U,t}(Q_{U,t}^\top Q_{U,t} - I_r)$. The NS error therefore perturbs the basis and feeds into subsequent gradient projections and momentum updates. We analyze this coupled recursion for the implemented algorithm. Under the assumptions

in Section 4, fixed finite NS achieves an $O(T^{-1/4})$ expected first-order stationarity rate while the expected maximum orthogonality error is $O(T^{-1/2})$.

Our main contributions are as follows:

- We identify a structural compatibility between Muon and orthonormal low-rank geometry. The msign update preserves horizontal directions and equalizes their nonzero singular values, reducing subspace motion and momentum transport to closed-form GEMM operations. Together with product-aware step allocation, this gives a lightweight optimizer without repeated retractions, nested optimization, or any additional hyperparameter tuning.
- We directly analyze the implemented fixed-NS dynamics and prove an $O(T^{-1/4})$ expected first-order stationarity rate together with an $O(T^{-1/2})$ bound on the expected maximum orthogonality error, showing that the practical NS approximation preserves the convergence order while keeping orthogonality drift controlled.
- We evaluate C4 pretraining and GLUE adaptation with reference models exceeding 1B parameters. PLUM achieves lower pretraining perplexity than prior low-rank methods while retaining compact parameter and optimizer storage, with favorable end-to-end training efficiency from its fully GEMM-based implementation.

2 Related Work

Low-rank pretraining and adaptation. LoRA (Hu et al., 2022) trains low-rank corrections to a frozen backbone, and QLoRA (Dettmers et al., 2023) combines this strategy with weight quantization. AdaLoRA (Zhang et al., 2023) adaptively allocates rank across weight matrices, while DoRA (Liu et al., 2024) decomposes weight updates into magnitude and directional components. Native factorized training instead reduces the parameters retained throughout optimization. LORO (Mo et al., 2025) uses low-rank Riemannian optimization for this setting, SLTrain (Han et al., 2024) adds sparse corrections, and ReLoRA (Lialin et al., 2024) periodically merges factors into a dense weight. GaLore (Zhao et al., 2024) reduces optimizer memory by projecting gradients, and GoLore (He et al., 2025) studies subspace optimization with convergence guarantees. Factor optimization has also motivated LoRA+ (Hayou et al., 2024), LoRA-RITE (Yen et al., 2025), and Riemannian preconditioned LoRA (Zhang & Pilanci, 2024). PLUM targets native low-rank training, jointly addressing compact factor storage and the control of updates to the represented weight.

Matrix optimization and Muon. Muon (Jordan et al., 2024) applies a spectral update to matrix momentum, and scalable implementations have demonstrated its utility for model training (Liu et al., 2025). Polar Express (Amsel et al., 2026) and Gram Newton–Schulz methods (Zhang et al., 2026a) improve matrix-sign computation. Recent theory characterizes Muon through spectral-norm and linear-minimization-oracle viewpoints (Chen et al., 2026; Shulgin et al., 2026b), while (Shulgin et al., 2026a) analyze convergence with inexact matrix-sign updates and (Riabinin et al., 2026) develop a layer-wise LMO analysis closer to practical LLM training. Low-rank extensions include Riemannion (Bogachev et al., 2026), LoRA-Muon (Cesista et al., 2026), and Spectron (Janson et al., 2026), which differ in their representation, preconditioning, and control of the induced weight update. PLUM instead leverages the spectral structure of the matrix sign to make orthonormal subspace updates and momentum transport inexpensive.

Orthogonal parameterizations and manifold optimization. The geometry of orthogonality constraints is developed by Edelman et al. (1998), Absil et al. (2008), and Boumal (2023) and surveyed for Grassmann manifolds by Bendokat et al. (2024). Fixed-rank quotient methods (Mishra et al., 2014; Luo et al., 2024) distinguish a represented matrix from its non-unique factor coordinates. Stiefel-LoRA (Park et al., 2025), PoLAR (Lion et al., 2025), and StelLA (Li et al., 2025) exploit orthogonal structure in low-rank adaptation, while LAVA (Ding et al., 2026) uses column-orthogonal regularization in language and vision model fine-tuning. Landing methods (Ablin & Peyré, 2022; Ablin et al., 2024) avoid repeated retractions through tangential and normal corrections, and Manifold-LoRA (Zhang et al., 2026b) develops a retraction-free Stiefel constraint method for low-rank adaptation. Rather than relying on generic retractions or penalty-based corrections, PLUM exploits the special Grassmann structure induced by the matrix sign and directly analyzes its finite-NS implementation.

3 METHOD

This section presents the parameterization, geometric updates, and factor-step allocation of PLUM.

3.1 LEARNING A SUBSPACE AND ITS COEFFICIENTS

PLUM represents each selected weight by $W = UY^\top$, where $U \in \mathrm{St}(m, r)$ and $Y \in \mathbb{R}^{n\times r}$. This uses $r(m+n)$ parameters and represents every matrix of rank at most r. The columns of U span an r-dimensional subspace containing the columns of W. At rank r these subspaces coincide, and all equivalent orthonormal representations are (UO, YO) for $O \in \mathrm{O}(r)$. Moreover, $W^\top W = YY^\top$, so Y carries the nonzero singular values of the weight. Its norm is therefore the relevant scale when a rotation of U is converted into a weight change.

The subspace $\mathrm{col}(U)$ belongs to $\mathrm{Gr}(m, r) = \mathrm{St}(m, r)/\mathrm{O}(r)$. At U, a Stiefel tangent has the orthogonal decomposition

$$T_U \,\mathrm{St}(m, r) = \mathcal{V}_U \oplus \mathcal{H}_U, \quad Z_U = \underbrace{U\Omega}_{\mathcal{V}_U} + \underbrace{(I - UU^\top)Z_U}_{\mathcal{H}_U}, \quad \Omega = U^\top Z_U \in \mathfrak{o}(r).$$

Here $\mathfrak{o}(r) = \{\Omega : \Omega^\top = -\Omega\}$ is the Lie algebra of $\mathrm{O}(r)$. The curve $U\exp(t\Omega)$ rotates the orthonormal frame within the same subspace, so the vertical space $\mathcal{V}_U = \{U\Omega\}$ consists of changes of frame. Its complement $\mathcal{H}_U = \{Z_U : U^\top Z_U = 0\}$, the horizontal space, represents changes of subspace and identifies the tangent space of the quotient (Absil et al., 2008; Boumal, 2023). PLUM uses this horizontal representation for the basis update.

Let $\widehat{G}_t$ denote the stochastic gradient with respect to the represented weight $W_t = U_t Y_t^\top$. Backpropagation gives the factor gradients $G_{U,t} = \widehat{G}_t Y_t$ and $G_{Y,t} = \widehat{G}_t^\top U_t$. The horizontal basis gradient is $G_{U,t} - U_t(U_t^\top G_{U,t})$, while the coefficient gradient is used directly. PLUM operates only on these low-rank factor gradients, so the ambient stochastic gradient matrix $\widehat{G}_t$ need not be explicitly formed or stored by the optimizer.

3.2 MATRIX-SIGN UPDATES SIMPLIFY GRASSMANN GEODESICS AND PARALLEL TRANSPORT

For a compact SVD $M = L\Sigma V^\top$, define $\mathrm{msign}(M) = LV^\top$ and $\mathrm{msign}(0) = 0$. A horizontal momentum has the spectral steepest-descent solution

$$\max_{U^\top D=0,\, \|D\|_2\le 1} \langle M, D\rangle_F = \|M\|_*, \qquad D = \mathrm{msign}(M). \tag{2}$$

The spectral ball has support function $\|M\|_*$, and the left singular vectors associated with nonzero singular values of M are orthogonal to U. Equivalently, $\mathrm{msign}(M) = M[(M^\top M)^\dagger]^{1/2}$ is a right preconditioning. Both this map and a finite NS polynomial preserve the column space of M. Thus a horizontal input remains horizontal, and no further projection is needed in exact arithmetic (Higham, 1986; Jordan et al., 2024).

For orthonormal bases $U, \widetilde{U}$, the principal angles satisfy $\cos\theta_j = \sigma_j(U^\top\widetilde{U})$, with $\theta_j \in [0, \pi/2]$. If $D = L\Sigma V^\top$ is horizontal, the Grassmann geodesic with initial velocity $-D$ and the parallel transport of a horizontal matrix Z along it are (Edelman et al., 1998; Bendokat et al., 2024)

$$\begin{cases} U(\eta) = U + UV[\cos(\eta\Sigma) - I]V^\top - L\sin(\eta\Sigma)V^\top, & (3)\\ \mathcal{T}_\eta(Z) = Z + [L(\cos(\eta\Sigma) - I) + UV\sin(\eta\Sigma)]L^\top Z. & (4)\end{cases}$$

Parallel transport preserves inner products while moving a tangent vector to the tangent space at the new point, allowing old momentum to be combined intrinsically with a new gradient. Both formulas depend on the singular vectors and on the separate angles $\eta\sigma_j(D)$. On Grassmann, QR and polar retractions recover the same subspace and agree with the geodesic through second order, but still require a per-step factorization (Absil et al., 2008; Boumal, 2023). Such decomposition-based operations are less amenable than GEMMs to efficient batching over many thin layerwise factors on GPUs.

For full-column-rank M_U, take $Q_U = \mathrm{msign}(M_U)$. Substitution of its unit singular values into (3) gives $U(\eta) = \cos\eta\, U - \sin\eta\, Q_U$, with $U^\top U(\eta) = \cos\eta\, I_r$. Hence every principal

angle equals η for $0 \le \eta < \pi/2$. In (4), the identity $Q_U Q_U^\top M_U = M_U$ yields $\mathcal{T}_\eta(M_U) = \cos\eta\, M_U + \sin\eta\, U Q_U^\top M_U$. Since $U(\eta)^\top M_U = -\sin\eta\, Q_U^\top M_U$, direct multiplication gives

$$[I - U(\eta)U(\eta)^\top]M_U = \cos\eta\,[\cos\eta\, M_U + \sin\eta\, U Q_U^\top M_U].$$

The basis update and the transport of its momentum thus simplify to

$$U_{t+1} = \cos\eta_{U,t} U_t - \sin\eta_{U,t} Q_{U,t}, \qquad \mathcal{T}_{\eta_{U,t}}(M_{U,t}) = 1/\cos\eta_{U,t}[M_{U,t} - U_{t+1}(U_{t+1}^\top M_{U,t})]. \quad (5)$$

At the next iteration, the transport correction and the new-gradient projection can therefore be combined into the same projection, yielding the simple matrix operations summarized in Algorithm 1. When M_U has smaller rank, the exact geodesic rotates only its nonzero singular directions, as derived in Appendix A.2.

3.3 Balancing the two factor updates

The two factors act on the weight at different scales. Their first-order contribution is $-\eta_U Q_U Y^\top - \eta_Y U Q_Y^\top$. At an orthonormal basis and a horizontal Q_U,

$$\|\eta_U Q_U Y^\top + \eta_Y U Q_Y^\top\|_F^2 = \eta_U^2 \|Q_U Y^\top\|_F^2 + \eta_Y^2 \|U Q_Y^\top\|_F^2. \quad (6)$$

For column-orthonormal updates, the two coefficients are $\|Y\|_F^2$ and $\|U\|_F^2$. For bounded NS updates, these quantities remain valid scale proxies up to a fixed factor. We therefore select the relative learning rates to maximize

$$[(\eta_U \langle M_U, Q_U\rangle_F + \eta_Y \langle M_Y, Q_Y\rangle_F)^2] \;/\; [\eta_U^2(\|Y\|_F^2 + \varepsilon) + \eta_Y^2(\|U\|_F^2 + \varepsilon)].$$

The numerator is the square of the momentum-predicted first-order decrease, while the denominator estimates the squared first-order change of the represented weight. The constant $\varepsilon > 0$ keeps the ratio well defined at $Y = 0$. Cauchy–Schwarz gives the optimal ratio $\eta_U : \eta_Y = \rho_U : \rho_Y$. We preserve the scheduled learning rate η_t as the maximum factor step by setting

$$\rho_U = \frac{\langle M_U, Q_U\rangle_F}{\|Y\|_F^2 + \varepsilon}, \;\; \rho_Y = \frac{\langle M_Y, Q_Y\rangle_F}{\|U\|_F^2 + \varepsilon}, \;\; \gamma_{i,t} = \frac{\rho_{i,t}}{\max_{j\in\{U,Y\}} \rho_{j,t}}, \;\; \eta_{i,t} = \eta_t \gamma_{i,t}, \;\; i \in \{U, Y\}. \quad (7)$$

With the factor steps determined by (7), the remaining operations are direct. For a matrix M, NS starts from $X_0 = M/(\|M\|_F + \nu)$ and applies $X_{j+1} = X_j[a_j I + b_j X_j^\top X_j + c_j (X_j^\top X_j)^2]$. The coefficients are fixed in advance, with the schedules and bounds used here given in Appendix E.

Algorithm 1 summarizes one PLUM iteration. We initialize $U_0^\top U_0 = I_r$, $Y_0 = 0$, $M_U = M_Y = 0$, and $\eta_{U,-1} = 0$, where the basis U_0 is randomly initialized and orthonormalized once before training. Its inputs are the factor gradients G_U, G_Y, scheduled learning rate η_t, momentum β, and constants $\varepsilon, \nu > 0$. The stored angle $\eta_{U,t-1}$ provides the transport factor in (5). Optional decay is applied only to Y, since scaling Y scales the represented product $UY^\top$ by the same factor without directly rescaling the basis.

Algorithm 1 PLUM: one iteration

$M_U \leftarrow \beta M_U / \cos\eta_{U,t-1} + (1-\beta) G_U$
$M_U \leftarrow M_U - U(U^\top M_U)$
$M_Y \leftarrow \beta M_Y + (1-\beta) G_Y$
$Q_U \leftarrow \mathrm{NS}_K(M_U; \nu)$
$Q_Y \leftarrow \mathrm{NS}_K(M_Y; \nu)$
$(\eta_U, \eta_Y) \leftarrow$ (7)
$U \leftarrow \cos\eta_U\, U - \sin\eta_U\, Q_U$
$Y \leftarrow (1 - \lambda\eta_Y) Y - \eta_Y Q_Y$

PLUM requires neither repeated re-orthogonalization nor an orthogonality penalty. Its matrix computations consist primarily of thin GEMMs, supplemented only by norm and inner-product reductions and scalar operations, thereby avoiding per-step QR, SVD, or matrix inversion. Each step costs $O(K(m+n)r^2 + Kr^3)$ arithmetic operations together with $O(mr^2)$ projection work. Same-shaped factor momenta from different layers can be stacked for batched GEMMs, as described in Section 5. The optimizer stores one momentum matrix per factor and requires no separately tuned factor learning rates or geometry-specific hyperparameters. In the next section, we show that despite avoiding repeated re-orthogonalization, the feasibility error induced by the fixed-NS updates remains controlled.

4 Theoretical Analysis

We analyze one factorized matrix block $\min_{U\in\mathrm{St}(m,r),\, Y\in\mathbb{R}^{n\times r}} \mathcal{L}(U, Y)$ with $\mathcal{L}(U, Y) = f(UY^\top)$. Both factor gradients are induced by the same mini-batch gradient of the represented weight, following the stochastic matrix-optimization setting considered in recent Muon and LMO analyses (Shulgin et al., 2026a; Riabinin et al., 2026). We will also use the standard assumptions below.

Assumption 1 (Lower boundedness and L-smoothness)**.** *The objective $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ satisfies $f \geq f_* > -\infty$ and $\|\nabla f(W') - \nabla f(W)\|_F \leq L\|W' - W\|_F$ for all W, W'.*

Assumption 2 (Unbiased gradients and bounded variance)**.** *Let $\mathcal{F}_t$ denote the optimization history before iteration t and $G_t = \nabla f(U_t Y_t^\top)$. For a mini-batch of size $B \geq 1$, $\mathbb{E}[\widehat{G}_t \mid \mathcal{F}_t] = G_t$ and $\mathbb{E}[\|\widehat{G}_t - G_t\|_F^2 \mid \mathcal{F}_t] \leq \sigma^2/B$.*

For $G = \nabla f(UY^\top)$, the factor derivatives are $\nabla_U \mathcal{L} = GY$ and $\nabla_Y \mathcal{L} = G^\top U$. At feasibility, the constrained first-order conditions are $GY = U\Lambda$ for symmetric Λ and $G^\top U = 0$, which give $\Lambda = U^\top GY = (G^\top U)^\top Y = 0$. Hence vanishing factor derivatives characterize constrained stationarity.

For a fixed NS sequence, let $p_j(x) = x(a_j + b_j x^2 + c_j x^4)$ and $\phi_K = p_{K-1} \circ \cdots \circ p_0$. We use $\alpha x \leq \phi_K(x) \leq \kappa x$ and $0 \leq \phi_K(x) \leq b$ for $x \in [0, 1]$, where $0 < \alpha \leq 1$ and $b, \kappa \geq 1$. Appendix E lists the NS schedules and the corresponding numerical values of α, b, κ.

We conduct the analysis under the boundedness condition $\|Y_t\|_F \leq R$ almost surely, where R is independent of T. At feasibility, $\|UY^\top\|_F = \|Y\|_F$, so this condition controls the scale of the represented weight rather than an arbitrary factor rescaling. Related boundedness assumptions have also appeared in analyses of low-rank factor optimization (Yen et al., 2025; Ghiasvand et al., 2025). Such scale control can also be induced by regularization: ℓ_2 regularization penalizes coefficient growth, while decoupled weight decay yields an explicit deterministic radius under mild step-size conditions. Further details are provided in Appendix B.3.

We now analyze the $\lambda = 0$ scenario and establish the stationarity and feasibility guarantees for PLUM with a fixed number of NS iterations.

Theorem 1 (Fixed finite NS)**.** *Suppose Assumptions 1–2 hold and initialize $U_0^\top U_0 = I_r$, $Y_0 = 0$, and both momenta at zero. Let $\Delta_b = \sqrt{r}\max\{1, b^2-1\}$ and $\Gamma_R = 1+\|G_0\|_F+\sqrt{3/2}LR+\sigma/\sqrt{B}$. Choose*

$$0 < \eta \leq (8b)^{-1}, \quad 2\eta^2 \leq 1-\beta, \quad 2\Delta_b T\eta^2 \leq \tfrac{1}{4}, \quad \nu \geq 16\kappa R\Gamma_R T\eta[4b\eta + \sqrt{3/2}(1-\beta)].$$

Define

$$\mathcal{B}_T = \frac{f(0) - f_*}{\eta T} + \frac{\beta\|G_0\|_F}{(1-\beta)T} + \frac{\eta}{1-\beta} + \sigma\sqrt{\frac{1-\beta}{B}} + T\eta^2 + \frac{1}{T}.$$

There exists a constant C, depending only on $L, R, r, \varepsilon, \alpha, b, \kappa, \|G_0\|_F$, such that

$$\begin{cases} \dfrac{1}{T}\displaystyle\sum_{t<T} \mathbb{E}[\|\nabla_U \mathcal{L}(U_t, Y_t)\|_F + \|\nabla_Y \mathcal{L}(U_t, Y_t)\|_F] \leq C\left[\mathcal{B}_T + \nu + \dfrac{(1+T\eta)^2}{T}\right], & (8) \\ \mathbb{E}\Big[\displaystyle\max_{0\leq t\leq T} \|U_t^\top U_t - I_r\|_F\Big] \leq C\left(T\eta^2 + \dfrac{1}{T}\right). & (9) \end{cases}$$

Moreover, the maximum orthogonality error admits the following high-probability bound

$$\mathbb{P}\left(\max_{0\leq t\leq T} \|U_t^\top U_t - I_r\|_F \leq 2\Delta_b T\eta^2\right) \geq 1 - \frac{\sigma^2}{B\Gamma_R^2 T}. \tag{10}$$

The proof and explicit constants can be found in Appendix C. With $\eta = \eta_0 T^{-3/4}$, $1-\beta = \beta_0 T^{-1/2}$, and $\nu = \nu_0 T^{-1/4}$ as specified there, (8) gives an $O(T^{-1/4})$ expected stationarity rate and (9) gives an $O(T^{-1/2})$ expected maximum orthogonality error. Equation (10) further bounds the maximum error along the entire trajectory by $O(T^{-1/2})$ with failure probability $O(T^{-1})$.

Finite NS produces slightly infeasible factors, whereas the stationarity conditions above are naturally defined on the constraint manifold. For full-column-rank U_t, set $\bar{U}_t = U_t(U_t^\top U_t)^{-1/2}$ and $\bar{Y}_t = Y_t(U_t^\top U_t)^{1/2}$. This restores $\bar{U}_t^\top \bar{U}_t = I_r$ while preserving $\bar{U}_t \bar{Y}_t^\top = U_t Y_t^\top$.

Corollary 1 (Constrained stationarity)**.** *Under Theorem 1 and the schedules above,*

$$\frac{1}{T}\sum_{t<T} \mathbb{E}\big[\|\operatorname{grad}_{\mathrm{St},U} \mathcal{L}(\bar{U}_t, \bar{Y}_t)\|_F + \|\nabla_Y \mathcal{L}(\bar{U}_t, \bar{Y}_t)\|_F\big] = O(T^{-1/4}).$$

Thus the $O(T^{-1/4})$ rate also holds for the standard constrained first-order measure at an exactly feasible representation of each computed weight. Appendix B.2 covers rank-deficient U_t, including the zero-coefficient initialization, and relates this criterion to the embedded fixed-rank gradient when Y has full column rank.

The exact matrix sign gives the idealized counterpart in which the basis remains feasible throughout and no NS approximation error is present.

Theorem 2 (Exact matrix sign)**.** *Under Assumptions 1–2 and the same initialization, let* $0 < \eta \le 1$, $0 \le \beta < 1$, *and define*

$$v_0 = \beta^2 \|G_0\|_F^2 + (1-\beta)^2 \frac{\sigma^2}{B}, \qquad \mathcal{B}_T^{\mathrm{ex}} = \frac{f(0) - f_*}{\eta T} + \frac{\sqrt{v_0}}{(1-\beta)T} + \frac{\eta}{1-\beta} + \sigma\sqrt{\frac{1-\beta}{B}}.$$

Then $U_t^\top U_t = I_r$ *for every* t, *and there exists a constant* C_{ex}, *independent of* $T, \eta, \beta, B, \sigma$, *such that*

$$\frac{1}{T}\sum_{t<T} \mathbb{E}[\|\nabla_U \mathcal{L}(U_t, Y_t)\|_F + \|\nabla_Y \mathcal{L}(U_t, Y_t)\|_F] \le C_{\mathrm{ex}} \mathcal{B}_T^{\mathrm{ex}}.$$

The proof and explicit constants are given in Appendix D. For $\eta = \eta_0 T^{-3/4}$ and $1-\beta = \beta_0 T^{-1/2}$, $\mathcal{B}_T^{\mathrm{ex}} = O(T^{-1/4})$, giving an $O(T^{-1/4})$ stationarity rate with exact orthogonality. Apart from the normalization residual ν, a fixed finite NS polynomial affects the leading stationarity bound only through T-independent constants. The induced feasibility terms scale as $T\eta^2$ and $(1+T\eta)^2/T$, both of which are $O(T^{-1/2})$ under $\eta = O(T^{-3/4})$, and are therefore higher order than the $O(T^{-1/4})$ stationarity term. Thus, PLUM with fixed finite NS preserves the $O(T^{-1/4})$ stationarity rate while its feasibility error vanishes at the faster $O(T^{-1/2})$ rate.

5 C4 Efficient Pretraining

Experiment setup. In this section, we evaluate the effectiveness of the proposed PLUM in training low-rank LLaMA-like models (including RoPE (Su et al., 2024), RMS Norm (Zhang & Sennrich, 2019), and SwiGLU (Shazeer, 2020) FFN) from scratch using the C4 corpus(Touvron et al., 2023). The reference sizes are 60M, 130M, 350M, and 1B, which share the same setting as GaLore(Zhao et al., 2024). For fair comparisons, all the models are trained in BF16 precision format. Detailed hyperparameter settings and search grids for all methods are reported in Appendix F.

Baselines. To compare PLUM against existing low-rank training methods, we consider GaLore (Zhao et al., 2024), SLTrain (Han et al., 2024), LORO (Mo et al., 2025), and Spectron (Janson et al., 2026). In our experiments, the first three methods use AdamW-based optimization, whereas Spectron uses a Muon-based optimizer with spectral renormalization to stabilize low-rank training. Overall, these baselines provide a representative comparison across several existing approaches to efficient low-rank training.

Table 1: C4 pretraining comparison at four dense reference sizes. PPL is validation perplexity, Param is the stored model parameter count in millions, and Mem is parameter-plus-optimizer storage in GB. Full-size denotes dense training with AdamW. The best result is marked in bold.

Model (Tokens)	LLaMA-60M (1.1B)			LLaMA-130M (2.2B)			LLaMA-350M (6.4B)			LLaMA-1B (13.1B)		
r / hidden / FFN	128 / 512 / 1376			256 / 768 / 2048			256 / 1024 / 2736			512 / 2048 / 5461		
Method	PPL ↓	Param	Mem	PPL ↓	Param	Mem	PPL ↓	Param	Mem	PPL ↓	Param	Mem
Full-size	**30.70**	58	0.35	24.37	134	0.80	18.76	367	2.21	16.52	1339	8.03
GaLore	34.90	58	0.28	25.37	134	0.61	19.38	367	1.39	15.57	1339	4.76
SLTrain	34.24	43	0.26	26.04	96	0.58	20.49	194	1.17	15.40	645	3.87
LORO	33.97	42	0.26	24.25	94	0.56	18.79	185	1.11	15.19	609	3.66
Spectron	33.00	42	0.24	<u>24.06</u>	94	0.47	<u>18.38</u>	185	0.87	<u>13.98</u>	609	2.70
PLUM (ours)	<u>32.76</u>	42	**0.24**	**23.93**	94	**0.47**	**18.17**	185	**0.87**	**13.75**	609	**2.70**

Quantitative results. Table 1 compares validation PPL, stored parameters, and parameter-plus-optimizer memory across four model scales. Among the memory-efficient methods, PLUM achieves the lowest PPL at all four scales. At 130M, 350M, and 1B, PLUM also outperforms full-size AdamW training by 0.44, 0.59, and 2.77 PPL, respectively. At the 1B scale, this is achieved with approximately 45% of the stored parameters and one-third of the parameter-plus-optimizer memory of the full-size model. The PPL gap relative to full-size training increases across these three larger scales. PLUM also consistently achieves lower PPL than Spectron across all four scales, while offering better training efficiency. As shown in Table 3, PLUM requires less time per step and achieves higher throughput than Spectron under the same setting.

Empirical dynamics of Y. Figure 1 tracks the mean RMS norm of the coefficient factor Y during C4 pretraining at four model scales. Given a matrix $Y \in \mathbb{R}^{n\times r}$, the metric RMS is defined as $\mathrm{RMS}(Y_t) = \sqrt{\frac{1}{nr}\sum_{i=1}^{n}\sum_{j=1}^{r}(Y_t)_{ij}^2} = \|Y_t\|_F/\sqrt{nr}$. As shown, $\mathrm{RMS}(Y_t)$ rises rapidly from its zero initialization but levels off during the later stages of training at every scale. This change in coefficient scale matters for optimization: for a column-orthonormal basis direction Q_U, a basis step induces a first-order weight change of magnitude $\|\eta_U Q_U Y^\top\|_F = \eta_U\|Y\|_F$. Thus, the same basis step size can produce different weight changes as Y evolves, motivating the product-aware step allocation in (7). The observed plateaus are consistent with the bounded scenario in Section 4.

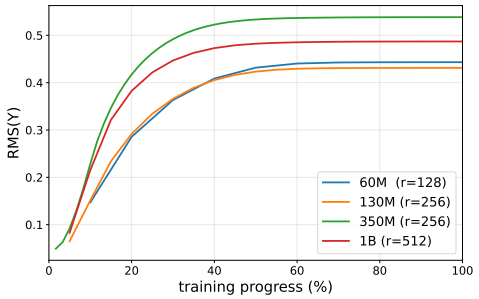

Figure 1: Visualization of mean $\mathrm{RMS}(Y_t)$.

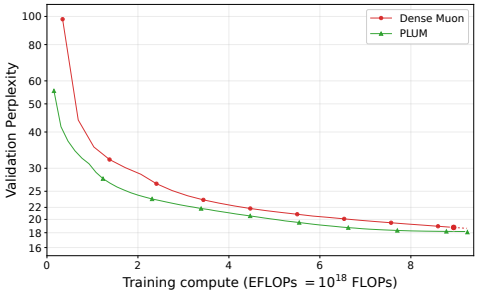

Figure 2: Comparisons with dense Muon under same FLOPs setting.

Additional comparison against dense Muon. We further compare PLUM with dense Muon on the LLaMA-350M reference setting. Since low-rank and full-rank training incur different costs per update, we plot validation PPL against cumulative training FLOPs rather than training steps. This provides a compute-normalized comparison while keeping the optimizer family aligned. As shown in Figure 2, PLUM achieves lower PPL throughout the overlapping compute range and reaches a given PPL with fewer FLOPs. Overall, these results further demonstrate the effectiveness of PLUM under a matched compute budget.

Ablation on the selection of rank r. Table 2 shows a clear trade-off between rank and model size. As expected, increasing r from 32 to 256 reduces PPL from 42.98 to 29.85, while increasing the parameter count from 35.28M to 52.77M. At $r = 256$, PLUM outperforms the full-size AdamW baseline in PPL (29.85 vs. 30.70) while using fewer parameters (52.77M vs. 58.07M).

The effect of allocation (7) in PLUM. Figure 3 compares PLUM with and without learning rate allocation (7) during LLaMA-60M pretraining. Enabling scaling yields consistently lower validation loss throughout training and retains advantage at the final step. This supports our theoretical analysis that updates to U and Y affect the weight $W = UY^\top$ at different scales and assign them the same numerical learning rate fail to balance their contributions to the weight update.

Implementation and efficiency The basis is initialized by a reduced QR factorization. The forward map is $(XY)U^\top$, and backpropagation provides both factor gradients. We stack same-shaped factor momenta from different layers. A group of J basis momenta forms a $J \times m \times r$ tensor, with momentum grouped separately. Norms and inner products reduce only the final two axes, and batched GEMMs apply NS independently to each matrix. Table 3 reports the training efficiency on LLaMA-350M. Further implementation and measurement details are provided in Appendix F.

Table 2: Rank ablation for LLaMA-60M.

Method	Rank	Eval Loss↓	PPL↓	Param (M)
Full-size	512	3.42	30.70	58.07
PLUM	256	3.40	29.85	52.77
PLUM	128	3.49	32.76	42.77
PLUM	64	3.60	36.76	37.77
PLUM	32	3.76	42.98	35.28

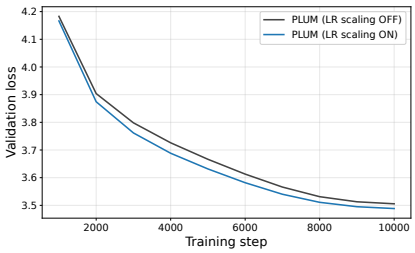

Figure 3: Effect of learning-rate scaling.

Table 3: Training efficiency comparison on LLaMA-350M.

Method	s / step ↓	Tokens / sec ↑	Total Time (h)	SpeedUp vs. Muon
Muon	0.81	161746	13.5	1.00×
Spectron	0.83	158093	13.8	0.98×
PLUM (ours)	0.62	211109	10.3	**1.31**×

6 FINETUNING ON DATASET GLUE

Experiment setup. Following the experiment setup in GaLore Zhao et al. (2024), we extend PLUM to finetune RoBERTa-base model (Liu et al., 2019) on downstreaming dataset GLUE (Wang et al., 2018). All the details of the experiments, including the descriptions of the dataset and the search grid of hyperparameters, are reported in Appendix F. For baselines, we choose LoRA Hu et al. (2022), and efficient pretraining methods including GaLore and Spectron Janson et al. (2026).

Quantitative results. Table 4 reports the finetuning performances on five GLUE tasks. At rank 4, PLUM achieves the highest score on CoLA, MRPC, and RTE, and ties Spectron on SST-2. At rank 8, PLUM obtains the highest score on all five reported tasks, exceeding the full fine-tuning AdamW baseline while training only 0.29M parameters. The comparison with Spectron is particularly relevant because both methods build on Muon: PLUM improves RTE by 2.88 points at rank 4 and 2.52 points at rank 8. These results suggest that PLUM's gains extend beyond pretraining to finetuning.

Table 4: RoBERTa-base adaptation on GLUE.

Method	Params	CoLA	STS-B	MRPC	RTE	SST-2	Avg.
Full-size	-	62.24	90.92	91.30	79.42	94.57	85.15
LoRA ($r = 4$)	0.15M	61.38	90.57	91.07	78.70	92.89	82.92
GaLore ($r = 4$)	-	60.35	90.73	92.25	79.42	94.04	83.34
Spectron ($r = 4$)	0.15M	62.12	**91.03**	93.06	76.90	94.72	83.57
PLUM ($r = 4$, ours)	0.15M	**62.99**	90.78	**93.26**	**79.78**	**94.72**	**85.80**
LoRA ($r = 8$)	0.29M	61.83	90.80	91.90	79.06	93.46	84.88
GaLore ($r = 8$)	-	60.06	90.82	92.01	79.78	94.38	83.34
Spectron ($r = 8$)	0.29M	62.61	90.83	91.99	77.62	94.38	83.49
PLUM ($r = 8$, ours)	0.29M	**63.24**	**91.18**	**93.54**	**80.14**	**94.61**	**85.94**

7 CONCLUSION

We studied native low-rank training from the perspective of optimizing the represented weight rather than its factors independently. PLUM shows that an orthonormal low-rank geometry need not incur the usual cost of manifold optimization: the matrix-sign structure of Muon simplifies the Grassmann geodesic and parallel transport, while product-aware learning rate allocation balances the two factor updates. This leads to a simple GEMM-only optimizer without repeated matrix factorizations or additional hyperparameter tuning. Our analysis further shows that fixed finite NS preserves the stationarity order while the orthogonality error remains controlled, and the experiments demonstrate effective low-rank pretraining and finetuning at scale.

AI USE STATEMENT

Generative AI tools were used to assist with language editing and to improve the clarity and presentation of the manuscript. All AI-assisted edits were reviewed by the authors.

REFERENCES

Pierre Ablin and Gabriel Peyré. Fast and Accurate Optimization on the Orthogonal Manifold without Retraction. In *International Conference on Artificial Intelligence and Statistics*, volume 151 of *Proceedings of Machine Learning Research*, pp. 5636–5657, 2022. URL https://proceedings.mlr.press/v151/ablin22a.html.

Pierre Ablin, Simon Vary, Bin Gao, and Pierre-Antoine Absil. Infeasible Deterministic, Stochastic, and Variance-Reduction Algorithms for Optimization under Orthogonality Constraints. *Journal of Machine Learning Research*, 25(389):1–38, 2024. URL https://jmlr.org/papers/v25/23-0451.html.

P.-A. Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization Algorithms on Matrix Manifolds*. Princeton University Press, 2008. doi: 10.1515/9781400830244.

Noah Amsel, David Persson, Christopher Musco, and Robert M. Gower. The polar express: Optimal matrix sign methods and their application to the muon algorithm. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/df88b275bef31ac96c85f0c4013734fc-Abstract-Conference.html.

Thomas Bendokat, Ralf Zimmermann, and P.-A. Absil. A Grassmann manifold handbook: Basic geometry and computational aspects. *Advances in Computational Mathematics*, 50(1):6, 2024. doi: 10.1007/s10444-023-10090-8.

Vladimir Bogachev, Vladimir Aletov, Alexander Molozhavenko, Denis Bobkov, Vera Soboleva, Aibek Alanov, and Maxim Rakhuba. Lora meets riemannion: Muon optimizer for parametrization-independent low-rank adapters. In *International Conference on Learning Representations*, 2026. URL https://proceedings.iclr.cc/paper_files/paper/2026/hash/8505a7e1da40320bf5692d7edec5aa45-Abstract-Conference.html.

Nicolas Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023. doi: 10.1017/9781009166164. URL https://www.nicolasboumal.net/book/.

Franz Louis Cesista, Katherine Crowson, Cédric Simal, and Stella Biderman. LoRA-Muon: Spectral Steepest Descent on the Low-Rank Manifold, 2026. URL https://arxiv.org/abs/2606.12921.

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *Transactions on Machine Learning Research*, 2026. URL https://openreview.net/forum?id=Blz4hjxLwU.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. QLoRA: Efficient Finetuning of Quantized LLMs. In *Advances in Neural Information Processing Systems*, 2023. URL https://arxiv.org/abs/2305.14314.

Daorui Ding, Fanhua Shang, Tiancan Feng, Junkang Liu, and Hongying Liu. LAVA: A unified framework for finetuning language and vision models. In *International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=xBKbXgaPwr.

Rick Durrett. *Probability: Theory and Examples*. Cambridge University Press, 5 edition, 2019. doi: 10.1017/9781108591034. URL https://sites.math.duke.edu/~rtd/PTE/PTE5_011119.pdf.

Alan Edelman, Tomás A. Arias, and Steven T. Smith. The Geometry of Algorithms with Orthogonality Constraints. *SIAM Journal on Matrix Analysis and Applications*, 20(2):303–353, 1998. doi: 10.1137/S0895479895290954.

Sajjad Ghiasvand, Mahnoosh Alizadeh, and Ramtin Pedarsani. Decentralized low-rank fine-tuning of large language models. In *Proceedings of the 1st Workshop for Research on Agent Language Models (REALM 2025)*, pp. 334–345. Association for Computational Linguistics, 2025. doi: 10.18653/v1/2025.realm-1.24. URL https://aclanthology.org/2025.realm-1.24/.

Xavier Glorot and Yoshua Bengio. Understanding the Difficulty of Training Deep Feedforward Neural Networks. In *International Conference on Artificial Intelligence and Statistics*, pp. 249–256, 2010. URL https://proceedings.mlr.press/v9/glorot10a.html.

Andi Han, Jiaxiang Li, Wei Huang, Mingyi Hong, Akiko Takeda, Pratik Jawanpuria, and Bamdev Mishra. SLTrain: A sparse plus low-rank approach for parameter and memory efficient pretraining. In *Advances in Neural Information Processing Systems*, volume 37, 2024. doi: 10.52202/079017-3755.

Soufiane Hayou, Nikhil Ghosh, and Bin Yu. LoRA+: Efficient low rank adaptation of large models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 17783–17806. PMLR, 2024.

Yutong He, Pengrui Li, Yipeng Hu, Chuyan Chen, and Kun Yuan. Subspace optimization for large language models with convergence guarantees. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 22468–22522, 2025. URL https://proceedings.mlr.press/v267/he25i.html.

Nicholas J. Higham. Computing the Polar Decomposition—with Applications. *SIAM Journal on Scientific and Statistical Computing*, 7(4):1160–1174, 1986. URL https://eprints.maths.manchester.ac.uk/694/1/high86p.pdf.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022.

Paul Janson, Edouard Oyallon, and Eugene Belilovsky. Stabilizing native low-rank llm pretraining. In *International Conference on Machine Learning*, 2026.

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/.

Zhizhong Li, Sina Sajadmanesh, Jingtao Li, and Lingjuan Lyu. StelLA: Subspace learning in low-rank adaptation using stiefel manifold. In *Advances in Neural Information Processing Systems*, volume 38, pp. 83603–83629, 2025. doi: 10.52202/085713-2521.

Vladislav Lialin, Sherin Muckatira, Namrata Shivagunde, and Anna Rumshisky. Relora: High-rank training through low-rank updates. In *International Conference on Learning Representations*, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/hash/d6fdc308a3b0f017f16059e5888f3346-Abstract-Conference.html.

Kai Lion, Liang Zhang, Bingcong Li, and Niao He. Polar: Polar-decomposed low-rank adapter representation. In *Advances in Neural Information Processing Systems*, volume 38, 2025. doi: 10.52202/085713-4996. URL https://papers.neurips.cc/paper_files/paper/2025/hash/db79b6c50430a52fbab9d63efb8433ca-Abstract-Conference.html.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is Scalable for LLM Training, 2025. URL https://arxiv.org/abs/2502.16982.

Shih-Yang Liu, Chien-Yi Wang, Hongxu Yin, Pavlo Molchanov, Yu-Chiang Frank Wang, Kwang-Ting Cheng, and Min-Hung Chen. DoRA: Weight-decomposed low-rank adaptation. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32100–32121. PMLR, 2024. URL https://proceedings.mlr.press/v235/liu24bn.html.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*, 2019.

Yuetian Luo, Xudong Li, and Anru R. Zhang. On geometric connections of embedded and quotient geometries in riemannian fixed-rank matrix optimization. *Mathematics of Operations Research*, 49(2):782–825, 2024. doi: 10.1287/moor.2023.1377. URL https://doi.org/10.1287/moor.2023.1377.

Bamdev Mishra, Gilles Meyer, Silvère Bonnabel, and Rodolphe Sepulchre. Fixed-rank matrix factorizations and riemannian low-rank optimization. *Computational Statistics*, 29(3–4):591–621, 2014. doi: 10.1007/s00180-013-0464-z.

Zhanfeng Mo, Long kai Huang, and Sinno Jialin Pan. Parameter and Memory Efficient Pretraining via Low-Rank Riemannian Optimization. In *International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=i0zzO7Hslk.

JuneYoung Park, Minjae Kang, Seongbae Lee, Haegang Lee, Seongwan Kim, and Jaeho Lee. Riemannian optimization for LoRA on the Stiefel manifold. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 20971–20985, 2025. doi: 10.18653/v1/2025.findings-emnlp.1143. URL https://aclanthology.org/2025.findings-emnlp.1143/.

Artem Riabinin, Egor Shulgin, Kaja Gruntkowska, and Peter Richtárik. From muon to gluon: Bridging theory and practice of LMO-based optimizers for LLMs. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=IelAHU5MVz.

Noam Shazeer. Glu variants improve transformer. *arXiv preprint arXiv:2002.05202*, 2020.

Egor Shulgin, Sultan AlRashed, Peter Richtárik, and Francesco Orabona. Beyond the ideal: Analyzing the inexact muon update. In *Proceedings of the 29th International Conference on Artificial Intelligence and Statistics*, volume 300 of *Proceedings of Machine Learning Research*, pp. 5167–5175. PMLR, 2026a. URL https://proceedings.mlr.press/v300/shulgin26a.html.

Egor Shulgin, Mohamed Awad, Peter Richtárik, and Eduard Gorbunov. General analysis of LMO-based optimizers: Beyond bounded variance. In *Proceedings of the 43rd International Conference on Machine Learning*, 2026b.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. Roformer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568:127063, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models, 2023. URL https://arxiv.org/abs/2302.13971.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Tal Linzen, Grzegorz Chrupała, and Afra Alishahi (eds.), *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-5446. URL https://aclanthology.org/W18-5446/.

Jui-Nan Yen, Si Si, Zhao Meng, Felix Yu, Sai Surya Duvvuri, Inderjit S. Dhillon, Cho-Jui Hsieh, and Sanjiv Kumar. Lora done rite: Robust invariant transformation equilibration for lora optimization. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/bcbc0f660d2dde42f9d1d0ecb14a6f9a-Abstract-Conference.html.

Biao Zhang and Rico Sennrich. Root mean square layer normalization. *Advances in neural information processing systems*, 32, 2019.

Fangzhao Zhang and Mert Pilanci. Riemannian preconditioned LoRA for fine-tuning foundation models. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 59641–59669. PMLR, 2024.

Jack Zhang, Noah Amsel, Berlin Chen, and Tri Dao. Gram Newton–Schulz: A fast, hardware-aware Newton–Schulz algorithm for Muon. Author technical report, 2026a. URL https://dao-lab.ai/blog/2026/gram-newton-schulz/.

Qingru Zhang, Minshuo Chen, Alexander Bukharin, Pengcheng He, Yu Cheng, Weizhu Chen, and Tuo Zhao. Adaptive budget allocation for parameter-efficient fine-tuning. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=lq62uWRJjiY.

Yuan Zhang, Jiang Hu, Zhijian Lai, Lin Lin, and Zaiwen Wen. Retraction-free optimization over the Stiefel manifold for the LoRA fine-tuning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2026b. doi: 10.1109/TPAMI.2026.3711595. URL https://doi.org/10.1109/TPAMI.2026.3711595.

Jiawei Zhao, Zhenyu Zhang, Beidi Chen, Zhangyang Wang, Anima Anandkumar, and Yuandong Tian. GaLore: Memory-efficient LLM training by gradient low-rank projection. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 61121–61143. PMLR, 2024.

APPENDIX

CONTENTS

A Background on Stiefel and Grassmann geometry **15**

A.1 Orthonormal bases, subspaces, and tangent vectors 15

A.2 Principal angles, geodesics, and parallel transport 16

A.3 Retractions and the horizontal spectral subproblem 17

B Factor stationarity and coefficient scale **17**

B.1 Ordinary and constrained factor gradients 17

B.2 Approximately orthogonal factors and fixed-rank weights 18

B.3 Coefficient scale and regularization 19

C Convergence with a fixed number of NS iterations **19**

C.1 Setup and orthogonality control 19

C.2 Momentum estimation 22

C.3 Proof of Theorem 1 24

D Convergence with the exact matrix sign **27**

E Finite Newton–Schulz computation **30**

F Experimental protocol **31**

F.1 Pretraining and adaptation settings 31

F.2 Efficiency and mechanism measurements 32

A Background on Stiefel and Grassmann geometry

This appendix collects the geometric background used by PLUM. We use the Frobenius inner product $\langle A, B\rangle_F = \operatorname{tr}(A^\top B)$, $\operatorname{sym}(A) = (A + A^\top)/2$, and $\operatorname{skew}(A) = (A - A^\top)/2$. Standard references are Edelman et al. (1998); Absil et al. (2008); Bendokat et al. (2024); Boumal (2023).

A.1 Orthonormal bases, subspaces, and tangent vectors

The Stiefel manifold $\operatorname{St}(m, r) = \{U \in \mathbb{R}^{m\times r} : U^\top U = I_r\}$ contains orthonormal bases. Differentiating $U(t)^\top U(t) = I_r$ at $t = 0$ gives

$$T_U \operatorname{St}(m, r) = \{Z : U^\top Z + Z^\top U = 0\}.$$

Every such tangent has the unique decomposition

$$Z = U\Omega + (I - UU^\top)Z, \qquad \Omega = U^\top Z \in \mathfrak{o}(r), \qquad \mathfrak{o}(r) = \{A : A^\top = -A\}.$$

The space $\mathfrak{o}(r) = T_I\mathrm{O}(r)$ is the Lie algebra of the orthogonal group, with bracket $[A, B] = AB - BA$. Its element Ω is an infinitesimal rotation of the orthonormal frame: $\frac{d}{dt}U\exp(t\Omega)|_{t=0} = U\Omega$. The curve changes the basis but not $\operatorname{col}(U)$ (Absil et al., 2008; Boumal, 2023). Its velocities form the vertical space $\mathcal{V}_U = \{U\Omega : \Omega^\top = -\Omega\}$. The Frobenius-orthogonal complement of $\mathcal{V}_U$ inside $T_U \operatorname{St}(m, r)$ is

$$\mathcal{H}_U = \{Z : U^\top Z = 0\}, \qquad T_U \operatorname{St}(m, r) = \mathcal{V}_U \oplus \mathcal{H}_U. \tag{11}$$

Indeed, for a tangent Z, $U^\top Z$ is skew-symmetric. Its inner product with every skew-symmetric Ω vanishes exactly when $U^\top Z = 0$.

A quotient identifies all representations related by a specified group action. Here the quotient map is $\pi(U) = \operatorname{col}(U)$ and its fiber over a subspace is $[U] = \{UO : O \in \mathrm{O}(r)\}$. Figure 4 illustrates the smallest example, where the two signs of a unit vector represent the same line.

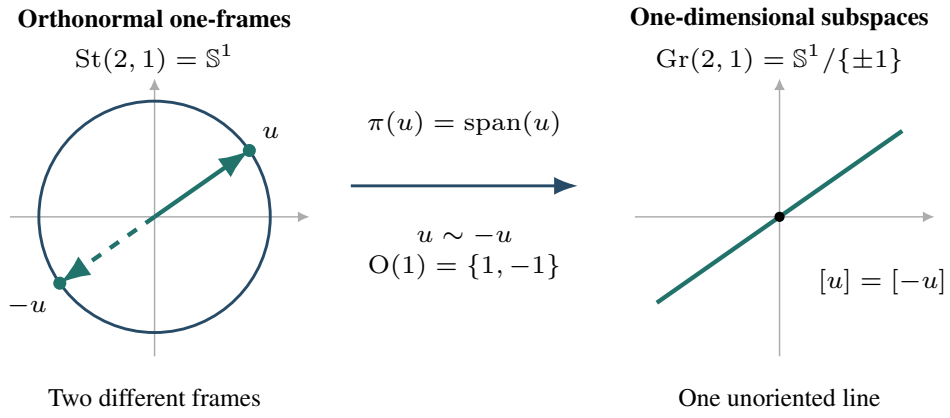

Figure 4: The quotient $\operatorname{St}(2, 1)/\mathrm{O}(1) = \operatorname{Gr}(2, 1)$. A Stiefel point is a unit vector, whereas a Grassmann point is an unoriented line. The quotient map identifies u and $-u$ because they span the same line.

The Grassmann manifold $\operatorname{Gr}(m, r)$ consists of r-dimensional subspaces of $\mathbb{R}^m$. The matrices U and UO, $O \in \mathrm{O}(r)$, represent the same subspace, giving $\operatorname{Gr}(m, r) = \operatorname{St}(m, r)/\mathrm{O}(r)$. A tangent to this quotient is represented uniquely at a chosen basis by a horizontal matrix in (11). For any ambient matrix D, its horizontal projection is $P_U D = (I - UU^\top)D$. Since P_U is an orthogonal projector,

$$\|P_U D\|_F \le \|D\|_F, \qquad \|P_U D\|_2 \le \|D\|_2.$$

The full Stiefel projection instead is $D - U\operatorname{sym}(U^\top D)$. These are different projections: the former removes all within-basis rotation, while the latter retains tangent rotations.

For $U = [e_1, e_2] \in \operatorname{St}(3, 2)$, the curve $UO(t)$ leaves its plane fixed, whereas $[\cos t\, e_1 + \sin t\, e_3, e_2]$ changes the plane. Their derivatives at zero belong to $\mathcal{V}_U$ and $\mathcal{H}_U$, respectively, as shown in Figure 5.

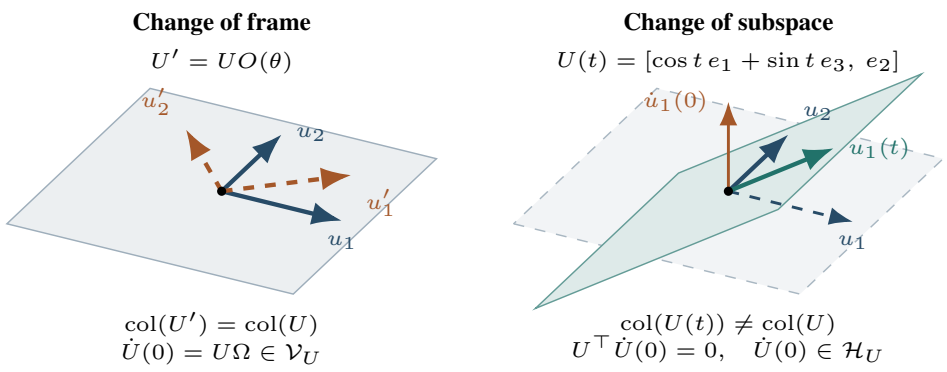

Figure 5: Vertical and horizontal motion at $U = [e_1, e_2] \in \mathrm{St}(3, 2)$. Left: both orthonormal pairs span one plane and project to the same point of $\mathrm{Gr}(3, 2)$. Right: the first basis vector moves out of the initial plane, so the represented subspace changes.

Quotient optimization represents a Grassmann tangent by the horizontal derivative and identifies the vertical derivative with zero.

A Riemannian metric is a smoothly varying positive-definite inner product on the tangent spaces. For a differentiable function F on a manifold, its gradient is characterized by $g_x(\mathrm{grad}\, F(x), Z) = dF_x[Z]$ for every tangent Z. For the Frobenius metric on Stiefel,

$$\mathrm{grad}\, F(U) = \nabla F(U) - U\, \mathrm{sym}(U^\top \nabla F(U)).$$

For a function of the subspace alone, its Grassmann gradient is represented by $(I - UU^\top)\nabla F(U)$.

A.2 Principal angles, geodesics, and parallel transport

For two orthonormal bases $U, \widetilde{U}$, the principal angles are defined by

$$\theta_j = \arccos \sigma_j(U^\top \widetilde{U}), \qquad 0 \leq \theta_j \leq \pi/2.$$

They are unchanged by replacing either basis by another orthonormal basis of its subspace.

Let $D = L\Sigma V^\top$ be the compact SVD of a horizontal tangent of rank q. Thus $L^\top L = V^\top V = I_q$ and $U^\top L = 0$. The geodesic with initial velocity $-D$ is (Edelman et al., 1998; Bendokat et al., 2024)

$$U(t) = U(I - VV^\top) + UV\cos(t\Sigma)V^\top - L\sin(t\Sigma)V^\top. \tag{12}$$

The matrices UV and L have orthonormal, mutually orthogonal columns. On the jth pair of columns, (12) is the ordinary rotation through $t\sigma_j(D)$. Consequently,

$$U(t)^\top U(t) = I, \quad U(t)^\top \dot{U}(t) = 0, \quad \ddot{U}(t) + U(t)\dot{U}(t)^\top \dot{U}(t) = 0.$$

Also $U^\top U(t) = I - VV^\top + V\cos(t\Sigma)V^\top$, which shows that the q nonzero principal angles are the values $t\sigma_j(D)$ when $0 \leq t\sigma_1(D) < \pi/2$. The other $r - q$ angles equal zero.

Tangents at different points belong to different vector spaces. Their ambient matrix sum is defined, but need not be tangent at either point. Parallel transport compares them along a specified curve by requiring a zero covariant derivative. Along (12), the horizontal transport of Z with $U^\top Z = 0$ is (Edelman et al., 1998; Bendokat et al., 2024)

$$\mathcal{T}_t(Z) = Z + \big[L(\cos(t\Sigma) - I_q) + UV\sin(t\Sigma)\big]L^\top Z. \tag{13}$$

To check the formula, write $Z = LA + Z_\perp$, where $A = L^\top Z$ and $L^\top Z_\perp = U^\top Z_\perp = 0$. Then

$$\mathcal{T}_t(Z) = [L\cos(t\Sigma) + UV\sin(t\Sigma)]A + Z_\perp.$$

The two summands are orthogonal, and the first bracket has Gram I_q. Hence $U(t)^\top \mathcal{T}_t(Z) = 0$ and $\|\mathcal{T}_t(Z)\|_F = \|Z\|_F$. Differentiation gives

$$\frac{d}{dt}\mathcal{T}_t(Z) = U(t)V\Sigma L^\top Z, \qquad (I - U(t)U(t)^\top)\frac{d}{dt}\mathcal{T}_t(Z) = 0,$$

which is the parallel-transport equation in this horizontal representation.

For the matrix sign, take $M = L\Sigma V^\top$, $Q = LV^\top$, and $\Pi = VV^\top = Q^\top Q$. All nonzero singular values of Q equal one, so (12) and (13) reduce to

$$U(t) = U(I - \Pi) + \cos t\, U\Pi - \sin t\, Q, \tag{14}$$

$$\mathcal{T}_t(Z) = Z + [(\cos t - 1)Q + \sin t\, U]Q^\top Z. \tag{15}$$

Here $\Pi^2 = \Pi$, $Q\Pi = Q$, and $\Pi Q^\top Z = Q^\top Z$. In particular,

$$U(t)^\top \mathcal{T}_t(Z) = 0, \qquad \|\mathcal{T}_t(Z)\|_F = \|Z\|_F, \qquad \|\mathcal{T}_t(Z) - Z\|_F \le |t|\|Z\|_F$$

for $|t| < \pi/2$. The last inequality follows from $2|\sin(t/2)| \le |t|$ on each rotating two-dimensional plane.

The momentum that produced Q satisfies $QQ^\top M = M$. Therefore its transport is especially simple:

$$\mathcal{T}_{\eta_U}(M) = \cos \eta_U\, M + \sin \eta_U\, UQ^\top M.$$

Using $U(\eta_U)^\top M = -\sin \eta_U\, Q^\top M$ and $\Pi Q^\top M = Q^\top M$,

$$\begin{aligned}(I - U(\eta_U)U(\eta_U)^\top)M &= M + \sin \eta_U\, U(\eta_U)Q^\top M \\ &= \cos^2 \eta_U\, M + \sin \eta_U \cos \eta_U\, UQ^\top M \\ &= \cos \eta_U\, \mathcal{T}_{\eta_U}(M).\end{aligned} \tag{16}$$

This proves the transport formula used for the stored momentum. The transport of any other horizontal matrix is given by (15). If $q = r$, then $\Pi = I$ and (14) is the scalar update $\cos \eta_U\, U - \sin \eta_U\, Q$. If $q < r$, the columns selected by $I - \Pi$ remain unchanged. If $M = 0$, $Q = \Pi = 0$ and the exact curve is constant.

A.3 Retractions and the horizontal spectral subproblem

A retraction R_U satisfies $R_U(0) = U$ and $DR_U(0)[Z] = Z$. It is a first-order local approximation of the exponential map. A second-order retraction additionally has zero covariant acceleration at the origin, and agrees with the geodesic through second order (Boumal, 2023). For a horizontal Z, the polar retraction is

$$R_U(Z) = (U + Z)(I + Z^\top Z)^{-1/2},$$

and a QR retraction uses an orthonormal factor of $U + Z$. On Grassmann these yield the same subspace. For $Z = -\eta L\Sigma V^\top$, it rotates the planes spanned by UV and L through $\arctan(\eta\Sigma)$, whereas (12) rotates them through $\eta\Sigma$. Since $\arctan x = x - x^3/3 + O(x^5)$, the two subspaces differ from the exponential by third-order terms. On Stiefel, polar is a second-order retraction under the embedded metric, whereas a QR choice need only be first order as a frame-valued map. The unit singular values of msign allow PLUM to evaluate the Grassmann exponential directly for its matrix-sign step.

The direction itself also has a closed-form horizontal subproblem. For any ambient M, every horizontal D satisfies $\langle M, D\rangle_F = \langle P_U M, D\rangle_F$. If $P_U M = \sum_j \sigma_j l_j v_j^\top$, then

$$\langle M, D\rangle_F = \sum_j \sigma_j l_j^\top D v_j \le \sum_j \sigma_j \|D\|_2.$$

The compact sign $D = \sum_j l_j v_j^\top$ is horizontal and attains equality when $\|D\|_2 \le 1$. This proves (2). Both compact msign and finite NS satisfy $\mathrm{col}(Q) \subseteq \mathrm{col}(M)$. Thus they preserve $\mathcal{H}_U$ under right multiplication. The optimal relative factor steps in the main text follow from Cauchy–Schwarz applied to the two scalar amplitudes.

B Factor stationarity and coefficient scale

B.1 Ordinary and constrained factor gradients

Let $\mathcal{L}(U, Y) = f(UY^\top)$ and $G = \nabla f(UY^\top)$. Its differential is

$$d\mathcal{L}(U,Y)[Z_U,Z_Y] = \langle G, Z_U Y^\top + U Z_Y^\top\rangle_F = \langle GY, Z_U\rangle_F + \langle G^\top U, Z_Y\rangle_F.$$

Thus $\nabla_U\mathcal{L} = GY$ and $\nabla_Y\mathcal{L} = G^\top U$. At $U^\top U = I_r$, let $H_U = (I - UU^\top)GY$ and $H_Y = G^\top U$. Under the standard Frobenius metric on $\mathrm{St}(m,r)\times\mathbb{R}^{n\times r}$ (Absil et al., 2008; Boumal, 2023), the Stiefel gradient is

$$\mathrm{grad}_{\mathrm{St},U}\,\mathcal{L} = GY - U\,\mathrm{sym}(U^\top GY) = H_U + U\,\mathrm{skew}(H_Y^\top Y), \qquad \nabla_Y\mathcal{L} = H_Y. \tag{17}$$

At feasibility, the first-order conditions are $GY = U\Lambda$ and $G^\top U = 0$, where Λ is symmetric. Since $\Lambda = U^\top GY = (G^\top U)^\top Y = 0$, these conditions are equivalent to vanishing ordinary factor gradients. Using the decomposition $GY = H_U + UH_Y^\top Y$, we obtain

$$\nabla_U\mathcal{L} = \nabla_Y\mathcal{L} = 0 \quad\Longleftrightarrow\quad H_U = H_Y = 0 \quad\Longleftrightarrow\quad \mathrm{grad}_{\mathrm{St},U}\,\mathcal{L} = \nabla_Y\mathcal{L} = 0. \tag{18}$$

The same decomposition is orthogonal and, when $\|Y\|_F \le R$, gives

$$\sqrt{\|GY\|_F^2 + \|G^\top U\|_F^2} \le \sqrt{1+R^2}\big(\|H_U\|_F + \|H_Y\|_F\big). \tag{19}$$

Since the Stiefel gradient is the Frobenius-orthogonal projection of GY onto $T_U\,\mathrm{St}(m,r)$, (19) also controls the Riemannian gradient norm on $\mathrm{St}(m,r)\times\mathbb{R}^{n\times r}$.

B.2 Approximately orthogonal factors and fixed-rank weights

For full-column-rank U, define $\Pi_U = I - U(U^\top U)^{-1}U^\top$ and write $H_U = \Pi_U GY$, $H_Y = G^\top U$. The decomposition

$$GY = H_U + U(U^\top U)^{-1}H_Y^\top Y$$

has orthogonal summands. If $\|U^\top U - I_r\|_F \le 1/2$, then $\|U(U^\top U)^{-1}\|_2 \le \sqrt{2}$, which implies

$$\sqrt{\|GY\|_F^2 + \|G^\top U\|_F^2} \le \sqrt{1+2R^2}\big(\|H_U\|_F + \|H_Y\|_F\big).$$

On this neighborhood, define $\bar U = U(U^\top U)^{-1/2}$ and $\bar Y = Y(U^\top U)^{1/2}$. Then $\bar U^\top\bar U = I_r$ and $\bar U\bar Y^\top = W$, so the represented weight is unchanged. Moreover, $G\bar Y = (GY)(U^\top U)^{1/2}$ and $G^\top\bar U = (G^\top U)(U^\top U)^{-1/2}$. The eigenvalue bounds and the nonexpansiveness of the Stiefel projection therefore give

$$\|\,\mathrm{grad}_{\mathrm{St},U}\,\mathcal{L}(\bar U,\bar Y)\|_F + \|\nabla_Y\mathcal{L}(\bar U,\bar Y)\|_F \le 2\sqrt{\|GY\|_F^2 + \|G^\top U\|_F^2}.$$

Outside this neighborhood, choose a measurable orthonormal r-frame $\bar U$ such that $\mathrm{col}(W)\subseteq\mathrm{col}(\bar U)$ using a fixed orthonormalization and completion convention, and set $\bar Y = W^\top\bar U$. Then $\bar U\bar Y^\top = W$ and $\|\bar Y\|_F = \|W\|_F$. Consequently,

$$\|\,\mathrm{grad}_{\mathrm{St},U}\,\mathcal{L}(\bar U,\bar Y)\|_F + \|\nabla_Y\mathcal{L}(\bar U,\bar Y)\|_F \le (1+\|W\|_F)\|G\|_F.$$

Together with the bound in the orthogonality neighborhood and the high-probability feasibility control from Theorem 1, this proves Corollary 1.

At a feasible pair with $\mathrm{rank}(Y) = r$, let $S = Y^\top Y$ and $V = YS^{-1/2}$. Then $V^\top V = I_r$ and the standard embedded fixed-rank gradient satisfies

$$\mathrm{grad}_{\mathcal{M}_r} f(W) = UU^\top G + (I-UU^\top)GVV^\top, \qquad \|\,\mathrm{grad}_{\mathcal{M}_r} f(W)\|_F^2 = \|H_Y\|_F^2 + \|H_U S^{-1/2}\|_F^2.$$

The two terms have orthogonal left supports. Hence the factor and fixed-rank stationary sets coincide at full rank. If $\sigma_r(Y) \ge \mu > 0$, then $\|\,\mathrm{grad}_{\mathcal{M}_r} f(W)\|_F \le \max\{1,\mu^{-1}\}(\|H_U\|_F + \|H_Y\|_F)$ (Mishra et al., 2014; Luo et al., 2024). When Y is rank deficient, the factor conditions in (18) remain well defined and are the stationarity conditions used in our analysis.

B.3 Coefficient scale and regularization

At an orthonormal basis, $\|W\|_F = \|Y\|_F$. More generally,

$$\left|\|UY^\top\|_F^2 - \|Y\|_F^2\right| = \left|\mathrm{tr}\left[(U^\top U - I_r)Y^\top Y\right]\right| \le \|U^\top U - I_r\|_2 \|Y\|_F^2.$$

Thus the coefficient norm and the represented-weight norm differ only by fixed factors in the orthogonality neighborhood.

For ℓ_2 regularization, let $\mathcal{L}_\lambda(U,Y) = f(UY^\top) + \lambda\|Y\|_F^2/2$, where $\lambda > 0$. Since $f \ge f_*$, any point satisfying $\mathcal{L}_\lambda(U,Y) \le C$ also satisfies

$$\|Y\|_F \le \sqrt{\frac{2(C - f_*)}{\lambda}}. \tag{20}$$

Hence any trajectory with uniformly bounded regularized objective values has a uniform coefficient bound. For a deterministic descent trajectory, one may take $C = \mathcal{L}_\lambda(U_0, Y_0)$.

On the constraint manifold, $\|UY^\top\|_F^2 = \|Y\|_F^2$, so the same regularization corresponds to $f_\lambda(W) = f(W) + \lambda\|W\|_F^2/2$. Its gradient is $G + \lambda W$, its smoothness constant is $L + \lambda$, and adding the deterministic term λW leaves the conditional variance unchanged. Define $H_Y^\lambda = G^\top U + \lambda Y$. Then

$$\frac{\lambda}{2}\left(\|Y - \eta_Y Q_Y\|_F^2 - \|Y\|_F^2\right) = -\lambda\eta_Y\langle Y, Q_Y\rangle_F + \frac{\lambda\eta_Y^2}{2}\|Q_Y\|_F^2.$$

The linear term is absorbed into H_Y^λ, while the quadratic term adds at most $\lambda b^2 r/2$ to C_f. Moreover, $\|H_{Y,t+1}^\lambda - H_{Y,t}^\lambda\|_F \le \|H_{Y,t+1} - H_{Y,t}\|_F + \lambda b\sqrt{r}\,\eta$, so C_H increases by at most $\lambda b\sqrt{r}$. At feasibility, $\mathrm{grad}_{\mathrm{St},U}\,\mathcal{L}_\lambda = H_U + U\,\mathrm{skew}\big((H_Y^\lambda)^\top Y\big)$ because $Y^\top Y$ is symmetric. With these changes, the exact-msign analysis gives the same $O(T^{-1/4})$ rate for the regularized constrained gradient on a controlled sublevel set, with the radius in (20).

Decoupled weight decay gives a direct bound on the coefficient trajectory. If $0 \le \lambda\eta_{Y,t} \le 1$, Algorithm 1 satisfies

$$\|Y_{t+1}\|_F \le (1 - \lambda\eta_{Y,t})\|Y_t\|_F + \eta_{Y,t} b\sqrt{r}.$$

By induction,

$$\sup_t \|Y_t\|_F \le \max\left\{\|Y_0\|_F,\ \frac{b\sqrt{r}}{\lambda}\right\}. \tag{21}$$

The same bound holds if decay uses the learning rate η_t instead of $\eta_{Y,t} \le \eta_t$, provided $0 \le \lambda\eta_t \le 1$. For fixed U, coefficient decay scales the represented weight exactly:

$$U\left[(1 - \lambda\eta_Y)Y\right]^\top = (1 - \lambda\eta_Y)UY^\top.$$

With simultaneous factor updates,

$$W_{t+1} = (1 - \lambda\eta_{Y,t})U_{t+1}Y_t^\top - \eta_{Y,t}U_{t+1}Q_{Y,t}^\top.$$

The ℓ_2-regularized update and decoupled weight decay are different optimization dynamics. The bound in (21) provides a coefficient bound for the latter. In the following analysis, we focus on the unregularized scenario $\lambda = 0$ and study the finite-NS dynamics directly.

C Convergence with a fixed number of NS iterations

C.1 Setup and orthogonality control

Let $\mathcal{L}(U,Y) = f(UY^\top)$, $G_t = \nabla f(U_tY_t^\top)$, and $P_t = I - U_tU_t^\top$. We analyze the following iteration with $M_{U,-1} = M_{Y,-1} = 0$ and $\eta_{U,-1} = 0$:

$$\begin{aligned}
M_{U,t} &= P_t\left[\frac{\beta M_{U,t-1}}{\cos\eta_{U,t-1}} + (1-\beta)\widehat{G}_tY_t\right], & M_{Y,t} &= \beta M_{Y,t-1} + (1-\beta)\widehat{G}_t^\top U_t,\\
Q_{U,t} &= \mathrm{NS}_K(M_{U,t};\nu), & Q_{Y,t} &= \mathrm{NS}_K(M_{Y,t};\nu),\\
U_{t+1} &= \cos\eta_{U,t}\,U_t - \sin\eta_{U,t}\,Q_{U,t}, & Y_{t+1} &= Y_t - \eta_{Y,t}Q_{Y,t}.
\end{aligned} \tag{22}$$

The steps are $\eta_{i,t} = \eta\gamma_{i,t}$, where

$$\rho_{U,t} = \frac{\langle M_{U,t}, Q_{U,t}\rangle_F}{\|Y_t\|_F^2 + \varepsilon}, \qquad \rho_{Y,t} = \frac{\langle M_{Y,t}, Q_{Y,t}\rangle_F}{\|U_t\|_F^2 + \varepsilon}, \qquad \gamma_{i,t} = \frac{\rho_{i,t}}{\max_j \rho_{j,t}}.$$

Both steps are zero when both inner products vanish. Assume $U_0^\top U_0 = I_r$, $Y_0 = 0$, and $\|Y_t\|_F \le R$ almost surely for $0 \le t \le T$, with a deterministic R independent of T. The decay coefficient is zero. The ambient objective and oracle satisfy Assumptions 1–2.

The filtration $\mathcal{F}_t$ is the information available before sample t. More precisely, it is the σ-algebra generated by the initialization and samples $0, \dots, t-1$, so U_t, Y_t and the previous momenta are $\mathcal{F}_t$-measurable. A square-integrable sequence X_t is a martingale-difference sequence for this indexing if X_t is $\mathcal{F}_{t+1}$-measurable and $\mathbb{E}[X_t \mid \mathcal{F}_t] = 0$. In particular, for $s < t$,

$$\mathbb{E}[X_s X_t] = \mathbb{E}[X_s \mathbb{E}(X_t \mid \mathcal{F}_t)] = 0, \qquad \mathbb{E}\Big(\sum_t X_t\Big)^2 = \sum_t \mathbb{E}X_t^2.$$

We use this orthogonality, conditional Jensen, and the tower property of conditional expectation (Durrett, 2019).

For the fixed NS polynomial ϕ_K, a compact SVD $M = L_M \Sigma_M V_M^\top$ gives

$$Q = L_M \phi_K\left(\frac{\Sigma_M}{\|M\|_F + \nu}\right) V_M^\top.$$

Thus the bounds on ϕ_K imply

$$\|Q\|_2 \le b, \quad \|Q\|_F \le b\sqrt{r}, \quad \langle M, Q\rangle_F \ge \frac{\alpha\|M\|_F^2}{\|M\|_F + \nu} \ge 0. \tag{23}$$

They also give the right-multiplication formula

$$Q = MS(M), \qquad S(M) \succeq 0, \qquad \|S(M)\|_2 \le \frac{\kappa}{\|M\|_F + \nu}. \tag{24}$$

For a positive singular value, the corresponding eigenvalue of $S(M)$ is $\phi_K(\sigma/(\|M\|_F + \nu))/\sigma$. At zero it is defined by the continuous polynomial extension. In particular, $\|AQ\|_F \le \kappa\|AM\|_F/(\|M\|_F + \nu)$ for every compatible A, including $M = 0$.

We will use the deterministic quantities

$$G_R = \|G_0\|_F + \sqrt{3/2}LR, \quad \Gamma_R = 1 + G_R + \sigma/\sqrt{B},$$
$$\Delta_b = \sqrt{r}\max\{1, b^2 - 1\}, \quad d_T = 2\Delta_b T\eta^2, \qquad p_T = \frac{\sigma^2}{B\Gamma_R^2 T}.$$

Define $\tau = \min\{t \le T : \|U_t^\top U_t - I_r\|_F > 1/2\}$, with $\tau = T + 1$ if the set is empty, and $I_t = \mathbf{1}_{\{t<\tau\}}$. Thus τ is the first violation of the chosen orthogonality neighborhood, and I_t retains exactly the steps before that violation. Both I_t and the event $\{\tau \le t\}$ are determined by $\mathcal{F}_t$. A sum $\sum_{t<T} I_t A_t$ therefore includes the updates $0, \dots, \min\{\tau, T\} - 1$.

Lemma 1 (Orthogonality control)**.** *If*

$$\eta \le \frac{1}{8b}, \qquad 2\eta^2 \le 1 - \beta, \qquad d_T \le \frac{1}{4}, \qquad \nu \ge 16\kappa R\Gamma_R T\eta[4b\eta + \sqrt{3/2}(1 - \beta)], \tag{25}$$

then (22) *satisfies*

$$\begin{cases} \mathbb{P}\Big(\max\limits_{t\le T} \|U_t^\top U_t - I_r\|_F \le d_T\Big) \ge 1 - p_T, & (26) \\ \mathbb{E}\max\limits_{t\le T} \|U_t^\top U_t - I_r\|_F \le d_T + \sqrt{r}\big[(1 + bT\eta)^2 + 1\big]p_T. & (27) \end{cases}$$

Proof. For $t < \tau$, the eigenvalues of $U_t^\top U_t$ are in $[1/2, 3/2]$. On $\mathrm{col}(U_t)$ the eigenvalues of P_t are $1 - \sigma_j(U_t)^2$, and on its orthogonal complement they are one. Consequently,

$$\|P_t\|_2 \le 1, \qquad \sigma_{\min}(U_t) \ge 1/\sqrt{2}, \qquad \|U_t\|_2 \le \sqrt{3/2}, \qquad \|G_t\|_F \le G_R. \tag{28}$$

The gradient bound follows from $\|G_t\|_F \le \|G_0\|_F + L\|U_t Y_t^\top\|_F \le \|G_0\|_F + \sqrt{3/2}LR$.

Let $S_T^{\text{in}} = \sum_{t<T} I_t \|\widehat{G}_t\|_F$. Before τ, the triangle inequality and conditional Jensen give

$$\begin{aligned}\mathbb{E}[\|\widehat{G}_t\|_F \mid \mathcal{F}_t] &\le \|G_t\|_F + \mathbb{E}[\|\widehat{G}_t - G_t\|_F \mid \mathcal{F}_t] \\ &\le G_R + \sqrt{\mathbb{E}[\|\widehat{G}_t - G_t\|_F^2 \mid \mathcal{F}_t]} \le G_R + \sigma/\sqrt{B}.\end{aligned}$$

For $X_t = I_t(\|\widehat{G}_t\|_F - \mathbb{E}[\|\widehat{G}_t\|_F \mid \mathcal{F}_t])$, conditional centering minimizes the mean-square distance to an $\mathcal{F}_t$-measurable scalar. Hence

$$\begin{aligned}\mathbb{E}[X_t^2 \mid \mathcal{F}_t] &\le I_t \mathbb{E}[(\|\widehat{G}_t\|_F - \|G_t\|_F)^2 \mid \mathcal{F}_t] \\ &\le I_t \mathbb{E}[\|\widehat{G}_t - G_t\|_F^2 \mid \mathcal{F}_t] \le \sigma^2/B.\end{aligned}$$

The sequence X_t is a martingale difference. For $Z_T = \sum_{t<T} X_t$, the preceding orthogonality calculation gives $\mathbb{E}Z_T^2 \le T\sigma^2/B$. Moreover, $S_T^{\text{in}} \le T\Gamma_R + Z_T$. Markov's inequality applied to Z_T^2 yields

$$\mathbb{P}(S_T^{\text{in}} > 2T\Gamma_R) \le \mathbb{P}(Z_T > T\Gamma_R) \le \mathbb{E}\frac{Z_T^2}{T^2\Gamma_R^2} \le p_T. \tag{29}$$

Since $z\mathbf{1}_{\{z>a\}} \le z^2/a$ for $a > 0$,

$$\begin{aligned}\mathbb{E}[S_T^{\text{in}}\mathbf{1}_{\{S_T^{\text{in}} > 2T\Gamma_R\}}] &\le T\Gamma_R p_T + \mathbb{E}[Z_T \mathbf{1}_{\{Z_T > T\Gamma_R\}}] \\ &\le T\Gamma_R p_T + \frac{\mathbb{E}Z_T^2}{T\Gamma_R} \le 2T\Gamma_R p_T.\end{aligned} \tag{30}$$

To control the basis momentum, put $N_t = U_t^\top M_{U,t}$ and $N_{-1} = 0$. From the basis update,

$$U_t^\top M_{U,t-1} = \cos\eta_{U,t-1} N_{t-1} - \sin\eta_{U,t-1} Q_{U,t-1}^\top M_{U,t-1}.$$

Combining this with $U_t^\top P_t = -(U_t^\top U_t - I_r)U_t^\top$ gives the recursion

$$\begin{aligned}N_t = \beta(U_t^\top U_t - I_r)\left[-N_{t-1} + \frac{\sin\eta_{U,t-1}}{\cos\eta_{U,t-1}} Q_{U,t-1}^\top M_{U,t-1}\right] \\ - (1-\beta)(U_t^\top U_t - I_r)U_t^\top \widehat{G}_t Y_t.\end{aligned} \tag{31}$$

For $0 \le x \le 1/8$, $1/\cos x \le 1 + x^2$ and $\sin x/\cos x \le 2x$. The step conditions imply $\beta/\cos\eta \le (1+\beta)/2$. Therefore, for $t < \tau$,

$$\|M_{U,t}\|_F \le \frac{1+\beta}{2}\|M_{U,t-1}\|_F + (1-\beta)R\|\widehat{G}_t\|_F.$$

Let $J = \min\{\tau, T\}$. Summing the recursion for $t = 0, \dots, J-1$, using $M_{U,-1} = 0$ and dropping the last nonnegative term, gives

$$\begin{aligned}\sum_{t<J}\|M_{U,t}\|_F &\le \frac{1+\beta}{2}\sum_{t<J}\|M_{U,t}\|_F + (1-\beta)RS_T^{\text{in}}, \\ \sum_{t<T} I_t\|M_{U,t}\|_F &\le 2RS_T^{\text{in}}.\end{aligned}$$

Write the maximum before J explicitly in (31). By (28), $\beta\|U_t^\top U_t - I_r\|_F \le 1/2$ and $\|Q_{U,t-1}^\top M_{U,t-1}\|_F \le b\|M_{U,t-1}\|_F$. Thus

$$\begin{aligned}\sum_{t<J}\|N_t\|_F \le \frac{1}{2}\sum_{t<J}\|N_t\|_F \\ + \max_{j<J}\|U_j^\top U_j - I_r\|_F\left[2b\eta\sum_{t<J}\|M_{U,t-1}\|_F + (1-\beta)\sqrt{3/2}RS_T^{\text{in}}\right].\end{aligned}$$

Moving the first sum to the left proves

$$\sum_{t<T} I_t \|N_t\|_F \le 2R[4b\eta + \sqrt{3/2}(1-\beta)] S_T^{\text{in}} \max_{j<J} \|U_j^\top U_j - I_r\|_F. \tag{32}$$

Direct multiplication of the actual basis update gives

$$\begin{aligned} U_{t+1}^\top U_{t+1} - I_r = \; & \cos^2 \eta_{U,t}(U_t^\top U_t - I_r) + \sin^2 \eta_{U,t}(Q_{U,t}^\top Q_{U,t} - I_r) \\ & - \sin \eta_{U,t} \cos \eta_{U,t}(U_t^\top Q_{U,t} + Q_{U,t}^\top U_t). \end{aligned} \tag{33}$$

The NS bounds imply $\|Q_{U,t}^\top Q_{U,t} - I_r\|_F \le \Delta_b$ and $\|U_t^\top Q_{U,t}\|_F \le \kappa \|N_t\|_F/\nu$. Summing (33) through the update that reaches J and using (32), we obtain

$$\max_{j\le J} \|U_j^\top U_j - I_r\|_F \le \Delta_b T\eta^2 + \frac{4\kappa R\eta}{\nu}[4b\eta + \sqrt{3/2}(1-\beta)] S_T^{\text{in}} \max_{j\le J} \|U_j^\top U_j - I_r\|_F.$$

On $\{S_T^{\text{in}} \le 2T\Gamma_R\}$, the coefficient of the last maximum is at most $1/2$ by (25). Hence that maximum is at most $d_T \le 1/4$. If $\tau \le T$, its endpoint would instead have error greater than $1/2$. This contradiction shows $\tau > T$ on the event, and proves (26).

On every trajectory, the update also satisfies $\|U_{t+1}\|_2 \le \|U_t\|_2 + b\eta$, so $\|U_t\|_2 \le 1 + bt\eta$ and $\max_{t\le T} \|U_t^\top U_t - I_r\|_F \le \sqrt{r}[(1+bT\eta)^2 + 1]$. Split the expectation into the event in (29) and its complement. The respective upper bounds are d_T and this deterministic quantity times p_T, proving (27). □

Two consequences of this proof are needed for the gradient analysis. On the event $S_T^{\text{in}} \le 2T\Gamma_R$, the maximum before J is at most d_T. It is always at most $1/2$. Since $\mathbb{E}S_T^{\text{in}} \le T\Gamma_R$, (30) implies

$$\begin{aligned} \mathbb{E}\left[S_T^{\text{in}} \max_{j<J} \|U_j^\top U_j - I_r\|_F \right] &\le d_T \mathbb{E} S_T^{\text{in}} + \tfrac{1}{2}\mathbb{E}[S_T^{\text{in}} \mathbf{1}_{\{S_T^{\text{in}} > 2T\Gamma_R\}}] \\ &\le T\Gamma_R(d_T + p_T). \end{aligned}$$

Consequently,

$$\begin{cases} \dfrac{1}{T}\displaystyle\sum_{t<T} \mathbb{E}[I_t \|U_t^\top U_t - I_r\|_F] \le d_T + p_T/2, \\ \dfrac{1}{T}\displaystyle\sum_{t<T} \mathbb{E}[I_t \|N_t\|_F] \le 2R\Gamma_R[4b\eta + \sqrt{3/2}(1-\beta)](d_T + p_T). \end{cases} \tag{34}$$

C.2 Momentum estimation

Before τ, define $\Pi_U = I - U(U^\top U)^{-1}U^\top$ and $H_{U,t} = \Pi_{U_t} G_t Y_t$, $H_{Y,t} = G_t^\top U_t$. If $U = L\Sigma V^\top$ is a thin SVD, then $P - \Pi_U = L(I - \Sigma^2)L^\top$ and $U^\top U - I_r = V(\Sigma^2 - I_r)V^\top$. Hence

$$\|P - \Pi_U\|_2 = \|U^\top U - I_r\|_2, \qquad \|P - \Pi_U\|_F = \|U^\top U - I_r\|_F, \qquad P\Pi_U = \Pi_U. \tag{35}$$

The error in the two momentum estimates is

$$\mu_t = \sqrt{\|M_{U,t} - H_{U,t}\|_F^2 + \|M_{Y,t} - H_{Y,t}\|_F^2} \qquad (t < \tau).$$

Whenever $I_t = 0$, every expression $I_t H_{U,t}$ or $I_t \mu_t$ below is defined as zero. For the next lemma, set

$$\begin{cases} C_W = 2b(R + \sqrt{r}), \\ C_H = [(12b+1)R + 2b\sqrt{r} + 2b]G_R + (2R + \sqrt{3/2})LC_W. \end{cases}$$

Lemma 2 (Momentum estimation)**.** *Under Lemma 1,*

$$\begin{aligned} \frac{1}{T}\sum_{t<T} \mathbb{E}[I_t \mu_t] \le \; & \frac{2\beta\|G_0\|_F}{(1-\beta)T} + \frac{2C_H\eta}{1-\beta} + 2RG_R(d_T + p_T/2) \\ & + \sigma\sqrt{\frac{2(1-\beta)(R^2 + 3/2)}{B}} = \mathcal{E}_T. \end{aligned} \tag{36}$$

Proof. The actual updates and (28) imply

$$\|U_{t+1} - U_t\|_2 \le \sqrt{3/2}\,\eta^2/2 + b\eta \le 2b\eta, \qquad \|Y_{t+1} - Y_t\|_F \le b\sqrt{r}\,\eta.$$

In particular, $\|U_{t+1}\|_2 < 2$ at the endpoint of a step starting before τ. Since $\Delta W_t = (U_{t+1} - U_t)Y_t^\top - \eta_{Y,t}U_{t+1}Q_{Y,t}^\top$,

$$\|\Delta W_t\|_F \le 2bR\eta + 2b\sqrt{r}\,\eta = C_W\eta. \tag{37}$$

For two consecutive points before τ, $\|\Pi_{U_{t+1}} - \Pi_{U_t}\|_2 \le 2\sqrt{2}\|U_{t+1} - U_t\|_2 \le 6b\eta$. To see this, decompose the difference into its actions between the two range spaces and their complements. For one action,

$$\Pi_{U_{t+1}}(I - \Pi_{U_t}) = \Pi_{U_{t+1}}(U_t - U_{t+1})(U_t^\top U_t)^{-1}U_t^\top,$$

whose norm is at most $\sqrt{2}\|U_{t+1} - U_t\|_2$. Interchanging the two points gives the same estimate for the other action.

The factor fields have the differences

$$\begin{cases} H_{U,t+1} - H_{U,t} = (\Pi_{U_{t+1}} - \Pi_{U_t})G_{t+1}Y_{t+1} + \Pi_{U_t}(G_{t+1} - G_t)Y_{t+1} + \Pi_{U_t}G_t(Y_{t+1} - Y_t), \\ H_{Y,t+1} - H_{Y,t} = (G_{t+1} - G_t)^\top U_{t+1} + G_t^\top(U_{t+1} - U_t). \end{cases}$$

Using $\|G_{t+1} - G_t\|_F \le LC_W\eta$ term by term gives

$$\begin{cases} \|H_{U,t+1} - H_{U,t}\|_F \le (6bRG_R + LRC_W + b\sqrt{r}\,G_R)\eta, \\ \|H_{Y,t+1} - H_{Y,t}\|_F \le (\sqrt{3/2}LC_W + 2bG_R)\eta. \end{cases}$$

Since $P_tH_{U,t} = H_{U,t}$,

$$\frac{P_tH_{U,t-1}}{\cos\eta_{U,t-1}} - H_{U,t} = \frac{P_t(H_{U,t-1} - H_{U,t})}{\cos\eta_{U,t-1}} + \left(\frac{1}{\cos\eta_{U,t-1}} - 1\right)H_{U,t}.$$

Apply $1/\cos\eta \le 2$, $1/\cos\eta - 1 \le \eta^2$, and $\|H_{U,t}\|_F \le RG_R$. The preceding three inequalities give

$$\left\|\frac{P_tH_{U,t-1}}{\cos\eta_{U,t-1}} - H_{U,t}\right\|_F + \|H_{Y,t-1} - H_{Y,t}\|_F \le C_H\eta. \tag{38}$$

Subtracting $H_{U,t}, H_{Y,t}$ from the actual momentum recursions gives

$$\begin{cases} M_{U,t} - H_{U,t} = \dfrac{\beta P_t(M_{U,t-1} - H_{U,t-1})}{\cos\eta_{U,t-1}} + \beta\left[\dfrac{P_tH_{U,t-1}}{\cos\eta_{U,t-1}} - H_{U,t}\right] \\ \qquad\qquad + (1-\beta)(P_t - \Pi_{U_t})G_tY_t + (1-\beta)P_t(\widehat{G}_t - G_t)Y_t, & (39) \\ M_{Y,t} - H_{Y,t} = \beta(M_{Y,t-1} - H_{Y,t-1}) + \beta(H_{Y,t-1} - H_{Y,t}) + (1-\beta)(\widehat{G}_t - G_t)^\top U_t. & (40) \end{cases}$$

The projection term in the second line of (39) has norm at most $RG_R\|U_t^\top U_t - I_r\|_F$, by (35). The two current noise matrices, denoted by $\xi_{U,t} = P_t(\widehat{G}_t - G_t)Y_t$ and $\xi_{Y,t} = (\widehat{G}_t - G_t)^\top U_t$, obey

$$\mathbb{E}[\xi_{U,t} \mid \mathcal{F}_t] = \mathbb{E}[\xi_{Y,t} \mid \mathcal{F}_t] = 0, \qquad \mathbb{E}[\|\xi_{U,t}\|_F^2 + \|\xi_{Y,t}\|_F^2 \mid \mathcal{F}_t] \le (R^2 + 3/2)\sigma^2/B. \tag{41}$$

This follows by bounding their norms by $R\|\widehat{G}_t - G_t\|_F$ and $\sqrt{3/2}\|\widehat{G}_t - G_t\|_F$, respectively.

Separate the noise and the remaining terms using four matrix sequences. For $t \ge 1$, define

$$\begin{cases} Z_{U,t} = I_t\left[\dfrac{\beta P_tZ_{U,t-1}}{\cos\eta_{U,t-1}} + (1-\beta)\xi_{U,t}\right], \\ Z_{Y,t} = I_t[\beta Z_{Y,t-1} + (1-\beta)\xi_{Y,t}], \\ V_{U,t} = I_t\left[\dfrac{\beta P_tV_{U,t-1}}{\cos\eta_{U,t-1}} + \beta\left(\dfrac{P_tH_{U,t-1}}{\cos\eta_{U,t-1}} - H_{U,t}\right)\right] + I_t(1-\beta)(P_t - \Pi_{U_t})G_tY_t, \\ V_{Y,t} = I_t[\beta V_{Y,t-1} + \beta(H_{Y,t-1} - H_{Y,t})]. \end{cases}$$

Initialize $Z_{i,0} = (1-\beta)\xi_{i,0}$, $V_{U,0} = 0$, and $V_{Y,0} = -\beta G_0^\top U_0$. After τ, set these four sequences to zero. By (39)–(40) and $I_t = 1 \Rightarrow I_{t-1} = 1$,

$$I_t(M_{U,t} - H_{U,t}) = Z_{U,t} + V_{U,t}, \qquad I_t(M_{Y,t} - H_{Y,t}) = Z_{Y,t} + V_{Y,t}.$$

For a pair of matrices, use $\|(A,B)\|_F^2 = \|A\|_F^2 + \|B\|_F^2$, and write $Z_t = (Z_{U,t}, Z_{Y,t})$, $V_t = (V_{U,t}, V_{Y,t})$.

All coefficients of the old states, including I_t, are $\mathcal{F}_t$-measurable. For example,

$$\mathbb{E}\left[I_t \left\langle \frac{\beta P_t Z_{U,t-1}}{\cos\eta_{U,t-1}}, \xi_{U,t} \right\rangle_F\right] = \mathbb{E}\left[I_t \left\langle \frac{\beta P_t Z_{U,t-1}}{\cos\eta_{U,t-1}}, \mathbb{E}[\xi_{U,t} \mid \mathcal{F}_t] \right\rangle_F\right] = 0.$$

The same identity holds in the coefficient block. With $q = (1+\beta)/2$, the old-state coefficient has norm at most q, so (41) yields

$$\begin{aligned}\mathbb{E}\|Z_t\|_F^2 &\le q^2\mathbb{E}\|Z_{t-1}\|_F^2 + (1-\beta)^2(R^2+3/2)\sigma^2/B \\ &\le (1-\beta)^2(R^2+3/2)\frac{\sigma^2}{B}\sum_{j=0}^{t} q^{2j} \\ &\le 2(1-\beta)(R^2+3/2)\sigma^2/B.\end{aligned}$$

The last step uses $1-q^2 \ge 1-q = (1-\beta)/2$.

For V_t, the triangle inequality and (38) imply

$$\|V_t\|_F \le q\|V_{t-1}\|_F + I_t C_H \eta + (1-\beta)RG_R I_t \|U_t^\top U_t - I_r\|_F.$$

Sum this inequality for $t = 1, \dots, T-1$, include $\|V_0\|_F \le \beta\|G_0\|_F$, and move the old-state sum to the left. Since $1-q = (1-\beta)/2$,

$$\frac{1}{T}\sum_{t<T}\mathbb{E}\|V_t\|_F \le \frac{2\beta\|G_0\|_F}{(1-\beta)T} + \frac{2C_H\eta}{1-\beta} + \frac{2RG_R}{T}\sum_{t<T}\mathbb{E}[I_t\|U_t^\top U_t - I_r\|_F].$$

Finally, $I_t\mu_t \le \|Z_t\|_F + \|V_t\|_F$ and $\mathbb{E}\|Z_t\|_F \le \sqrt{\mathbb{E}\|Z_t\|_F^2}$ by Jensen. Substitute the preceding two bounds and (34) to obtain (36). □

C.3 Proof of Theorem 1

For an explicit version of the bounds, use C_W, C_H from Lemma 2 and set

$$\begin{aligned} C_f &= G_R[(\sqrt{3/2}/2 + b/6)R + 2b^2\sqrt{r}] + \tfrac{L}{2}C_W^2, \quad & c_R &= \sqrt{\frac{\varepsilon}{R^2+3r/2+2\varepsilon}}, \\ J &= \max\left\{1, \frac{\kappa(3r/2+\varepsilon)}{\alpha\varepsilon}\right\}, & C_\perp &= 2R\sqrt{3/2}\sqrt{\kappa^2J^2+b^2r}. \end{aligned} \tag{42}$$

With $\mathcal{E}_T$ defined in (36) and the constants in (42), put

$$\mathcal{D}_T = \frac{1}{c_R}\left[\frac{f(0)-f_*}{\eta T} + C_f\eta + (b\sqrt{2r} + C_\perp)\mathcal{E}_T\right]. \tag{43}$$

We prove the following explicit form of (8):

$$\begin{aligned}&\frac{1}{T}\sum_{t<T}\mathbb{E}\big[\|\nabla_U\mathcal{L}(U_t,Y_t)\|_F + \|\nabla_Y\mathcal{L}(U_t,Y_t)\|_F\big] \\ &\quad\le \sqrt{2(1+2R^2)}\left[\frac{\mathcal{D}_T}{\alpha} + \min\left\{2\nu, \sqrt{\frac{2\nu\mathcal{D}_T}{\alpha}}\right\} + \sqrt{2}\mathcal{E}_T\right] \\ &\qquad + (R+1+bT\eta)[\|G_0\|_F + LR(1+bT\eta)]p_T.\end{aligned} \tag{44}$$

The corresponding explicit orthogonality bounds are (26)–(27).

For this calculation only, abbreviate $g_i = \langle M_i, Q_i\rangle_F$, $w_U = \|Y\|_F^2 + \varepsilon$, and $w_Y = \|U\|_F^2 + \varepsilon$. If at least one inner product is positive, then

$$\begin{aligned}\gamma_U g_U + \gamma_Y g_Y &= \frac{g_U^2/w_U + g_Y^2/w_Y}{\max\{g_U/w_U, g_Y/w_Y\}} \ge \sqrt{\min\{w_U,w_Y\}}\sqrt{g_U^2/w_U + g_Y^2/w_Y} \\ &\ge \sqrt{\frac{\min\{w_U,w_Y\}}{w_U+w_Y}}(g_U+g_Y) \ge c_R(g_U+g_Y).\end{aligned} \tag{45}$$

The first inequality uses $\max_i(g_i/w_i) \le (\sum_i g_i^2/w_i)^{1/2}/\sqrt{\min_i w_i}$. The second is Cauchy–Schwarz, and the last uses $w_i \ge \varepsilon$ and $w_U + w_Y \le R^2 + 3r/2 + 2\varepsilon$ before τ. If both inner products are zero, the same bound holds with zero on both sides. Consequently,

$$\gamma_U \langle H_U, Q_U\rangle_F + \gamma_Y \langle H_Y, Q_Y\rangle_F \ge c_R(g_U + g_Y) - b\sqrt{2r}\,\mu_t. \tag{46}$$

Here the difference is bounded by $b\sqrt{r}(\|M_U - H_U\|_F + \|M_Y - H_Y\|_F) \le b\sqrt{2r}\mu_t$. This inequality holds for the current sample, even though the steps depend on that sample.

To apply it to the loss, first expand the actual weight update:

$$\begin{aligned}\Delta W_t + \eta_{U,t} Q_{U,t} Y_t^\top + \eta_{Y,t} U_t Q_{Y,t}^\top = [(\cos\eta_{U,t} - 1)U_t + (\eta_{U,t} - \sin\eta_{U,t})Q_{U,t}]Y_t^\top \\ - \eta_{Y,t}(U_{t+1} - U_t)Q_{Y,t}^\top.\end{aligned}$$

Its norm is at most $[(\sqrt{3/2}/2 + b/6)R + 2b^2\sqrt{r}]\eta^2$. Indeed, the three terms are bounded by $\sqrt{3/2}R\eta^2/2$, $bR\eta^3/6$, and $2b^2\sqrt{r}\eta^2$. The smoothness inequality and (37) therefore give

$$f(W_{t+1}) - f(W_t) \le -\eta_{U,t}\langle G_t Y_t, Q_{U,t}\rangle_F - \eta_{Y,t}\langle H_{Y,t}, Q_{Y,t}\rangle_F + C_f\eta^2. \tag{47}$$

The first inner product uses G_tY_t, rather than $H_{U,t}$. Their difference is

$$G_tY_t - H_{U,t} = U_t(U_t^\top U_t)^{-1}H_{Y,t}^\top Y_t.$$

The effect of this component can be bounded together with the step-size ratio. Write $m_i = \|M_i\|_F$ and $e_i = \|M_i - H_i\|_F$ for this calculation. The NS estimates give

$$\frac{\alpha m_i^2}{m_i + \nu} \le g_i \le \frac{\kappa m_i^2}{m_i + \nu}.$$

If $m_Y \le m_U$, then $\gamma_U m_Y/(m_U + \nu) \le 1$. If $m_Y > m_U > 0$, the definition of γ_U implies

$$\frac{\gamma_U m_Y}{m_U + \nu} \le \frac{w_Y}{w_U}\frac{g_U}{g_Y}\frac{m_Y}{m_U + \nu} \le \frac{w_Y\kappa}{w_U\alpha}\frac{m_U^2(m_Y + \nu)}{m_Y(m_U + \nu)^2} \le \frac{w_Y\kappa}{w_U\alpha}\frac{m_U}{m_U + \nu} \le J.$$

The third line uses $(m_Y + \nu)/m_Y \le (m_U + \nu)/m_U$. If $m_U = 0$, then $Q_U = 0$ and the same conclusion follows directly. Thus

$$\frac{\gamma_U\|M_Y\|_F}{\|M_U\|_F + \nu} \le J. \tag{48}$$

Because $U^\top H_U = 0$, we have $\|U^\top M_U\|_F \le \sqrt{3/2}\,e_U$. Moreover, $\|H_Y\|_F \le m_Y + e_Y$ and

$$\|U^\top Q_U\|_F \le \min\left\{\frac{\kappa\sqrt{3/2}\,e_U}{m_U + \nu},\, b\sqrt{3r/2}\right\}.$$

The first bound is useful for the m_Y term, and the second for the e_Y term. In detail,

$$\begin{aligned}\gamma_U|\langle GY - H_U, Q_U\rangle_F| &\le 2R\gamma_U\|H_Y\|_F\|U^\top Q_U\|_F \\ &\le 2R\sqrt{3/2}\left[\kappa J e_U + b\sqrt{r}\,e_Y\right] \le C_\perp\mu_t.\end{aligned} \tag{49}$$

This estimate retains the two-block normalization instead of bounding its factors separately. Combining (46), (47), and (49) gives

$$f(W_{t+1}) - f(W_t) \le -\eta c_R(g_{U,t} + g_{Y,t}) + \eta(b\sqrt{2r} + C_\perp)\mu_t + C_f\eta^2.$$

Multiply by I_t and sum. On every sample path,

$$\sum_{t<T} I_t[f(W_t) - f(W_{t+1})] = f(0) - f(W_{\min\{\tau,T\}}) \le f(0) - f_*.$$

Taking expectations and applying Lemma 2 therefore yields

$$\frac{1}{T}\sum_{t<T}\mathbb{E}[I_t(g_{U,t} + g_{Y,t})] \le \frac{1}{c_R}\left[\frac{f(0) - f_*}{\eta T} + C_f\eta + (b\sqrt{2r} + C_\perp)\mathcal{E}_T\right] = \mathcal{D}_T. \tag{50}$$

We now convert the inner-product estimate into a gradient bound using (23). If $m = \|M\|_F$ and $g = \langle M, Q\rangle_F$, then $m^2 - (g/\alpha)m - \nu g/\alpha \le 0$. Solving this inequality gives

$$m \le \frac{g/\alpha + \sqrt{(g/\alpha)^2 + 4\nu g/\alpha}}{2} \le \frac{g}{\alpha} + \sqrt{\frac{\nu g}{\alpha}}. \tag{51}$$

The same quadratic inequality gives

$$\frac{g}{\alpha} \ge \frac{m^2}{m+\nu} = m - \nu + \frac{\nu^2}{m+\nu}, \qquad m \le \frac{g}{\alpha} + \nu.$$

Apply (51) to the two blocks and use $\sqrt{g_U} + \sqrt{g_Y} \le \sqrt{2(g_U + g_Y)}$ to obtain

$$\|H_U\|_F + \|H_Y\|_F \le \frac{g_U + g_Y}{\alpha} + \sqrt{\frac{2\nu(g_U + g_Y)}{\alpha}} + \sqrt{2}\mu_t.$$

Cauchy–Schwarz on the joint time and probability average gives

$$\left(\frac{1}{T}\sum_{t<T} \mathbb{E}[I_t \sqrt{g_{U,t} + g_{Y,t}}]\right)^2 \le \left(\frac{1}{T}\sum_{t<T} \mathbb{E} I_t\right)\left(\frac{1}{T}\sum_{t<T} \mathbb{E}[I_t(g_{U,t} + g_{Y,t})]\right) \le \mathcal{D}_T.$$

The alternative bound $m \le g/\alpha + \nu$ gives the same average with 2ν in place of the square-root term. Taking the smaller of the two valid estimates, we obtain

$$\frac{1}{T}\sum_{t<T} \mathbb{E}[I_t(\|H_{U,t}\|_F + \|H_{Y,t}\|_F)] \le \frac{\mathcal{D}_T}{\alpha} + \min\left\{2\nu, \sqrt{\frac{2\nu\mathcal{D}_T}{\alpha}}\right\} + \sqrt{2}\mathcal{E}_T. \tag{52}$$

At points before τ, the orthogonal decomposition $G_t Y_t = H_{U,t} + U_t(U_t^\top U_t)^{-1} H_{Y,t}^\top Y_t$ implies

$$\begin{aligned}\|G_t Y_t\|_F^2 + \|G_t^\top U_t\|_F^2 &\le \|H_{U,t}\|_F^2 + (1 + 2R^2)\|H_{Y,t}\|_F^2 \\ &\le (1 + 2R^2)(\|H_{U,t}\|_F + \|H_{Y,t}\|_F)^2.\end{aligned}$$

Therefore their sum is at most $\sqrt{2(1+2R^2)}$ times the sum in (52). After a possible exit, the global bounds $\|U_t\|_2 \le 1 + bT\eta$ and $\|Y_t\|_F \le R$ still imply

$$\begin{cases}\|G_t\|_F \le \|G_0\|_F + LR(1 + bT\eta), \\ \|G_t Y_t\|_F + \|G_t^\top U_t\|_F \le (R + 1 + bT\eta)[\|G_0\|_F + LR(1 + bT\eta)].\end{cases}$$

For each t, $\{I_t = 0\} \subseteq \{\tau \le T\}$ and its probability is at most p_T. Splitting the ordinary gradient average into I_t and $1 - I_t$ terms proves (44).

To obtain the concise statement in the main text, note that $p_T \le 1/T$ and $d_T = 2\Delta_b T\eta^2$. From (36),

$$\mathcal{E}_T \le C_E \mathcal{B}_T, \qquad C_E = 2 + 2C_H + 4RG_R\Delta_b + RG_R + \sqrt{2(R^2 + 3/2)}.$$

Since $\eta \le \eta/(1-\beta)$, (43) gives

$$\mathcal{D}_T \le C_D \mathcal{B}_T, \qquad C_D = \frac{1 + C_f + (b\sqrt{2r} + C_\perp)C_E}{c_R}.$$

The term contributed by $I_t = 0$ is bounded by

$$(R + 1 + bT\eta)[\|G_0\|_F + LR(1 + bT\eta)]p_T \le b^2(R+1)(\|G_0\|_F + LR)\frac{(1+T\eta)^2}{T}.$$

Also,

$$\min\left\{2\nu, \sqrt{\frac{2\nu\mathcal{D}_T}{\alpha}}\right\} \le \max\left\{2, \sqrt{\frac{2C_D}{\alpha}}\right\} \min\{\nu, \sqrt{\nu\mathcal{B}_T}\}.$$

These bounds yield (8). For (9), use $2b\eta \le b(T\eta^2 + 1/T)$ to obtain

$$d_T + \sqrt{r}[(1 + bT\eta)^2 + 1]p_T \le [2\Delta_b + \sqrt{r}(b + b^2)]T\eta^2 + \frac{\sqrt{r}(2+b)}{T}.$$

For example, the same C in both main bounds can be chosen as the sum of

$$\sqrt{2(1+2R^2)}\left[\frac{C_D}{\alpha}+\sqrt{2}C_E+\max\left\{2,\sqrt{\frac{2C_D}{\alpha}}\right\}\right]+b^2(R+1)(\|G_0\|_F+LR)$$

and $2\Delta_b+\sqrt{r}(b^2+2b+2)$. This choice is independent of T,η,β,ν,B, and σ. Finally, (26) is exactly (10).

For the budget specialization, choose fixed positive constants satisfying

$$\begin{gathered}0<\beta_0\le 1,\quad \eta_0\le(8b)^{-1},\quad 2\eta_0^2\le\beta_0,\quad 2\Delta_b\eta_0^2\le\tfrac{1}{4},\\ \nu_0\ge 16\kappa R\Gamma_R\eta_0[4b\eta_0+\sqrt{3/2}\beta_0].\end{gathered}\tag{53}$$

Set $\eta=\eta_0T^{-3/4}$, $1-\beta=\beta_0T^{-1/2}$, and $\nu=\nu_0T^{-1/4}$. Then $d_T=2\Delta_b\eta_0^2T^{-1/2}$ and

$$T\eta[4b\eta+\sqrt{3/2}(1-\beta)]=4b\eta_0^2T^{-1/2}+\sqrt{3/2}\eta_0\beta_0T^{-1/4},$$

which verifies (25). In the momentum bound,

$$\frac{1}{(1-\beta)T}=O(T^{-1/2}),\qquad \frac{\eta}{1-\beta}=O(T^{-1/4}),$$
$$\sqrt{1-\beta}=O(T^{-1/4}),\qquad d_T+p_T=O(T^{-1/2}).$$

Thus $\mathcal{E}_T=O(T^{-1/4})$. By (50), we have $\mathcal{D}_T=O(T^{-1/4})$ because $(\eta T)^{-1}=O(T^{-1/4})$. Consequently, $\sqrt{\nu\mathcal{D}_T}=O(T^{-1/4})$. The terms contributed by $\{\tau\le T\}$ in (8) and (27) are $O(T^{-1/2})$, because $1+bT\eta=O(T^{1/4})$ and $p_T=O(T^{-1})$. This proves the two stated orders.

Proof of Corollary 1

At points before τ, the orthonormal representation in Appendix B satisfies

$$\|\operatorname{grad}_{\mathrm{St},U}\mathcal{L}(\bar U_t,\bar Y_t)\|_F+\|\nabla_Y\mathcal{L}(\bar U_t,\bar Y_t)\|_F\le 2\sqrt{1+2R^2}(\|H_{U,t}\|_F+\|H_{Y,t}\|_F).$$

On every other point, $\|\bar Y_t\|_F=\|W_t\|_F\le R(1+bT\eta)$, so the same sum is bounded by $[1+R(1+bT\eta)][\|G_0\|_F+LR(1+bT\eta)]$. Using (52) and the probability p_T yields

$$\begin{aligned}&\frac{1}{T}\sum_{t<T}\mathbb{E}\big[\|\operatorname{grad}_{\mathrm{St},U}\mathcal{L}(\bar U_t,\bar Y_t)\|_F+\|\nabla_Y\mathcal{L}(\bar U_t,\bar Y_t)\|_F\big]\\ &\quad\le 2\sqrt{1+2R^2}\left(\frac{\mathcal{D}_T}{\alpha}+\min\left\{2\nu,\sqrt{\frac{2\nu\mathcal{D}_T}{\alpha}}\right\}+\sqrt{2}\mathcal{E}_T\right)\\ &\qquad+[1+R(1+bT\eta)][\|G_0\|_F+LR(1+bT\eta)]p_T.\end{aligned}$$

The budget orders established above prove the corollary.

D Convergence with the exact matrix sign

We prove Theorem 2 under the assumptions and initialization in Appendix C.

At a feasible pair define $H_{U,t}=(I-U_tU_t^\top)G_tY_t$ and $H_{Y,t}=G_t^\top U_t$, with corresponding stochastic estimates $\widehat H_{U,t}$ and $\widehat H_{Y,t}$. The msign iteration is

$$\begin{aligned}M_{U,t}&=\beta\mathcal{T}_{t-1}(M_{U,t-1})+(1-\beta)\widehat H_{U,t}, & M_{Y,t}&=\beta M_{Y,t-1}+(1-\beta)\widehat H_{Y,t},\\ Q_{i,t}&=\operatorname{msign}(M_{i,t}), & \Pi_t&=Q_{U,t}^\top Q_{U,t},\\ U_{t+1}&=U_t+(\cos\eta_{U,t}-1)U_t\Pi_t-\sin\eta_{U,t}Q_{U,t}, & Y_{t+1}&=Y_t-\eta_{Y,t}Q_{Y,t}.\end{aligned}\tag{54}$$

The learning rates use (7). Since $\|U_t\|_F^2=r$, the second denominator is $r+\varepsilon$. The map $\mathcal{T}_{t-1}$ is the Grassmann parallel transport along the preceding subspace step. Its action on the stored momentum

equals the compensated projection in (16). All old states start at zero. When the basis momentum has rank less than r, Π_t retains precisely the right singular directions in which the basis moves.

For $0 < \eta \le 1$ and $0 \le \beta < 1$, put

$$G_R^{\mathrm{ex}} = \|G_0\|_F + LR, \qquad c_R^{\mathrm{ex}} = \sqrt{\frac{\varepsilon}{R^2 + r + 2\varepsilon}}, \qquad v_0 = \beta^2 \|G_0\|_F^2 + (1-\beta)^2 \sigma^2 / B.$$

For the changes of the weight and its factor gradients, define

$$\begin{aligned} C_f^{\mathrm{ex}} &= G_R^{\mathrm{ex}}(2R/3 + \sqrt{r}) + \tfrac{L}{2}(R + \sqrt{r})^2, \\ C_H^{\mathrm{ex}} &= (3R+2) G_R^{\mathrm{ex}} + L(R+1)(R+\sqrt{r}). \end{aligned} \tag{55}$$

For any such η, β, we will prove

$$\begin{aligned} &\frac{1}{T} \sum_{t<T} \mathbb{E}\big[\|\operatorname{grad}_{\mathrm{St},U} \mathcal{L}(U_t, Y_t)\|_F + \|\nabla_Y \mathcal{L}(U_t, Y_t)\|_F\big] \\ &\quad \le \frac{\sqrt{2(1+R^2)}}{c_R^{\mathrm{ex}}} \left\{ \frac{f(0) - f_*}{\eta T} + C_f^{\mathrm{ex}} \eta \right. \\ &\qquad \left. + \sqrt{2}(c_R^{\mathrm{ex}} + \sqrt{r}) \left[\frac{2\sqrt{v_0}}{(1-\beta)T} + \frac{\beta C_H^{\mathrm{ex}} \eta}{1-\beta} + \sigma \sqrt{\frac{(1-\beta)(R^2+1)}{B}} \right] \right\}. \end{aligned} \tag{56}$$

The same right-hand side bounds the two ordinary factor-gradient norms, and $U_t^\top U_t = I_r$ at every step.

Proof. Write the compact SVD of a horizontal momentum as $M_U = L\Sigma V^\top$. Then $Q_U = LV^\top$, $\Pi = VV^\top$, $U^\top Q_U = 0$, $\Pi^2 = \Pi$, and $Q_U \Pi = Q_U$. When $M_U = 0$, set $Q_U = \Pi = 0$. The update can be written

$$U_+ = U(I - \Pi) + \cos\eta_U \, U\Pi - \sin\eta_U \, Q_U.$$

The three components give $U_+^\top U_+ = (I - \Pi) + \Pi = I$. For any horizontal Z, the exact transport is

$$\mathcal{T}(Z) = Z + [(\cos\eta_U - 1) Q_U + \sin\eta_U \, U] Q_U^\top Z.$$

Set $A = Q_U^\top Z$ and $Z_\perp = Z - Q_U A$. Since $\Pi A = A$, the two terms in

$$\mathcal{T}(Z) = (\cos\eta_U \, Q_U + \sin\eta_U \, U\Pi) A + Z_\perp$$

are orthogonal, and the first left factor has Gram Π. Therefore

$$U_+^\top \mathcal{T}(Z) = 0, \qquad \|\mathcal{T}(Z)\|_F = \|Z\|_F, \qquad \|\mathcal{T}(Z) - Z\|_F \le \eta_U \|Z\|_F. \tag{57}$$

The last estimate follows from the change of a two-dimensional unit rotation, $2|\sin(\eta_U/2)| \le \eta_U$. The geodesic and parallel-transport equations are verified in Appendix A.2 (Edelman et al., 1998; Bendokat et al., 2024). Since each stochastic basis gradient is horizontal, induction from $U_0^\top U_0 = I$ proves exact feasibility and horizontal momenta throughout (54).

For every $x \in \mathbb{R}^r$, the preceding orthogonal decomposition gives

$$\|(U_+ - U)x\|_2^2 = 4\sin^2(\eta_U/2) \|\Pi x\|_2^2 \le \eta_U^2 \|x\|_2^2.$$

Since $\|Q_Y\|_2 \le 1$ and $\|Q_Y\|_F \le \sqrt{r}$,

$$\|\Delta W_t\|_F \le \eta(R + \sqrt{r}), \qquad \|\Delta Y_t\|_2 \le \eta, \qquad \|G_t\|_F \le G_R^{\mathrm{ex}}.$$

The last inequality uses $\|W_t\|_F = \|Y_t\|_F \le R$ and ambient smoothness. Expanding the actual weight difference gives

$$\Delta W + \eta_U Q_U Y^\top + \eta_Y U Q_Y^\top = [(\cos\eta_U - 1) U\Pi + (\eta_U - \sin\eta_U) Q_U] Y^\top - \eta_Y \Delta U Q_Y^\top.$$

Its Frobenius norm is bounded by $\eta^2(2R/3 + \sqrt{r})$, using $1 - \cos x \le x^2/2$, $|x - \sin x| \le x^3/6$, and $\eta \le 1$. Hence the L-smoothness inequality gives

$$f(W_{t+1}) - f(W_t) \le -\eta \sum_{i \in \{U,Y\}} \gamma_{i,t} \langle H_{i,t}, Q_{i,t} \rangle_F + C_f^{\mathrm{ex}} \eta^2. \tag{58}$$

In its first-order term, $\langle GY, Q_U\rangle_F = \langle H_U, Q_U\rangle_F$ because $U^\top Q_U = 0$.

Write $P_t = I - U_t U_t^\top$. From $P_{t+1} - P_t = -\Delta U_t U_{t+1}^\top - U_t \Delta U_t^\top$, $\|P_{t+1} - P_t\|_2 \le 2\eta$. Expanding the two fields gives

$$\begin{cases} H_{U,t+1} - H_{U,t} = (P_{t+1} - P_t) G_{t+1} Y_{t+1} + P_t (G_{t+1} - G_t) Y_{t+1} + P_t G_t (Y_{t+1} - Y_t), \\ H_{Y,t+1} - H_{Y,t} = (G_{t+1} - G_t)^\top U_{t+1} + G_t^\top (U_{t+1} - U_t). \end{cases}$$

The norm of each summand is bounded using $\|G_{t+1} - G_t\|_F \le L\eta(R + \sqrt{r})$ and $\|G_t\|_F \le G_R^{\mathrm{ex}}$. Thus

$$\begin{cases} \|H_{U,t+1} - H_{U,t}\|_F \le \eta[(2R+1)G_R^{\mathrm{ex}} + LR(R+\sqrt{r})], \\ \|H_{Y,t+1} - H_{Y,t}\|_F \le \eta[G_R^{\mathrm{ex}} + L(R+\sqrt{r})]. \end{cases}$$

In addition, (57) implies $\|\mathcal{T}_t(H_{U,t}) - H_{U,t}\|_F \le \eta R G_R^{\mathrm{ex}}$. The sum of the three displayed coefficients is exactly C_H^{ex} in (55). Thus

$$\|\mathcal{T}_t(H_{U,t}) - H_{U,t+1}\|_F^2 + \|H_{Y,t} - H_{Y,t+1}\|_F^2 \le (C_H^{\mathrm{ex}})^2 \eta^2.$$

Let $\mu_t^2 = \|M_{U,t} - H_{U,t}\|_F^2 + \|M_{Y,t} - H_{Y,t}\|_F^2$. Subtract the true fields from each momentum update:

$$\begin{cases} M_{U,t} - H_{U,t} = \beta \mathcal{T}_{t-1}(M_{U,t-1} - H_{U,t-1}) + \beta[\mathcal{T}_{t-1}(H_{U,t-1}) - H_{U,t}] + (1-\beta)(\widehat{H}_{U,t} - H_{U,t}), \\ M_{Y,t} - H_{Y,t} = \beta(M_{Y,t-1} - H_{Y,t-1}) + \beta(H_{Y,t-1} - H_{Y,t}) + (1-\beta)(\widehat{H}_{Y,t} - H_{Y,t}). \end{cases}$$

The first two terms are measurable before sample t. The current noise has zero conditional mean and conditional squared norm at most $(R^2+1)\sigma^2/B$ for the matrix pair. Using the isometry in (57) and

$$\|\beta A + \beta B\|_F^2 = \left\| \beta A + (1-\beta)\frac{\beta B}{1-\beta} \right\|_F^2 \le \beta \|A\|_F^2 + \frac{\beta^2}{1-\beta}\|B\|_F^2,$$

we obtain

$$\mathbb{E}\mu_t^2 \le \beta \mathbb{E}\mu_{t-1}^2 + \frac{\beta^2 (C_H^{\mathrm{ex}})^2 \eta^2}{1-\beta} + (1-\beta)^2 (R^2+1)\sigma^2/B. \tag{59}$$

At initialization $H_{U,0} = M_{U,0} = 0$ and $M_{Y,0} - H_{Y,0} = -\beta G_0^\top U_0 + (1-\beta)(\widehat{G}_0 - G_0)^\top U_0$. By conditional unbiasedness,

$$\begin{aligned} \mathbb{E}\mu_0^2 &= \beta^2 \mathbb{E}\|G_0^\top U_0\|_F^2 + (1-\beta)^2 \mathbb{E}\|(\widehat{G}_0 - G_0)^\top U_0\|_F^2 \\ &\le \beta^2 \|G_0\|_F^2 + (1-\beta)^2 \sigma^2/B = v_0. \end{aligned}$$

Iterating (59),

$$\mathbb{E}\mu_t^2 \le \beta^t v_0 + \frac{\beta^2 (C_H^{\mathrm{ex}})^2 \eta^2}{(1-\beta)^2} + (1-\beta)(R^2+1)\sigma^2/B.$$

Take square roots, sum, and use $\sum_{t<T} \beta^{t/2} \le (1-\sqrt{\beta})^{-1} \le 2/(1-\beta)$. This gives

$$\frac{1}{T}\sum_{t<T} \sqrt{\mathbb{E}\mu_t^2} \le \frac{2\sqrt{v_0}}{(1-\beta)T} + \frac{\beta C_H^{\mathrm{ex}} \eta}{1-\beta} + \sigma \sqrt{\frac{(1-\beta)(R^2+1)}{B}}. \tag{60}$$

For compact msign, $\langle M_i, Q_i\rangle_F = \|M_i\|_* \ge \|M_i\|_F$ and $\|Q_i\|_F \le \sqrt{r}$. The Cauchy–Schwarz allocation argument in (45), now with $w_Y = r + \varepsilon$, gives

$$\begin{aligned} \sum_i \gamma_i \langle H_i, Q_i\rangle_F \ge c_R^{\mathrm{ex}} (\|H_U\|_F + \|H_Y\|_F) \\ - \sqrt{2}(c_R^{\mathrm{ex}} + \sqrt{r}) \sqrt{\|M_U - H_U\|_F^2 + \|M_Y - H_Y\|_F^2}. \end{aligned}$$

Substitute this inequality in (58), take expectations, and sum from $t = 0$ to $T-1$. Since $f(W_0) = f(0)$ and $f(W_T) \ge f_*$,

$$\frac{1}{T}\sum_{t<T} \mathbb{E}(\|H_{U,t}\|_F + \|H_{Y,t}\|_F) \le \frac{1}{c_R^{\mathrm{ex}}} \left[\frac{f(0) - f_*}{\eta T} + C_f^{\mathrm{ex}} \eta + \frac{\sqrt{2}(c_R^{\mathrm{ex}} + \sqrt{r})}{T} \sum_{t<T} \sqrt{\mathbb{E}\mu_t^2} \right].$$

By (19), the sum of the two ordinary factor-gradient norms is at most $\sqrt{2(1+R^2)}(\|H_U\|_F + \|H_Y\|_F)$. Stiefel projection can only decrease the first norm. Applying (60) proves (56).

Finally, for $\eta = \eta_0 T^{-3/4}$ and $1-\beta = \beta_0 T^{-1/2}$, $(\eta T)^{-1}$, $\eta/(1-\beta)$, and $\sqrt{1-\beta}$ are $O(T^{-1/4})$, while $[(1-\beta)T]^{-1} = O(T^{-1/2})$. The constants in (55) and the radius are independent of T. This proves Theorem 2. □

E FINITE NEWTON–SCHULZ COMPUTATION

Numerical realization. PLUM computes each matrix-sign direction with a fixed Newton–Schulz polynomial (Jordan et al., 2024; Amsel et al., 2026). For an input $M \in \mathbb{R}^{d\times r}$, we initialize $X_0 = M/(\|M\|_F + \nu)$ and apply

$$A_j = X_j^\top X_j, \qquad X_{j+1} = X_j\big(a_j I_r + b_j A_j + c_j A_j^2\big), \qquad j = 0, \dots, K-1.$$

The returned direction is $Q = X_K$. Since the factor momenta in PLUM are thin matrices, each iteration forms an $r \times r$ Gram matrix and then applies matrix multiplications only. No QR, SVD, matrix inverse, or data-dependent stopping criterion is required.

Let $p_j(x) = a_j x + b_j x^3 + c_j x^5$ and $\phi_K = p_{K-1} \circ \cdots \circ p_0$. If $M = L\Sigma V^\top$ is a compact SVD, the iteration above gives

$$Q = L\,\phi_K\!\left(\frac{\Sigma}{\|M\|_F + \nu}\right) V^\top.$$

Thus the polynomial acts only on the singular values while preserving the singular subspaces. In particular, $\mathrm{col}(Q) \subseteq \mathrm{col}(M)$, so applying NS to a horizontal basis momentum preserves horizontality in exact arithmetic (Higham, 1986; Amsel et al., 2026).

Fixed polynomial schedules. Table 5 lists the fixed schedules considered in this work together with the global polynomial certificates used in the analysis. KJ5 applies Keller Jordan's quintic coefficients $(3.4445, -4.7750, 2.0315)$ for five iterations. PE8 uses the fixed eight-step Polar Express sequence listed in Table 6 (Jordan et al., 2024; Amsel et al., 2026; Cesista et al., 2026). We use PE8 for all the Muon-based methods in our experiments.

Table 5: Fixed NS schedules and global certificates $\alpha x \le \phi_K(x) \le \kappa x$ and $0 \le \phi_K(x) \le b$ for all $x \in [0,1]$.

Schedule	α	b	κ
Classical cubic NS, K steps	1	1	$(3/2)^K$
Classical quintic NS, K steps	1	1	$(15/8)^K$
KJ5	0.69	1.203	485
PE8	1	1.0004	20776

Table 6: Fixed eight-step Polar Express coefficients used by PE8.

j	a_j	b_j	c_j
0	7.2086	−15.5131	9.0178
1	3.9623	−2.5813	0.4542
2	3.9466	−2.5765	0.4544
3	3.8991	−2.5671	0.4566
4	3.7186	−2.5308	0.4653
5	3.1390	−2.3073	0.4733
6	2.1715	−1.5246	0.3885
7	1.8648	−1.2224	0.3577

For classical cubic NS, $p(x) = x(3-x^2)/2$ maps $[0,1]$ into itself and satisfies $x \le p(x) \le 3x/2$. The classical quintic $p(x) = x(15-10x^2+3x^4)/8$ similarly satisfies $x \le p(x) \le 15x/8$. Composition gives the corresponding certificates in Table 5. For KJ5 and PE8, we verify the bounds over the full interval $[0,1]$ using interval arithmetic with the printed coefficients. Here, κ is the uniform linear-growth constant of ϕ_K on $[0,1]$.

Finite-NS direction residual. For one NS call,

$$\|Q^\top Q - I_r\|_F^2 = \sum_{j=1}^{r} \left[\phi_K\left(\frac{\sigma_j(M)}{\|M\|_F + \nu} \right)^2 - 1 \right]^2 \le r \max\{1, b^2 - 1\}^2. \tag{61}$$

Zero singular values contribute one to the corresponding terms. Hence (61) gives $\Delta_b = \sqrt{r}\max\{1, b^2 - 1\}$ used in Theorem 1. The three certificate constants enter the finite-NS analysis in different ways. The lower bound α controls the momentum–direction alignment in (23), b bounds the direction norm and its Gram residual, and κ controls the propagation of nonhorizontal components through (24). Since the NS schedule and iteration count are fixed, these quantities are independent of the training horizon.

Batched implementation. The coefficient sequence and iteration count are fixed before training. Same-shaped basis momenta and coefficient momenta from different layers can therefore be stacked separately and processed by the same sequence of batched GEMMs. For a batch of matrices, the normalization and Gram products are computed independently along the last two dimensions, and each NS iteration applies the corresponding $r \times r$ polynomial on the right. This fixed-depth computation is the NS primitive used for both Q_U and Q_Y in Algorithm 1.

F Experimental protocol

F.1 Pretraining and adaptation settings

The C4 reference protocol follows LORO Sections 5.1 and A.1, which in turn use the GaLore architecture family (Mo et al., 2025; Zhao et al., 2024). The maximum sequence length is 256. The query, key, value, and output attention projections and the feed-forward gate, up, and down projections are factorized. The embedding, output head, and normalization modules use the regular optimizer group. Table 7 records the reference dimensions and budgets. Its token counts are the counts reported by the source protocol.

Table 7: C4 reference architectures and training settings. "Reference" follows the benchmark model naming. "Dense params" is the corresponding unfactorized parameter count reported by LORO.

Reference	Dense params	Hidden	Intermediate	Layers	Rank	Updates	Tokens
60M	58M	512	1376	8	128	10K	1.1B
130M	134M	768	2048	12	256	20K	2.2B
350M	368M	1024	2736	24	256	60K	6.4B
1B	1339M	2048	5461	32	512	100K	13.1B

The benchmark learning-rate schedule uses linear warmup over the first 10% of updates, followed by cosine decay to 10% of the peak. Peak rates are tuned separately for each optimizer. These practical schedules are distinct from the constant-horizon choices used to state the stochastic complexity theorem. The factor initialization and any layer-specific multipliers are recorded with the run configuration. PLUM initializes the basis with the orthonormal factor of a reduced QR factorization computed in fp32. The coefficient correction starts at zero for adapter training. A given nonzero factorized weight can instead be converted once without changing the represented function: if $B = UR$ is a thin QR factorization, set $Y = A^\top R^\top$ so that $BA = UY^\top$. This is an initialization conversion, not an iterative retraction. Xavier initialization of the reference low-rank baselines follows the source protocol (Glorot & Bengio, 2010; Mo et al., 2025).

For fair comparison, we use the same learning rate search grid $\{1e-4, 5e-4, 1e-3, 5e-3, 1e-2, 5e-2, 1e-1\}$ for all the methods. We deliberately keep the range large as the best learning rates for Muon-based methods are typically larger than those of AdamW-based ones. We choose the best LR based on the final validation perplexity. All the other hyperparameters are kept the same as reported in Table 8. For PLUM, we find that the best LR is consistent across different model scales ($5e-2$).

Table 8: Training configuration for C4 pretraining.

Setting	**60M**	**130M**	**350M**	**1B**
Schedule (identical across methods)				
Training steps	10,000	20,000	60,000	100,000
Warmup steps	1,000	2,000	6,000	10,000
LR scheduler	cosine, min-LR ratio 0.1			
Weight decay	0			
Grad clipping	none			
Batch layout (effective batch size $= 512$ *in all cases)*				
Micro-batch $\times$ GPUs	256×1	256×1	128×4	64×8
Gradient accumulation	2	2	1	1
Low-rank factorization				
Rank r (attn + mlp)	128	256	256	512
Data and evaluation (shared)				
Dataset / tokenizer	C4-en / T5-base			
Sequence length	256			
Precision	bfloat16			
Evaluation	every 1,000 steps ($\sim$10M val. tokens); PPL $= \exp$(eval loss)			

For GLUE, the pretrained RoBERTa-base weights are frozen while the prescribed low-rank corrections are trained. Each selected query or value weight takes the form $W = W_0 + \gamma UY^\top$, with fixed adapter scale γ. The factor gradients include this multiplier through autograd. We conduct comparisons at ranks $r = 4$ and $r = 8$. We sweep the learning rate using the grid $\{1e-4, 2e-4, 4e-4, 6e-4, 8e-4, 1e-3, 2e-3\}$ on each task for AdamW-based methods (GaLore, LoRA), and multiply each element by 10 for Muon-based methods (Spectron, PLUM).

Table 9: Task budgets used by the LORO GLUE protocol (Mo et al., 2025). These entries are settings, not benchmark outcomes.

	CoLA	STS-B	MRPC	RTE	SST-2
Batch size	32	16	16	16	16
Epochs	30	30	30	30	30

The reported metrics are Matthews correlation for CoLA, Pearson correlation for STS-B, F1 for MRPC, and accuracy for the other tasks, following LORO. Correlation scores use the same percentage scale as the aggregate score.

F.2 Efficiency and mechanism measurements

For pretraining, parameter storage counts the matrices actually retained by each method, not only the number of trainable entries. In adaptation, the frozen backbone is counted separately from the trainable correction and optimizer states. Parameter-plus-state memory and peak allocated memory are reported as different quantities. In the PLUM pretraining implementation, factors can use bf16 storage, while momentum, normalization, inner products, and geometric operations use fp32.

Timing uses identical hardware, accumulation, precision, compilation, and synchronization settings within a comparison. Compilation and warmup are excluded from steady-state timing, while stacking, gathering, and writeback during the optimizer step are included. Coefficient norms, orthogonality residuals, and update diagnostics are recorded outside the timed steps.

The step-allocation ablation compares PLUM with an otherwise identical implementation using equal factor learning rates. Recorded quantities include both factor step sizes, the factor norms, the two momentum–update inner products, and the Frobenius norm of the actual weight difference. Zero-momentum updates use a zero step in both implementations.

Table 10: Non-learning-rate hyperparameters used for GLUE fine-tuning. A dash denotes a parameter that is not applicable to the corresponding method.

Hyperparameter	GaLore	LoRA	Spectron	PLUM
Shared training configuration				
Backbone	`roberta-base`			
Epochs	30			
Maximum sequence length	512			
Training batch size per GPU	32 for CoLA; 16 for all other tasks			
Evaluation batch size per GPU	128			
Gradient accumulation steps	1			
Processes per run	1 (one GPU per run)			
Scheduler	Linear			
Warmup steps	0			
Global weight decay	0			
Random seed	0			
Training precision	FP32; mixed precision disabled			
Padding	Dynamic padding			
Training sampler	Shuffled			
Evaluation frequency	Once per epoch			
Model selection	Highest validation metric over all 30 epochs			
Evaluation metrics	CoLA: Matthews correlation; STS-B: Pearson; MRPC: F1; RTE, SST-2, and QNLI: accuracy			
Parameterization and optimization				
Task coverage	CoLA, STS-B, MRPC, RTE, SST-2	CoLA, STS-B, MRPC, RTE, SST-2	CoLA, STS-B, MRPC, RTE, SST-2	CoLA, STS-B, MRPC, RTE, SST-2
Low-rank mechanism	Low-rank projection of dense gradients	PEFT LoRA adapters	Native low-rank adapters with Spectron updates	Native manifold low-rank adapters with PLUM updates
Target modules	Self-attention `query.weight` in all 12 layers	Self-attention `query` and `value` in all 12 layers	Self-attention `query` and `value` in all 12 layers	Self-attention `query` and `value` in all 12 layers
Rank	4; 8	4; 8	4; 8	4; 8
Alpha	–	4 at rank 4; 2 at rank 8	4 at rank 4; 2 at rank 8	4 at rank 4; 2 at rank 8
Adapter scale α/r	–	1 at rank 4; 0.25 at rank 8	1 at rank 4; 0.25 at rank 8	1 at rank 4; 0.25 at rank 8
Initialization	Pretrained dense weights; projection constructed from gradients	PEFT default (A: Kaiming; B: zero)	Top-r SVD with function-preserving residual adjustment	Random orthogonal U, and $Y = 0$
Trainable backbone parameters	Target `query.weight` matrices only	LoRA A/B factors only	Low-rank factors only	Low-rank factors only
Classification head	Fully trainable	Fully trainable	Fully trainable	Fully trainable
Low-rank optimizer	GaLoreAdamW	PyTorch AdamW	Spectron	PLUM
Classification-head optimizer	Regular AdamW parameter group in GaLoreAdamW	AdamW	AdamW	AdamW
Adam betas	$(0.9, 0.999)$	$(0.9, 0.999)$	$(0.9, 0.999)$ for the head	$(0.9, 0.999)$ for the head
Adam epsilon	10^{-6}	10^{-8}	10^{-6} for the head	10^{-6} for the head
Adam bias correction	Enabled	Enabled	Enabled	Enabled
Adapter dropout	0	0	0	0
Target-layer bias	Frozen	Frozen	Frozen	Frozen