BIAX: FIXED-RADIUS MATRIX OPTIMIZATION WITH FACTORED PRECONDITIONING FOR TRANSFORMERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Matrix optimizers for Transformers must reconcile spectral direction, anisotropic gradient scale, and parameter-norm control. We present Biax, which combines Muon^2-Factorized-style temporal row/column profiling before polar extraction with Hyperball's per-matrix fixed-initial-radius projection, adding $O(m+n)$ persistent state. The idealized update is a profiled-covector spectral linear oracle followed by exact sphere projection. We evaluate Muon, Hyperball-Muon, and Biax with the same AdamW fallback rule in four-seed, 20B-token studies at 500M and 1B parameters. The primary results are paired contrasts between fixed complete configurations. At 500M, where Hyperball-Muon and Biax use inherited unequal global learning rates, the Biax-minus-Hyperball-Muon held-out NLL difference is -0.0012 with 95% CI $[-0.0021, -0.0002]$, and Biax wins 4/4 seeds. At 1B, where both use learning rate 0.01, the difference is $+0.0001$ with 95% CI $[-0.0004, +0.0006]$, and Biax wins 2/4 seeds. Thus the 500M comparison favors Biax, whereas the 1B comparison is null; neither contrast identifies the contribution of profiling in isolation.

1 INTRODUCTION

Training optimizers jointly determine final loss, persistent state, memory pressure, update cost, and tokens-to-quality. Named optimizers often differ in several operations and hyperparameters at once. Decomposing them into ordered computations clarifies their structure, but an end-to-end training comparison still measures the difference between complete configurations rather than the causal contribution of any one internal operation.

We compare fixed Biax and Hyperball-Muon configurations at 500M and 1B parameters, using four matched seeds, a common 20B-token data protocol, and the same AdamW fallback rule for ineligible tensors. Muon is retained as a reference baseline. The 500M Biax and Hyperball-Muon configurations use inherited unequal global learning rates, whereas the 1B configurations both use 0.01.

The two scales give different outcomes: Biax is lower on all four paired seeds at 500M, while the 1B comparison shows no resolved difference. The evidence therefore does not support a scale-consistent advantage.

The paper makes three contributions. We (i) instantiate Biax, combining Muon^2-Factorized-style temporal pre-polar profiling with Hyperball's per-matrix fixed-initial-radius update while adding $O(m+n)$ persistent state; (ii) characterize its idealized direction as a profiled spectral linear oracle followed by fixed-radius angular dynamics; and (iii) report paired, fixed-configuration comparisons at two model scales under an explicit learning-rate, fallback, and statistical contract.

2 RELATED WORK

Muon maps matrix momentum toward its polar factor using finite Newton–Schulz iterations (Jordan et al., 2024). This direction also admits the norm-constrained linear-minimization-oracle view established for matrix optimizers (Pethick et al., 2025). Biax retains this direction-extraction template.

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 closest directional predecessor is Muon^2-Factorized (Liu et al., 2026). It maintains Adafactor-style row and column gradient-square EMAs (Shazeer & Stern, 2018), reconstructs a separable profile, applies it to momentum, and then polarizes. Under matched histories, inactive clamps, $\epsilon = 0$, and full-rank profiled momentum, Lemma 1 shows that Biax's bias correction changes the reconstructed profile only by a positive scalar, so the ideal polar directions agree; this does not assert equality of the finite-precision implementations. MALT likewise applies lightweight two-sided diagonal preconditioning before polarization, but then reapplies diagonal factors and norm-grafts the result (Wu et al., 2026). Biax performs neither post-polar operation.

Hyperball supplies the scale-control component: a generic optimizer wrapper that keeps each eligible matrix on its frozen initial Frobenius-norm sphere (Wen et al., 2026). Biax instantiates that update with Muon^2-Factorized-style profiling. We therefore do not claim a new factored estimator, pre-polar adaptation, fixed-radius projection, or novelty for wrapping this direction in Hyperball's generic construction. Our contribution is the concrete Biax instantiation, its profiled-oracle and fixed-radius characterization, and paired held-out fixed-configuration comparisons with Hyperball-Muon.

Table 1: Closest operator-level relationships. "Pre-polar input" is passed to the polar/Newton–Schulz map; added state excludes matrix momentum and scalar step counters.

Method	Operator sequence, scale rule, and added state
Muon	Momentum $\to$ polar direction; additive Muon scaling; no added state.
Muon^2-Factorized	Factored-G^2-profiled momentum $\to$ polar direction; additive Muon scaling; $O(m+n)$ state.
MALT	Two-sided diagonally profiled momentum $\to$ polar direction $\to$ reapply diagonal factors; norm-grafted step; $O(m+n)$ state.
Hyperball-Muon	Momentum $\to$ polar direction; fixed $\|W_0\|_F$ sphere; one-radius state.
Biax	Factored-G^2-profiled momentum $\to$ polar direction; fixed $\|W_0\|_F$ sphere; $O(m+n)$ plus one-radius state.

3 An ordered view of matrix optimization

Index updates by $t \geq 1$. For an eligible matrix parameter $W_{t-1} \in \mathbb{R}^{m\times n}$, let G_t denote its gradient, M_{t-1} its previous momentum, and $r = \|W_0\|_F$ its frozen initial radius, with $M_0 = u_0 = v_0 = 0$. We use the following operator calculus:

$$\mathsf{N}_F(X) = X/\max(\|X\|_F, \epsilon), \qquad \mathsf{H}_\beta(M, G) = \beta M + (1-\beta)G, \tag{1}$$

$$\mathsf{P}(X) = \mathsf{N}_F(\mathrm{NS}_5(X)), \qquad \mathsf{B}_V(X) = X \odot (V+\epsilon)^{-1/2}. \tag{2}$$

Finite Newton–Schulz implementation. For the Newton–Schulz map, the implementation transposes a tall input, casts to BF16, sets $X_0 = X/\max(\|X\|_F, \epsilon)$, and applies five steps of

$$X_{k+1} = 3.4445X_k - 4.775A_kX_k + 2.0315A_k^2X_k, \qquad A_k = X_kX_k^\top, \tag{3}$$

before transposing back. The outer N_F is computed in FP32. B_V profiles a direction before the polar map.

Hyperball-Muon takes an ambient step and radially projects it to the initial Frobenius sphere,

$$\mathsf{R}_r(W, D; \eta) = r\frac{W - \eta r D}{\|W - \eta r D\|_F}. \tag{4}$$

3.1 Biax: biaxial profiling before direction extraction

Relative to Hyperball-Muon, Biax keeps the fixed-radius update and replaces the raw momentum input to the polar map with a temporal row/column profile. Relative to Muon^2-Factorized, it keeps that pre-polar profiling pattern but replaces the additive parameter update with Hyperball's radial projection.

Biax maintains row and column gradient-energy EMAs $u_t \in \mathbb{R}^m$ and $v_t \in \mathbb{R}^n$ and reconstructs only a transient matrix profile,

$$u_t = \beta_2 u_{t-1} + (1-\beta_2)\,\mathrm{mean}_j(G_t^2), \tag{5}$$

$$v_t = \beta_2 v_{t-1} + (1-\beta_2)\,\mathrm{mean}_i(G_t^2), \tag{6}$$

$$\widehat{u}_t = u_t/(1-\beta_2^t), \qquad \widehat{v}_t = v_t/(1-\beta_2^t), \tag{7}$$

$$(V_t)_{ij} = \frac{(\widehat{u}_t)_i(\widehat{v}_t)_j}{\max(\mathrm{mean}(\widehat{u}_t), \epsilon)}. \tag{8}$$

It then computes

$$M_t = \mathsf{H}_\beta(M_{t-1}, G_t), \qquad D_t = \mathsf{P}(\mathsf{B}_{V_t}(M_t)), \tag{9}$$

$$W_t = \mathsf{R}_r(W_{t-1}, D_t; \eta_t). \tag{10}$$

Thus the persistent addition over Hyperball-Muon is $m+n$ scalars, not an $m \times n$ second-moment tensor.

Variational interpretation. Let $Z_t = \mathsf{B}_{V_t}(\mathsf{H}_\beta(M_{t-1}, G_t))$ and $p = \min(m,n)$. Define the ideal operator $\mathsf{P}_\star(Z) = \mathrm{polar}(Z)/\sqrt{p}$, keeping it distinct from the implemented five-step Newton–Schulz approximation.

Lemma 1 (Factored shape–scale separation)**.** *Assume the row and column accumulators start at zero and $q_t = \mathrm{mean}(\widehat{u}_t) > 0$. Then $\mathrm{mean}(\widehat{v}_t) = q_t$. For $a_t = \widehat{u}_t/q_t$, $b_t = \widehat{v}_t/q_t$, and $s_t = q_t^2/\max(q_t, \epsilon)$,*

$$V_t = s_t a_t b_t^\top, \tag{11}$$

$$Z_t = s_t^{-1/2} M_t \odot \left(a_t b_t^\top + \lambda_t \mathbf{1}\mathbf{1}^\top\right)^{-1/2}, \qquad \lambda_t = \epsilon/s_t. \tag{12}$$

If $q_t \geq \epsilon$, V_t exactly reproduces both axis marginals.

Proof of Lemma 1. Averaging either moment recurrence gives the same scalar EMA of $\|G_t\|_F^2/(mn)$, proving equality of the means. Substituting $\widehat{u}_t = q_t a_t$ and $\widehat{v}_t = q_t b_t$ gives both identities. When $q_t \geq \epsilon$, averaging V_t along either axis recovers the corresponding moment vector. □

With $\epsilon = 0$, positive axis moments, zero-initialized momentum and axis accumulators, and full-rank profiled momentum, the scalar $s_t^{-1/2}$ does not affect $\mathsf{P}_\star$. Consequently, a constant positive rescaling of the complete gradient history leaves the ideal Biax step unchanged, and a scalar profile reduces its direction to Hyperball-Muon. The same calculation relates Biax to Muon^2-Factorized: if $\gamma_t = 1-\beta_2^t$, their reconstructed profiles satisfy $V_t^{\mathrm{Biax}} = V_t^{\mathrm{Muon^2\text{-}F}}/\gamma_t$ under matched histories, inactive clamps, and $\epsilon = 0$. Their ideal polar directions therefore agree, although Muon^2-Factorized subsequently takes a differently scaled unconstrained additive update.

Proposition 1 (Profiled spectral oracle)**.** *If $\mathrm{rank}(Z_t) = p$, then $D_t^\star = \mathsf{P}_\star(Z_t)$ satisfies*

$$-D_t^\star \in \operatorname*{arg\,min}_{\sqrt{p}\|S\|_2 \leq 1} \langle Z_t, S\rangle_F, \qquad \|D_t^\star\|_F = 1. \tag{13}$$

Proof of Proposition 1. Let $Z_t = U\Sigma V^\top$ be a thin singular value decomposition. For every feasible S, nuclear–spectral duality gives

$$\langle Z_t, S\rangle_F \geq -\|Z_t\|_*\|S\|_2 \geq -\frac{\|Z_t\|_*}{\sqrt{p}}. \tag{14}$$

The lower bound is attained by $S = -UV^\top/\sqrt{p} = -D_t^\star$. Full rank implies that the p singular values of $D_t^\star$ all equal $1/\sqrt{p}$, so $\|D_t^\star\|_F = 1$. □

Pullback-norm distinction. Conditionally fix the strictly positive self-adjoint profiling operator $\mathcal{B} = \mathsf{B}_{V_t}$. Under the same full-rank condition and the pullback spectral norm $\|\Delta\|_{\mathcal{B},\mathrm{sp}} = \sqrt{p}\|\mathcal{B}^{-1}(\Delta)\|_2$,

$$-\mathcal{B}[\mathsf{P}_\star(\mathcal{B}(M_t))] \in \operatorname*{arg\,min}_{\|\Delta\|_{\mathcal{B},\mathrm{sp}}\le 1} \langle M_t, \Delta\rangle_F. \tag{15}$$

Biax instead executes $-\mathsf{P}_\star(\mathcal{B}(M_t))$ before sphere projection. It therefore uses the profile to select orientation without mapping the direction back through the profile; it is not the steepest step under this pullback norm.

Proposition 2 (Fixed-radius angular dynamics)**.** *Let $\|W\|_F = r > 0$, $\|D\|_F = 1$, and $0 < \eta < 1$. The endpoint $W^+ = \mathsf{R}_r(W, D; \eta)$ is the unique solution of*

$$\operatorname*{arg\,min}_{\|X\|_F = r} \left\{ \langle D, X - W\rangle_F + \frac{\|X - W\|_F^2}{2\eta r} \right\}. \tag{16}$$

Writing $w = W/r$, $\rho = \langle w, D\rangle_F$, and $d^\perp = (D - \rho w)/\sqrt{1-\rho^2}$ for a nonzero tangent component, the same endpoint obeys

$$\frac{W^+}{r} = \cos(\theta) w - \sin(\theta) d^\perp, \tag{17}$$

$$\theta = \operatorname{atan2}\Big(\eta\sqrt{1-\rho^2},\, 1 - \eta\rho\Big). \tag{18}$$

Moreover, $0 \le \theta \le \arcsin(\eta)$; if D is tangent, $\theta = \arctan(\eta)$.

Thus idealized Biax separates adaptive spectral orientation selection from fixed-radius parameter dynamics. The proof of Proposition 2 appears in Appendix A.

4 Experimental protocol

The study evaluates Muon, Hyperball-Muon, and Biax at 500M and 1B parameters. Within each tier, the arms share architecture, tokenizer, data order, initialization, training budget, validation split, and evaluation code. Training consumes the same 20B-token FineWeb-Edu prefix, and held-out NLL is computed from each terminal saved checkpoint. Four seeds are run per arm; the Biax and Hyperball-Muon contrasts pair runs with the same seed.

4.1 Models, routing, and evidence separation

Both architectures are decoder-only Llama-style transformers with vocabulary 8,192, context length 2,048, tied embeddings, head dimension 64, and no attention or MLP bias. Eligible two-dimensional attention and MLP tensors use the matrix optimizer. Embeddings, normalization parameters, and other ineligible tensors use AdamW with the arm's global learning rate, betas $(0.9, 0.95)$, $\epsilon = 10^{-8}$, and weight decay 0.1. Thus the fallback algorithm and its non-learning-rate hyperparameters are common, but its learning rate follows the arm and is unequal between the 500M Biax and Hyperball-Muon configurations. Eligible fixed-radius matrices have zero weight decay.

Biax and Hyperball-Muon use momentum $\beta = 0.95$; Biax additionally uses second-moment decay $\beta_2 = 0.95$. Both use five Newton–Schulz steps and $\epsilon = 10^{-8}$. Matrix momentum follows the parameter dtype, while radii and axis moments are stored in FP32.

Table 2: Model architectures represented in the reported results. Routed is the fraction of trainable parameters receiving the matrix update.

Tier	Parameters	Layers	Width	FFN	KV heads	Routed
500M	500,435,200	26	1,280	3,840	5	97.9%
1B	1,000,045,056	31	1,536	5,632	6	98.7%

Parameter routing. For a tied-embedding, bias-free decoder with L layers, hidden width h, intermediate width i, key/value width $k_{\mathrm{kv}}d_h$, and vocabulary $\mathcal{V}$, the matrix and fallback counts are

$$N_{\mathrm{matrix}} = L(2h^2 + 2hk_{\mathrm{kv}}d_h + 3hi), \tag{19}$$

$$N_{\mathrm{fallback}} = |\mathcal{V}|h + (2L+1)h. \tag{20}$$

Coverage tests freeze the same routing membership for every matrix-family comparison. These counts yield the routed fractions in Table 2; the remaining parameters use the common AdamW fallback.

Reported training contract. The 500M and 1B comparisons use BF16, FlashAttention 2, gradient checkpointing, sequence length 2,048, deterministic initialization, and a fixed token order. Each arm performs 9,540 optimizer updates over 20,006,830,080 tokens, equivalent to 1,024 full sequences per global update. The cosine schedule has 477 warmup updates and decays to 0.1 times its peak learning rate; checkpoint 9,540 is the terminal checkpoint used for held-out evaluation.

Training, validation, and terminal-checkpoint held-out evaluation are treated as distinct evidence classes. The held-out split is disjoint from training and validation and does not influence the trained checkpoints.

4.2 Learning-rate and statistical contract

At 500M, the fixed global learning rates are 0.002 for Muon, 0.01091 for Hyperball-Muon, and 0.01189 for Biax. The latter two rates were inherited from prior exploratory qualification and were not retuned after adopting the common AdamW fallback. At 1B, Hyperball-Muon and Biax both use 0.01000, while the Muon reference uses 0.001682. Consequently, the 500M paired contrast includes a learning-rate difference, whereas the 1B contrast holds global learning rate fixed. Both remain comparisons of complete optimizer configurations. Learning rates are rounded to four significant digits for presentation.

For each tier, let L_i^A denote terminal held-out NLL for configuration A at seed i. The paired observation and reported contrast are

$$d_i^{A-B} = L_i^A - L_i^B, \qquad \widehat{\Delta}_{A-B} = \frac{1}{4}\sum_{i=1}^{4} d_i^{A-B}. \tag{21}$$

We report the sample SD of the four paired differences, the number of seeds with $d_i^{\mathrm{Biax-Hyperball}} < 0$, and the two-sided Student 95% interval $\widehat{\Delta} \pm t_{3,0.975}s_d/2$. Muon is reported as a reference baseline rather than as part of the primary paired contrast.

5 Results

Table 3: Checkpoint-only held-out NLL for the evaluated AdamW-fallback implementations. Method rows report the four-seed mean and sample SD. The contrasts are paired by seed; negative favors Biax. Displayed statistics are rounded to four decimal places; paired statistics use unrounded seed-level values.

Tier	Method	LR	Mean NLL $\pm$ SD
500M	Muon	.002000	2.2539 ± 0.0004
500M	Hyperball-Muon + AdamW	.01091	2.1814 ± 0.0007
500M	Biax + AdamW	.01189	2.1803 ± 0.0005
1B	Muon	.001682	2.1720 ± 0.0003
1B	Hyperball-Muon + AdamW	.01000	2.0981 ± 0.0008
1B	Biax + AdamW	.01000	2.0981 ± 0.0005

Tier	Paired Biax $-$ Hyperball-Muon	95% CI	Wins
500M	-0.0012 ± 0.0006	$[-0.0021, -0.0002]$	4/4
1B	$+0.0001 \pm 0.0003$	$[-0.0004, +0.0006]$	2/4

At 500M, Biax records mean held-out NLL 2.1803, compared with 2.1814 for Hyperball-Muon. The paired Biax-minus-Hyperball-Muon difference is -0.0012 with 95% CI $[-0.0021, -0.0002]$, and Biax is lower on 4/4 seeds. This favors the evaluated Biax configuration, but the contrast includes the inherited learning-rate difference between the two arms.

At 1B, Biax records mean held-out NLL 2.0981, compared with 2.0981 for Hyperball-Muon. With the same 0.01 global learning rate in both arms, the paired difference is $+0.0001$ with 95% CI $[-0.0004, +0.0006]$, and Biax is lower on 2/4 seeds. This is a null result: the data do not resolve a difference between the two fixed configurations.

The Muon reference has mean held-out NLL 2.2539 at 500M and 2.1720 at 1B, above both fixed-radius configurations at each tier. Because Muon also differs in update rule and learning rate, these rows provide baseline context rather than an isolated comparison of radius control or profiling.

6 Limitations

The study covers two sizes of one model family, one corpus and token budget, and four paired seeds per contrast. It therefore does not establish cross-corpus or data-regime invariance, and four seeds limit precision for small differences.

The two tiers do not constitute a controlled scale interaction. At 500M, Biax and Hyperball-Muon use inherited unequal global learning rates, which also propagate to their AdamW fallback parameters. At 1B they share a global learning rate, but the comparison still concerns complete optimizer implementations and does not attribute the null result to any single stage. The Muon rows likewise provide a reference baseline rather than a hyperparameter-matched mechanistic control.

The structural propositions describe an idealized polar operator. Finite Newton–Schulz iterations, BF16 arithmetic, clamps, and routing through AdamW separate the implementation from that idealization; the propositions do not establish convergence or descent of the training objective.

7 Conclusion

Biax instantiates Hyperball's generic per-matrix fixed-initial-radius wrapper using Muon^2-Factorized-style temporal pre-polar profiling with $O(m+n)$ additional persistent state. Its idealized construction separates profiled spectral orientation selection from fixed-radius parameter dynamics.

The empirical comparisons do not show a scale-consistent advantage. The 500M comparison favors the evaluated Biax configuration under inherited unequal learning rates, whereas the same-learning-rate 1B comparison is null. Together with the disclosed AdamW fallback and Muon reference, these results characterize the tested fixed configurations without assigning their differences to the profiling mechanism alone.

AI use statement

In this work, we used generative AI tools to propose and refine optimizer hypotheses; help formulate and check mathematical claims and proof arguments; design and critique methods and experiments; implement and test software; orchestrate experiments; analyze and interpret results; search and summarize literature; verify citations; prepare figures; and draft and edit the manuscript. Human authors defined the experimental protocol, authorized executions, reviewed source changes, and verified reported measurements against source results. AI-generated code was reviewed and tested, and AI-suggested scientific claims and citations were checked against primary evidence. We have reviewed all AI-assisted work and take responsibility for the final content, claims, code, and artifacts.

Ethics statement

This study introduces no human-subject experiment and releases no new personal data. Language-model training can nevertheless reproduce biases and harmful content present in its corpus. Any external artifact release remains subject to dataset-license, privacy, security, and responsible-release review.

Reproducibility statement

The main text specifies model, routing, training-schedule, learning-rate, statistical, and evidence-separation contracts; the appendix records the inherited spherical-step proof and finite-arithmetic caveats.

References

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

Ziyue Liu, Ruijie Zhang, Zhengyang Wang, Yequan Zhao, Yupeng Su, Zi Yang, and Zheng Zhang. Muon^2: Boosting Muon via adaptive second-moment preconditioning, 2026. URL https://arxiv.org/abs/2604.09967.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 49069–49104. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/pethick25a.html.

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 4596–4604. PMLR, 2018. URL https://proceedings.mlr.press/v80/shazeer18a.html.

Kaiyue Wen, Xingyu Dang, Kaifeng Lyu, Tengyu Ma, and Percy Liang. Fantastic pretraining optimizers and where to find them II: Hyperball optimization, 2026. URL https://arxiv.org/abs/2606.16899.

Tongle Wu, Huanyu Dong, Ying Sun, and Ziye Ma. MALT: Lightweight curvature-aware Muon via diagonal preconditioning, 2026. URL https://arxiv.org/abs/2608.05088.

A Structural proof and auxiliary details

Proof of Proposition 2. Set $Y = W - \eta r D$. Completing the square shows that the displayed objective differs from $\|X - Y\|_F^2/(2\eta r)$ only by a constant. For $Y \neq 0$, its unique minimizer on the radius-r sphere is $rY/\|Y\|_F$, the ideal real-arithmetic endpoint corresponding to the implemented FP32 projection before parameter-dtype writeback.

If $|\rho| = 1$, the normalized endpoint stays on the same ray as W and the angular bound holds with zero turn. Otherwise, decompose $D = \rho w + \tau d^{\perp}$, where $\tau = \sqrt{1-\rho^2}$. Normalizing $(1-\eta\rho)w - \eta\tau d^{\perp}$ gives the stated cosine, sine, and $\mathrm{atan2}$ expressions. Furthermore,

$$\sin^2\theta = \frac{\eta^2(1-\rho^2)}{1+\eta^2-2\eta\rho} \leq \eta^2, \tag{22}$$

because the denominator minus $1-\rho^2$ is $(\rho-\eta)^2$. Hence $\theta \leq \arcsin(\eta)$, with equality at $\rho = \eta$. Expanding the normalization at $\eta = 0$ also yields

$$W^+ = W - \eta r(D - \rho w) + O(r\eta^2), \tag{23}$$

so the first-order motion is the negative tangent projection. When $\rho = 0$, the exact angle reduces to $\arctan(\eta)$. □

These statements separate the ideal structural map from its implementation. Finite ϵ, denominator and norm clamps, and BF16 arithmetic weaken profile-scale invariance. The implemented direction has unit norm only when the outer normalization clamp is inactive. Input normalization preserves positive scale invariance of finite Newton–Schulz iterations in exact arithmetic when its clamp is inactive, but five iterations only approximate the polar oracle. None of the propositions establishes descent of the instantaneous loss or a convergence rate.