GRAM–SCHMIDT BREAKS THE GRADIENT STRUCTURE OF RIEMANNIAN GRADIENT DESCENT: FIRST-ORDER RETRACTIONS ON THE STIEFEL MANIFOLD

Anonymous authors
Paper under double-blind review

ABSTRACT

Orthogonality constraints are commonly found in deep learning, appearing in applications such as unitary recurrent networks, orthogonal convolutions, and Stiefel-constrained layers. These constraints are typically enforced by retracting each update using a thin QR factorization, which involves Gram-Schmidt orthogonalization. However, QR factorization is known to be only a first-order retraction. We demonstrate that this has structural implications for the optimizer itself. Specifically, for Riemannian gradient descent on the Stiefel manifold, we derive, in closed form, the tangential defect introduced by QR factorization at second order in the step size. Notably, this defect disappears when using the polar retraction. Furthermore, we prove that using explicit objectives validated through exact rational arithmetic, the 1-form dual to this defect is not closed. As a result, gradient descent employing QR factorization does not support any modified loss at first order in the step size, in contrast to methods based on the exponential map and other second-order retractions. Since the defect vanishes at critical points, it goes unnoticed in standard local convergence and stability analyses. Additionally, because Gram-Schmidt processes columns sequentially, the optimization based on QR is not equivariant under permutations of the weight columns. This means that in a Stiefel-constrained convolutional network and an orthogonal recurrent network, permuted copies of the same model diverge within the first few updates, while updates based on polar retraction remain consistent to machine precision. We cannot determine whether this lack of equivariance affects final accuracy since, over the course of full training, even the polar control eventually loses equivariance due to accumulated rounding errors. Our primary concern is with well-posedness and reproducibility. Lastly, we show that the polar retraction can be computed by a short Newton–Schulz iteration built entirely from matrix multiplications. This method restores equivariance at the tested sizes and performs faster than thin QR factorization at larger dimensions.

1 INTRODUCTION

Orthogonality constraints are a standard component in modern deep learning. Unitary and orthogonal recurrent networks constrain the recurrent weight matrix to prevent vanishing and exploding gradients across long sequences (Arjovsky et al., 2016; Wisdom et al., 2016); orthogonal convolutions and orthogonality regularizers improve conditioning, training stability, and representation robustness (Bansal et al., 2018; Wang et al., 2020); and Stiefel-constrained layers are trained via specialized Riemannian optimizers (Li et al., 2020). In nearly every scenario, the constraint is enforced in the same way: take an unconstrained or tangent-gradient step, then project back onto the constraint set via thin QR factorization, i.e., Gram-Schmidt orthogonalization of the columns.

This paper demonstrates that this choice of projection is not innocuous. Fundamentally, it alters the geometry of what is being optimized in a manner invisible to conventional local analysis, while coupling the optimization trajectory to an arbitrary labeling of the weight columns.

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.

Let $\mathcal{M}$ be a Riemannian submanifold embedded in an ambient Euclidean space, and let $f : \mathcal{M} \to \mathbb{R}$ be a smooth objective function. Riemannian gradient descent (RGD) updates parameters according to

$$X_{k+1} = R_{X_k}(-\eta \operatorname{grad} f(X_k)), \tag{1}$$

where $R_X : T_X\mathcal{M} \to \mathcal{M}$ is a retraction map that satisfies $R_X(0) = X$ and $\mathrm{D}R_X(0) = \mathrm{id}_{T_X\mathcal{M}}$ (Absil et al., 2008; Adler et al., 2002; Boumal, 2023). Retractions circumvent the Riemannian exponential map, which is computationally prohibitive or analytically unavailable on most manifolds of practical interest. Among the available retractions, the thin QR retraction $R_X(V) = \mathrm{qf}(X + V)$ is by far the most widely adopted due to the low per-step overhead.

Backward error analysis (BEA) seeks a modified vector field whose continuous flow interpolates the discrete updates. In Euclidean space $\mathbb{R}^n$, Barrett & Dherin (2021) demonstrated that gradient descent tracks to $\mathcal{O}(\eta^2)$, the gradient flow of an augmented energy functional $f + \frac{\eta}{4}\|\operatorname{grad} f\|^2$. The key property is structural: the $\mathcal{O}(\eta)$ correction is itself a conservative gradient field, ensuring that a modified loss exists and preserving the standard descent dynamics. Recently, Halmos & Hanin (2025) established the Riemannian analog for the geodesic update $X_{k+1} = \exp_{X_k}(-\eta \operatorname{grad} f(X_k))$, deriving a modified Riemannian energy.

This paper investigates what occurs when $\exp$ is replaced by retractions actually executed on computational hardware. We show that the existence of a modified loss is an artifact of the exponential map and of second-order retractions, and is false for first-order retractions such as thin QR. Because Gram-Schmidt processes columns sequentially, the resulting defect fundamentally breaks gradient structure, induces non-conservative rotational drift, and destroys permutation equivariance across weight columns in deep vision and recurrent sequence models.

Contributions.

1. **Closed-Form Tangential Defect.** We derive the order-η^2 tangential defect τ_{qf} of the thin QR retraction on the Stiefel manifold in closed form (Theorem 1), validate it against Richardson-extrapolated finite differences on six Stiefel manifolds, and confirm it by predicting an independently measured loop circulation within a relative 4×10^{-5}.

2. **Critical-Point Invariance.** We prove that the tangential defect vanishes at critical points, making the Jacobian of the RGD iteration retraction-independent there and explaining why standard local convergence and saddle-escape analyses do not detect the defect.

3. **Breakdown of Gradient Structure.** We show through exterior calculus and an exact rational witness that the induced order-η correction 1-form is non-conservative ($d\omega \neq 0$), ruling out an order-η modified continuous loss for RGD under thin QR.

4. **Breakdown of the Permutation Equivariance.** We show that sequential column processing in thin QR breaks the column-permutation equivariance, whereas the polar factor preserves it (Proposition 6). We confirm this in a Stiefel-constrained CNN and in an orthogonal RNN — the latter in the square case $p = n$, which lies outside the scope of our non-closedness theory — under exact double-precision controls.

5. **A GEMM-Only Remedy.** We show that the input to a retraction is already nearly orthonormal (Lemma 7). Hence, an *unscaled* Newton–Schulz iteration computes the polar factor with matrix multiplications alone and converges quadratically. We show that the norm-bound scaling standard for general matrices breaks this iteration. On the GPU, it is faster than the thin QR in the largest dimensions we tested (Table 1).

Scope and Theoretical Regime. Our core non-integrability analysis focuses on the regime $p \leq n - 2$, where the Stiefel manifold $\mathrm{St}(p, n)$ is simply connected, ensuring the failure of closedness ($d\omega \neq 0$) yields non-zero circulation along contractible loops via Stokes' theorem. In Section 7, we delimit these boundaries by contrasting this setting with the Lie-group case $p = n$ (where QR reduces to a left-invariant group action) and the non-simply connected boundary $p = n - 1$. Our orthogonal-RNN experiment (Section 6.3) lies in the case $p = n$: it illustrates the permutation breakdown of Proposition 6, which holds for every $p \geq 2$, but not the non-closedness result.

2 Related Work

Retractions on Matrix Manifolds. Retractions serve as a computationally efficient alternative to substitute for computationally intensive exponential map optimization (Absil et al., 2008; Adler et al., 2002; Boumal, 2023; Edelman et al., 1998). Absil & Malick (2012) characterized second-order retractions by the condition that tangential acceleration vanishes ($\tau_R \equiv 0$). Although the polar factor satisfies this property, the thin QR is only a first-order retraction. Higher-order retractions have been constructed systematically (Gawlik & Leok, 2018; Barbero-Liñán & Martín de Diego, 2023), but standard deep learning libraries predominantly implement first-order QR factorizations due to per-step efficiency.

Geometric Backward Error Analysis. Backward error analysis bridges discrete numerical integrators and modified continuous-time systems (Hairer et al., 2006; Reich, 1999). Although intrinsic formulations exist for Lie groups and homogeneous spaces (Faltinsen, 2000; Lundervold & Munthe-Kaas, 2013), and general classification theorems exist for transformation semigroups that preserve geometric invariants (Hansen, 2011), gradient flows fail to form closed semigroups under composition. Consequently, general manifold BEA results do not imply the preservation of gradient structures.

Implicit Regularization. Modified continuous-time equations explain the inductive bias in the Euclidean gradient descent and SGD (Barrett & Dherin, 2021; Smith et al., 2021). Extending these frameworks to manifolds using geodesics captures modified losses for idealized exponential updates (Halmos & Hanin, 2025), but these guarantees fail when using first-order retractions.

3 Preliminaries

Let $\mathrm{St}(p,n) = \{X \in \mathbb{R}^{n\times p} : X^\top X = I_p\}$ denote the compact Stiefel manifold of orthonormal p-frames in $\mathbb{R}^n$, endowed with the Riemannian metric induced by the ambient Euclidean/Frobenius inner product, $\langle A, B\rangle = \mathrm{tr}(A^\top B)$. The tangent space at a point $X \in \mathrm{St}(p,n)$ is defined as

$$T_X\mathrm{St}(p,n) = \{Z \in \mathbb{R}^{n\times p} : \mathrm{sym}(X^\top Z) = 0\}, \tag{2}$$

where $\mathrm{sym}(M) = \frac{1}{2}(M + M^\top)$ denotes the symmetric part of a square matrix. The orthogonal projection of an ambient matrix $Z \in \mathbb{R}^{n\times p}$ onto the tangent space $T_X\mathrm{St}(p,n)$ is given by

$$P_X(Z) = Z - X\,\mathrm{sym}(X^\top Z). \tag{3}$$

For any matrix $M \in \mathbb{R}^{p\times p}$, let $\mathrm{up}(M)$ denote its strictly upper triangular part and $\mathrm{diag}(M)$ its diagonal part, so that $M = \mathrm{diag}(M) + \mathrm{up}(M) + \mathrm{up}(M)^\top$ whenever M is symmetric. For an ambient matrix $M \in \mathbb{R}^{n\times p}$ with full column rank, let $\mathrm{qf}(M)$ denote the orthogonal factor Q from its thin QR decomposition normalized to have strictly positive diagonal entries in R, and let $\mathrm{pf}(M) = UV^\top$ denote the polar factor derived from its thin singular value decomposition $M = U\Sigma V^\top$.

A smooth mapping $R : T\mathcal{M} \to \mathcal{M}$ is a *retraction* if for every $X \in \mathcal{M}$, the restriction $R_X : T_X\mathcal{M} \to \mathcal{M}$ satisfies the zeroth and first-order consistency conditions: $R_X(0) = X$ and $\mathrm{D}R_X(0) = \mathrm{id}_{T_X\mathcal{M}}$.

Definition 1 (Tangential Defect)**.** For a retraction R on a Riemannian submanifold $\mathcal{M}$ and a tangent vector $V \in T_X\mathcal{M}$, let $a_R(X)[V,V] := \frac{d^2}{dt^2}\Big|_{t=0} R_X(tV)$ denote the ambient second-order acceleration. The tangential defect of R at (X,V) is the projection of this acceleration onto the tangent space:

$$\tau_R(X)[V,V] := P_{T_X\mathcal{M}}\left(\frac{d^2}{dt^2}\Big|_{t=0} R_X(tV)\right) = P_{T_X\mathcal{M}}\big(a_R(X)[V,V]\big). \tag{4}$$

A retraction is second order if and only if $\tau_R \equiv 0$ identically across $T\mathcal{M}$ (Absil & Malick, 2012). Taylor-expanding the Riemannian gradient update $X_{k+1} = R_X(-\eta\,\mathrm{grad}\,f(X))$ in step size η yields

$$R_X(-\eta\,\mathrm{grad}\,f(X)) = X - \eta\,\mathrm{grad}\,f(X) + \frac{\eta^2}{2}a_R(X)[\mathrm{grad}\,f(X), \mathrm{grad}\,f(X)] + \mathcal{O}(\eta^3). \tag{5}$$

Projecting onto the tangent space $T_X\mathcal{M}$ isolates the tangential contribution at order η^2:

$$P_X(X_{k+1} - X) = -\eta\,\mathrm{grad}\,f(X) + \frac{\eta^2}{2}\tau_R(X)[\mathrm{grad}\,f(X), \mathrm{grad}\,f(X)] + \mathcal{O}(\eta^3). \tag{6}$$

The tangential defect τ_R is precisely the structural perturbation that enters the order-η term of the continuous modified vector field in backward error analysis. For any second-order retraction—such as the Riemannian exponential map or the polar factor—this term vanishes identically, whereas for first-order retractions it drives the non-gradient drift investigated in this work.

4 The Tangential Defect and Breakdown of Gradient Structure

4.1 Closed-Form Defect of the QR Retraction

Theorem 1 (Tangential Defect of QR). *Let $X \in \mathrm{St}(p,n)$, $V \in T_X\mathrm{St}(p,n)$, and $S = V^\top V \in \mathbb{R}^{p\times p}$. The tangential defect of the thin QR retraction $R_X(V) = \mathrm{qf}(X+V)$ is*

$$\tau_{\mathrm{qf}}(X)[V,V] = X\big(\mathrm{up}(S)^\top - \mathrm{up}(S)\big). \tag{7}$$

Because $\mathrm{up}(S)^\top - \mathrm{up}(S)$ is skew-symmetric, $\tau_{\mathrm{qf}}(X)[V,V] \in T_X\mathrm{St}(p,n)$. In contrast, polar retraction $R_X(V) = \mathrm{pf}(X+V)$ satisfies $\tau_{\mathrm{pf}} \equiv 0$ identically across $T\mathrm{St}(p,n)$.

When $p = 1$, S is a scalar, so $\mathrm{up}(S) = 0$ and $\tau_{\mathrm{qf}} \equiv 0$, recovering the known result that vector normalization is a second-order retraction on spheres. For $p \geq 2$, $\tau_{\mathrm{qf}} = 0$ if and only if $S = V^\top V$ is diagonal.

Numerical Verification. We evaluate Theorem 1 against empirical accelerations extracted by the finite-difference quotient

$$\widehat{\tau}(X,V) = P_X\left(\frac{R_X(-\eta g) - X + \eta g}{\frac{1}{2}\eta^2}\right), \qquad g = \mathrm{grad}\, f(X), \tag{8}$$

Richardson-extrapolated as $2\widehat{\tau}(\eta/2) - \widehat{\tau}(\eta)$ to cancel its $\mathcal{O}(\eta)$ error. Across 18 test problems spanning $\mathrm{St}(2,4)$, $\mathrm{St}(2,5)$, $\mathrm{St}(3,5)$, $\mathrm{St}(3,6)$, $\mathrm{St}(3,7)$ and $\mathrm{St}(4,8)$, the relative discrepancy $\|\widehat{\tau} - \tau_{\mathrm{qf}}\|_F / \|\tau_{\mathrm{qf}}\|_F$ is at most 1.1×10^{-4}, the truncation floor of the scheme. Under the same protocol the polar defect has norms between 6×10^{-9} and 3×10^{-4}, pure roundoff, consistent with $\tau_{\mathrm{pf}} \equiv 0$.

4.2 Critical Point Invariance

Lemma 2 (Invariance at Critical Points). *Let $X^* \in \mathrm{St}(p,n)$ be a critical point of a smooth function f ($\mathrm{grad}\, f(X^*) = 0$), and let $\Phi(X) = R_X(-\eta\, \mathrm{grad}\, f(X))$ denote the discrete RGD iteration map. Then for every retraction R,*

$$\mathrm{D}\Phi(X^*) = I_{T_{X^*}\mathcal{M}} - \eta\, \mathrm{Hess}\, f(X^*). \tag{9}$$

Lemma 2 explains why local asymptotic convergence rates, saddle-point escape dynamics, and linear stability thresholds are invariant under the choice of retraction. On $\mathrm{St}(2,4)$, empirical Jacobians satisfy $\|J_{\mathrm{qf}} - J_{\mathrm{pf}}\|_F = 0$ to machine precision, with both schemes losing stability at the identical threshold $\eta_c = 2/\lambda_{\max}(\mathrm{Hess}\, f(X^*)) = 0.119536$. The defect is strictly a non-equilibrium phenomenon operating along transient trajectories where $\mathrm{grad}\, f \neq 0$.

4.3 Non-Conservativity of the Defect Field

Let ω denote the differential 1-form on $\mathrm{St}(p,n)$ dual to the tangential defect field:

$$\omega_X(Z) := \big\langle \tau_{\mathrm{qf}}(X)[\mathrm{grad}\, f(X), \mathrm{grad}\, f(X)],\, Z \big\rangle, \quad Z \in T_X\mathrm{St}(p,n). \tag{10}$$

If τ_{qf} were a gradient vector field, ω would be exact, requiring $d\omega = 0$. For $p = 2$, $S = g^\top g$ is 2×2 symmetric with scalar off-diagonal entry $s(X) := S_{12} = \langle g_1, g_2\rangle$. Here, $\mathrm{up}(S)^\top - \mathrm{up}(S) = s(X)J$, where $J = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$. The defect field and its dual 1-form simplify to

$$\tau_{\mathrm{qf}}(X) = s(X)XJ, \qquad \omega = s\,\alpha, \quad \text{where } \alpha_X(Z) = \langle XJ, Z\rangle. \tag{11}$$

Lemma 3. *For $Y, Z \in T_X\mathrm{St}(2,n)$, let $Y^\top Z - Z^\top Y = m(Y,Z)J$. Then $d\alpha(Y,Z) = 2\,m(Y,Z)$. Thus, α is not closed.*

Applying the exterior Leibniz product rule $d\omega = ds \wedge \alpha + s\, d\alpha$ yields

$$d\omega(Y, Z) = (\mathrm{D}_Y s)\alpha(Z) - (\mathrm{D}_Z s)\alpha(Y) + 2s(X)m(Y, Z). \tag{12}$$

Theorem 4 (Breakdown of Gradient Structure)**.** *The defect 1-form ω is not closed on $\mathrm{St}(2, 4)$ and $\mathrm{St}(3, 6)$. Specifically, there exist configurations where $d\omega(Y, Z) \neq 0$ in exact rational arithmetic. Consequently, τ_{qf} is not a gradient vector field, and RGD with QR admits no modified energy functional at order η.*

The exact rational evaluation gives $d\omega(Y, Z) = 352$ on $\mathrm{St}(2, 4)$ and $d\omega(Y, Z) = 1253/2$ on $\mathrm{St}(3, 6)$; both witnesses are given in the Appendix A.

Corollary 5. *For $2 \leq p \leq n - 2$, there exist contractible loops $\gamma \subset \mathrm{St}(p, n)$ with non-zero circulation: $\oint_\gamma \omega \neq 0$. Thus, backward error analysis cannot generate an order-η modified loss for RGD under QR, while the modified energy framework of Halmos & Hanin (2025) extends verbatim to any second-order retraction.*

Numerically, on the closed loop $X(t) = e^{tK} X_0$ in $\mathrm{St}(2, 4)$, whose velocity $KX(t)$ is exact, the closed form of Theorem 1 gives a circulation of -15.3289 against a finite-difference value of -15.3295, a relative difference of 4×10^{-5}. A control field that is a gradient by construction returns -1.53×10^{-13}, confirming that the measurement can vanish, and the polar defect returns -2.43×10^{-5}, the finite-difference noise level.

5 Structural and Dynamical Consequences

5.1 Breakdown of Permutation Equivariance

Let $\Pi \in \mathbb{R}^{p \times p}$ be a column permutation matrix. An objective f is permutation-invariant if $f(X\Pi) = f(X)$.

Proposition 6 (Permutation Equivariance)**.** *Polar retraction satisfies $\mathrm{pf}(M\Pi) = \mathrm{pf}(M)\Pi$. QR retraction satisfies $\mathrm{qf}(M\Pi) \neq \mathrm{qf}(M)\Pi$ in general.*

Gram-Schmidt computes projections sequentially from left to right, making updates dependent on the arbitrary column indexing. On $\mathrm{St}(3, 6)$ with $\eta = 0.02$, the discrepancy $\|\Phi^k(X\Pi) - \Phi^k(X)\Pi\|_F$ under QR is 3.5×10^{-3} after one step and reaches 1.8×10^{-2} after 20, while under polar it remains below 10^{-13} throughout.

5.2 Asymmetry on Periodic Orbits

Along a period-two orbit $X_a \rightleftarrows X_b$, where $\mathrm{grad}\, f \neq 0$, the Floquet matrix $J = D\Phi^2(X_a)$ is markedly less symmetric under QR. For step sizes between 1.6 and 3.4 times the stability threshold η_c, across ten problems on $\mathrm{St}(2, 4)$, $\mathrm{St}(2, 5)$, $\mathrm{St}(3, 5)$, $\mathrm{St}(2, 6)$ and $\mathrm{St}(3, 6)$, the relative asymmetry $\|J - J^\top\|_F / \|J\|_F$ averages 1.340 (range 0.867–1.576) under QR versus 0.547 (range 0.090–1.027) under polar, larger for QR in all ten problems (paired Wilcoxon signed-rank test, $p = 0.002$).

6 Empirical Validation on Deep Architectures

The proposition 6 establishes that the Riemannian gradient descent under the thin QR retraction is not equivariant under column permutations of the parameter matrix. Here, we investigate whether this theoretical symmetry breaking produces measurable consequences during the optimization of deep vision and recurrent architectures.

6.1 Protocol and Exact Numerical Controls

Evaluating permutation equivariance in deep networks presents a numerical hurdle: floating-point non-associativity. Permuting the columns of $W \in \mathrm{St}(p, n)$ while permuting the downstream classifier's inputs reorders a p-term sum in the forward pass. That reordering alone changes the logits by roughly 10^{-5} in `float32` and 10^{-14} in `float64` at initialization. In a `float32` pilot run, this seed was

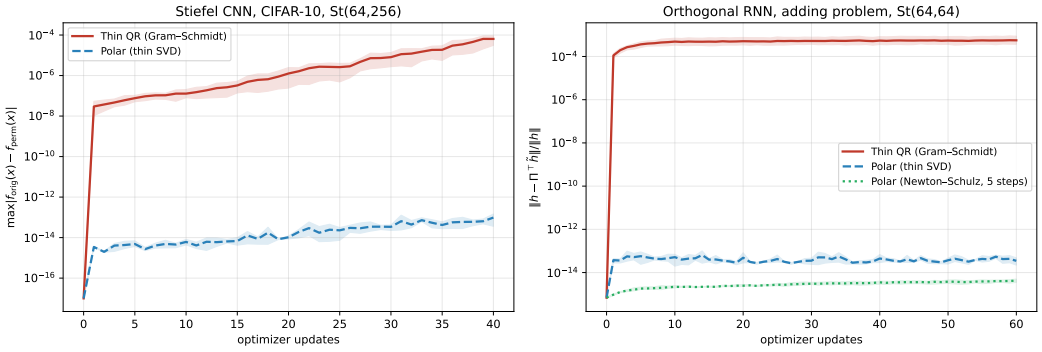


Figure 1: Permutation-equivariance breakdown in `float64`, mean over three seeds with min–max bands; curves start at update 1, since at update 0 the two copies are bitwise identical and the divergence is exactly zero. **Left:** Stiefel CNN on CIFAR-10, $W \in \mathrm{St}(64, 256)$. Thin QR separates the copies by 2.97×10^{-8} after one update and 6.42×10^{-5} after 40, while the polar retraction never exceeds 1.45×10^{-13}. **Right:** Orthogonal RNN on the adding problem, $W_{\mathrm{rec}} \in \mathrm{St}(64, 64)$, the square case $p = n$. Thin QR reaches 1.96×10^{-4} after two updates; SVD polar (3.49×10^{-14}) and unscaled Newton-Schulz (4.22×10^{-15}) remain at machine precision after sixty.

amplified chaotically to $O(1)$ within two epochs under *both* retractions, completely masking the retraction.

To isolate the retraction defect from numerical noise, our protocol introduces two methodological safeguards:

1. **Internal Representation Tracking:** Let $\pi \in S_p$ denote a bijection in the column indices $\{1, \ldots, p\}$, and let $\Pi \in \mathbb{R}^{p \times p}$ be its associated permutation matrix defined by $\Pi_{ij} = \delta_{i,\pi(j)}$ (such that post-multiplication $(W\Pi)_{:,j} = W_{:,\pi(j)}$ reorders the columns according to π). The permuted model directly stores and optimizes the physically permuted parameter matrix $\widetilde{W} = W\Pi$ on the Stiefel manifold. It compensates for column reordering by applying the inverse permutation π^{-1} on the latent representation coordinates ($z = x\widetilde{W}$, followed by indexing $z[:, \pi^{-1}]$) immediately before the linear head, eliminating the reordering of the summation.

2. **Double-Precision Execution:** All forward activations, backward gradients, and manifold retraction updates are executed in `float64`.

In this design, the functional discrepancy $|f_{\mathrm{orig}}(x) - f_{\mathrm{perm}}(x)|$ in step 0 is identically zero across all seeds. All runs use three seeds, deterministic cuDNN and cuBLAS settings, and a single-process data loader. We record the divergence between the two copies *after* the optimizer updates of k for $k = 0, 1, \ldots$; at $k = 0$ it is exactly zero in every seed, confirming that the forward passes are bitwise identical.

6.2 Stiefel Convolutional Networks (CIFAR-10)

A convolutional backbone feeds a layer $W \in \mathrm{St}(64, 256)$ followed by a linear classifier. We run SGD with momentum 0.9 on the unconstrained parameters and RGD on W, both with step size 0.02, on CIFAR-10 batches, and measure $\max_x |f_{\mathrm{orig}}(x) - f_{\mathrm{perm}}(x)|$ on each batch during the first 40 updates (Figure 1, left). The two copies are bitwise identical at initialization. After a single update, the thin QR has already separated them by 2.97×10^{-8}, and the gap increases to 6.42×10^{-5} after 40 updates. Under polar retraction, the divergence after one update is 3.46×10^{-15} and grows only through accumulated `float64` rounding, never exceeding 1.45×10^{-13} across all updates and seeds, more than eight orders of magnitude below thin QR at the end of the window. Values are means over three seeds; per-seed ranges are in the supplement. We deliberately confine the measurement to this early window. Over longer horizons, the rounding asymmetry inside the polar update itself is

chaotically amplified, and the control degrades, after which the comparison no longer isolates the retraction (Section 7).

6.3 Orthogonal Recurrent Networks

We train an orthogonal RNN on the addition problem ($T = 100$, hidden dimension $d = 64$), updating the recurrent matrix $W_{\mathrm{rec}} \in \mathrm{St}(64, 64)$ by RGD with step size 0.01 and the input and output maps by SGD, over three seeds (Figure 1, right). Two differences from CNN matter. The relabeling here is a conjugation $W_{\mathrm{rec}} \mapsto \Pi^\top W_{\mathrm{rec}} \Pi$ of the hidden state rather than a column permutation, and $\mathrm{St}(64, 64) \cong O(64)$ is the square case $p = n$, which lies outside the scope of Theorem 4 (Section 7). Nevertheless, the breakdown follows from Proposition 6: since $\mathrm{qf}(\Pi^\top M \Pi) = \Pi^\top \mathrm{qf}(M\Pi)$ and $\mathrm{qf}(M\Pi) \neq \mathrm{qf}(M)\Pi$ in general, the QR retraction also does not commute with conjugation. Under thin QR the relative hidden-state divergence $\|h - \Pi^\top \tilde{h}\|/\|h\|$ reaches 1.96×10^{-4} after two updates and 5.60×10^{-4} after sixty. The SVD polar retraction remains at 3.49×10^{-14}, and the Newton–Schulz retraction of Section 6.4 at 4.22×10^{-15}, with its orthogonality residual $\|W^\top W - I\|_F$ never exceeding 2.22×10^{-15} during training, below the 4.61×10^{-13} of the SVD polar retraction.

6.4 A GEMM-Only Polar Retraction

The polar factor is usually computed using an SVD, which is the main reason that practitioners default to QR. Instead, it can be computed by the Newton–Schulz iteration (Higham, 2008) using matrix multiplications only, and in the retraction setting, a simple identity removes the normalization step that these iterations usually require.

Lemma 7. *Let $X \in \mathrm{St}(p, n)$, $Z \in T_X \mathrm{St}(p, n)$ and $Y = X + Z$. Then*

$$Y^\top Y = I_p + Z^\top Z,$$

so that every singular value satisfies $\sigma_i(Y) = \sqrt{1 + \sigma_i(Z)^2} \geq 1$. If $\|Z\|_2 < \sqrt{2}$, the unscaled iteration

$$Y_0 = Y, \qquad Y_{k+1} = \tfrac{1}{2} Y_k \big(3I_p - Y_k^\top Y_k\big)$$

converges quadratically to $\mathrm{pf}(Y)$: write ε_k for the deviation of a singular value of Y_k from 1, $\varepsilon_{k+1} = -\frac{3}{2}\varepsilon_k^2 + O(\varepsilon_k^3)$.

Proof. Because $X^\top Z$ is skew-symmetric, $Y^\top Y = X^\top X + 2\,\mathrm{sym}(X^\top Z) + Z^\top Z = I_p + Z^\top Z$. The iteration preserves singular vectors, and acts on singular values by $\sigma \mapsto \frac{1}{2}\sigma(3 - \sigma^2)$, which maps $(0, \sqrt{3})$ to $(0, 1]$ and has 1 as an attracting fixed point; expanding to $\sigma = 1 + \varepsilon$ gives the recursion. Finally, $\sigma_{\max}(Y) < \sqrt{3}$ is equivalent to $\|Z\|_2 < \sqrt{2}$. □

In RGD, $Z = -\eta\, \mathrm{grad}\, f(X)$, so the condition is $\eta \|\mathrm{grad}\, f(X)\|_2 < \sqrt{2}$, which was true throughout our experiments. The initial error is $\varepsilon_0 = \sqrt{1 + \eta^2 \|\mathrm{grad}\, f\|_2^2} - 1 \approx \frac{1}{2}\eta^2 \|\mathrm{grad}\, f\|_2^2$, so a handful of iterations reach machine precision. The iteration also commutes with right multiplication by any orthogonal Π in exact arithmetic, since $Y \mapsto Y\Pi$ maps every iteration Y_k to $Y_k\Pi$; therefore, it inherits the permutation equivariance of the polar factor.

Scaling breaks the iteration. Newton–Schulz implementations for general matrices first rescale the input by a cheap bound on its spectral norm, such as $\sqrt{\|Y\|_1 \|Y\|_\infty}$. For nearly orthonormal Y, this bound overestimates $\|Y\|_2$ by a factor that grows with p — about 19 at 2048×512 — so the rescaled iterate starts with singular values near 0.05 rather than 1. Then five iterations leave $\|Y^\top Y - I\|_F$ between 2.7 and 19.4 between the sizes we tested: the output is not on the manifold and the method is not a retraction. The Lemma 7 shows that the rescaling is unnecessary.

Latency. Table 1 reports per-call latency alongside the orthogonality residual of each output, so that no speed is claimed for a result that does not lie on the manifold. At 1024×256 and 2048×512, NS-4 is faster than thin QR by factors of 1.85 and 1.53, with residuals of 2.66×10^{-6} and 7.53×10^{-6}, at the `float32` floor and comparable to thin QR. At smaller sizes, the thin QR remains faster; there, the Newton–Schulz time is essentially constant from 256×64 to 1024×256, indicating

Table 1: Per-call retraction time (μs, `float32`, averaged over 500 calls on NVIDIA Tesla T4) and orthogonality residual $\|Y^\top Y - I\|_F$ of the Newton–Schulz output; fastest method per size in bold. Newton–Schulz is unscaled (Lemma 7). Thin QR residuals lie at the `float32` floor, from 1.9×10^{-6} to 7.4×10^{-6}; `float32` SVD polar residuals grow with size, from 9.3×10^{-5} to 3.5×10^{-3}.

	time (μs)				residual	
$n \times p$	Thin QR	SVD polar	NS-4	NS-5	NS-4	NS-5
256×64	**475.3**	2138.6	985.4	1170.0	6.95×10^{-7}	6.40×10^{-7}
512×128	**622.7**	4952.3	978.8	1211.2	1.15×10^{-6}	9.99×10^{-7}
1024×256	1848.8	10228.3	**997.3**	1163.7	2.66×10^{-6}	1.98×10^{-6}
2048×512	5412.0	31570.9	**3539.1**	4182.4	7.53×10^{-6}	4.17×10^{-6}

that per-call overhead rather than arithmetic dominates so that the crossover point will depend on hardware. Against the SVD polar retraction, Newton–Schulz is between 2.2 and 10.3 times faster at every size and, in `float32`, two orders of magnitude more accurate: SVD residuals grow to 3.5×10^{-3} while NS-4 stays below 7.6×10^{-6}. Four iterations suffice in `float32` at every tested size.

7 Discussion, Delimitations, and Scope

Final accuracy is not measured. We make no claim about whether the QR defect changes the final accuracy, and our protocol cannot resolve it. In a ten-epoch run of a related convolutional architecture, the polar control did not survive: the contraction over p inside $W \operatorname{sym}(W^\top G)$ is reordered by permutation, injecting $\sim 10^{-16}$ per step, which accumulated to $\sim 10^{-11}$ within one epoch and was chaotically amplified to $O(1)$ by epoch 5 and epoch 2 in two of three seeds. The accuracy gap between column orderings was then 0.0094 for QR against 0.0195 for polar, with a seed-to-seed standard deviation of 0.0134 — the control noisier than the treatment, the signature of noise rather than signal. The obstruction is structural rather than a matter of tuning: isolating the retraction's effect at convergence would require holding two chaotic trajectories bitwise identical for thousands of steps, which no finite precision permits. Our claims concern well-posedness and reproducibility, measured within the window where the control holds.

Bifurcations and Floquet multipliers. On one problem, we located a Neimark–Sacker bifurcation of the period-two orbit under QR, suggesting that the non-conservative defect might uniquely produce invariant circles. It did not replicate: across ten problems, the polar retraction also produced complex Floquet multipliers in six, and Euclidean gradient descent did so in one of fourteen non-separable problems; the Floquet matrix of a period-two orbit is a product of symmetric matrices at *different* points and need not be symmetric. Only the asymmetry gap of Section 5.2 survives replication.

Topological and Lie-group boundaries. For $p \leq n - 2$ the Stiefel manifold is simply connected, so non-closedness of ω yields nonzero circulation along contractible loops. When $p = n$, $\mathrm{St}(n, n) \cong O(n)$ and $\mathrm{qf}(XM) = X\ \mathrm{qf}(M)$ for orthogonal X, so the QR retraction becomes a left translation and RGD a Lie-group integrator, within the scope of the Lie-group backward error analysis of Faltinsen (2000) and Lundervold & Munthe-Kaas (2013); Theorem 4 is not claimed there. The orthogonal RNN of Section 6.3 lies in this case: it demonstrates the equivariance breakdown of Proposition 6, which holds for every $p \geq 2$, but not the non-closedness result. The boundary case $p = n - 1$, where $\mathrm{St}(n - 1, n) \cong SO(n)$ is not simply connected, is untested.

Reproducibility Statement

The proofs of all results appear in Appendix A. The anonymous supplement contains four components, each mapped in its README to the values it produces. (i) A notebook reproducing Section 6, Table 1, Figure 1, the circulation of Section 4.3, and the $\mathrm{St}(2, 4)$ witness of Theorem 4; it runs top to bottom on a fresh GPU session, refuses to report the circulation unless its gradient-field control returns zero, and runs its network experiments in `float64` with deterministic kernels over three seeds. (ii) CPU scripts reproducing Sections 4.1, 4.2 and 5, the $\mathrm{St}(3, 6)$ witness, and the Floquet and Euclidean

results of Section 7, in about twelve minutes on one core. (iii) The ten-epoch run cited in Section 7. (iv) The `float32` pilot cited in Section 6.1. Both witnesses are exact rational computations and reproduce bit-for-bit.

Ethics Statement

This theoretical work investigates the dynamics of optimization in matrix manifolds. It does not pose direct negative ethical, privacy or social concerns.

AI Assistance Disclosure

A large language model was used for LaTeX-style formatting, and as a language-writing assistant to refine grammatical structures in sections.

References

P.-A. Absil and Jérôme Malick. Projection-like retractions on matrix manifolds. *SIAM Journal on Optimization*, 22(1):135–158, 2012.

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

Roy L. Adler, Jean-Pierre Dedieu, Joseph Y. Margulies, Marco Martens, and Mike Shub. Newton's method on Riemannian manifolds and a geometric model for the human spine. *IMA Journal of Numerical Analysis*, 22(3):359–390, 2002.

Martin Arjovsky, Amar Shah, and Yoshua Bengio. Unitary evolution recurrent neural networks. In *International Conference on Machine Learning (ICML)*, pp. 1120–1128, 2016.

Nitin Bansal, Xiaohan Chen, and Zhangyang Wang. Can we gain more from orthogonality regularizations in training deep networks? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 31, pp. 4261–4271, 2018.

María Barbero-Liñán and David Martín de Diego. Retraction maps: A seed of geometric integrators. *Foundations of Computational Mathematics*, 23(3):801–839, 2023.

David G. T. Barrett and Benoit Dherin. Implicit gradient regularization. In *International Conference on Learning Representations (ICLR)*, 2021.

Nicolas Boumal. *An Introduction to Optimization on Smooth Manifolds*. Cambridge University Press, 2023.

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.

Stig Faltinsen. Backward error analysis for Lie-group methods. *BIT Numerical Mathematics*, 40(4): 652–670, 2000.

Evan S. Gawlik and Melvin Leok. High-order retractions on matrix manifolds using projected polynomials. *SIAM Journal on Matrix Analysis and Applications*, 39(2):801–828, 2018.

Ernst Hairer, Christian Lubich, and Gerhard Wanner. *Geometric Numerical Integration*. Springer, 2nd edition, 2006.

Peter Halmos and Boris Hanin. Implicit bias of the JKO scheme. *arXiv preprint arXiv:2511.14827*, 2025.

Anders C. Hansen. A theoretical framework for backward error analysis on manifolds. *Journal of Geometric Mechanics*, 3(1):81–111, 2011.

Nicholas J. Higham. *Functions of Matrices: Theory and Computation*. SIAM, 2008.

Jun Li, Fuxin Li, and Sinisa Todorovic. Efficient Riemannian optimization on the Stiefel manifold via the Cayley transform. In *International Conference on Learning Representations (ICLR)*, 2020.

Alexander Lundervold and Hans Z. Munthe-Kaas. Backward error analysis and the substitution law for Lie group integrators. *Foundations of Computational Mathematics*, 13(2):161–186, 2013.

Sebastian Reich. Backward error analysis for numerical integrators. *SIAM Journal on Numerical Analysis*, 36(5):1549–1570, 1999.

Samuel L. Smith, Benoit Dherin, David G. T. Barrett, and Soham De. On the origin of implicit regularization in stochastic gradient descent. In *International Conference on Learning Representations (ICLR)*, 2021.

Jiayun Wang, Yubei Chen, Rudrasis Chakraborty, and Stella X. Yu. Orthogonal convolutional neural networks. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 11505–11515, 2020.

Scott Wisdom, Thomas Powers, Jonathan Hershey, Jonathan Le Roux, and Les Atlas. Full-capacity unitary recurrent neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 29, pp. 4880–4888, 2016.

A PROOFS

Proof of Theorem 1 (Tangential Defect of QR). Consider the curve $Y(t) = X + tV \in \mathbb{R}^{n\times p}$ for $X \in \mathrm{St}(p,n)$ and $V \in T_X\mathrm{St}(p,n)$. Its thin QR factorization is $Y(t) = Q(t)R(t)$ with $Q(0) = X$ and $R(0) = I_p$. Expand $Q(t)$ and $R(t)$ in powers of t:

$$Q(t) = X + tQ_1 + \frac{t^2}{2}Q_2 + \mathcal{O}(t^3), \qquad R(t) = I_p + tR_1 + \frac{t^2}{2}R_2 + \mathcal{O}(t^3), \tag{13}$$

where $R_1, R_2 \in \mathbb{R}^{p\times p}$ are upper triangular matrices. Because $Y(t)$ has no quadratic term in t, expanding $Y(t) = Q(t)R(t)$ and equating coefficients of identical powers yields:

$$V = Q_1 + XR_1, \qquad \frac{1}{2}Q_2 + Q_1R_1 + \frac{1}{2}XR_2 = 0. \tag{14}$$

From the orthogonality constraint $Q(t)^\top Q(t) = I_p$, the first-order term in t gives $\mathrm{sym}(X^\top Q_1) = 0$, meaning that $X^\top Q_1$ is skew-symmetric. Left-multiplying the first relation by $X^\top$ gives

$$X^\top V = X^\top Q_1 + R_1. \tag{15}$$

Because $V \in T_X\mathrm{St}(p,n)$, $X^\top V$ is skew-symmetric. By the uniqueness of the decomposition of a square matrix into a skew-symmetric and an upper-triangular matrix, we conclude that $R_1 = 0$ and $Q_1 = V$.

Substituting $R_1 = 0$ into the second relation yields $Q_2 = -XR_2$. At order t^2, the orthogonality condition $Q(t)^\top Q(t) = I_p$ requires $Q_1^\top Q_1 + \mathrm{sym}(X^\top Q_2) = 0$. Substituting $Q_1 = V$ and $X^\top Q_2 = -R_2$, we obtain

$$\mathrm{sym}(R_2) = V^\top V =: S. \tag{16}$$

Because R_2 is upper triangular, the relation $\mathrm{sym}(R_2) = S$ uniquely determines R_2 (the difference between any two upper-triangular solutions is both upper-triangular and skew-symmetric, hence zero). By inspection,

$$R_2 = \mathrm{diag}(S) + 2\,\mathrm{up}(S). \tag{17}$$

Using $Q_2 = -XR_2$, we have $\mathrm{sym}(X^\top Q_2) = -\mathrm{sym}(R_2) = -S$. The tangential defect is the orthogonal projection of the ambient acceleration Q_2 onto $T_X\mathrm{St}(p,n)$:

$$\begin{aligned}\tau_{\mathrm{qf}}(X)[V,V] &= P_X(Q_2) = Q_2 - X\,\mathrm{sym}(X^\top Q_2)\\ &= -XR_2 + XS = X(S - R_2).\end{aligned} \tag{18}$$

Recalling that $S = \mathrm{diag}(S) + \mathrm{up}(S) + \mathrm{up}(S)^\top$ for any symmetric matrix S, we obtain

$$S - R_2 = \big(\mathrm{diag}(S) + \mathrm{up}(S) + \mathrm{up}(S)^\top\big) - \big(\mathrm{diag}(S) + 2\,\mathrm{up}(S)\big) = \mathrm{up}(S)^\top - \mathrm{up}(S). \tag{19}$$

Multiplying on the left by X yields $\tau_{\mathrm{qf}}(X)[V,V] = X(\mathrm{up}(S)^\top - \mathrm{up}(S))$, completing the proof. □

Proof of Polar Retraction Order (Theorem 1). Let $V \in T_X\mathrm{St}(p,n)$. Since $X^\top V + V^\top X = 0$, we have

$$(X+tV)^\top(X+tV) = X^\top X + t(X^\top V + V^\top X) + t^2 V^\top V = I_p + t^2 S. \tag{20}$$

The inverse symmetric square root factor gives the polar retraction:

$$R_X(tV) = (X+tV)(I_p+t^2S)^{-1/2} = (X+tV)\left(I_p - \frac{t^2}{2}S + \mathcal{O}(t^4)\right) = X+tV-\frac{t^2}{2}XS+\mathcal{O}(t^3). \tag{21}$$

The second-order acceleration is therefore $a_{\mathrm{pf}}(X)[V,V] = -XS$. Because $S = V^\top V$ is symmetric, $\mathrm{sym}(X^\top a_{\mathrm{pf}}) = \mathrm{sym}(-X^\top XS) = -S$. The tangential projection satisfies

$$\tau_{\mathrm{pf}}(X)[V,V] = P_X(a_{\mathrm{pf}}) = -XS - X(-S) = 0. \tag{22}$$

Thus, $\tau_{\mathrm{pf}} \equiv 0$ identically, confirming that the polar retraction is second order. □

Proof of Lemma 2 (Critical Point Invariance). Let $\Phi(X) = R_X(-\eta \operatorname{grad} f(X))$. For any tangent direction $\xi \in T_{X^*}\mathrm{St}(p,n)$ at a stationary point X^* ($\operatorname{grad} f(X^*) = 0$), the differential $\mathrm{D}\Phi(X^*)[\xi]$ decomposes via the chain rule into variations through the base point and variations through the vector argument:

$$\mathrm{D}\Phi(X^*)[\xi] = \mathrm{D}_1 R_{X^*}(0)[\xi] + \mathrm{D}_2 R_{X^*}(0)[-\eta \mathrm{D}(\operatorname{grad} f)(X^*)[\xi]]\,. \tag{23}$$

Because $R_X(0) = X$ for all $X \in \mathrm{St}(p,n)$, differentiating with respect to X yields $\mathrm{D}_1 R_{X^*}(0)[\xi] = \xi = \mathrm{id}(\xi)$. Because R is a first-order retraction, $\mathrm{D}_2 R_{X^*}(0)[V] = V$ for any tangent vector V. The Riemannian derivative of the gradient field is the Riemannian Hessian, $\mathrm{D}(\operatorname{grad} f)(X^*)[\xi] = \operatorname{Hess} f(X^*)[\xi]$. Thus,

$$\mathrm{D}\Phi(X^*)[\xi] = \xi - \eta \operatorname{Hess} f(X^*)[\xi] = (\mathrm{id} - \eta \operatorname{Hess} f(X^*))\xi. \tag{24}$$

All second- and higher-order terms in the expansion of R scale with powers of the tangent vector $-\eta \operatorname{grad} f(X^*)$, which vanishes at X^*. Hence, $\mathrm{D}\Phi(X^*)$ is completely retraction-independent. □

Proof of Lemma 3 (Exterior Derivative of α). For a 1-form defined on an embedded Riemannian submanifold by $\alpha_X(Z) = \langle W(X), Z\rangle$, the exterior derivative is given by

$$d\alpha(Y,Z) = \langle \mathrm{D}_Y W, Z\rangle - \langle \mathrm{D}_Z W, Y\rangle, \tag{25}$$

as the normal components of the ambient directional derivatives are annihilated by tangent vectors $Y, Z \in T_X\mathrm{St}(2,n)$. For $W(X) = XJ$, the directional derivative is $\mathrm{D}_Y W = YJ$. Substituting into the formula:

$$\begin{aligned} d\alpha(Y,Z) &= \langle YJ, Z\rangle - \langle ZJ, Y\rangle = \mathrm{tr}\big((YJ)^\top Z\big) - \mathrm{tr}\big((ZJ)^\top Y\big) \\ &= \mathrm{tr}\big(J^\top Y^\top Z\big) - \mathrm{tr}\big(J^\top Z^\top Y\big) = \mathrm{tr}\big(J^\top (Y^\top Z - Z^\top Y)\big). \end{aligned} \tag{26}$$

Writing the 2×2 skew-symmetric matrix $Y^\top Z - Z^\top Y = m(Y,Z)J$, we compute:

$$d\alpha(Y,Z) = m(Y,Z)\,\mathrm{tr}(J^\top J). \tag{27}$$

Because $J = \left(\begin{smallmatrix} 0 & -1 \\ 1 & 0 \end{smallmatrix}\right)$, we have $J^\top J = I_2$, and $\mathrm{tr}(J^\top J) = \mathrm{tr}(I_2) = 2$. Therefore, $d\alpha(Y,Z) = 2m(Y,Z)$. □

Proof of Theorem 4 (Non-Closed Defect 1-Form). Applying the exterior Leibniz product rule to $\omega = s\,\alpha$ gives $d\omega = ds \wedge \alpha + s\,d\alpha$. Expanding on tangent vectors Y, Z:

$$d\omega(Y,Z) = (\mathrm{D}_Y s)\alpha(Z) - (\mathrm{D}_Z s)\alpha(Y) + 2s(X)m(Y,Z). \tag{28}$$

To exhibit an explicit counterexample where $d\omega \neq 0$, let $n = 4$ and set $X_0 = [e_1, e_2] \in \mathrm{St}(2,4)$. Consider the quadratic objective $f(X) = \mathrm{tr}(X^\top AXN) + \mathrm{tr}(B^\top X)$ with

$$A = \begin{pmatrix} 2 & 1 & 0 & -1 \\ 1 & 3 & 1 & 0 \\ 0 & 1 & 1 & 2 \\ -1 & 0 & 2 & 1 \end{pmatrix}, \quad N = \begin{pmatrix} 1 & 0 \\ 0 & 2 \end{pmatrix}, \quad B = \begin{pmatrix} 0 & 1 \\ 2 & 0 \\ 1 & 1 \\ 0 & -3 \end{pmatrix}, \tag{29}$$

and the valid 4×2 tangent vectors satisfying $\mathrm{sym}(X_0^\top Y) = 0$ and $\mathrm{sym}(X_0^\top Z) = 0$:

$$Y = \begin{pmatrix} 0 & -1 \\ 1 & 0 \\ 2 & 0 \\ -1 & 1 \end{pmatrix}, \qquad Z = \begin{pmatrix} 0 & 2 \\ -2 & 0 \\ 1 & 3 \\ 0 & 1 \end{pmatrix}. \tag{30}$$

Exact rational evaluation yields:

$$s(X_0) = 11, \quad \mathrm{D}_Y s = -52, \quad \mathrm{D}_Z s = -127, \quad \alpha(Y) = 2, \quad \alpha(Z) = -4, \quad m(Y,Z) = -5. \tag{31}$$

Substituting these values into $d\omega(Y,Z)$ gives

$$d\omega(Y,Z) = (-52)(-4) - (-127)(2) + 11 \cdot (-10) = 208 + 254 - 110 = 352 \neq 0. \tag{32}$$

Directly evaluating the ambient formulation $\langle \mathrm{D}_Y W, Z\rangle - \langle \mathrm{D}_Z W, Y\rangle = 153 - (-199) = 352$ corroborates the calculation. Because $d\omega$ is not identically zero, the 1-form ω is not closed, which proves that τ_{qf} is not a gradient vector field even on contractible open neighbourhoods. Thus, RGD under thin QR admits no modified energy functional.

Second witness ($p = 3$). On $\mathrm{St}(3,6)$ take $X_0 = \left(\begin{smallmatrix} I_3 \\ 0 \end{smallmatrix}\right)$, the objective $f(X) = \mathrm{tr}(X^\top A X N) + \mathrm{tr}(B^\top X)$ with $N = \mathrm{diag}(1,2,3)$,

$$A = \begin{pmatrix} 2 & 1 & 0 & -1 & 1 & 0 \\ 1 & 3 & 1 & 0 & 0 & 2 \\ 0 & 1 & 1 & 2 & -1 & 1 \\ -1 & 0 & 2 & 1 & 1 & 0 \\ 1 & 0 & -1 & 1 & 2 & 1 \\ 0 & 2 & 1 & 0 & 1 & 3 \end{pmatrix}, \qquad B = \begin{pmatrix} 0 & 1 & 1 \\ 2 & 0 & -1 \\ 1 & 1 & 0 \\ 0 & -3 & 2 \\ 1 & 0 & 1 \\ -1 & 2 & 0 \end{pmatrix},$$

and the tangent vectors

$$Y = \begin{pmatrix} 0 & -1 & 2 \\ 1 & 0 & -1 \\ -2 & 1 & 0 \\ 2 & 0 & -1 \\ 1 & 1 & 0 \\ -1 & 2 & 1 \end{pmatrix}, \qquad Z = \begin{pmatrix} 0 & 2 & -1 \\ -2 & 0 & -3 \\ 1 & 3 & 0 \\ 1 & 3 & 0 \\ 0 & 1 & 2 \\ 2 & -1 & 1 \end{pmatrix},$$

whose upper 3×3 blocks are skew-symmetric, so both are tangent at X_0. Here $W(X) = X\big(\mathrm{up}(S)^\top - \mathrm{up}(S)\big)$ with $S = g^\top g$, whose off-diagonal entries at X_0 are $s_{12} = -4$, $s_{13} = -49$, $s_{23} = 18$. Exact rational evaluation gives $\langle D_Y W, Z\rangle = 1098$ and $\langle D_Z W, Y\rangle = 943/2$, hence

$$d\omega(Y,Z) = 1098 - \tfrac{943}{2} = \tfrac{1253}{2} \neq 0.$$

□

Proof of Corollary 5 (Circulation along Closed Loops). The Stiefel manifold $\mathrm{St}(p,n)$ is $(n-p-1)$-connected; hence, for $p \leq n-2$, $\pi_1(\mathrm{St}(p,n)) = 0$ and the manifold is simply connected. Therefore, any closed loop $\gamma \subset \mathrm{St}(p,n)$ bounds a 2-dimensional smooth disc Σ. By Stokes' theorem on manifolds:

$$\oint_\gamma \omega = \int_\Sigma d\omega. \tag{33}$$

By Theorem 4, $d\omega$ is continuous and non-vanishing on an open neighbourhood around the witness point X_0. Choosing a loop γ bounding a small disc in this neighbourhood gives $\oint_\gamma \omega \neq 0$. □

Proof of Proposition 6 (Permutation Equivariance). Let $\Pi \in \mathbb{R}^{p \times p}$ be a permutation matrix. For any matrix $M \in \mathbb{R}^{n \times p}$ of full column rank with thin singular value decomposition $M = U\Sigma V^\top$, the permuted matrix factors as $M\Pi = U\Sigma(\Pi^\top V)^\top$. By definition of the polar factor:

$$\mathrm{pf}(M\Pi) = U(\Pi^\top V)^\top = UV^\top \Pi = \mathrm{pf}(M)\Pi. \tag{34}$$

Thus, polar RGD updates commute with column permutations.

For the QR retraction, Theorem 1 expresses the tangential defect in terms of $\mathrm{up}(S)$ where $S = V^\top V$. Under $X \mapsto X\Pi$, the tangent velocity transforms to $V \mapsto V\Pi$, conjugating the Gram matrix: $S \mapsto \Pi^\top S\Pi$. Because the operator $\mathrm{up}(\cdot)$ extracts entries strictly above the diagonal according to the canonical index ordering $1, \ldots, p$, it does not commute with permutation conjugation: $\mathrm{up}(\Pi^\top S\Pi) \neq \Pi^\top \mathrm{up}(S)\Pi$ in general. Thus, $\mathrm{qf}(M\Pi) \neq \mathrm{qf}(M)\Pi$, breaking permutation equivariance. □