A Quadratic Lens on Muon: Orthogonalization, Invariance, and Implicit Preconditioning

Egor Shulgin* EGOR.SHULGIN@KAUST.EDU.SA
King Abdullah University of Science and Technology (KAUST), Saudi Arabia

Sam Laing* SLAING155@GMAIL.COM
ELLIS Institute Tübingen, Tübingen AI Center, Germany
OpenEuroLLM

Antonio Orvieto ANTONIO@TUE.ELLIS.EU
ELLIS Institute Tübingen, Tübingen AI Center, Germany

Peter Richtárik PETER.RICHTARIK@KAUST.EDU.SA
King Abdullah University of Science and Technology (KAUST), Saudi Arabia

Abstract

Muon and related optimizers improve training by approximately orthogonalizing matrix-valued updates, but the geometry behind their empirical gains remains partially understood. We study an idealized polar version of Muon on matrix-quadratic objectives, using exact line search to isolate update direction from step-size effects. This exposes a gain–curvature tradeoff with three exactness regimes: curvature isotropy (GD), coordinate-aligned vertex structure (SignGD), and row-isotropic iterates (PolarGD), the last holding independently of Hessian conditioning. We show PolarGD is equivariant under orthogonal basis changes, unlike SignGD, explaining its robustness to rotations of ill-conditioned structures in Zipf-style regression experiments. Beyond one step, PolarGD admits an implicit iterate-dependent preconditioner that whitens iterate covariance in the full-row-rank regime, yielding an exact progress coefficient separating the fast row-isotropic regime from the pessimistic tiny-singular-value regime. Rank-deficient extensions replace global by active-subspace curvature, and finite Newton–Schulz orthogonalization is shown to act as a bounded spectral filter that damps exact polar updates near rank deficiency.

1. Introduction

Muon [8] has recently emerged as a highly efficient optimizer for large neural networks, often outperforming AdamW [7, 12, 16, 17]. Its distinguishing feature is the matrix-aware approximate orthogonalization of momentum updates. However, the exact dynamics of this orthogonalization step remains only partially understood.

Prior frameworks connecting Muon to spectral-norm descent and normalized methods [1, 9, 15] typically yield worst-case bounds. They confirm Muon is a valid descent direction but fail to concretely explain when and why orthogonalization improves over standard Gradient Descent (GD).

© E. Shulgin*, S. Laing*, A. Orvieto & P. Richtárik.

This gap is already present in the simplest deterministic setting where matrix geometry matters: matrix quadratics. Quadratics are the classical testbed for understanding GD, yet their matrix-valued form already contains the main complications that arise for orthogonalized updates: singular structure, curvature alignment, initialization dependence, and coordinate-system effects.
By employing exact line search, we compare the direction rays of GD, SignGD, and PolarGD independently of learning-rate choices, framing their comparison as an explicit gain–curvature tradeoff.

Our analysis reveals that PolarGD's behavior is governed by the current iterate and coordinate system, not just the Hessian condition number. While GD favors curvature isotropy and SignGD favors coordinate alignment, PolarGD favors different structures: curvature isotropy, coordinate alignment, and row-isotropy of the iterate.
Beyond one-step analysis, we model PolarGD trajectories as implicitly preconditioned methods driven by iterate covariance, identifying both fast and pessimistic regimes. Finally, we extend these insights to rank-deficient active subspaces, connecting exact polar updates to rank truncation, damping, and Newton–Schulz filtering. We summarize our main **contributions** below:

- Exact line-search identities for GD, SignGD, and PolarGD, with a local condition for PolarGD outperforming GD.
- Three exactness regimes: curvature isotropy (GD), coordinate alignment (SignGD), row-isotropy (PolarGD).
- Proving an orthogonal equivariance property for GD and PolarGD that fails for SignGD.
- Developing an implicit-preconditioning view of PolarGD that separates fast and pessimistic regimes based on iterate covariance.
- Extending our theory to active subspaces, rank truncation, damping, and finite Newton–Schulz filters.

2. Notation and Setup

Let $f : \mathbb{R}^{d\times m} \to \mathbb{R}$ be differentiable. We write $G := \nabla f(X)$ and use the Frobenius inner product $\langle A, B\rangle := \mathrm{tr}(A^\top B)$. We compare three direction rays generated from G:

$$D_{\mathrm{GD}} := G, \qquad D_{\mathrm{Pol}} := Q := \mathrm{polar}(G), \qquad D_{\mathrm{Sign}} := S := \mathrm{sign}(G),$$

where S is the entrywise sign matrix, with $\mathrm{sign}(0) := 0$. If $G = U\Sigma V^\top$ is a thin SVD, then $Q = UV^\top$. The polar direction is the spectral-norm LMO direction, since it maximizes $\langle G, D\rangle$ over $\|D\|_2 \le 1$; the sign direction analogously arises from the entrywise ℓ_∞-ball.
Our comparisons use line search along a chosen direction. This removes learning-rate tuning from the local comparison. Moreover, after optimizing over the scalar step, the direction is scale invariant. We specialize to the matrix quadratic case

$$f(X) = \frac{1}{2}\mathrm{tr}((X - X^\star)^\mathsf{T} H(X - X^\star)), \qquad \mathbb{R}^{d\times d} \ni H = H^\top \succeq 0, X \in \mathbb{R}^{d\times n} \tag{1}$$

Remark 1 *Our results can be partially generalized under matrix smoothness (see J).*

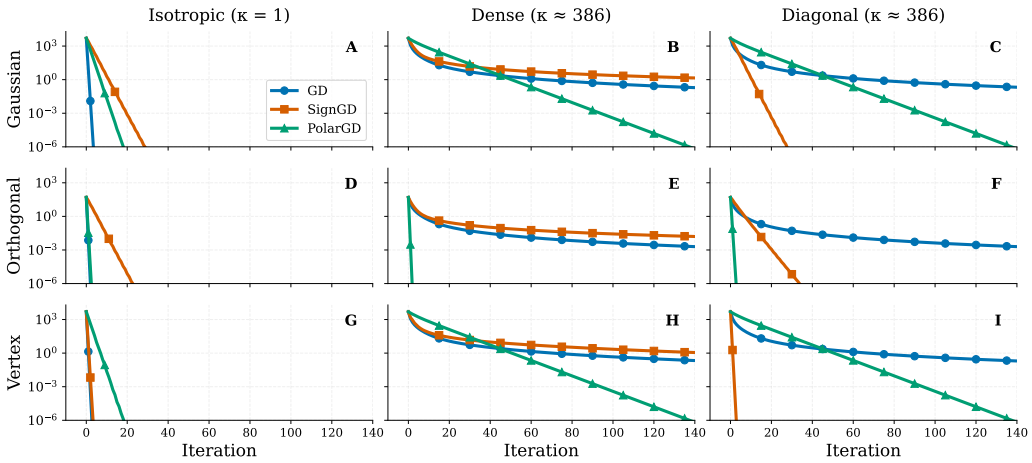

Figure 1: **Convergence Across Initialization-Hessian Regimes.** Performance of dual-adapted **GD**, **SignGD** and **PolarGD** (idealized Muon) on matrix quadratic objectives $f(X) = \frac{1}{2}\mathrm{tr}(X^\top HX)$ with extensively tuned step sizes (see Section A.1). To isolate structural dependencies, we test three initializations: *Gaussian* ($X_0 \sim \mathcal{N}(0,1)$), *Orthogonal* and *Vertex* ($X_0 \sim \{-1,1\}$). We evaluate these against three Hessian profiles: *Isotropic* ($H = I_d$), *Dense* Marchenko Pastur ($\kappa = 386$) and *Diagonal* (sharing the same spectrum as the dense case). Consistent with the analysis above, each optimizer favors a distinct structural regime: GD is robust to initialization but degrades under ill-conditioning, SignGD benefits from coordinate-aligned curvature, and PolarGD is rotation-invariant and benefits from row-isotropic initialization. No single optimizer dominates across regimes.

3. One-step geometry on matrix quadratics

We formulate the central theorem which elucidates the initialization-Hessian regimes in which each optimizer benefits. Generally the directional curvatures for SignGD and PolarGD depend on both the Hessian (H) and the initialization X_0

Theorem 2 (One-step convergence regimes under exact line search on quadratics) *Consider* $f(X) = \frac{1}{2}\mathrm{tr}\big((X - X^\star)^\top H(X - X^\star)\big)$ *with* $E := X - X^\star \in \mathbb{R}^{d\times m}$, $H = H^\top \succeq 0$, *and* $G = HE$. *Let* $Q = \mathrm{polar}(G)$, $S = \mathrm{sign}(G)$ *(entrywise,* $\mathrm{sign}(0) := 0$*), and* $X^+(D) = X - \gamma^\star(D)D$ *the exact line-search update.*

(i) ***GD.*** *One-step exact at every* X *iff* $H = \alpha I_d$*; at a fixed* X *iff* $HE = \lambda E$ *for some* $\lambda > 0$.

(ii) ***SignGD.*** *One-step exact at* X *iff* $E = cS$ *for some* $c > 0$ *and* $\mathrm{sign}(HS) = S$ *on* $\mathrm{supp}(S)$. *In particular, strict diagonal dominance* ($H_{ii} > \sum_{j\neq i} |H_{ij}|$) *ensures* $\mathrm{sign}(HS) = S$ *for every* $S \in \{\pm 1\}^d$*, so any vertex initialization* $E = cS$ *(columnwise) converges in one step.*

(iii) ***PolarGD.*** *Assume* $H \succ 0$ *and* $\mathrm{rank}(G) = d \leq m$. *One-step exact at* X *iff* $EE^\top = cI_d$ *for some* $c > 0$, *i.e.* $E = \sqrt{c}UV^\top$ *with* $U \in \mathbb{R}^{d\times d}$ *orthogonal and* $V \in \mathbb{R}^{m\times d}$ *having orthonormal columns (reducing to scaled orthogonal when* $m = d$*).*

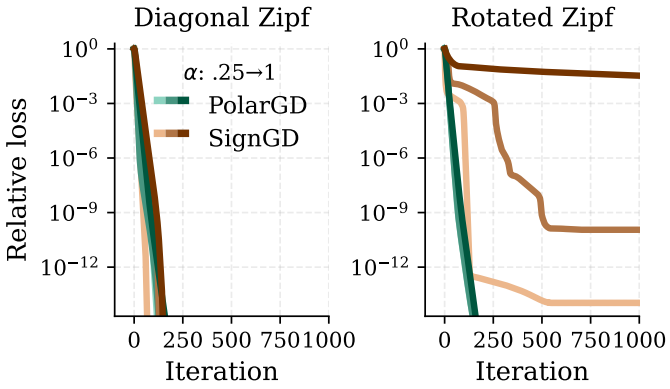

Figure 2: Dynamics for Zipf spectrum $\lambda_k(H) = k^{\alpha}$ for exponents $\alpha \in \{0.25, 0.5, 1.0\}$ with both diagonal and rotated Hessians. We report the best-tuned dual-adapted learning rate. PolarGD is unaffected by the rotation, while SignGD degrades significantly. Details in App. A.3.

What remains unexplained. Figure 3 shows an interesting phenomenon: PolarGD remains fast even when H is dense and ill-conditioned. Theorem 2 hints at this – row-isotropic iterates can remove the effect of H in one step – but does not explain what happens beyond the first step, or why high condition number need not dominate the trajectory. This requires understanding how the polar direction reshapes the iterate geometry over time, the trajectory-level preconditioning view developed in Section 4.

Zipf Laws experiment. Kunstner and Bach [11] models next-token prediction as linear regression over one-hot inputs, with squared loss $L(X) = \frac{1}{2n}\|ZX - Y\|_F^2$. Since tokens are modeled as orthogonal 1-hot vectors, $Z^\top Z$ is diagonal with entries given by token counts; under Zipf frequencies $\pi_k \propto k^{-\alpha}$. The loss can easily be put in direct relation with $f(Z) = \frac{1}{2}\operatorname{tr}\big((X - X^\star)^\top H (X - X^\star)\big)$, with $h_k = k^{\alpha}$. In Figure 3 we show that PolarGD has similar convergence properties, yet unlocks performance beyond the 1-hot assumption since it is invariant to orthogonal transformations. This insight can motivate to some extent the exceptional performance of Muon in language modeling.

On our quadratic loss, PolarGD is empirically as robust as SignGD to ill-conditioning. We study this in Section 4. However, in contrast to SignGD, PolarGD is invariant to orthogonal transformations and hence converges faster on a larger class of ill-conditioned functions.

4. Trajectory-level view: implicit preconditioning of PolarGD

The previous section explains local progress at a fixed iterate. What remains is the trajectory-level question: why can PolarGD remain effective on dense, ill-conditioned Hessians? The key is that the polar direction can be rewritten as an iterate-dependent preconditioned gradient step. On quadratics, this exposes a mechanism invisible from Hessian condition number alone: the conditioning of the implicit preconditioner is governed by the row covariance ($XX^\top$) of the current iterate.

4.1. Why understanding needs to go beyond one step

Local preferences need not be consistent along a trajectory. Even on convex quadratics, the sign of $\Delta_{\mathrm{Pol}}(X) - \Delta_{\mathrm{GD}}(X)$ can change with X, so pointwise one-step comparisons don't imply global behaviour. Appendix C gives a 2×2 example where the local winner alternates.

Vanishing returns of greedy PolarGD near solution One step analysis does not describe how the optimal step size γ^* evolves along a trajectory:

Proposition 3 *For $f(X) = \frac{1}{2}\mathrm{tr}(X^T H X)$, the exact line search step-size $\gamma_D^*(X)$ satisfies*

$$\gamma_{GD}^*(cX) = \gamma_{GD}^*(X), \qquad \gamma_{Pol}^*(cX) = c\gamma_{Pol}^*(X)$$

Thus as $X_k \to 0$, PolarGD struggles to make progress while GD does not (proof in D).

4.2. PolarGD as an implicit preconditioned method

Let $G = U\Sigma_G V^\top$ be a thin SVD and $Q = \mathrm{polar}(G) := UV^\top$. Then

$$Q = (GG^\top)^{\dagger/2}G, \qquad QQ^\top = \Pi_{\mathrm{range}(G)}, \tag{2}$$

so polar normalization whitens the row covariance of the direction on the active row-space. Since $G = HX$ and $\Sigma := XX^\top$,

$$Q = \mathrm{polar}(HX) = (H\Sigma H)^{\dagger/2}HX =: A(X)X, \qquad A(X) := (H\Sigma H)^{\dagger/2}H, \tag{3}$$

so PolarGD applies a left preconditioner determined by the current iterate. The cleanest identities occur in the full-row-rank regime $d \le m$, $\mathrm{rank}(X) = d$.

Proposition 4 (Exact progress coefficient for PolarGD) *Assume $H \succ 0$ and $\mathrm{rank}(X) = d$. Let $Q = \mathrm{polar}(HX)$ and $X^+ = X - \gamma^\star(Q)Q$. Then*

$$\gamma^\star(Q) = \frac{\|HX\|_*}{\mathrm{tr}(H)}, \qquad f(X^+) = f(X) - \frac{\|HX\|_*^2}{2\mathrm{tr}(H)}, \tag{4}$$

and equivalently

$$\frac{f(X^+)}{f(X)} = 1 - p(X), \qquad p(X) := \frac{\|HX\|_*^2}{\mathrm{tr}(H)\mathrm{tr}(HXX^\top)}. \tag{5}$$

4.3. Near-rank-deficiency and softened polar updates

The representation $Q = (GG^\top)^{\dagger/2}G$ exposes a failure mode: tiny nonzero singular values are inverted. Two simple variants reduce this sensitivity: rank-truncated polar $Q^{(s)} := U_{:,1:s}V_{:,1:s}^\top$, which drops small modes, and damped polar

$$Q_\varepsilon := (GG^\top + \varepsilon I_d)^{-1/2}G,$$

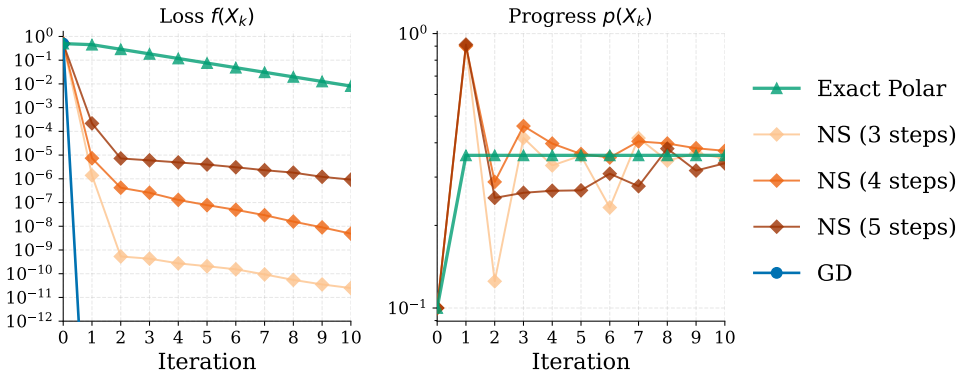

Figure 3: *(Left)* Loss $f(X_k)$, *(Right)* Exact progress coefficient $p(X_k)$. Initialized at $X_0 = Q^\top \mathrm{diag}(1, \delta, \ldots, \delta)Q$ for $\delta = 10^{-4}$, Q random orthogonal, $H = I_d$, $d = 10$, with exact line search throughout. Newton–Schulz iterations recover from near-singular initialization while exact PolarGD struggles; GD shown for reference. Details in Appendix A.

mapping each singular value to $\sigma_i/\sqrt{\sigma_i^2 + \varepsilon}$. Both are analyzed in Appendix F. In full row rank, $QQ^\top = I_d$, hence $\mathrm{tr}(Q^\top HQ) = \mathrm{tr}(H)$. Combined with the line-search identity from Section 3:

Finite Newton–Schulz as a soft spectral filter. Practical Muon computes approximate orthogonalization via a small number of Newton–Schulz iterations. Any fixed finite polynomial iteration preserves singular vectors and applies a scalar filter $\sigma_i \mapsto \phi_q(\sigma_i)$. Exact polar applies the hard filter $\sigma_i \mapsto 1$ with unbounded amplification $1/\sigma_i$. In contrast, finite Newton–Schulz satisfies

$$\phi_q(\sigma) = c_q\sigma + O(\sigma^3) \qquad \text{as } \sigma \downarrow 0,$$

behaving like a damped polar update on tiny singular modes. This explains why approximate orthogonalization can be more robust near rank deficiency than exact polar; see Appendix G and Figure 4.3.

5. Conclusion

We isolated the geometric mechanisms driving Muon updates by analyzing an exact-polar idealization, PolarGD, on matrix quadratics. PolarGD benefits from orthogonal equivariance and trajectory-level self-preconditioning, but is not uniformly superior to GD or SignGD. Its fast regime arises near row-isotropic iterates; near-rank deficiency can produce pessimistic behavior because exact polar normalization overweights tiny singular modes. Our analysis abstracts away stochasticity, momentum, nonlinear network effects, and approximate Newton–Schulz orthogonalization; extending the theory to these ingredients remains an important next step.

Acknowledgements

The research reported in this publication was supported by funding from King Abdullah University of Science and Technology (KAUST): i) KAUST Baseline Research Scheme, ii) Center of Excellence for Generative AI, under award number 5940, iii) SDAIA-KAUST Center of Excellence in Artificial Intelligence and Data Science. This research was partially funded by the European Commission under the grant No. 10119523 (OpenEuroLLM). Antonio Orvieto acknowledges the financial support of the Hector Foundation.

References

[1] Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. In *OPT 2024: Optimization for Machine Learning*, 2024. 1

[2] Jeremy M. Cohen, Simran Kaur, Yuanzhi Li, J. Zico Kolter, and Ameet Talwalkar. Gradient descent on neural networks typically occurs at the edge of stability, 2022. URL https://arxiv.org/abs/2103.00065. 24

[3] Damek Davis and Dmitriy Drusvyatskiy. When do spectral gradient updates help in deep learning? *arXiv preprint arXiv:2512.04299*, 2025. 53

[4] Antoine Gonon, Andreea-Alexandra Muşat, and Nicolas Boumal. Insights on muon from simple quadratics. *arXiv preprint arXiv:2602.11948*, 2026. 53, 54

[5] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Delving deep into rectifiers: Surpassing human-level performance on imagenet classification, 2015. URL https://arxiv.org/abs/1502.01852. 24

[6] Xiaowen Jiang, Andrei Semenov, and Sebastian U Stich. Enhancing llm training via spectral clipping. *arXiv preprint arXiv:2603.14315*, 2026. 53, 55

[7] Keller Jordan, Jeremy Bernstein, Ben Rappazzo, Vlado Boža, Jiacheng You, Franz Cesista, and Braden Koszarsky. Modded-nanoGPT: Speedrunning the nanoGPT baseline, 2024. URL https://github.com/KellerJordan/modded-nanogpt. GitHub repository; additional contributors: @fern-bear.bsky.social, @Grad62304977. 1

[8] Keller Jordan, Yuchen Jin, Vlado Boža, 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/. Technical blog post. 1

[9] Dmitry Kovalev. Understanding gradient orthogonalization for deep learning via non-Euclidean trust-region optimization, 2025. URL https://arxiv.org/abs/2503.12645. 1

[10] Alex Krizhevsky, Vinod Nair, and Geoffrey Hinton. Cifar-10 (canadian institute for advanced research), 2009. URL http://www.cs.toronto.edu/~kriz/cifar.html. 24

[11] Frederik Kunstner and Francis Bach. Scaling laws for gradient descent and sign descent for linear bigram models under zipf's law, 2025. URL https://arxiv.org/abs/2505.19227. 4, 11

[12] Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, Yanru Chen, Huabin Zheng, Yibo Liu, Shaowei Liu, Bohong Yin, Weiran He, Han Zhu, Yuzhi Wang, Jianzhou Wang, Mengnan Dong, Zheng Zhang, Yongsheng Kang, Hao Zhang, Xinran Xu, Yutao Zhang, Yuxin Wu, Xinyu Zhou, and Zhilin Yang. Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*, 2025. URL https://arxiv.org/abs/2502.16982. 1

[13] Jianhao Ma, Yu Huang, Yuejie Chi, and Yuxin Chen. Preconditioning benefits of spectral orthogonalization in muon. *arXiv preprint arXiv:2601.13474*, 2026. 53, 55

[14] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019. URL https://arxiv.org/abs/1912.01703. 11

[15] Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs. In *Forty-second International Conference on Machine Learning*, 2025. 1

[16] Ishaan Shah, Anthony M. Polloreno, Karl Stratos, Philip Monk, Adarsh Chaluvaraju, Andrew Hojel, Andrew Ma, Anil Thomas, Ashish Tanwer, Darsh J. Shah, Khoi Nguyen, Kurt Smith, Michael Callahan, Michael Pust, Mohit Parmar, Peter Rushton, Platon Mazarakis, Ritvik Kapila, Saurabh Srivastava, Somanshu Singla, Tim Romanski, Yash Vanjani, and Ashish Vaswani. Practical efficiency of Muon for pretraining. *arXiv preprint arXiv:2505.02222*, 2025. URL https://arxiv.org/abs/2505.02222. 1

[17] Kaiyue Wen, David Leo Wright Hall, Tengyu Ma, and Percy Liang. Fantastic pretraining optimizers and where to find them. In *International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=2J51qUZ0iG. 1

[18] Tianyue H. Zhang, Lucas Maes, Alan Milligan, Alexia Jolicoeur-Martineau, Ioannis Mitliagkas, Damien Scieur, Simon Lacoste-Julien, and Charles Guille-Escuret. Understanding adam requires better rotation dependent assumptions, 2025. URL https://arxiv.org/abs/2410.19964. 12

Contents

1 Introduction **1**

2 Notation and Setup **2**

3 One-step geometry on matrix quadratics **3**

4 Trajectory-level view: implicit preconditioning of PolarGD **4**

4.1 Why understanding needs to go beyond one step 5

4.2 PolarGD as an implicit preconditioned method 5

4.3 Near-rank-deficiency and softened polar updates 5

5 Conclusion **6**

A Experimental Details **10**

A.1 Learning rate selection. 10

A.2 Initialization-Hessian Regimes (Figure 3) 11

A.3 Experiment 2: Zipfian Language Modeling Proxy (Figure 3) 11

A.4 Spectral Recovery under Rank Deficiency (Figure 4.3) 12

B Proofs for Section 3 **13**

B.1 Exact line-search quantities for GD and SignGD (diagonal H) 15

B.1.1 Local comparison in the diagonal case (explicit but initialization-dependent) 17

B.2 Proof of Theorem 2 17

B.3 Orthogonal equivariance and coordinate dependence 19

C An analytic 2×2 example where the local winner alternates **21**

D Diminishing Returns of PolarGD and SignGD under Line Search **23**

D.1 Proof of Proposition 3 23

D.2 Insights from Cifar5k 24

E Proofs and details for Section 4 **24**

E.1 Proofs for the implicit-preconditioning identities 24

E.2 A near-rank-deficient instance attaining the pessimistic PolarGD bound 27

E.3 Generic multiplicative recursion under exact line search 28

E.4 GD with exact line search: classical rate 28

E.5 PolarGD with exact line search: trajectory dependence through Σ_k 29

E.6 Rank-deficient and active-subspace extensions 30

E.6.1 Active-subspace preconditioning 31

E.6.2 Exact line search and progress in the rank-deficient case 33
E.6.3 Rank monotonicity and rank drops 35
E.6.4 A rank-r one-step exactness regime 36
E.7 Active-subspace root dynamics 37
E.8 Local comparison with active curvature 38

F Rank-truncated and damped polar directions 39
F.1 Rank-truncated polar direction 39
F.2 Damped polar direction 40

G Finite Newton–Schulz as a bounded spectral filter 40
G.1 A general polynomial Newton–Schulz filter 40
G.2 Comparison with exact polar, damped polar, and truncation 41
G.3 Near-zero behavior and damping equivalence 42
G.4 Bounded amplification 43

H Exact dynamics through the square-rooted gradient covariance 45
H.1 Optimizer Trajectories in Singular Value Space 53

I Comparison to concurrent work on spectral/polar methods 53

J Matrix Smoothness: A Generalization of this Work 55

Appendix A. Experimental Details

Below we provide implementation details of each experiment with a figure presented in the main text.

A.1. Learning rate selection.

Whenever we report results for a "best-tuned" dual-adapted learning rate (update $X_{k+1} = X_k - \eta\|G_t\|_*\mathrm{LMO}(G_t)$), the following two-stage grid search has been applied. In the first stage, we evaluate 50 learning rates drawn from an evenly-spaced logarithmic grid over $\{10^{-8}, 10^2\}$. We select the candidate η_c that minimizes the terminal loss $f(X_k)$ among runs which converged. In the second stage, we evaluate a further 50 points on a finer logarithmic interval $[\frac{\eta_c}{10}, 10\eta_c]$. The final reported learning rate is the one whose run first crosses a threshold of ε (in the least number of iterates). If no run reaches this threshold, we select the run which minimizes the terminal loss $f(X_K)$.

A.2. Initialization-Hessian Regimes (Figure 3)

Hessian Structures. We test three specific curvature profiles to isolate the effects of condition number and coordinate alignment. To model the Hessian dynamics arising in ill-conditioned, multi-output regression problems, we draw random matrices from the Marchenko Pastur (MP) ensemble:

$$H = \frac{1}{N} Z^\top Z, \qquad Z \in \mathbb{R}^{N \times d}, \quad Z_{ij} \sim \mathcal{N}(0,1)$$

Here the aspect ratio $\gamma := \frac{d}{N} \in (0, 1]$ (the ratio of input features to samples) dictates the conditioning. Using this framework, we construct the following Hessian regimes

- **Isotropic Hessian:** $H = I_d$. This regime has curvature $\kappa = 1$
- **Dense:** $H = \frac{1}{N} Z^\top Z$ with $\gamma = 0.8$. Precisely the raw MP matrix H defined above.
- **Diagonal:** $H = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$, where λ_i are the exact eigenvalues of the dense MP matrix.

Initialization Strategies.

- **Gaussian:** $X_0 \sim \mathcal{N}(0, 1)$ element-wise. This serves as a standard (analogous to neural network conditions) baseline.
- **Orthogonal:** X_0 is a random orthogonal matrix.
- **Vertex:** $X_0 \in \{-1, 1\}$ element-wise. Each entry is drawn randomly.

Within each Hessian-Initialization setting, the best learning rate for each optimizer is selected using the two stage learning rate grid (A.1) with threshold $\varepsilon = 10^{-8}$. We choose $d = 100$ and 150 iterations for each setting.

To empirically validate the full-rank assumption of Proposition 4, we track $\mathrm{rank}(HX_k)$ at each iterate, computed in float32 precision via PyTorch [14].

A.3. Experiment 2: Zipfian Language Modeling Proxy (Figure 3)

To evaluate how these optimizers perform under the heavy-tailed geometries typical of language modeling, we construct a multi-output linear regression proxy inspired by [11].

Linear Bigram Model. Consider fitting a linear next-token prediction model $X \in \mathbb{R}^{d \times d}$ on n token pairs $z_i, y_i \in \{0, 1\}^d$ representing one-hot encodings. The squared loss is given by

$$L(X) = \frac{1}{2n} \|ZX - Y\|_F^2,$$

where X_{ij} is an estimate for $p(\text{token}_i \mid \text{token}_j)$ for each $i, j \in [d]$. To simulate the heavy-tailed nature of language modeling, token frequencies follow a Zipfian distribution $\pi_k \propto k^{-\alpha}$ where $\alpha > 0$

controls the heaviness of the tail. Following their framework, we consider the equivalent quadratic problem

$$f(X) = \frac{1}{2}\mathrm{tr}((X - X^*)^\top H(X - X^*)),$$

initialized at $X_0 = 0$ and the solution $X^*_{ij} = \pi_i \pi_j$ is the rank 1 matrix representing the token frequencies and the Hessian is $H = \mathrm{diag}(\pi_1, \ldots, \pi_d)$.

Validating Rotational invariance for PolarGD. To explicitly test the rotational invariance of the optimizers, we additionally evaluate a rotated setting: $H = R^\top \mathrm{diag}(k^{-\alpha})R$, where $R \in \mathbb{R}^{d\times d}$ is a random orthogonal matrix. This rotation preserves the spectrum while completely destroying axis alignment. This approach follows recent work by Zhang et al. [18], which demonstrates that Adam is highly sensitive to random rotations.
Crucially, when applying this rotation to the Hessian, we leave the target solution X^* fixed in the standard coordinate basis. Rotating the target alongside the Hessian ($X^* \leftarrow R^T X^*$) would make the problem axis-aligned again – defeating the purpose of the rotation. We set $d = 100$ and for each $\alpha \in \{0.25, 0.5, 1.0\}$ and Hessian regime $\in \{\text{Diagonal}, \text{Rotated}\}$, we tune the best dual-adapted learning rate with the procedure in A.1 with threshold $\varepsilon = 10^{-15}$ on the relative frequency $f(X_k)/f(X_0)$.

A.4. Spectral Recovery under Rank Deficiency (Figure 4.3)

To empirically validate the theoretical claim that Newton-Schulz iterations can act as a soft spectral regularizer, we evaluate its robustness near rank deficiency compared to exact polar normalization.

Near singular Initialization. We isolate the spectral recovery dynamics by setting the Hessian to the identity matrix $H = I$ in dimension $d = 10$. We initialize the weights at a near rank-deficient state:

$$X_0 = Q^\top \mathrm{diag}(1, \delta, \ldots, \delta)Q, \qquad Q \in \mathbb{R}^{10\times 10} \text{ random orthogonal }, \delta = 10^{-5}$$

Newton-Schulz as a bounded Spectral Filter. We compare the *exact* PolarGD update against the practical Muon update with Newton-Schulz iterations to observe how tiny singular modes are handled. We further track the progress coefficient $p(X_k) := \frac{\|HX\|_*^2}{\mathrm{tr}(H)\mathrm{tr}(HXX^\top)}$ as introduced in 4.2.

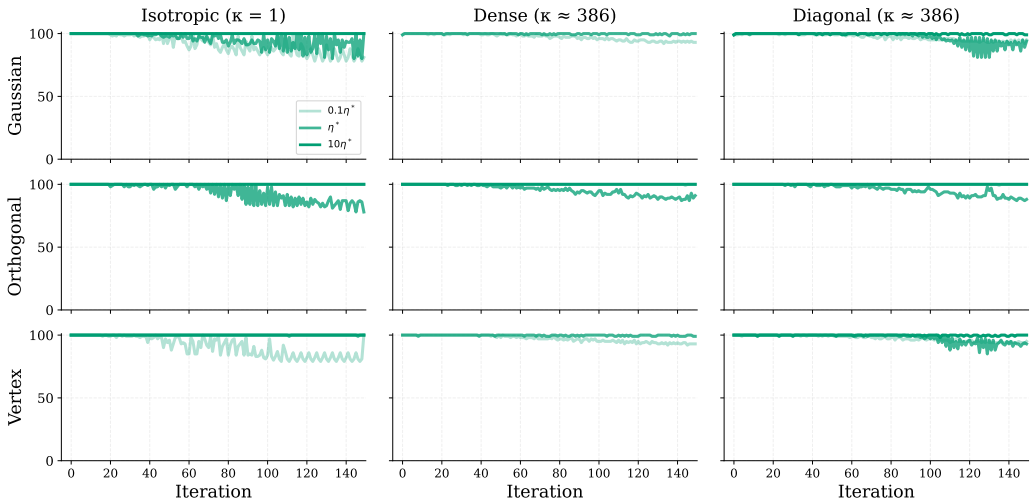

Figure 4: **Tracking the Rank for PolarGD:** For the experiment A.2 displayed in Figure 3, we track the rank of the gradient $G = HX_k$ for the best performing PolarGD learning rate $\eta^\star$ along with $0.1\eta^\star$ and $10\eta^\star$. We observe that, the rank remains near full across the nine regimes, validating the full-rank assumption used in Proposition 4.

Appendix B. Proofs for Section 3

Proposition 5 (Conditioning of the implicit preconditioner) *Assume $H \succ 0$ and $\mathrm{rank}(X) = d$. Then*

$$A(X)\Sigma A(X)^\top = I_d, \qquad A(X)^\top A(X) = \Sigma^{-1}, \tag{6}$$

and consequently $\kappa(A(X)) = \kappa(X) = \sqrt{\kappa(\Sigma)}$.

This is the first qualitative difference from GD. For GD on a quadratic, the linear rate is governed by $\kappa(H)$. For PolarGD, the conditioning of the implicit preconditioner is governed by the iterate covariance: a badly conditioned H can be harmless near row-isotropic iterates, while a well-conditioned H can be paired with an ill-conditioned preconditioner if X has tiny singular values.

Remark 6 *The full-rank assumption is empirically supported: Figure 4 shows that $\mathrm{rank}(HX)$ remains near full across all nine regimes of Figure 3. Appendix E.6 extends the theory to the non-full-rank case.*

The coefficient $p(X)$ is scale-invariant, measuring the shape of the iterate rather than its magnitude. Row-isotropic iterates give $p(X) = 1$, recovering one-step exactness. The pessimistic lower bound $\lambda_{\min}(H)/\mathrm{tr}(H)$ can be as small as $1/(\kappa d)$—potentially d times worse than GD—and is not merely a proof artifact: Appendix E.2 constructs a near-rank-deficient family attaining it, while GD is nearly one-step exact there. Thus exact line search alone does not make PolarGD uniformly better than GD; its advantage depends on how the iterate covariance evolves along the trajectory.

On our quadratic loss, PolarGD acts like a self-preconditioned GD step: the polar map whitens the gradient direction, so the effective preconditioner is controlled by the iterate covariance $XX^\top$, not by $\kappa(H)$ alone. This helps near row-isotropic iterates, but can be pessimistic near rank deficiency, where exact polar normalization gives tiny singular modes the same weight as large ones.

Proof [Proof of Proposition 38] Fix X, D, and write

$$g := \langle G, D\rangle, \qquad c_L := c_L(D) = \langle LD, D\rangle, \qquad c_M := c_M(D) = \langle MD, D\rangle.$$

The model-based step is $\gamma_L(D) = [g]_+/c_L$ when $c_L > 0$.

First, by the upper model in (55), for any $\gamma \geq 0$,

$$f(X - \gamma D) \leq f(X) - \gamma g + \frac{\gamma^2}{2} c_L.$$

If $g \leq 0$, then $\gamma_L(D) = 0$, so $X^+ = X$ and the claimed decrease bound is trivial. If $g > 0$, substituting $\gamma = g/c_L$ gives

$$f(X - \gamma_L(D)D) \leq f(X) - \frac{g^2}{c_L} + \frac{g^2}{2c_L} = f(X) - \frac{g^2}{2c_L}.$$

This proves (57).

We now prove the exact line-search bracket. Let

$$\phi(\gamma) := f(X - \gamma D), \qquad \gamma \geq 0.$$

The assumption (54), applied along the line $X - \gamma D$, implies the one-dimensional curvature bounds

$$-g + \gamma c_M \leq \phi'(\gamma) \leq -g + \gamma c_L.$$

Indeed, these inequalities follow by applying (54) at $X - \gamma D$ in direction $-tD$, dividing by $t > 0$, and taking $t \downarrow 0$.

If $g \leq 0$, then the lower bound gives $\phi'(\gamma) \geq 0$ for all $\gamma \geq 0$, so $\gamma^\star(D) = 0$ and $\Delta^\star(D) = 0$. The stated bracket is therefore immediate.

Assume now that $g > 0$ and $c_M > 0$. For $\gamma < g/c_L$, the upper derivative bound gives $\phi'(\gamma) < 0$, so the minimizer cannot occur before g/c_L. For $\gamma > g/c_M$, the lower derivative bound gives $\phi'(\gamma) > 0$, so the minimizer cannot occur after g/c_M. Hence

$$\frac{g}{c_L} \leq \gamma^\star(D) \leq \frac{g}{c_M}.$$

This proves the step-size bracket.

Finally, the exact decrease satisfies

$$\Delta^\star(D) = f(X) - \phi(\gamma^\star(D)).$$

The lower bound on $\Delta^\star(D)$ follows because exact line search does at least as well as the model-based step:

$$\Delta^\star(D) \geq f(X) - f(X - \gamma_L(D)D) \geq \frac{g^2}{2c_L}.$$

For the upper bound, the lower quadratic model gives, for every $\gamma \geq 0$,

$$\phi(\gamma) \geq f(X) - \gamma g + \frac{\gamma^2}{2} c_M \geq f(X) - \frac{g^2}{2c_M}.$$

Evaluating at $\gamma = \gamma^\star(D)$ yields

$$\Delta^\star(D) \leq \frac{g^2}{2c_M}.$$

Combining the cases $g \leq 0$ and $g > 0$ gives exactly (58), with $[g]_+$ in place of g. ■

Remark 7 (Scale invariance of direction rays) *For any $a > 0$, replacing D by aD leaves the optimized update unchanged. Indeed,*

$$\gamma^\star(aD) = \frac{\langle G, aD \rangle}{\operatorname{tr}((aD)^\top H(aD))} = \frac{1}{a}\gamma^\star(D)$$

on a quadratic, and similarly for the model step γ_L. Hence

$$X - \gamma^\star(aD)(aD) = X - \gamma^\star(D)D.$$

The same cancellation holds for the certified decrease because both the numerator and denominator scale quadratically in a. Thus line-search comparisons depend only on the direction ray, not on the normalization of the direction.

B.1. Exact line-search quantities for GD and SignGD (diagonal H)

GD (direction $D = G$). Using $\langle G, G \rangle = \|G\|_{\mathrm{F}}^2$ and $\operatorname{tr}(G^\mathsf{T} HG)$, the diagonal structure yields

$$\|G\|_{\mathrm{F}}^2 = \sum_{i=1}^{d} \|g_i\|_2^2 = \sum_{i=1}^{d} \lambda_i^2 \|e_i\|_2^2, \qquad \operatorname{tr}(G^\mathsf{T} HG) = \operatorname{tr}(E^\mathsf{T} H^3 E) = \sum_{i=1}^{d} \lambda_i^3 \|e_i\|_2^2.$$

Therefore the exact line-search step size and one-step decrease are

$$\gamma_{\mathrm{GD}}^\star = \frac{\|G\|_{\mathrm{F}}^2}{\operatorname{tr}(G^\mathsf{T} HG)} = \frac{\sum_i \lambda_i^2 \|e_i\|_2^2}{\sum_i \lambda_i^3 \|e_i\|_2^2}, \qquad \Delta_{\mathrm{GD}} = \frac{\|G\|_{\mathrm{F}}^4}{2\operatorname{tr}(G^\mathsf{T} HG)} = \frac{\left(\sum_i \lambda_i^2 \|e_i\|_2^2\right)^2}{2\sum_i \lambda_i^3 \|e_i\|_2^2}. \tag{7}$$

Moreover, the GD update remains rowwise-scaling (coupled only through the shared $\gamma_{\mathrm{GD}}^\star$):

$$e_i^+ = (1 - \gamma_{\mathrm{GD}}^\star \lambda_i)\, e_i, \qquad i = 1, \ldots, d.$$

SignGD (direction $D = S = \mathrm{sign}(G)$). Let $S = \mathrm{sign}(G)$ be the entrywise sign matrix with $\mathrm{sign}(0) = 0$. For diagonal H with nonnegative diagonal, S is coordinate-aligned with E on the active rows: if $\lambda_i > 0$ then $\mathrm{sign}(g_i) = \mathrm{sign}(e_i)$ entrywise. The gain term is

$$\langle G, S\rangle = \|G\|_1 = \sum_{i=1}^{d} \lambda_i \|e_i\|_1,$$

and the curvature term reduces to a weighted count of nonzeros:

$$\mathrm{tr}(S^\top H S) = \sum_{i=1}^{d} \lambda_i \|s_i\|_2^2, \qquad \|s_i\|_2^2 = \#\{j : \ e_{ij} \neq 0\}.$$

Hence

$$\gamma^\star_{\mathrm{Sign}} = \frac{\|G\|_1}{\mathrm{tr}(S^\top H S)} = \frac{\sum_i \lambda_i \|e_i\|_1}{\sum_i \lambda_i \|s_i\|_2^2}, \qquad \Delta_{\mathrm{Sign}} = \frac{\|G\|_1^2}{2\,\mathrm{tr}(S^\top H S)} = \frac{\left(\sum_i \lambda_i \|e_i\|_1\right)^2}{2\sum_i \lambda_i \|s_i\|_2^2}. \tag{8}$$

Vertex regime (one-step exactness, diagonal H). If $E = cS$ for some $c > 0$ and a sign pattern $S \in \{\pm 1, 0\}^{d\times m}$, then $G = HE = c(HS)$ and, since H is diagonal with nonnegative diagonal, $\mathrm{sign}(HS) = S$ on the support of S. Consequently $\gamma^\star_{\mathrm{Sign}} = c$ and $E^+ = 0$ in one step, consistent with Theorem 1(ii).

Exact line-search quantities for PolarGD (diagonal H) Let $G = U\Sigma V^\top$ be the thin SVD of G with rank $r = \mathrm{rank}(G)$ and define $Q = \mathrm{polar}(G) = UV^\top$. The gain remains $\langle G, Q\rangle = \|G\|_*$, while the diagonal structure makes the curvature term particularly interpretable:

$$\mathrm{tr}(Q^\top H Q) = \mathrm{tr}(U^\top H U) = \sum_{i=1}^{d} \lambda_i\, w_i, \qquad w_i := \|u_{i,:}\|_2^2, \quad \sum_{i=1}^{d} w_i = r, \quad 0 \le w_i \le 1. \tag{9}$$

Thus PolarGD "samples" the diagonal spectrum $\{\lambda_i\}$ through the *row-energy weights* of the left singular vectors of $G = HE$, which depend on the iterate (and therefore on initialization). The exact line-search step and one-step decrease are

$$\gamma^\star_{\mathrm{Pol}} = \frac{\|G\|_*}{\mathrm{tr}(Q^\top H Q)}, \qquad \Delta_{\mathrm{Pol}} = \frac{\|G\|_*^2}{2\,\mathrm{tr}(Q^\top H Q)} = \frac{\|G\|_*^2}{2\sum_i \lambda_i w_i}. \tag{10}$$

Full row rank simplification (shape condition). If $\mathrm{rank}(G) = d$ (full row rank), then necessarily $d \le m$ and $U \in \mathbb{R}^{d\times d}$ is orthogonal, so $QQ^\top = I_d$. In this case *for any* H (in particular for diagonal H),

$$\mathrm{tr}(Q^\top H Q) = \mathrm{tr}(HQQ^\top) = \mathrm{tr}(H) = \sum_{i=1}^{d} \lambda_i,$$

so the PolarGD curvature is constant (iterate-independent) and

$$\gamma^\star_{\mathrm{Pol}} = \frac{\|G\|_*}{\mathrm{tr}(H)}, \qquad \Delta_{\mathrm{Pol}} = \frac{\|G\|_*^2}{2\,\mathrm{tr}(H)}. \tag{11}$$

B.1.1. LOCAL COMPARISON IN THE DIAGONAL CASE (EXPLICIT BUT INITIALIZATION-DEPENDENT)

In the diagonal setting, the exact local comparison $\Delta_{\mathrm{Pol}} > \Delta_{\mathrm{GD}}$ is equivalent to

$$\frac{\|G\|_*^2}{\operatorname{tr}(Q^\top HQ)} > \frac{\|G\|_{\mathrm{F}}^4}{\operatorname{tr}(G^\top HG)} \quad \Longleftrightarrow \quad \|G\|_*^2 \operatorname{tr}(G^\top HG) > \|G\|_{\mathrm{F}}^4 \operatorname{tr}(Q^\top HQ).$$

Using (7) and (9), this becomes

$$\|G\|_*^2 \Big(\sum_{i=1}^d \lambda_i^3 \|e_i\|_2^2\Big) > \Big(\sum_{i=1}^d \lambda_i^2 \|e_i\|_2^2\Big)^2 \Big(\sum_{i=1}^d \lambda_i w_i\Big), \tag{12}$$

which makes explicit the two sources of iterate-dependence: (i) the singular spectrum of $G = HE$ through $\|G\|_*$, and (ii) the alignment of the left singular subspace of G with the coordinate axes through $(w_i)_i$.

Diagonal iterates (PolarGD coincides with SignGD). If $d = m$ and E is diagonal, then $G = HE$ is also diagonal, and $Q = \operatorname{polar}(G) = \operatorname{sign}(G)$ (diagonal $\pm 1/0$). Thus PolarGD and SignGD coincide on this invariant submanifold, and both act as an *additive shrinkage* of the active diagonal entries with a shared step. This regime is useful for explicit phase diagrams in 2×2 (see below).

B.2. Proof of Theorem 2

Proof [Proof of Theorem 2] Let

$$E := X - X^\star, \qquad G = HE.$$

The exact line-search update along a direction D is one-step exact precisely when there exists a line-search minimizer $\gamma^\star(D) \geq 0$ such that

$$E - \gamma^\star(D) D = 0.$$

Since the quadratic is nonnegative and minimized at $E = 0$, it is enough to characterize when the direction ray contains E.

GD. For GD, the direction is $D = G = HE$. A one-step exact update at a fixed $E \neq 0$ requires

$$E - \gamma HE = 0$$

for some $\gamma > 0$. Equivalently,

$$HE = \lambda E \qquad \text{with} \qquad \lambda = \frac{1}{\gamma} > 0.$$

Conversely, if $HE = \lambda E$ with $\lambda > 0$, then the line-search step $\gamma = 1/\lambda$ gives $E^+ = 0$. Since this point is the global minimizer, it is selected by exact line search.

It remains to characterize when this holds from every E. If $H = \alpha I_d$ with $\alpha > 0$, then $HE = \alpha E$ for all E, so GD is one-step exact from every point. Conversely, suppose GD is one-step exact from every E. Then every vector in $\mathbb{R}^d$ must be an eigenvector of H. Taking two nonzero vectors u, v with $Hu = \lambda_u u$ and $Hv = \lambda_v v$, the vector $u + v$ must also be an eigenvector:

$$H(u + v) = \lambda(u + v).$$

Thus $\lambda_u = \lambda_v = \lambda$. Since u, v were arbitrary, $H = \lambda I_d$. The positivity $\lambda > 0$ is necessary for convergence to $X^\star$ from every point.

SignGD. Let $S = \mathrm{sign}(G)$. The SignGD direction is $D = S$. A one-step exact update at E is possible if and only if

$$E = \gamma S$$

for some $\gamma > 0$. Writing $c := \gamma$, this gives $E = cS$. Since $G = HE = cHS$, the direction S must be consistent with the sign of the gradient:

$$S = \mathrm{sign}(G) = \mathrm{sign}(HS).$$

Thus SignGD is one-step exact if and only if there is a scalar $c > 0$ and a sign matrix S such that

$$E = cS \qquad \text{and} \qquad S = \mathrm{sign}(HS).$$

Under this condition, choosing $\gamma = c$ gives $E - \gamma S = 0$, so exact line search reaches a global minimizer.

If S has full support, this is exactly the condition stated in the theorem. If sparse sign matrices are allowed, the precise statement is the global equality $S = \mathrm{sign}(HS)$, including zeros. Equivalently, on the support of S, the signs must match, and off the support HS must be zero.

Now suppose H is symmetric, has positive diagonal, and is strictly diagonally dominant:

$$H_{ii} > \sum_{j \neq i} |H_{ij}| \qquad \text{for all } i.$$

Let $s \in \{\pm 1\}^d$. Then for each coordinate,

$$s_i(Hs)_i = H_{ii} + \sum_{j \neq i} H_{ij} s_i s_j \geq H_{ii} - \sum_{j \neq i} |H_{ij}| > 0.$$

Therefore $\mathrm{sign}(Hs) = s$. Applying this columnwise gives $\mathrm{sign}(HS) = S$ for every full-support sign matrix S. Hence any vertex initialization $E = cS$ with $S \in \{\pm 1\}^{d \times m}$ is solved in one exact line-search SignGD step.

PolarGD. Assume $H \succ 0$ and $\mathrm{rank}(G) = d$. Then $d \leq m$, and the polar factor satisfies

$$QQ^\top = I_d.$$

If PolarGD is one-step exact, then for some $\gamma > 0$,

$$E = \gamma Q.$$

Multiplying by the transpose gives

$$EE^\top = \gamma^2 QQ^\top = \gamma^2 I_d.$$

Thus $EE^\top = cI_d$ with $c = \gamma^2 > 0$.

Conversely, suppose $EE^\top = cI_d$ for some $c > 0$. Then E is full row rank and

$$GG^\top = HEE^\top H = cH^2.$$

Using the whitening representation of the polar factor,

$$Q = (GG^\top)^{-1/2} G,$$

we obtain

$$Q = (cH^2)^{-1/2} HE = \frac{1}{\sqrt{c}} H^{-1} HE = \frac{1}{\sqrt{c}} E,$$

where we used $H \succ 0$. Therefore $E = \sqrt{c}\, Q$, and the line-search step $\gamma = \sqrt{c}$ reaches $E^+ = 0$. Since $E = 0$ is the global minimizer, exact line search is one-step exact.

Finally, $EE^\top = cI_d$ is equivalent to E being a scaled partial isometry. Indeed, if $E = U\Sigma V^\top$ is a thin SVD and $EE^\top = cI_d$, then all d singular values are $\sqrt{c}$, so $E = \sqrt{c}\, UV^\top$, with $U \in \mathbb{R}^{d\times d}$ orthogonal and $V \in \mathbb{R}^{m\times d}$ having orthonormal columns. In the square case $m = d$, this is a scaled orthogonal matrix. ■

B.3. Orthogonal equivariance and coordinate dependence

This subsection formalizes which of the three canonical direction rules (GD, PolarGD, SignGD) are invariant (equivariant) under orthogonal changes of coordinates. These invariances justify working in a diagonal eigenbasis for H when analyzing GD and PolarGD, and clarify why the same reduction is *not* lossless for SignGD.

Orthogonal actions on matrices. For $P \in \mathbb{R}^{d\times d}$ and $R \in \mathbb{R}^{m\times m}$ orthogonal, define the left–right orthogonal action

$$\mathcal{T}_{P,R}(X) := PXR^\mathsf{T}.$$

The Frobenius inner product and the Schatten norms are invariant under $\mathcal{T}_{P,R}$: for any $A, B \in \mathbb{R}^{d\times m}$,

$$\langle \mathcal{T}_{P,R}(A), \mathcal{T}_{P,R}(B)\rangle = \langle A, B\rangle,\ \|\mathcal{T}_{P,R}(A)\|_{\mathrm{F}} = \|A\|_{\mathrm{F}},\ \|\mathcal{T}_{P,R}(A)\|_* = \|A\|_*,\ \|\mathcal{T}_{P,R}(A)\|_2 = \|A\|_2.$$

Lemma 8 (Bi-orthogonal equivariance of the polar factor) *Let $G \in \mathbb{R}^{d\times m}$ have thin SVD $G = U\Sigma V^\mathsf{T}$ and define its thin polar factor $\operatorname{polar}(G) := UV^\mathsf{T}$. Then for any orthogonal $P \in \mathbb{R}^{d\times d}$ and $R \in \mathbb{R}^{m\times m}$,*

$$\operatorname{polar}(PGR^\mathsf{T}) = P \operatorname{polar}(G)\, R^\mathsf{T}.$$

Proof Write $G = U\Sigma V^\mathsf{T}$. Then $PGR^\mathsf{T} = (PU)\Sigma(RV)^\mathsf{T}$ is a thin SVD because PU and RV have orthonormal columns. Hence $\operatorname{polar}(PGR^\mathsf{T}) = (PU)(RV)^\mathsf{T} = P(UV^\mathsf{T})R^\mathsf{T} = P \operatorname{polar}(G)\, R^\mathsf{T}$. ■

Quadratic objectives and change of basis. Consider the (shifted) quadratic model

$$f_H(E) = \frac{1}{2}\operatorname{tr}(E^\top H E), \qquad H = H^\top \succeq 0, \qquad E \in \mathbb{R}^{d\times m}.$$

Under the change of variables $E = \mathcal{T}_{P,R}(Y) = PYR^\top$, the objective becomes

$$f_H(PYR^\top) = \frac{1}{2}\operatorname{tr}\big(Y^\top (P^\top H P)\, Y\big) =: f_{\tilde{H}}(Y), \qquad \tilde{H} := P^\top H P.$$

Moreover, the gradients transform equivariantly:

$$\nabla_E f_H(PYR^\top) = H(PYR^\top) = P(\tilde{H}Y)R^\top = P\,\nabla_Y f_{\tilde{H}}(Y)\,R^\top.$$

Proposition 9 (Rotation equivariance of GD and PolarGD on quadratics) *Fix $H = H^\top \succeq 0$. Let $\tilde{H} = P^\top H P$ for some orthogonal $P \in \mathbb{R}^{d\times d}$, and let $R \in \mathbb{R}^{m\times m}$ be orthogonal. Consider corresponding iterates related by $E_k = PY_kR^\top$.*

(i) (GD direction) *If $D_{\mathrm{GD}}(E) := \nabla f_H(E) = HE$, then*

$$D_{\mathrm{GD}}(E_k) = P\,D_{\mathrm{GD}}^{(\tilde{H})}(Y_k)\,R^\top, \qquad D_{\mathrm{GD}}^{(\tilde{H})}(Y) := \nabla f_{\tilde{H}}(Y) = \tilde{H}Y.$$

(ii) (PolarGD direction) *If $D_{\mathrm{Pol}}(E) := \operatorname{polar}(\nabla f_H(E))$, then*

$$D_{\mathrm{Pol}}(E_k) = P\,D_{\mathrm{Pol}}^{(\tilde{H})}(Y_k)\,R^\top, \qquad D_{\mathrm{Pol}}^{(\tilde{H})}(Y) := \operatorname{polar}(\nabla f_{\tilde{H}}(Y)),$$

by Lemma 8.

(iii) (Line search invariance) *For either direction choice $D \in \{D_{\mathrm{GD}}, D_{\mathrm{Pol}}\}$, the exact line-search step size*

$$\gamma^\star(D;E) := \arg\min_{\gamma\ge 0} f_H(E - \gamma D(E))$$

satisfies

$$\gamma^\star(D;E_k) = \gamma^\star(D;Y_k),$$

and the one-step decrease in objective is preserved:

$$f_H(E_k) - f_H(E_{k+1}) = f_{\tilde{H}}(Y_k) - f_{\tilde{H}}(Y_{k+1}).$$

Consequently, the GD and PolarGD trajectories are identical up to orthogonal change of basis.

Proof Parts (i) and (ii) follow from the gradient transformation and Lemma 8. For (iii), the exact line-search formulas on a quadratic depend only on $\langle G, D\rangle$ and $\operatorname{tr}(D^\top H D)$. Under $E = PYR^\top$ and $D = P\tilde{D}R^\top$, invariance of $\langle\cdot,\cdot\rangle$ and cyclicity of trace give $\langle G, D\rangle = \langle \tilde{G}, \tilde{D}\rangle$ and $\operatorname{tr}(D^\top H D) = \operatorname{tr}(\tilde{D}^\top \tilde{H}\tilde{D})$, hence the minimizing γ and the decrease coincide. ■

Corollary 10 (Diagonalization is lossless for GD and PolarGD) *Let $H = Q\Lambda Q^\top$ be an eigendecomposition with Q orthogonal and Λ diagonal. For GD and PolarGD (with any scalar step size rule, including exact line search), analyzing $f_H(E)$ is equivalent to analyzing $f_\Lambda(Y)$ under the change of variables $Y = Q^\top E$ (up to the same objective values and the same one-step decreases).*

Remark 11 (Why SignGD is not rotation-invariant) *SignGD uses the entrywise direction $D_{\mathrm{Sign}}(E) := \mathrm{sign}(\nabla f_H(E))$. In general, for a dense orthogonal P,*

$$\mathrm{sign}(PG) \neq P\,\mathrm{sign}(G),$$

so SignGD is not *equivariant under arbitrary orthogonal changes of basis and therefore diagonalizing H changes the algorithm. SignGD is equivariant only under signed permutations (coordinate relabelings and sign flips), i.e., orthogonal matrices with exactly one nonzero entry ± 1 per row and per column. Consequently, diagonal (or nearly diagonal) H should be interpreted as a* coordinate-aligned regime *for SignGD rather than a lossless reduction.*

Experimental implication. Because of Proposition 9, "rotating" a diagonal Hessian by an orthogonal change of basis should not change GD or PolarGD behavior (up to the corresponding rotation of the initialization), while it can substantially change SignGD behavior. This separation helps design experiments that isolate truly spectral effects (PolarGD) from coordinate effects (SignGD).

Appendix C. An analytic 2×2 example where the local winner alternates

This subsection gives a closed-form example showing that the local one-step preference between GD and PolarGD can *swap repeatedly* along a trajectory, even on a convex quadratic. This strengthens the message that one-step comparisons are informative but not compositional.

Consider $d = m = 2$, $X^\star = 0$, and a diagonal Hessian

$$H = \begin{pmatrix} \kappa & 0 \\ 0 & 1 \end{pmatrix}, \qquad \kappa > 1.$$

Restrict attention to diagonal iterates $X = \mathrm{diag}(p, q)$ with $pq \neq 0$. Then $G = HX = \mathrm{diag}(\kappa p, q)$ and the exact one-step decreases under line search are

$$\Delta_{\mathrm{GD}}(X) = \frac{(\kappa^2 p^2 + q^2)^2}{2(\kappa^3 p^2 + q^2)}, \qquad \Delta_{\mathrm{Pol}}(X) = \frac{(\kappa|p| + |q|)^2}{2(\kappa + 1)}, \tag{13}$$

where we used that for full-rank 2×2 gradients the polar factor $Q = \mathrm{polar}(G)$ is orthogonal and $\mathrm{tr}(Q^\top H Q) = \mathrm{tr}(H) = \kappa + 1$.

Define the (scale-invariant) anisotropy parameter

$$s(X) := \frac{\kappa|p|}{|q|}.$$

A direct simplification of $\Delta_{\mathrm{Pol}}(X) > \Delta_{\mathrm{GD}}(X)$ using (13) shows that, for diagonal X, the local winner is determined by the sign of the scalar polynomial

$$g_\kappa(s) := (s+1)^2(\kappa s^2 + 1) - (\kappa + 1)(s^2 + 1)^2. \tag{14}$$

In particular, $\Delta_{\mathrm{Pol}}(X) > \Delta_{\mathrm{GD}}(X)$ if and only if $g_\kappa(s(X)) > 0$.

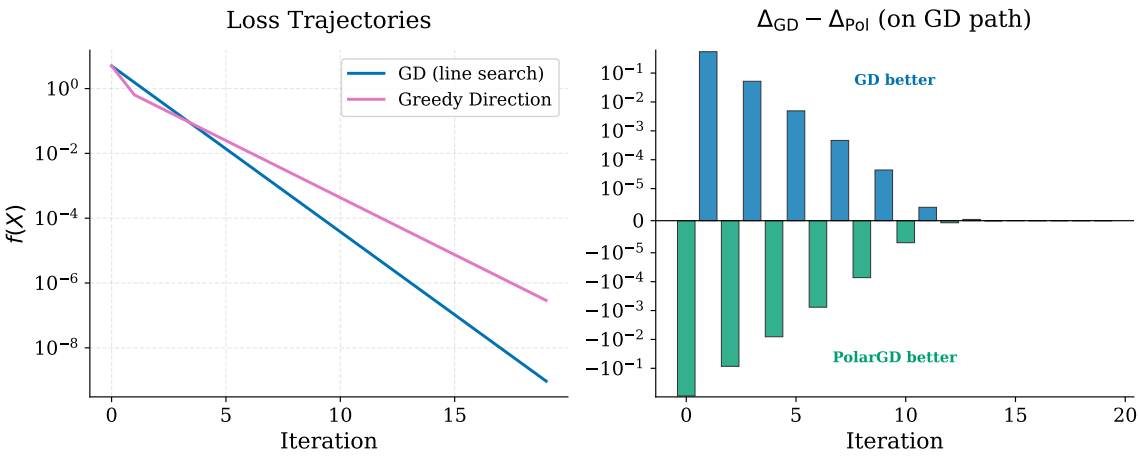

Figure 5: **Local Advantage Oscillates along GD Path**
2×2 diagonal quadratic with $H = \mathrm{diag}(\kappa, 1)$, $\kappa = 10$, initialized at $X_0 = \mathrm{diag}(1, \sqrt{\kappa})$. *(Left)* Loss trajectories for **GD (exact line search)** and a **Greedy** comparison. At each iterate X_k, Greedy selects the direction $D_k \in \{G_k, Q_k\}$ which yields the largest one step decrease. Despite always choosing the locally better step, Greedy underperforms GD globally. *(Right)* The one-step advantage $\Delta_{\mathrm{GD}} - \Delta_{\mathrm{Pol}}$ along the GD trajectory, visualizing the example developed in Appendix C

.

A two-cycle under GD with exact line search. Now run *GD with exact line search* from the diagonal initialization $X_0 = I_2$. Since H and X_0 are diagonal, GD preserves diagonality: $X_k = \mathrm{diag}(p_k, q_k)$ for all k. Moreover, one can compute from the exact GD step size $\gamma_k^\star = \|G_k\|_{\mathrm{F}}^2 / \mathrm{tr}(G_k^\top H G_k)$ that the anisotropy parameter satisfies

$$s(X_{k+1}) = \frac{1}{s(X_k)}. \tag{15}$$

Since $s(X_0) = \kappa$ and $s(X_1) = 1/\kappa$, alternation follows once we evaluate g_κ at these two values. A direct computation gives

$$g_\kappa(\kappa) = \kappa(\kappa-1)^2(\kappa+1) > 0, \qquad g_\kappa(1/\kappa) = -\frac{(\kappa-1)^2(\kappa+1)(\kappa^2+\kappa+1)}{\kappa^4} < 0,$$

for every $\kappa > 1$. Therefore, along the exact-line-search GD trajectory starting from $X_0 = \mathrm{diag}(1, \sqrt{\kappa})$, the local comparison alternates at each step: $\Delta_{\mathrm{Pol}}(X_{2t}) > \Delta_{\mathrm{GD}}(X_{2t})$ and $\Delta_{\mathrm{Pol}}(X_{2t+1}) < \Delta_{\mathrm{GD}}(X_{2t+1})$ for all $t \geq 0$.

Strict alternation of the local winner. Evaluating (14) at $s = \kappa$ and $s = 1/\kappa$ yields

$$g_\kappa(\kappa) = \kappa(\kappa+1)(\kappa-1)^2 > 0, \qquad g_\kappa(1/\kappa) = -\frac{(\kappa+1)(\kappa-1)^2(\kappa^2+\kappa+1)}{\kappa^4} < 0.$$

Therefore, along the GD trajectory $\{X_k\}$,

$$\Delta_{\mathrm{Pol}}(X_{2t}) > \Delta_{\mathrm{GD}}(X_{2t}), \qquad \Delta_{\mathrm{GD}}(X_{2t+1}) > \Delta_{\mathrm{Pol}}(X_{2t+1}), \qquad \forall\, t \geq 0.$$

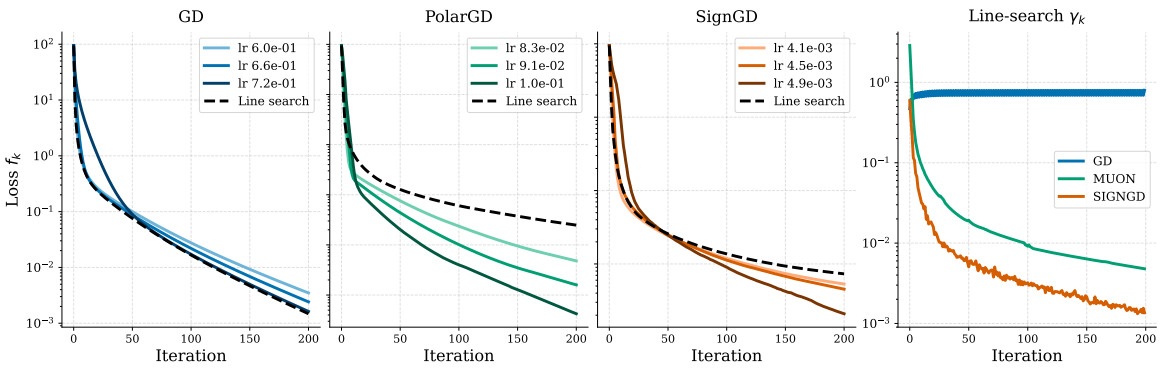

Figure 6: **Diminishing Returns of Line Search** Comparing three fixed learning rates (best from grid A.1) to line search for GD, SignGD and PolarGD on a quadratic with dense MP Hessian ($d = 20, \gamma = 0.8$), and Gaussian initialization ($X_0 \sim \mathcal{N}(0, 1)$). We see that while line search typically outperforms a tuned learning rate for GD, this is not true for PolarGD and SignGD, where a fixed learning rate can significantly outperform line search. The rightmost panel displays the step sizes selected by each line search method. Consistent with Proposition 3, this step size decays to zero as $X_k \to 0$ for PolarGD and SignGD whereas the GD step size remains stable throughout. This further emphasizes the necessity to analyze beyond exact line search.

That is, even in this simple convex quadratic, the local one-step preference between GD and PolarGD swaps *every iteration* along a natural trajectory. This illustrates why local comparisons should be interpreted as diagnostics at a point, not as a global ranking.

Appendix D. Diminishing Returns of PolarGD and SignGD under Line Search

We begin by restating and proving Proposition 3 from the main text.

D.1. Proof of Proposition 3

Proposition 12 *For $f(X) = \frac{1}{2}\mathrm{tr}(X^T HX)$, the exact line search step-size $\gamma^*_D(X)$ satisfies*

$$\gamma^*_{GD}(cX) = \gamma^*_{GD}(X), \qquad \gamma^*_{Pol}(cX) = c\gamma^*_{Pol}(X)$$

Proof Consider the quadratic objective $f(X) = \frac{1}{2}\mathrm{tr}(X^\top HX)$ with $H \succ 0$. For any search direction D, the exact step size γ^* is found by minimizing $f(X - \gamma D)$, yielding:

Gradient Descent. For GD, the direction is the gradient $G = \nabla f(X) = HX$. The step size is:

$$\gamma^*_{\mathrm{GD}}(X) = \frac{\mathrm{tr}((HX)^\top HX)}{\mathrm{tr}((HX)^\top H(HX))} = \frac{\|HX\|_F^2}{\mathrm{tr}(X^\top H^3 X)}.$$

Replacing X by cX multiplies both numerator and denominator by c^2, leaving the ratio unchanged.

PolarGD. For PolarGD, the direction is $Q = \mathrm{polar}(HX)$. By the properties of the Polar decomposition, $\mathrm{polar}(cHX) = \mathrm{polar}(HX) = Q$ for any $c > 0$. The step size is:

$$\gamma^*_{\mathrm{Pol}}(X) = \frac{\mathrm{tr}((HX)^\top Q)}{\mathrm{tr}(Q^\top HQ)}.$$

Since $\|H(cX)\|_* = c\|HX\|_*$ while $\mathrm{tr}(Q^\top HQ)$ does not depend on c, we obtain $\gamma^*_{\mathrm{Pol}}(cX) = c\,\gamma^*_{\mathrm{Pol}}(X)$. ■

Remark 13 (SignGD) *The same logic applies to SignGD, where the direction is $D = \mathrm{sign}(HX)$. Since $\mathrm{sign}(c \cdot HX) = \mathrm{sign}(HX)$ for $c > 0$, the direction remains unchanged under scaling. Consequently, we have $\gamma^*_{Sign}(cX) = c\gamma^*_{Sign}(X)$ and the exact line search step size vanishes as $X \to 0$.*

D.2. Insights from Cifar5k

We evaluate on a two-layer fully-connected network with ReLU activations, following the CIFAR-5K testbed of Cohen et al. [2] – a subsampled collection of the CIFAR 10 [10] dataset consisting of 5 classes, each with $1,000$ samples. The architecture has input dimension $d = 3{,}072$ (flattened $32 \times 32 \times 3$ images), hidden dimension $h = 4{,}608$ ($= 1.5\,d$), output dimension 5, and no bias terms; weights are initialized via Kaiming normal initialization [5]. We train on a balanced 5-class, 5,000-sample subset of CIFAR-10 using cross-entropy loss. The updates are dual-adapted, *full-batch* and with momentum set to zero throughout.

We compare standard gradient descent (GD) and idealized Muon (PolarGD). Since we are in the full batch setting with no momentum, the only hyperparameter is the step size. For each optimizer we evaluate (i) a grid of fixed learning rates and (ii) an exact line search implemented via Brent's method on the interval $[0, 4]$ with tolerance 10^{-5}; each line-search step requires approximately 10–15 forward passes. All runs share a fixed random seed and are trained for 100 iterations.

Appendix E. Proofs and details for Section 4

E.1. Proofs for the implicit-preconditioning identities

Proof [Proof of Proposition 5] Let $\Sigma := XX^\top$ and $A := A(X) = (H\Sigma H)^{-1/2}H$. Since $H \succ 0$ and $\mathrm{rank}(X) = d$, we have $\Sigma \succ 0$, so $H\Sigma H \succ 0$ and all inverse square roots below are ordinary inverse square roots.

First,

$$A\Sigma A^\top = (H\Sigma H)^{-1/2} H\Sigma H (H\Sigma H)^{-1/2} = I_d.$$

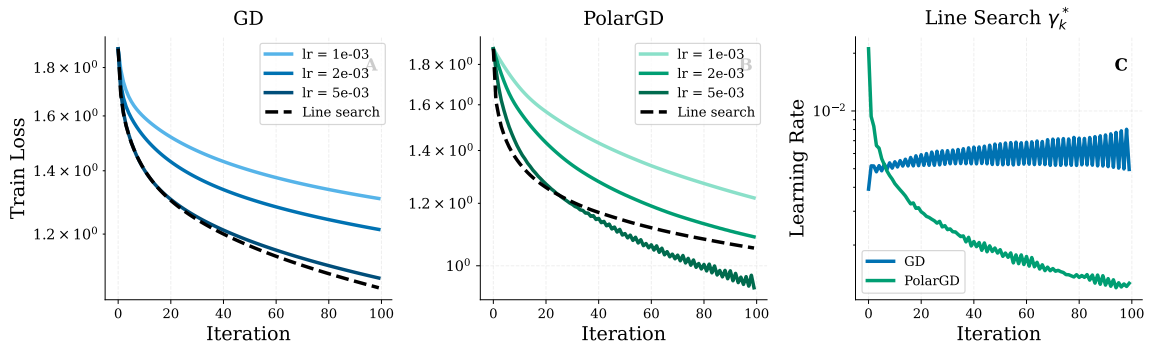

Figure 7: **Line Search on MLPs.**
Training loss on CIFAR-5K under fixed dual-adapted learning rates and line search (dashed). The three fixed rates shown are the best-performing from a grid search. For GD, the line search is optimal; for Muon, a fixed learning rate outperforms it. This is consistent with the local-global gap observed in the quadratic case (as discussed in Section 3). The rightmost panel plots the exact line search step size: for GD it oscillates, whereas for PolarGD it decays to zero. Further implementation details are given in D.2

.

Second, using $(H\Sigma H)^{-1} = H^{-1}\Sigma^{-1}H^{-1}$, we obtain

$$A^\top A = H(H\Sigma H)^{-1}H = H(H^{-1}\Sigma^{-1}H^{-1})H = \Sigma^{-1}.$$

Thus the squared singular values of A are the eigenvalues of Σ^{-1}. Since the eigenvalues of $\Sigma = XX^\top$ are $\sigma_i(X)^2$, the singular values of A are $1/\sigma_i(X)$. Therefore

$$\kappa(A) = \frac{1/\sigma_{\min}(X)}{1/\sigma_{\max}(X)} = \kappa(X) = \sqrt{\kappa(\Sigma)}.$$

■

Proof [Proof of Proposition 4] Let $G = HX$ and $Q = \mathrm{polar}(G)$. Since $H \succ 0$ and $\mathrm{rank}(X) = d$, also $\mathrm{rank}(G) = d$. Hence $QQ^\top = I_d$, and the line-search curvature along the polar direction is

$$\mathrm{tr}(Q^\top HQ) = \mathrm{tr}(HQQ^\top) = \mathrm{tr}(H).$$

Using the exact line-search identity from Section 3 and the polar duality identity $\langle G, Q\rangle = \|G\|_*$, we get

$$\gamma^\star(Q) = \frac{\langle G, Q\rangle}{\mathrm{tr}(Q^\top HQ)} = \frac{\|G\|_*}{\mathrm{tr}(H)}, \qquad f(X) - f(X^+) = \frac{\|G\|_*^2}{2\mathrm{tr}(H)}.$$

Since $f(X) = \frac{1}{2}\mathrm{tr}(X^\top HX) = \frac{1}{2}\mathrm{tr}(H\Sigma)$, this is equivalently

$$\frac{f(X^+)}{f(X)} = 1 - \frac{\|G\|_*^2}{\mathrm{tr}(H)\mathrm{tr}(H\Sigma)} = 1 - p(X).$$

It remains to prove the bounds on $p(X)$. For the upper bound, write $G = H^{1/2}(H^{1/2}X)$. The Schatten inequality $\|AB\|_* \le \|A\|_F \|B\|_F$ gives

$$\|G\|_*^2 \le \|H^{1/2}\|_F^2 \|H^{1/2}X\|_F^2 = \mathrm{tr}(H)\mathrm{tr}(X^\top HX) = \mathrm{tr}(H)\mathrm{tr}(H\Sigma),$$

so $p(X) \le 1$.

For the first lower bound, use $\|G\|_* \ge \|G\|_F$. Since

$$\|G\|_F^2 = \mathrm{tr}(X^\top H^2 X) = \mathrm{tr}(H^2\Sigma) \ge \lambda_{\min}(H)\mathrm{tr}(H\Sigma),$$

we have

$$p(X) = \frac{\|G\|_*^2}{\mathrm{tr}(H)\mathrm{tr}(H\Sigma)} \ge \frac{\lambda_{\min}(H)}{\mathrm{tr}(H)}.$$

For the second lower bound, let $s_{\min}$ and $s_{\max}$ be the smallest and largest eigenvalues of Σ. Since $s_{\min} I_d \preceq \Sigma \preceq s_{\max} I_d$, we have

$$H\Sigma H \succeq s_{\min} H^2, \qquad \mathrm{tr}(H\Sigma) \le s_{\max}\mathrm{tr}(H).$$

By operator monotonicity of the matrix square root,

$$(H\Sigma H)^{1/2} \succeq \sqrt{s_{\min}}\, H.$$

Therefore

$$\|G\|_* = \mathrm{tr}\big((GG^\top)^{1/2}\big) = \mathrm{tr}\big((H\Sigma H)^{1/2}\big) \ge \sqrt{s_{\min}}\mathrm{tr}(H).$$

Combining this with $\mathrm{tr}(H\Sigma) \le s_{\max}\mathrm{tr}(H)$ yields

$$p(X) \ge \frac{s_{\min}\mathrm{tr}(H)^2}{\mathrm{tr}(H)\, s_{\max}\mathrm{tr}(H)} = \frac{1}{\kappa(\Sigma)}.$$

Together, the two lower bounds prove

$$\max\left\{\frac{\lambda_{\min}(H)}{\mathrm{tr}(H)}, \frac{1}{\kappa(\Sigma)}\right\} \le p(X) \le 1.$$

Finally, if $\Sigma = cI_d$, then $H\Sigma H = cH^2$. Hence

$$\|G\|_* = \mathrm{tr}\big((H\Sigma H)^{1/2}\big) = \sqrt{c}\,\mathrm{tr}(H), \qquad \mathrm{tr}(H\Sigma) = c\,\mathrm{tr}(H),$$

and therefore $p(X) = 1$. Thus $f(X^+) = 0$. ■

E.2. A near-rank-deficient instance attaining the pessimistic PolarGD bound

Proposition 4 gives the lower bound

$$p(X) \geq \frac{\lambda_{\min}(H)}{\operatorname{tr}(H)}.$$

We now show that this bound is essentially tight. The example also illustrates why exact polar normalization can behave poorly near rank deficiency.

Consider the square case $m = d$. Let

$$H = \operatorname{diag}(\mu, L, \ldots, L), \qquad 0 < \mu \ll L, \qquad \kappa := \frac{L}{\mu},$$

and take the full-rank diagonal iterate

$$X_\varepsilon = \operatorname{diag}(1, \varepsilon, \ldots, \varepsilon), \qquad \varepsilon > 0.$$

Then

$$\Sigma_\varepsilon = X_\varepsilon X_\varepsilon^\top = \operatorname{diag}(1, \varepsilon^2, \ldots, \varepsilon^2), \qquad G_\varepsilon = H X_\varepsilon = \operatorname{diag}(\mu, L\varepsilon, \ldots, L\varepsilon).$$

Since G_ε is diagonal with nonnegative entries, its nuclear norm is $\|G_\varepsilon\|_* = \mu + (d-1)L\varepsilon$. Moreover,

$$\operatorname{tr}(H) = \mu + (d-1)L, \qquad \operatorname{tr}(H\Sigma_\varepsilon) = \mu + (d-1)L\varepsilon^2.$$

Therefore the exact PolarGD progress coefficient is

$$p(X_\varepsilon) = \frac{\big(\mu + (d-1)L\varepsilon\big)^2}{\big(\mu + (d-1)L\big)\big(\mu + (d-1)L\varepsilon^2\big)}.$$

Taking $\varepsilon \downarrow 0$ gives

$$\lim_{\varepsilon \downarrow 0} p(X_\varepsilon) = \frac{\mu}{\mu + (d-1)L} = \frac{1}{1 + (d-1)\kappa} \approx \frac{1}{\kappa d}.$$

Thus the lower bound $\lambda_{\min}(H)/\operatorname{tr}(H)$ is attained in the near-rank-deficient limit.

Comparison with GD. On the same family, GD with exact line search is nearly one-step exact. Its progress coefficient is

$$p_{\mathrm{GD}}(X) := \frac{\Delta_{\mathrm{GD}}(X)}{f(X)} = \frac{\|G\|_F^4}{\operatorname{tr}(G^\top H G)\operatorname{tr}(H\Sigma)}.$$

For $X = X_\varepsilon$, we have

$$\|G_\varepsilon\|_F^2 = \mu^2 + (d-1)L^2\varepsilon^2, \qquad \operatorname{tr}(G_\varepsilon^\top H G_\varepsilon) = \mu^3 + (d-1)L^3\varepsilon^2.$$

Hence

$$p_{\mathrm{GD}}(X_\varepsilon) = \frac{\big(\mu^2 + (d-1)L^2\varepsilon^2\big)^2}{\big(\mu^3 + (d-1)L^3\varepsilon^2\big)\big(\mu + (d-1)L\varepsilon^2\big)} \longrightarrow 1 \qquad \text{as } \varepsilon \downarrow 0.$$

Thus this family separates the two methods: GD is nearly one-step exact, while exact PolarGD decreases only a $1/(1 + (d-1)\kappa)$-fraction of the loss per step.

Interpretation. For every $\varepsilon > 0$, the iterate is full rank, so full-row-rank PolarGD pays curvature $\mathrm{tr}(H) = \mu + (d-1)L$. However, as $\varepsilon \downarrow 0$, almost all gradient gain is concentrated in the first, low-curvature direction. Exact polar normalization still maps the tiny singular modes $L\varepsilon$ to unit-size directions, so the line-search denominator pays for all directions even though most of the gain comes from only one.

At the limiting point $\varepsilon = 0$, the gradient is rank one. Then PolarGD and GD are collinear, and the active-subspace formula from Appendix E.6 uses the active curvature $\mathrm{tr}(HP) = \mu$ instead of $\mathrm{tr}(H)$. In that rank-one limit, PolarGD again reaches the minimizer in one step. Hence the pathology is a near-rank-deficient full-rank effect: exact polar is discontinuous at rank drops because it hard-normalizes every nonzero singular mode. This is the failure mode targeted by rank truncation, damping, and finite Newton–Schulz filters.

E.3. Generic multiplicative recursion under exact line search

Consider a quadratic objective $f(X) = \frac{1}{2}\mathrm{tr}(X^\top HX)$ with $H \succeq 0$. Let a generic method produce updates

$$X_{k+1} = X_k - \gamma_k D_k,$$

where γ_k is chosen by exact line search along D_k. For any $D_k \neq 0$ with $\mathrm{tr}(D_k^\top HD_k) > 0$, exact line search yields

$$\gamma_k = \frac{\langle G_k, D_k\rangle}{\mathrm{tr}(D_k^\top HD_k)}, \qquad G_k = \nabla f(X_k) = HX_k,$$

and the one-step decrease equals

$$\Delta_k := f(X_k) - f(X_{k+1}) = \frac{\langle G_k, D_k\rangle^2}{2\,\mathrm{tr}(D_k^\top HD_k)}.$$

Whenever $f(X_k) > 0$, define the progress coefficient

$$p_k := \frac{\Delta_k}{f(X_k)} \in [0,1],$$

so that

$$f(X_{k+1}) = (1-p_k)\, f(X_k), \qquad f(X_T) = f(X_0)\prod_{k=0}^{T-1}(1-p_k) \le f(X_0)\exp\left(-\sum_{k=0}^{T-1} p_k\right). \quad (16)$$

Thus any *trajectory-wise* lower bound on p_k implies a global linear (geometric) rate.

E.4. GD with exact line search: classical rate

For GD, $D_k = G_k$, hence

$$\Delta_k^{\mathrm{GD}} = \frac{\|G_k\|_{\mathrm{F}}^4}{2\,\mathrm{tr}(G_k^\top HG_k)}.$$

For $H \succ 0$, it is classical that exact line search yields the sharp contraction

$$f(X_{k+1}) \leq \left(\frac{\kappa(H)-1}{\kappa(H)+1}\right)^2 f(X_k), \qquad \kappa(H) = \frac{\lambda_{\max}(H)}{\lambda_{\min}(H)}. \tag{17}$$

Equivalently, the progress coefficient satisfies $p_k^{\mathrm{GD}} \geq 1 - \left(\frac{\kappa(H)-1}{\kappa(H)+1}\right)^2 = \frac{4\kappa(H)}{(\kappa(H)+1)^2}$.

E.5. PolarGD with exact line search: trajectory dependence through Σ_k

In the full-row-rank regime ($d \leq m$ and $\mathrm{rank}(X_k) = d$), Proposition 4 gives an explicit $p(X_k)$:

$$p(X_k) = \frac{\|HX_k\|_*^2}{\mathrm{tr}(H)\,\mathrm{tr}(HX_kX_k^\top)}.$$

Combining this with (16) shows that global behavior depends on how $p(X_k)$ evolves along the trajectory. In particular, the bound $p(X_k) \geq 1/\kappa(X_kX_k^\top)$ implies that if $\kappa(X_kX_k^\top) \leq K$ uniformly along the path, then

$$f(X_k) \leq \left(1 - \frac{1}{K}\right)^k f(X_0).$$

This makes precise the intuition from Section 4: PolarGD can be fast if the iterate geometry stays close to row-isotropic, but can slow down if $\Sigma_k = X_kX_k^\top$ becomes highly anisotropic.

E.6. Rank-deficient and active-subspace extensions

This appendix records the versions of the PolarGD identities that remain valid when the gradient is not full row rank. The main text focuses on the full-row-rank regime because it yields the cleanest formulas, in particular $QQ^\top = I_d$ and $\mathrm{tr}(Q^\top HQ) = \mathrm{tr}(H)$. Outside this regime, the same mechanisms persist after replacing I_d by the active projector

$$P := QQ^\top = \Pi_{\mathrm{range}(G)}.$$

Throughout this section, let

$$f(X) = \frac{1}{2}\mathrm{tr}(X^\top HX), \qquad H = H^\top \succ 0, \qquad G = HX, \qquad \Sigma = XX^\top.$$

Let $G = USV^\top$ be the thin SVD of G, where

$$U \in \mathbb{R}^{d\times r}, \qquad V \in \mathbb{R}^{m\times r}, \qquad S = \mathrm{diag}(\sigma_1, \dots, \sigma_r), \qquad r = \mathrm{rank}(G) = \mathrm{rank}(X),$$

with $\sigma_i > 0$. The thin polar factor is

$$Q := \mathrm{polar}(G) := UV^\top.$$

Then

$$P := QQ^\top = UU^\top = \Pi_{\mathrm{range}(G)}.$$

Lemma 14 (Active projector identities) *Let $G = USV^\top$ and $Q = \mathrm{polar}(G) = UV^\top$ as above. Then*

$$Q = (GG^\top)^{\dagger/2}G, \qquad QQ^\top = P = \Pi_{\mathrm{range}(G)}, \qquad Q^\top Q = VV^\top = \Pi_{\mathrm{range}(G^\top)}.$$

In particular,

$$\mathrm{tr}(Q^\top HQ) = \mathrm{tr}(HQQ^\top) = \mathrm{tr}(HP).$$

Proof Since $GG^\top = US^2U^\top$, we have

$$(GG^\top)^{\dagger/2} = US^{-1}U^\top.$$

Thus

$$(GG^\top)^{\dagger/2}G = US^{-1}U^\top USV^\top = UV^\top = Q.$$

Furthermore,

$$QQ^\top = UV^\top VU^\top = UU^\top, \qquad Q^\top Q = VU^\top UV^\top = VV^\top.$$

The trace identity follows by cyclicity:

$$\mathrm{tr}(Q^\top HQ) = \mathrm{tr}(HQQ^\top) = \mathrm{tr}(HP).$$

■

E.6.1. ACTIVE-SUBSPACE PRECONDITIONING

Recall the implicit preconditioner

$$A(X) := (H\Sigma H)^{\dagger/2}H.$$

Then

$$Q = A(X)X$$

even when Σ is singular. The full-rank identity $A\Sigma A^\top = I_d$ becomes an active-subspace whitening identity.

Proposition 15 (Active-subspace whitening) *Let $H \succ 0$ and let $A(X) = (H\Sigma H)^{\dagger/2}H$. Then*

$$Q = A(X)X, \qquad A(X)\Sigma A(X)^\top = P.$$

Moreover,

$$A(X)^\top A(X) = H(H\Sigma H)^\dagger H.$$

Consequently, in the rank-deficient case $A(X)^\top A(X)$ is generally not equal to $\Sigma^\dagger$.

Proof Since $G = HX$, we have

$$H\Sigma H = HXX^\top H = GG^\top.$$

Therefore

$$A(X)X = (H\Sigma H)^{\dagger/2}HX = (GG^\top)^{\dagger/2}G = Q.$$

Also,

$$A\Sigma A^\top = (H\Sigma H)^{\dagger/2}H\Sigma H(H\Sigma H)^{\dagger/2}.$$

Because $H\Sigma H = GG^\top$ has range $\mathrm{range}(G)$, the last expression is the orthogonal projector onto $\mathrm{range}(G)$, namely P. Finally,

$$A^\top A = H(H\Sigma H)^{\dagger/2}(H\Sigma H)^{\dagger/2}H = H(H\Sigma H)^\dagger H.$$

■

Remark 16 (Why the full-rank conditioning identity is special) *In the full-row-rank case, $H\Sigma H \succ 0$ and*

$$(H\Sigma H)^{-1} = H^{-1}\Sigma^{-1}H^{-1}.$$

Thus

$$A^\top A = H(H\Sigma H)^{-1}H = \Sigma^{-1},$$

which gives the main-text identity $\kappa(A) = \kappa(X)$. If Σ is singular, this inverse factorization no longer holds with Moore–Penrose pseudoinverses in general, and the conditioning of A depends not only on the positive spectrum of Σ, but also on the alignment of the active subspace with H.

Proposition 17 (A general conditioning bound for the active preconditioner) *Let $\kappa_+(\cdot)$ denote the condition number restricted to the positive singular spectrum. Then*

$$\sigma_{\max}(A) \le \frac{\lambda_{\max}(H)}{\sigma^+_{\min}(G)}, \qquad \sigma^+_{\min}(A) \ge \frac{\lambda_{\min}(H)}{\sigma_{\max}(G)}.$$

Consequently,

$$\kappa_+(A) \le \kappa(H)\,\kappa_+(G) \le \kappa(H)^2\,\kappa_+(X).$$

Proof Let $B := GG^\top = H\Sigma H$. Then

$$A = B^{\dagger/2}H.$$

The positive singular values of A are the square roots of the positive eigenvalues of

$$AA^\top = B^{\dagger/2}H^2B^{\dagger/2}$$

on $\mathrm{range}(B)$. Since

$$\lambda_{\min}(H)^2 I_d \preceq H^2 \preceq \lambda_{\max}(H)^2 I_d,$$

we get, on $\mathrm{range}(B)$,

$$\lambda_{\min}(H)^2 B^\dagger \preceq B^{\dagger/2}H^2B^{\dagger/2} \preceq \lambda_{\max}(H)^2 B^\dagger.$$

The positive eigenvalues of $B^\dagger$ are

$$\frac{1}{\sigma_i(G)^2}, \qquad i = 1, \dots, r.$$

Therefore

$$\sigma_{\max}(A) \le \frac{\lambda_{\max}(H)}{\sigma^+_{\min}(G)}, \qquad \sigma^+_{\min}(A) \ge \frac{\lambda_{\min}(H)}{\sigma_{\max}(G)}.$$

This gives

$$\kappa_+(A) \le \kappa(H)\kappa_+(G).$$

Finally, since $G = HX$ and $H \succ 0$,

$$\sigma_{\max}(G) \le \lambda_{\max}(H)\sigma_{\max}(X), \qquad \sigma^+_{\min}(G) \ge \lambda_{\min}(H)\sigma^+_{\min}(X),$$

so

$$\kappa_+(G) \le \kappa(H)\kappa_+(X).$$

■

Proposition 18 (Invariant active subspaces recover the full-rank identity) *Let $\mathcal{S} := \mathrm{range}(\Sigma) = \mathrm{range}(X)$ and suppose that $\mathcal{S}$ is H-invariant:*

$$H\mathcal{S} \subseteq \mathcal{S}.$$

Then $\mathrm{range}(G) = \mathcal{S}$, $P = \Pi_{\mathcal{S}}$, and

$$A\Sigma A^\top = P, \qquad A^\top A = \Sigma^\dagger.$$

Consequently,

$$\kappa_+(A) = \kappa_+(X).$$

Proof Choose an orthogonal basis adapted to the decomposition $\mathbb{R}^d = \mathcal{S} \oplus \mathcal{S}^\perp$. Since H is symmetric and $\mathcal{S}$ is H-invariant, the matrix H is block diagonal in this decomposition. Likewise,

$$\Sigma = \begin{pmatrix} \Sigma_{\mathcal{S}} & 0 \\ 0 & 0 \end{pmatrix}, \qquad \Sigma_{\mathcal{S}} \succ 0.$$

Writing

$$H = \begin{pmatrix} H_{\mathcal{S}} & 0 \\ 0 & H_{\mathcal{S}^\perp} \end{pmatrix},$$

we have

$$H\Sigma H = \begin{pmatrix} H_{\mathcal{S}}\Sigma_{\mathcal{S}}H_{\mathcal{S}} & 0 \\ 0 & 0 \end{pmatrix}.$$

Therefore

$$(H\Sigma H)^\dagger = \begin{pmatrix} (H_{\mathcal{S}}\Sigma_{\mathcal{S}}H_{\mathcal{S}})^{-1} & 0 \\ 0 & 0 \end{pmatrix}.$$

Hence

$$A^\top A = H(H\Sigma H)^\dagger H = \begin{pmatrix} H_{\mathcal{S}}(H_{\mathcal{S}}\Sigma_{\mathcal{S}}H_{\mathcal{S}})^{-1}H_{\mathcal{S}} & 0 \\ 0 & 0 \end{pmatrix} = \begin{pmatrix} \Sigma_{\mathcal{S}}^{-1} & 0 \\ 0 & 0 \end{pmatrix} = \Sigma^\dagger.$$

The identity $A\Sigma A^\top = P$ follows from Proposition 15. The singular-value conclusion follows from $A^\top A = \Sigma^\dagger$. ■

E.6.2. EXACT LINE SEARCH AND PROGRESS IN THE RANK-DEFICIENT CASE

Proposition 19 (Exact active-subspace line-search formula) *Assume $H \succ 0$ and $G = HX \neq 0$. Let $Q = \mathrm{polar}(G)$ and*

$$P := QQ^\top = \Pi_{\mathrm{range}(G)}.$$

Then for every $\gamma \geq 0$,

$$f(X - \gamma Q) = f(X) - \gamma\|G\|_* + \frac{\gamma^2}{2}\mathrm{tr}(HP).$$

Consequently,

$$\gamma^\star(Q) = \frac{\|G\|_*}{\mathrm{tr}(HP)}$$

and

$$f(X - \gamma^\star(Q)Q) = f(X) - \frac{\|G\|_*^2}{2\mathrm{tr}(HP)}.$$

Equivalently, whenever $f(X) > 0$,

$$\frac{f(X - \gamma^\star(Q)Q)}{f(X)} = 1 - \widetilde{p}(X), \qquad \widetilde{p}(X) := \frac{\|G\|_*^2}{\mathrm{tr}(HP)\mathrm{tr}(H\Sigma)}.$$

Proof Using the quadratic expansion along Q,

$$f(X - \gamma Q) = f(X) - \gamma\langle G, Q\rangle + \frac{\gamma^2}{2}\mathrm{tr}(Q^\top HQ).$$

Since $Q = \mathrm{polar}(G)$,

$$\langle G, Q\rangle = \|G\|_*.$$

Moreover, by Lemma 14,

$$\mathrm{tr}(Q^\top HQ) = \mathrm{tr}(HP).$$

Thus

$$f(X - \gamma Q) = f(X) - \gamma\|G\|_* + \frac{\gamma^2}{2}\mathrm{tr}(HP).$$

Minimizing this scalar quadratic over $\gamma \geq 0$ gives

$$\gamma^\star(Q) = \frac{\|G\|_*}{\mathrm{tr}(HP)}.$$

The decrease formula follows by substitution. Finally, since

$$f(X) = \frac{1}{2}\mathrm{tr}(H\Sigma),$$

the multiplicative form follows immediately. ■

Corollary 20 (Rank-dependent pessimistic bound) *Let $r = \mathrm{rank}(G)$ and $P = \Pi_{\mathrm{range}(G)}$. Then*

$$\widetilde{p}(X) \geq \frac{\lambda_{\min}(H)}{\mathrm{tr}(HP)} \geq \frac{1}{r\,\kappa(H)}.$$

Proof Since $\|G\|_*^2 \geq \|G\|_F^2$, we have

$$\|G\|_*^2 \geq \|G\|_F^2 = \mathrm{tr}(G^\top G) = \mathrm{tr}(X^\top H^2 X) = \mathrm{tr}(H^2\Sigma).$$

Because $H^2 \succeq \lambda_{\min}(H)H$,

$$\mathrm{tr}(H^2\Sigma) \geq \lambda_{\min}(H)\mathrm{tr}(H\Sigma).$$

Therefore

$$\widetilde{p}(X) = \frac{\|G\|_*^2}{\mathrm{tr}(HP)\mathrm{tr}(H\Sigma)} \geq \frac{\lambda_{\min}(H)}{\mathrm{tr}(HP)}.$$

Since P is a rank-r orthogonal projector,

$$\mathrm{tr}(HP) \leq r\,\lambda_{\max}(H).$$

Thus

$$\widetilde{p}(X) \geq \frac{\lambda_{\min}(H)}{r\,\lambda_{\max}(H)} = \frac{1}{r\,\kappa(H)}.$$

■

Remark 21 (Interpretation of the rank-dependent bound) *The full-row-rank bound in the main text has a dimension factor d through $\lambda_{\min}(H)/\mathrm{tr}(H) \gtrsim 1/(d\kappa(H))$. In the rank-$r$ active-subspace case, this becomes $1/(r\kappa(H))$. Thus rank deficiency is not automatically pessimistic for one-step progress: the curvature denominator is reduced from $\mathrm{tr}(H)$ to the active curvature $\mathrm{tr}(HP)$. The competing effect is that low rank also reduces the spectral distinction between PolarGD and GD; at rank one the two direction rays coincide.*

E.6.3. RANK MONOTONICITY AND RANK DROPS

Proposition 22 (Rank is non-increasing under exact PolarGD directions) *Assume $H \succ 0$ and let*

$$X^+ = X - \gamma Q$$

for any $\gamma \geq 0$, where $Q = \mathrm{polar}(G)$ and $G = HX$. Then

$$\mathrm{rank}(X^+) = \mathrm{rank}(HX^+) \leq \mathrm{rank}(G) = \mathrm{rank}(X).$$

In particular, exact PolarGD cannot create new active gradient directions.

Proof Let $R := (GG^\top)^{1/2}$. Since $G = RQ$, we have

$$HX^+ = H(X - \gamma Q) = G - \gamma HQ = (R - \gamma H)Q.$$

Therefore

$$\mathrm{rank}(HX^+) = \mathrm{rank}((R - \gamma H)Q) \leq \mathrm{rank}(Q) = \mathrm{rank}(G).$$

Since H is invertible,

$$\mathrm{rank}(X^+) = \mathrm{rank}(HX^+).$$

Also $\mathrm{rank}(G) = \mathrm{rank}(HX) = \mathrm{rank}(X)$. ■

Proposition 23 (Characterizing rank drops) *Let $G = USV^\top$ be the thin SVD and $Q = UV^\top$. For the update*

$$X^+ = X - \gamma Q,$$

we have

$$\mathrm{rank}(X^+) = \mathrm{rank}\big((R - \gamma H)U\big), \qquad R = (GG^\top)^{1/2}.$$

Equivalently, rank drops if and only if there exists a nonzero vector $u \in \mathrm{range}(G)$ such that

$$(R - \gamma H)u = 0.$$

In the full-row-rank case, this reduces to

$$\mathrm{rank}(X^+) < d \quad \Longleftrightarrow \quad \det(R - \gamma H) = 0.$$

Equivalently, with

$$S_H := H^{-1/2} R H^{-1/2},$$

rank drops if and only if

$$\gamma \in \mathrm{spec}(S_H).$$

Proof Since $Q = UV^\top$ and $V^\top$ has rank r,

$$HX^+ = (R - \gamma H)UV^\top$$

has the same rank as $(R - \gamma H)U$. Therefore

$$\mathrm{rank}(X^+) = \mathrm{rank}(HX^+) = \mathrm{rank}((R - \gamma H)U).$$

The matrix $(R - \gamma H)U$ loses column rank if and only if there exists a nonzero $a \in \mathbb{R}^r$ such that

$$(R - \gamma H)Ua = 0.$$

Writing $u = Ua \in \mathrm{range}(G)$ gives the stated condition.

If $r = d$, then U is square orthogonal, so rank drops if and only if $R - \gamma H$ is singular. Since

$$R - \gamma H = H^{1/2}\big(H^{-1/2}RH^{-1/2} - \gamma I_d\big)H^{1/2},$$

this is equivalent to

$$\gamma \in \mathrm{spec}(H^{-1/2}RH^{-1/2}).$$

■

E.6.4. A RANK-r ONE-STEP EXACTNESS REGIME

Proposition 24 (One-step exactness on an invariant active subspace) *Let $H \succ 0$. Suppose*

$$X = \sqrt{c}\,UV^\top, \qquad c > 0,$$

where $U \in \mathbb{R}^{d\times r}$ and $V \in \mathbb{R}^{m\times r}$ have orthonormal columns. Assume the active subspace $\mathrm{range}(U)$ is H-invariant:

$$H\mathrm{range}(U) \subseteq \mathrm{range}(U).$$

Then one exact-line-search PolarGD step reaches the minimizer:

$$X^+ = 0.$$

Conversely, if a nonzero rank-r point X is one-step exact under PolarGD, then

$$X = \sqrt{c}\,UV^\top$$

for some partial isometries U, V and $c > 0$, and the active left subspace $\mathrm{range}(U)$ is H-invariant.

Proof First assume $X = \sqrt{c}\,UV^\top$ and $\mathrm{range}(U)$ is H-invariant. Then there exists an $r\times r$ symmetric positive definite matrix B such that

$$HU = UB.$$

Therefore

$$G = HX = \sqrt{c}\,HUV^\top = \sqrt{c}\,UBV^\top.$$

Since $B \succ 0$, the polar factor of $UBV^\top$ is $UV^\top$. Hence

$$Q = \mathrm{polar}(G) = UV^\top.$$

The exact line-search step size is

$$\gamma^\star(Q) = \frac{\langle G, Q\rangle}{\mathrm{tr}(Q^\top HQ)} = \frac{\sqrt{c}\,\mathrm{tr}(B)}{\mathrm{tr}(B)} = \sqrt{c}.$$

Thus

$$X^+ = X - \gamma^\star Q = \sqrt{c}\,UV^\top - \sqrt{c}\,UV^\top = 0.$$

Conversely, suppose one exact-line-search PolarGD step reaches zero:

$$X - \gamma^\star Q = 0.$$

Then $X = \gamma^\star Q$. Since $Q = UV^\top$ is a partial isometry, this implies

$$X = \sqrt{c}\,UV^\top$$

with $\sqrt{c} = \gamma^\star$. Moreover,

$$Q = \mathrm{polar}(G) = \mathrm{polar}(HX) = \mathrm{polar}(HUV^\top).$$

Since $Q = UV^\top$, this requires $\mathrm{polar}(HU) = U$. This is possible only if $\mathrm{range}(HU) = \mathrm{range}(U)$, i.e. $\mathrm{range}(U)$ is H-invariant. This proves the converse. ■

Remark 25 (Relation to the full-row-rank theorem) *When $r = d$, the active subspace is all of $\mathbb{R}^d$, so the invariance condition is automatic. Proposition 24 then reduces to the full-row-rank statement: PolarGD is one-step exact exactly when*

$$XX^\top = cI_d.$$

Thus the correct general slogan is isotropy on an invariant active subspace.

E.7. Active-subspace root dynamics

The square-rooted gradient covariance view also extends outside full row rank, but the identity contains the active projector P.

Proposition 26 (Root dynamics with an active projector) *Let*

$$R := (GG^\top)^{1/2}, \qquad P := QQ^\top = \Pi_{\mathrm{range}(G)}.$$

For $X(\gamma) = X - \gamma Q$ and $\Sigma(\gamma) = X(\gamma)X(\gamma)^\top$,

$$H\Sigma(\gamma)H = (R - \gamma H)P(R - \gamma H).$$

Consequently,

$$(H\Sigma(\gamma)H)^{1/2} = \Big((R-\gamma H)P(R-\gamma H)\Big)^{1/2}.$$

In the full-row-rank case, $P = I_d$, and this reduces to

$$H\Sigma(\gamma)H = (R-\gamma H)^2, \qquad (H\Sigma(\gamma)H)^{1/2} = |R-\gamma H|.$$

Proof Since $G = RQ$, we have

$$HX(\gamma) = G - \gamma HQ = (R-\gamma H)Q.$$

Thus

$$H\Sigma(\gamma)H = HX(\gamma)X(\gamma)^\top H = (R-\gamma H)QQ^\top(R-\gamma H) = (R-\gamma H)P(R-\gamma H).$$

The square-root identity follows by taking the PSD square root. ■

E.8. Local comparison with active curvature

The exact local comparison between PolarGD and GD remains valid in the rank-deficient case after writing the PolarGD curvature as $\mathrm{tr}(HP)$:

$$\Delta_{\mathrm{Pol}} > \Delta_{\mathrm{GD}} \quad \Longleftrightarrow \quad \frac{\mathrm{tr}(G^\top HG)}{\mathrm{tr}(HP)} > \left(\frac{\|G\|_F^2}{\|G\|_*}\right)^2.$$

Rank deficiency affects this comparison in two competing ways. The denominator $\mathrm{tr}(HP)$ can be smaller than $\mathrm{tr}(H)$, which helps PolarGD by reducing active curvature. On the other hand, lower rank decreases the spectral separation between $\|G\|_*$ and $\|G\|_F$. At rank one, Q is collinear with G, and PolarGD coincides with GD under exact line search.

Appendix F. Rank-truncated and damped polar directions

This appendix formalizes simple modifications of the polar direction that address the "tiny singular values" sensitivity suggested by the whitening representation $Q = (GG^\top)^{\dagger 1/2}G$.

F.1. Rank-truncated polar direction

Let $G \in \mathbb{R}^{d\times m}$ with thin SVD $G = USV^\top$, $S = \mathrm{diag}(\sigma_1, \dots, \sigma_r)$, $\sigma_1 \geq \cdots \geq \sigma_r > 0$. For an integer $1 \leq s \leq r$, define the rank-s truncation

$$Q^{(s)} := U_{[:,1:s]}V_{[:,1:s]}^\top. \tag{18}$$

Then $\mathrm{rank}(Q^{(s)}) = s$, $\|Q^{(s)}\|_2 = 1$, and

$$Q^{(s)}(Q^{(s)})^\top = U_{[:,1:s]}U_{[:,1:s]}^\top \tag{19}$$

is the projector onto the span of the top-s left singular vectors.

LMO interpretation. $Q^{(s)}$ is the maximizer of $\langle G, Q\rangle$ over spectral-norm-bounded matrices with rank at most s:

$$Q^{(s)} \in \arg\max\Big\{\langle G, Q\rangle : \|Q\|_2 \leq 1,\ \mathrm{rank}(Q) \leq s\Big\}. \tag{20}$$

Moreover, the attained gain is the Ky Fan s-nuclear quantity

$$\langle G, Q^{(s)}\rangle = \sum_{i=1}^{s}\sigma_i =: \|G\|_{*,s}. \tag{21}$$

Effect on exact line search for quadratics. For the quadratic $f(X) = \frac{1}{2}\mathrm{tr}(X^\top HX)$ with gradient $G = HX$, the exact line-search decrease along $Q^{(s)}$ is

$$\Delta_{\mathrm{Pol}}^{(s)}(X) = \frac{\langle G, Q^{(s)}\rangle^2}{2\,\mathrm{tr}\big((Q^{(s)})^\top HQ^{(s)}\big)} = \frac{\|G\|_{*,s}^2}{2\,\mathrm{tr}\big((Q^{(s)})^\top HQ^{(s)}\big)}. \tag{22}$$

In the isotropic case $H = \alpha I_d$, we have $\mathrm{tr}((Q^{(s)})^\top HQ^{(s)}) = \alpha\|Q^{(s)}\|_F^2 = \alpha s$, hence

$$f(X^+) = f(X) - \frac{\alpha\,\|X\|_{*,s}^2}{2s}, \qquad \frac{f(X^+)}{f(X)} = 1 - \frac{\|X\|_{*,s}^2}{s\,\|X\|_F^2}. \tag{23}$$

When X has rapidly decaying singular values, choosing $s \ll d$ can significantly improve the fractional decrease by ignoring tiny singular modes that otherwise degrade the isotropic coefficient $p(X) = \|X\|_*^2/(d\|X\|_F^2)$.

F.2. Damped polar direction

A soft alternative is Tikhonov damping: for $\varepsilon > 0$, define

$$Q_\varepsilon := (GG^\mathsf{T} + \varepsilon I_d)^{-1/2} G. \tag{24}$$

If $G = USV^\mathsf{T}$, then

$$Q_\varepsilon = U \operatorname{diag}\left(\frac{\sigma_i}{\sqrt{\sigma_i^2 + \varepsilon}}\right) V^\mathsf{T}, \tag{25}$$

so each singular mode is shrunk by a factor in $(0, 1)$. In particular, the whitening operator is uniformly bounded: $\|(GG^\mathsf{T} + \varepsilon I)^{-1/2}\|_2 = 1/\sqrt{\varepsilon}$.

Discussion. Both truncation (18) and damping (24) preserve the "spectral geometry" idea of polar normalization while reducing sensitivity to near-rank-deficiency (tiny singular values). These modifications do not make PolarGD uniformly better than GD in the worst case, but they directly target a specific failure mode where near-zero singular directions destabilize the polar factor and reduce effective progress. In practice, one may choose s (or ε) as a function of the observed singular spectrum of G (e.g. keep modes above a relative threshold), yielding a spectrum-adaptive polar direction.

Appendix G. Finite Newton–Schulz as a bounded spectral filter

This appendix explains how finite Newton–Schulz orthogonalization relates to the exact polar, damped polar, and rank-truncated polar directions. The key point is that finite Newton–Schulz is not identical to Tikhonov damping or rank truncation, but it is a *bounded polynomial spectral filter*. In particular, unlike the exact polar map, it does not apply an unbounded multiplier $1/\sigma_i$ to tiny nonzero singular values.

G.1. A general polynomial Newton–Schulz filter

Consider an iteration of the form

$$Z_{t+1} = a_t Z_t + b_t Z_t Z_t^\top Z_t + c_t (Z_t Z_t^\top)^2 Z_t, \qquad t = 0, \dots, q-1. \tag{26}$$

This includes the quintic Newton–Schulz form commonly used in Muon implementations. Suppose

$$Z_0 = U \operatorname{diag}(s_1, \dots, s_r) V^\top$$

is a thin SVD, with $s_i \geq 0$. Then Z_t has the same singular vectors for every t, and only the singular values evolve.

Proposition 27 (Finite Newton–Schulz preserves singular vectors) *Let $Z_0 = U \operatorname{diag}(s_i) V^\top$ and define Z_t by (26). Then*

$$Z_t = U \operatorname{diag}(\phi_t(s_i)) V^\top,$$

where the scalar functions ϕ_t are defined recursively by

$$\phi_0(s) = s, \qquad \phi_{t+1}(s) = a_t\phi_t(s) + b_t\phi_t(s)^3 + c_t\phi_t(s)^5.$$

In particular, after q steps,

$$Z_q = U\mathrm{diag}(\phi_q(s_i))V^\top.$$

Proof Assume

$$Z_t = U\mathrm{diag}(\phi_t(s_i))V^\top.$$

Then

$$Z_tZ_t^\top = U\mathrm{diag}(\phi_t(s_i)^2)U^\top.$$

Therefore

$$Z_tZ_t^\top Z_t = U\mathrm{diag}(\phi_t(s_i)^3)V^\top$$

and

$$(Z_tZ_t^\top)^2Z_t = U\mathrm{diag}(\phi_t(s_i)^5)V^\top.$$

Substituting into (26) gives

$$Z_{t+1} = U\mathrm{diag}\big(a_t\phi_t(s_i) + b_t\phi_t(s_i)^3 + c_t\phi_t(s_i)^5\big)V^\top,$$

which proves the recursion. ■

G.2. Comparison with exact polar, damped polar, and truncation

Let

$$G = U\mathrm{diag}(\sigma_1, \ldots, \sigma_r)V^\top.$$

The exact polar factor is

$$Q_{\mathrm{pol}} = UV^\top = U\mathrm{diag}(1, \ldots, 1)V^\top.$$

Thus exact polar applies the singular-value filter

$$\phi_{\mathrm{pol}}(\sigma) = 1 \qquad \text{for } \sigma > 0.$$

Equivalently,

$$Q_{\mathrm{pol}} = (GG^\top)^{\dagger/2}G,$$

so its whitening multiplier on a singular mode is

$$m_{\mathrm{pol}}(\sigma) = \frac{1}{\sigma}.$$

This multiplier is unbounded as $\sigma \downarrow 0$.

The damped polar direction

$$Q_\varepsilon = (GG^\top + \varepsilon I_d)^{-1/2}G$$

has the filter

$$\phi_\varepsilon(\sigma) = \frac{\sigma}{\sqrt{\sigma^2 + \varepsilon}},$$

and multiplier

$$m_\varepsilon(\sigma) = \frac{1}{\sqrt{\sigma^2 + \varepsilon}}.$$

Thus

$$\sup_{\sigma \geq 0} m_\varepsilon(\sigma) = \varepsilon^{-1/2}.$$

Near zero,

$$\phi_\varepsilon(\sigma) = \varepsilon^{-1/2}\sigma + O(\sigma^3).$$

A rank-truncated polar direction keeps only the top s modes:

$$Q^{(s)} = U_{:,1:s} V_{:,1:s}^\top,$$

corresponding to the hard filter

$$\phi_s(\sigma_i) = \begin{cases} 1, & i \leq s, \\ 0, & i > s. \end{cases}$$

Finite Newton–Schulz instead applies the polynomial filter

$$\phi_q(\sigma)$$

generated by the scalar recursion above. Thus finite Newton–Schulz is neither exactly damping nor exactly truncation: it is a soft polynomial filter.

G.3. Near-zero behavior and damping equivalence

The finite Newton–Schulz filter has a simple Taylor expansion near zero.

Proposition 28 (Finite Newton–Schulz is damping-like near zero) *Suppose ϕ_q is generated by the recursion in Proposition 27. Then*

$$\phi_q(\sigma) = c_q \sigma + O(\sigma^3) \qquad \text{as } \sigma \to 0,$$

where

$$c_q = \prod_{t=0}^{q-1} a_t.$$

In particular, near zero, finite Newton–Schulz matches the local behavior of a damped polar filter with

$$\varepsilon_q = c_q^{-2}$$

in normalized singular-value units.

Proof We prove the claim by induction. Clearly $\phi_0(\sigma) = \sigma$, so the linear coefficient is 1. Suppose

$$\phi_t(\sigma) = c_t^{\text{lin}}\sigma + O(\sigma^3).$$

Then

$$\phi_t(\sigma)^3 = O(\sigma^3), \qquad \phi_t(\sigma)^5 = O(\sigma^5),$$

and therefore

$$\phi_{t+1}(\sigma) = a_t\phi_t(\sigma) + b_t\phi_t(\sigma)^3 + c_t\phi_t(\sigma)^5 = a_t c_t^{\text{lin}}\sigma + O(\sigma^3).$$

Thus the linear coefficient after q steps is

$$c_q = \prod_{t=0}^{q-1} a_t.$$

The damped polar filter satisfies

$$\frac{\sigma}{\sqrt{\sigma^2+\varepsilon}} = \varepsilon^{-1/2}\sigma + O(\sigma^3).$$

Matching slopes gives $\varepsilon_q = c_q^{-2}$. ■

G.4. Bounded amplification

Practical Newton–Schulz orthogonalization is typically applied after scaling the input so that its singular values lie in a bounded interval, e.g. $s_i \in [0, 1]$. The next proposition formalizes the key robustness distinction relative to exact polar.

Proposition 29 (Finite Newton–Schulz has bounded singular-mode amplification) *Assume the input is scaled as*

$$Z_0 = \frac{G}{\alpha}, \qquad \alpha > 0,$$

so that the singular values $s_i = \sigma_i(G)/\alpha$ lie in $[0, 1]$. Let Z_q be the result of q Newton–Schulz steps, and define

$$L_q := \sup_{s\in[0,1]} \left|\frac{\phi_q(s)}{s}\right|,$$

where the value at $s = 0$ is defined by continuity. Then

$$Z_q = P_q(GG^\top)G$$

for a matrix polynomial P_q, and

$$\|P_q(GG^\top)\|_2 \le \frac{L_q}{\alpha} < \infty.$$

Thus finite Newton–Schulz applies a uniformly bounded multiplier to all singular modes. By contrast, exact polar applies the multiplier $1/\sigma_i(G)$ to the ith singular mode, which is unbounded as $\sigma_i(G) \downarrow 0$.

Proof Let

$$G = U\mathrm{diag}(\sigma_i)V^\top, \qquad s_i = \frac{\sigma_i}{\alpha}.$$

By Proposition 27,

$$Z_q = U\mathrm{diag}(\phi_q(s_i))V^\top.$$

For $\sigma_i > 0$,

$$\phi_q(s_i) = \frac{\phi_q(\sigma_i/\alpha)}{\sigma_i}\,\sigma_i.$$

Hence

$$Z_q = U\mathrm{diag}\left(\frac{\phi_q(\sigma_i/\alpha)}{\sigma_i}\right)U^\top G.$$

Define the scalar polynomial multiplier

$$p_q(t) := \frac{\phi_q(\sqrt{t}/\alpha)}{\sqrt{t}},$$

with the value at $t = 0$ defined by continuity. Then

$$Z_q = p_q(GG^\top)G.$$

Moreover,

$$\left|\frac{\phi_q(\sigma_i/\alpha)}{\sigma_i}\right| = \frac{1}{\alpha}\left|\frac{\phi_q(s_i)}{s_i}\right| \le \frac{L_q}{\alpha}.$$

Therefore

$$\|P_q(GG^\top)\|_2 \le \frac{L_q}{\alpha}.$$

For exact polar,

$$Q_{\mathrm{pol}} = (GG^\top)^{\dagger/2}G,$$

whose multiplier on a positive singular value $\sigma_i(G)$ is $1/\sigma_i(G)$. ■

Remark 30 (What finite Newton–Schulz can and cannot fix) *The bounded-amplification property explains why finite Newton–Schulz can be more robust than exact polar in near-rank-deficient regimes: tiny nonzero singular modes are not all immediately mapped to unit size. Instead, they are passed through a bounded polynomial filter and, near zero, behave like a damped polar update. However, finite Newton–Schulz does not create new singular directions: exact zero singular values remain zero. Thus it softens near-rank-deficiency, but does not undo exact low rank.*

Appendix H. Exact dynamics through the square-rooted gradient covariance

The identities in Section 4 simplify further after introducing the *square-rooted gradient covariance*

$$R := (H\Sigma H)^{1/2} = (GG^\mathsf{T})^{1/2}, \qquad \Sigma := XX^\mathsf{T}, \qquad G := HX.$$

The next proposition shows that, in the full-row-rank regime, the exact PolarGD trajectory is most transparent in terms of R rather than Σ.

Proposition 31 (Exact state recursion for PolarGD in the full-row-rank regime) *Assume $H \succ 0$ and $\operatorname{rank}(X) = d$ (hence $d \le m$ and $\operatorname{rank}(G) = d$). Let*

$$\Sigma := XX^\mathsf{T}, \qquad G := HX, \qquad Q := \operatorname{polar}(G), \qquad R := (H\Sigma H)^{1/2} = (GG^\mathsf{T})^{1/2}.$$

For any $\gamma \ge 0$, define

$$X(\gamma) := X - \gamma Q, \qquad \Sigma(\gamma) := X(\gamma)X(\gamma)^\mathsf{T}.$$

Then

$$\Sigma(\gamma) = \Sigma - \gamma\Big(\Sigma H(H\Sigma H)^{-1/2} + (H\Sigma H)^{-1/2}H\Sigma\Big) + \gamma^2 I_d, \tag{27}$$

$$H\Sigma(\gamma)H = (R - \gamma H)^2, \tag{28}$$

$$\begin{aligned} f(X(\gamma)) &= \frac{1}{2}\operatorname{tr}\Big(H^{-1}(R-\gamma H)^2\Big) \\ &= \frac{1}{2}\Big[\operatorname{tr}(H^{-1}R^2) - 2\gamma\operatorname{tr}(R) + \gamma^2\operatorname{tr}(H)\Big]. \end{aligned} \tag{29}$$

Consequently, the exact line-search step along Q is

$$\gamma^\star(Q) = \frac{\operatorname{tr}(R)}{\operatorname{tr}(H)} = \frac{\|G\|_*}{\operatorname{tr}(H)}. \tag{30}$$

Moreover, if $X^+ := X - \gamma^\star(Q)Q$, $\Sigma^+ := X^+X^{+\top}$, and

$$R^+ := (H\Sigma^+H)^{1/2},$$

then

$$R^+ = \big|R - \gamma^\star(Q)H\big|, \qquad \Sigma^+ = H^{-1}(R - \gamma^\star(Q)H)^2H^{-1}, \tag{31}$$

where $|M| := (M^2)^{1/2}$ for symmetric M.

Proof Since $\operatorname{rank}(X) = d$ and $H \succ 0$, we also have $\operatorname{rank}(G) = d$, so $QQ^\mathsf{T} = I_d$ and

$$Q = (GG^\mathsf{T})^{-1/2}G = (H\Sigma H)^{-1/2}HX.$$

Hence

$$\Sigma(\gamma) = (X - \gamma Q)(X - \gamma Q)^\mathsf{T} = \Sigma - \gamma(XQ^\mathsf{T} + QX^\mathsf{T}) + \gamma^2 QQ^\mathsf{T},$$

and substituting the expression for Q together with $QQ^\mathsf{T} = I_d$ gives (27).

Next, because

$$Q = (GG^\top)^{-1/2}G = R^{-1}G,$$

we have

$$G = RQ.$$

Therefore, with $G(\gamma) := HX(\gamma)$,

$$G(\gamma) = G - \gamma HQ = (R - \gamma H)Q.$$

Using $QQ^\top = I_d$, we obtain

$$H\Sigma(\gamma)H = G(\gamma)G(\gamma)^\top = (R - \gamma H)QQ^\top(R - \gamma H) = (R - \gamma H)^2,$$

which proves (28).

Finally,

$$f(X(\gamma)) = \frac{1}{2}\mathrm{tr}\big(X(\gamma)^\top HX(\gamma)\big) = \frac{1}{2}\mathrm{tr}\big(H^{-1}G(\gamma)G(\gamma)^\top\big) = \frac{1}{2}\mathrm{tr}\Big(H^{-1}(R - \gamma H)^2\Big).$$

Expanding the square and using cyclicity of the trace gives

$$\mathrm{tr}\Big(H^{-1}(R - \gamma H)^2\Big) = \mathrm{tr}(H^{-1}R^2) - 2\gamma\,\mathrm{tr}(R) + \gamma^2\mathrm{tr}(H).$$

Minimizing this quadratic in γ yields

$$\gamma^\star(Q) = \frac{\mathrm{tr}(R)}{\mathrm{tr}(H)}.$$

Since $\mathrm{tr}(R) = \|G\|_*$, this proves (30). The identities in (31) follow immediately from (28) with $\gamma = \gamma^\star(Q)$. ■

Corollary 32 (A scalar decomposition of the one-step decrease) *Under the assumptions of Proposition 31,*

$$2f(X) = \mathrm{tr}(H^{-1}R^2), \qquad 2f(X^+) = \mathrm{tr}(H^{-1}R^2) - \frac{\mathrm{tr}(R)^2}{\mathrm{tr}(H)}.$$

Equivalently,

$$\mathrm{tr}(H^{-1}R^2) = \min_{\gamma\geq 0}\mathrm{tr}\Big(H^{-1}(R - \gamma H)^2\Big) + \frac{\mathrm{tr}(R)^2}{\mathrm{tr}(H)}. \tag{32}$$

Hence the progress coefficient from Proposition 4 can be rewritten as

$$p(X) = \frac{\|G\|_*^2}{\mathrm{tr}(H)\mathrm{tr}(H\Sigma)} = \frac{\mathrm{tr}(R)^2}{\mathrm{tr}(H)\mathrm{tr}(H^{-1}R^2)} = 1 - \frac{\min_{\gamma\geq 0}\mathrm{tr}\big(H^{-1}(R - \gamma H)^2\big)}{\mathrm{tr}(H^{-1}R^2)}. \tag{33}$$

Proof Since $R^2 = H\Sigma H$, we have

$$\mathrm{tr}(H^{-1}R^2) = \mathrm{tr}(\Sigma H) = \mathrm{tr}(H\Sigma) = 2f(X).$$

The expression for $2f(X^+)$ follows by evaluating (29) at $\gamma = \gamma^\star(Q)$. Equation (32) is the same identity rewritten, and (33) follows from Proposition 4. ■

Corollary 33 (Commuting case: additive shrinkage of the singular values) *Assume $H \succ 0$, $\mathrm{rank}(X) = d$, and $H\Sigma = \Sigma H$. Let*

$$H = U\mathrm{diag}(\lambda_1, \dots, \lambda_d)U^\top, \qquad \Sigma = U\mathrm{diag}(s_1^2, \dots, s_d^2)U^\top,$$

with $\lambda_i > 0$ and $s_i > 0$. Then

$$R = (H\Sigma H)^{1/2} = U\mathrm{diag}(\lambda_1 s_1, \dots, \lambda_d s_d)U^\top, \qquad \gamma^\star(Q) = \bar{s}_H := \frac{\sum_{i=1}^d \lambda_i s_i}{\sum_{i=1}^d \lambda_i}. \tag{34}$$

Moreover,

$$\Sigma^+ = (\Sigma^{1/2} - \bar{s}_H I_d)^2 = U\mathrm{diag}\big((s_1 - \bar{s}_H)^2, \dots, (s_d - \bar{s}_H)^2\big)U^\top. \tag{35}$$

Equivalently, the singular values of X evolve according to

$$s_i^+ = |s_i - \bar{s}_H|, \qquad i = 1, \dots, d. \tag{36}$$

Finally,

$$f(X) = \frac{1}{2}\sum_{i=1}^d \lambda_i s_i^2, \qquad f(X^+) = \frac{1}{2}\sum_{i=1}^d \lambda_i (s_i - \bar{s}_H)^2, \tag{37}$$

and

$$p(X) = \frac{\big(\sum_{i=1}^d \lambda_i s_i\big)^2}{\big(\sum_{i=1}^d \lambda_i\big)\big(\sum_{i=1}^d \lambda_i s_i^2\big)}. \tag{38}$$

Proof Since $H\Sigma = \Sigma H$ and both matrices are symmetric, they are simultaneously diagonalizable, and therefore

$$R = (H\Sigma H)^{1/2} = H\Sigma^{1/2} = \Sigma^{1/2}H.$$

Define

$$W := \Sigma^{-1/2}X.$$

Then $WW^\top = I_d$ and $X = \Sigma^{1/2}W$. Using $G = HX = H\Sigma^{1/2}W = RW$, we get

$$Q = R^{-1}G = W = \Sigma^{-1/2}X.$$

Hence

$$X^+ = X - \gamma^\star(Q)Q = (\Sigma^{1/2} - \gamma^\star(Q)I_d)W,$$

which implies

$$\Sigma^+ = X^+X^{+\top} = (\Sigma^{1/2} - \gamma^\star(Q)I_d)^2.$$

This proves (35). The formula for $\gamma^\star(Q)$ follows from (30) and

$$\mathrm{tr}(R) = \mathrm{tr}(H\Sigma^{1/2}) = \sum_{i=1}^d \lambda_i s_i.$$

The expressions for $f(X)$, $f(X^+)$, and $p(X)$ follow by simultaneous diagonalization. ■

Remark 34 (This is not monotone whitening of Σ_k) *Even in the commuting case with $H = I_d$, PolarGD does* not *monotonically improve the conditioning of Σ_k. For example, if $d = 3$ and the singular values of X are*

$$(s_1, s_2, s_3) = (4, 2, 1),$$

then $\bar{s} = (4 + 2 + 1)/3 = 7/3$, so

$$(s_1^+, s_2^+, s_3^+) = \left(\frac{5}{3}, \frac{1}{3}, \frac{4}{3}\right).$$

Hence

$$\kappa(\Sigma) = \left(\frac{4}{1}\right)^2 = 16, \qquad \kappa(\Sigma^+) = \left(\frac{5/3}{1/3}\right)^2 = 25.$$

Thus the row-covariance can become more *ill-conditioned after a PolarGD step.*

Remark 35 (The progress coefficient $p(X_k)$ is not monotone) *The exact progress coefficient $p(X_k)$ from Proposition 4 need not be monotone along a PolarGD trajectory. In the commuting case with $H = I_3$,*

$$p(X) = \frac{(s_1 + s_2 + s_3)^2}{3(s_1^2 + s_2^2 + s_3^2)}.$$

If $(s_1, s_2, s_3) = (3, 3, 2)$, then

$$p(X) = \frac{32}{33}, \qquad (s_1^+, s_2^+, s_3^+) = \left(\frac{1}{3}, \frac{1}{3}, \frac{2}{3}\right), \qquad p(X^+) = \frac{8}{9} < \frac{32}{33}.$$

On the other hand, if $(s_1, s_2, s_3) = (4, 1, 1)$, then

$$p(X) = \frac{2}{3}, \qquad (s_1^+, s_2^+, s_3^+) = (2, 1, 1), \qquad p(X^+) = \frac{8}{9} > \frac{2}{3}.$$

So $p(X_k)$ may either decrease or increase from one step to the next.

Proposition 36 (One-step collapse to gradient isotropy in the 2×2 full-rank case) *Assume $d = 2$, $H \succ 0$, and $\operatorname{rank}(X) = 2$. Let $X^+ := X - \gamma^\star(Q)Q$ be one PolarGD step with exact line search, and define*

$$\Sigma := XX^\mathsf{T}, \qquad \Sigma^+ := X^+X^{+\top}, \qquad R := (H\Sigma H)^{1/2}, \qquad R^+ := (H\Sigma^+ H)^{1/2}.$$

Then there exists $c \geq 0$ such that

$$R^+ = cI_2. \tag{39}$$

Equivalently,

$$G^+G^{+\top} = c^2 I_2, \qquad \Sigma^+ = c^2 H^{-2}. \tag{40}$$

In particular, after one exact-line-search PolarGD step, the new iterate automatically satisfies

$$H\Sigma^+ = \Sigma^+ H.$$

Proof By Proposition 31,

$$R^+ = \left|R - \gamma^\star(Q)H\right|, \qquad \gamma^\star(Q) = \frac{\mathrm{tr}(R)}{\mathrm{tr}(H)}.$$

Hence the symmetric matrix

$$M := R - \gamma^\star(Q)H$$

has zero trace:

$$\mathrm{tr}(M) = \mathrm{tr}(R) - \gamma^\star(Q)\mathrm{tr}(H) = 0.$$

Any symmetric 2×2 trace-zero matrix has eigenvalues $\pm c$ for some $c \geq 0$, and therefore $|M| = cI_2$. Thus (39) holds. Since

$$G^+G^{+\top} = (R^+)^2, \qquad H\Sigma^+H = (R^+)^2,$$

we obtain (40). ■

Corollary 37 (Exact post-collapse dynamics in 2×2) *Under the assumptions of Proposition 36, let X_1 denote the iterate after the first PolarGD step, and let $0 < \lambda_1 \leq \lambda_2$ be the eigenvalues of H. Then for every $k \geq 1$,*

$$(HX_kX_k^\mathsf{T}H)^{1/2} = c_kI_2$$

for some $c_k \geq 0$, and these scalars satisfy

$$c_{k+1} = \rho\, c_k, \qquad \rho := \frac{\lambda_2 - \lambda_1}{\lambda_1 + \lambda_2} = \frac{\kappa(H) - 1}{\kappa(H) + 1}. \tag{41}$$

Consequently,

$$f(X_{k+1}) = \rho^2 f(X_k), \qquad k \geq 1. \tag{42}$$

Moreover, for every $k \geq 1$ with $c_k > 0$, the PolarGD direction ray coincides with the GD direction ray.

Proof Proposition 36 gives

$$(HX_1X_1^\mathsf{T}H)^{1/2} = c_1I_2$$

for some $c_1 \geq 0$. Suppose now that for some $k \geq 1$,

$$R_k := (HX_kX_k^\mathsf{T}H)^{1/2} = c_kI_2.$$

Then

$$\gamma_k^\star(Q) = \frac{\mathrm{tr}(R_k)}{\mathrm{tr}(H)} = \frac{2c_k}{\lambda_1 + \lambda_2}.$$

Using Proposition 31,

$$R_{k+1} = \left|c_kI_2 - \frac{2c_k}{\lambda_1 + \lambda_2}H\right|.$$

Diagonalizing $H = U\mathrm{diag}(\lambda_1, \lambda_2)U^\mathsf{T}$, we obtain

$$R_{k+1} = U\mathrm{diag}\left(\left|c_k\left(1 - \frac{2\lambda_1}{\lambda_1 + \lambda_2}\right)\right|, \left|c_k\left(1 - \frac{2\lambda_2}{\lambda_1 + \lambda_2}\right)\right|\right)U^\mathsf{T} = \rho c_kI_2.$$

Thus $c_{k+1} = \rho c_k$, proving (41).

Since $R_k = c_k I_2$, we have

$$G_k = R_k Q_k = c_k Q_k,$$

so whenever $c_k > 0$, the PolarGD and GD directions are collinear and therefore define the same direction ray.

Finally, because

$$H X_k X_k^\top H = R_k^2 = c_k^2 I_2,$$

we get

$$X_k X_k^\top = c_k^2 H^{-2}.$$

Hence

$$f(X_k) = \frac{1}{2}\mathrm{tr}(H X_k X_k^\top) = \frac{c_k^2}{2}\mathrm{tr}(H^{-1}),$$

and therefore

$$\frac{f(X_{k+1})}{f(X_k)} = \left(\frac{c_{k+1}}{c_k}\right)^2 = \rho^2.$$

This proves (42). ■

Specialization to the 2×2 case

We now specialize the line-search comparison between GD and Muon to the case $d = m = 2$, after diagonalizing H. Without loss of generality assume

$$H = \begin{bmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{bmatrix}, \qquad 0 < \lambda_1 \le \lambda_2,$$

and let

$$X = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}, \qquad G = HX = \begin{bmatrix} \lambda_1 x_{11} & \lambda_1 x_{12} \\ \lambda_2 x_{21} & \lambda_2 x_{22} \end{bmatrix}.$$

Write the rows of G as

$$g_1^\top = (\lambda_1 x_{11},\, \lambda_1 x_{12}), \qquad g_2^\top = (\lambda_2 x_{21},\, \lambda_2 x_{22}),$$

and let

$$r_1 := \|g_1\|_2^2, \quad r_2 := \|g_2\|_2^2, \quad c := \langle g_1, g_2 \rangle.$$

Then

$$\|G\|_F^2 = r_1 + r_2 =: a, \qquad \det(GG^\top) = r_1 r_2 - c^2.$$

Since $GG^\top$ is 2×2, its eigenvalues are the squared singular values σ_1^2, σ_2^2 of G, and

$$\sigma_1^2 + \sigma_2^2 = \|G\|_F^2 = a, \qquad \sigma_1 \sigma_2 = \sqrt{\det(GG^\top)} = \sqrt{r_1 r_2 - c^2}.$$

Hence the nuclear norm satisfies

$$\|G\|_*^2 = (\sigma_1 + \sigma_2)^2 = \sigma_1^2 + \sigma_2^2 + 2\sigma_1\sigma_2 = a + 2\sqrt{r_1 r_2 - c^2}. \tag{43}$$

The curvature term in the GD line-search denominator is

$$\mathrm{tr}(G^\top HG) = \lambda_1 \|g_1\|_2^2 + \lambda_2 \|g_2\|_2^2 = \lambda_1 r_1 + \lambda_2 r_2 =: K, \tag{44}$$

and the objective value can be expressed in terms of G as

$$f(X) = \frac{1}{2}\mathrm{tr}(X^\top HX) = \frac{1}{2}\mathrm{tr}(G^\top H^{-1} G) = \frac{1}{2}\left(\frac{r_1}{\lambda_1} + \frac{r_2}{\lambda_2}\right). \tag{45}$$

GD (line search) in the 2×2 case. For GD with line search, the direction is $D_{\mathrm{GD}} = G$, and the optimal step is

$$\gamma^{\mathrm{GD}} = \frac{\|G\|_F^2}{\mathrm{tr}(G^\top HG)} = \frac{a}{K}.$$

The one-step decrease is

$$\Delta_{\mathrm{GD}} := f(X) - f(X - \gamma^{\mathrm{GD}} G) = \frac{\|G\|_F^4}{2\,\mathrm{tr}(G^\top HG)} = \frac{a^2}{2K}. \tag{46}$$

Muon (line search) in the 2×2 case. Let $G = USV^\top$ be the SVD and $Q := \mathrm{polar}(G) = UV^\top$. We distinguish two cases.

Case 1: $\mathrm{rank}(G) = 2$. Then $U, V \in \mathbb{R}^{2\times 2}$ are orthogonal and $Q = UV^\top$ is orthogonal, so

$$Q^\top HQ = VU^\top HUV^\top$$

is similar to H and has the same eigenvalues. In particular,

$$\tau := \mathrm{tr}(Q^\top HQ) = \mathrm{tr}(H) = \lambda_1 + \lambda_2. \tag{47}$$

Line search along $-Q$ yields

$$\gamma^{\mathrm{Mu}} = \frac{\langle G, Q\rangle}{\mathrm{tr}(Q^\top HQ)} = \frac{\|G\|_*}{\tau},$$

and the one-step decrease

$$\Delta_{\mathrm{Mu}} := f(X) - f(X - \gamma^{\mathrm{Mu}} Q) = \frac{\|G\|_*^2}{2\,\mathrm{tr}(Q^\top HQ)} = \frac{\|G\|_*^2}{2(\lambda_1 + \lambda_2)}. \tag{48}$$

Using (43), this becomes

$$\Delta_{\mathrm{Mu}} = \frac{a + 2\sqrt{r_1 r_2 - c^2}}{2(\lambda_1 + \lambda_2)}. \tag{49}$$

Case 2: $\mathrm{rank}(G) = 1$. Then $G = \sigma uv^\top$ with one nonzero singular value σ, $\|u\|_2 = \|v\|_2 = 1$, and

$$Q = \mathrm{polar}(G) = uv^\top$$

spans the same direction as G. It is straightforward to check that GD and Muon directions coincide in this case and that with line search

$$\Delta_{\mathrm{Mu}} = \Delta_{\mathrm{GD}}, \tag{50}$$

so Muon and GD are equivalent whenever G is rank one.

When is Muon better than GD in 2×2? In the full-rank case, combining (46) and (49), Muon produces a larger decrease than GD if and only if

$$\Delta_{\mathrm{Mu}} > \Delta_{\mathrm{GD}} \quad \Longleftrightarrow \quad \frac{a + 2\sqrt{r_1 r_2 - c^2}}{2(\lambda_1 + \lambda_2)} > \frac{a^2}{2K},$$

i.e.

$$\big(a + 2\sqrt{r_1 r_2 - c^2}\big)\big(\lambda_1 r_1 + \lambda_2 r_2\big) > a^2(\lambda_1 + \lambda_2). \tag{51}$$

Recall

$$r_1 = \|g_1\|_2^2, \quad r_2 = \|g_2\|_2^2, \quad c = \langle g_1, g_2 \rangle, \quad a = r_1 + r_2.$$

Thus Muon with line search strictly outperforms GD with line search on the 2×2 quadratic (in the sense of one-step decrease) if and only if (51) holds and G is full rank ($r_1 r_2 > c^2$).

It is often convenient to parametrize the geometry of G via:

$$\alpha := \frac{r_1}{r_1 + r_2} \in [0, 1], \qquad \rho := \frac{c}{\sqrt{r_1 r_2}} \in [-1, 1] \text{ if } r_1 r_2 > 0,$$

so that

$$r_1 = \alpha a, \quad r_2 = (1 - \alpha) a, \quad \sqrt{r_1 r_2 - c^2} = a\sqrt{\alpha(1 - \alpha)}\sqrt{1 - \rho^2}.$$

In terms of (α, ρ), condition (51) becomes

$$\big[\lambda_1 \alpha + \lambda_2 (1 - \alpha)\big]\big[1 + 2\sqrt{\alpha(1 - \alpha)}\sqrt{1 - \rho^2}\big] > \lambda_1 + \lambda_2. \tag{52}$$

Here:

- α encodes how much of the gradient energy lies in the first eigen-direction (λ_1) versus the second (λ_2),
- ρ encodes the correlation between the two gradient rows (with $\rho = 0$ corresponding to orthogonal rows and $\rho = \pm 1$ to colinear rows).

Condition (52) makes explicit how Muon's advantage depends on the anisotropy of H (through λ_1, λ_2), the distribution of gradient energy across eigendirections (through α), and the mutual orientation of the gradient rows (through ρ).

In particular:

- If G is rank one ($r_1 r_2 = c^2$), then $\sqrt{r_1 r_2 - c^2} = 0$ and Muon coincides with GD.
- If H is isotropic ($\lambda_1 = \lambda_2$), then (52) cannot hold strictly, and GD is always at least as good as Muon in terms of one-step decrease (with equality only in the highly symmetric case where the singular values of G coincide).
- For anisotropic H with $\lambda_1 \neq \lambda_2$ and full-rank G, the region defined by (52) is nonempty: there exist configurations where Muon achieves a strictly larger line-search decrease than GD, reflecting a favorable alignment of the gradient singular structure with the spectrum of H.

H.1. Optimizer Trajectories in Singular Value Space

By symmetry of the trace operator and von Neumann's identity we have

$$f(X) = \frac{1}{2}\mathrm{tr}(HXX^T) \leq \sum_{i=1}^{\min(d,m)} \sigma_i(X)^2 \lambda_i(H), \tag{53}$$

where $\lambda_i(H)$, $\sigma_i(X)$ are in decreasing order. Crucially, this upper bound depends only on the singular values of X and the eigenvalues of H. In the 2×2 case we can therefore overlay this upper bound on the trajectories of singular values. This visualization (Figure 8) makes the interplay between the spectrum of X and the structure of H immediately apparent in two dimensions.

Appendix I. Comparison to concurrent work on spectral/polar methods

Concurrent perspectives on spectral/polar updates. Several recent works study Muon/SpecGD-type spectral directions from complementary angles.Davis and Drusvyatskiy [3] develop a local (one-step) condition for when spectral/polar-normalized updates can outperform Euclidean GD in neural-network training, relating the advantage to singular-value structure (nuclear vs Frobenius geometry) and to rank-type properties of activations. In contrast, Gonon et al. [4] show that even on simple strongly convex quadratics, Muon-style dynamics with *constant* step sizes and exact polar projection can fail to reach the minimizer ("grid confinement"), that approximation/inexactness in the polar step can *improve* reachability and finite-time performance, and that greedy one-step superiority need not translate into faster end-to-end convergence. On the more "mechanism" side, Ma et al. [13] prove problem-specific global guarantees showing that spectral orthogonalization can act as a genuine preconditioner (e.g. in matrix factorization and linear-transformer in-context learning), yielding rates that do not degrade with the Hessian condition number. Finally, Jiang et al. [6] interpret post-spectral clipping through a Frank–Wolfe lens and demonstrate its practical viability for LLM pretraining, further emphasizing the relevance of spectral geometry in large-scale optimization. Our work is complementary: we isolate *direction geometry* via exact line search, derive *exact* gain–curvature identities on matrix quadratics, and use an *implicit-preconditioning* view to explain both fast (row-isotropic) and pessimistic (tiny-singular-value) regimes and to motivate rank-truncated/damped polar variants.

More detailed comparison.

Davis and Drusvyatskiy [3]: local conditions for SpecGD gains in deep learning. Our one-step quadratic identities (e.g. the exact gain–curvature tradeoff and the inequality $\Delta_{\mathrm{Pol}} > \Delta_{\mathrm{GD}}$) provide a clean "laboratory" for interpreting more general one-step criteria in nonconvex training. A key message common to both works is that *singular-value structure of the gradient matters*: spectral/polar directions behave differently from GD depending on how spread the singular values are. A key difference is scope: Davis and Drusvyatskiy [3] target neural networks and develop a condition involving activation structure, whereas we keep the objective class minimal (matrix quadratics) but obtain exactness regimes and a trajectory-level preconditioning mechanism.

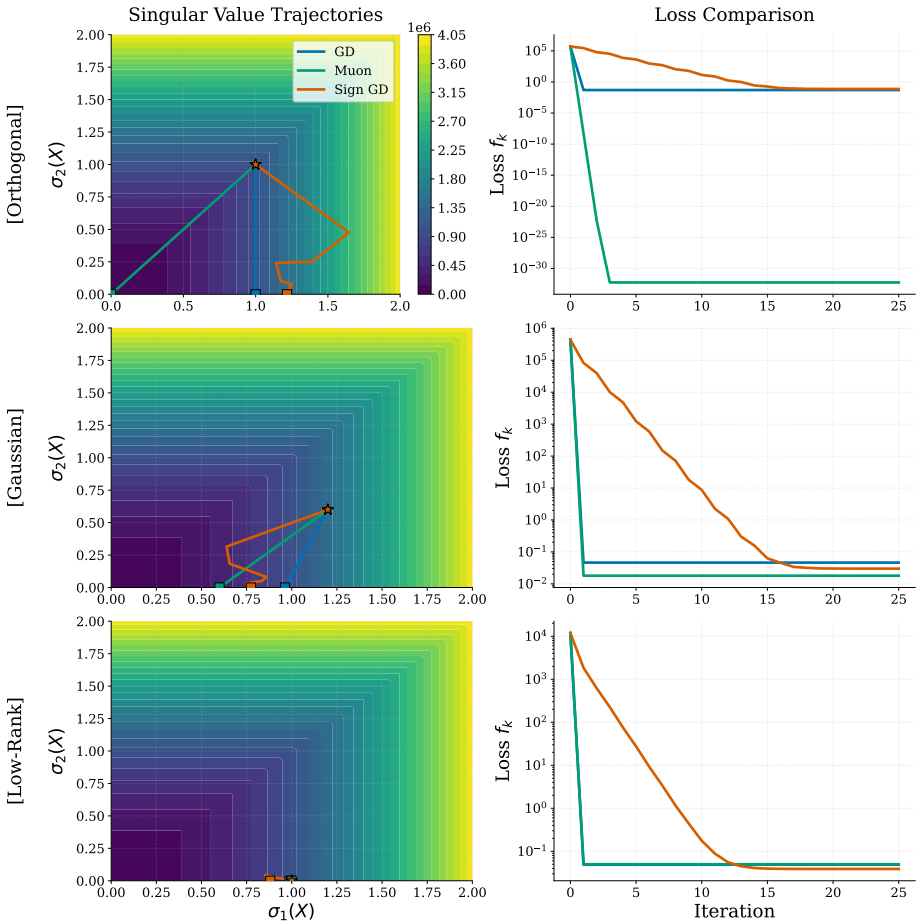

Figure 8: **Illustration of Singular Value Trajectories in $\mathbb{R}^2$**
Rows represent different initialization regimes and the right hand column we display the corresponding loss curves. For each row, optimizer trajectories in singular value space, all start from the same initialization ($\star$), with the von Neumann bound (53) overlaid. Terminal iterates are marked with a square ($\square$). Here $H = Q\text{diag}(1{,}10{,}000)Q^T$ for $Q_{ij} \sim \mathcal{N}(0,1)$ We see that Muon's singular value trajectory maintains a 45° direction in singular value space. By contrast GD is seen to optimize the larger eigen-direction first. SignGD does not obviously exploit singular geometry here.

Gonon et al. [4]: discrete-time pathologies and constructive inexactness on quadratics. This work is the closest "quadratic stress test" to ours, but it targets different questions. They show that *constant* step sizes can prevent Muon-style dynamics from reaching the minimizer (grid confinement), and that (stochastic or approximate) polar steps can restore reachability and even accelerate finite-time progress. This directly complements our choice to emphasize exact line search when isolating geometry: line search avoids constant-step-size reachability obstructions and cleanly separates (i) direction geometry from (ii) step-size scheduling. Their "greedy one-step" counterexample reinforces our message that local certificates do not induce a global method ordering, and motivates our analytic alternation example and our trajectory-level preconditioning view.

Ma et al. [13]: global preconditioning benefits in structured problems. Where we derive an *implicit preconditioner* for PolarGD and show how its conditioning is governed by iterate geometry, Ma et al. [13] prove that in certain structured objectives this mechanism yields global rates independent of Hessian conditioning. This provides a strong external validation of the "preconditioning story" and suggests a useful narrative bridge: our analysis explains *how* the preconditioner arises and what it depends on (e.g. row-covariance / conditioning of iterates), while their work shows *when* this can translate into global improvements in specific learning models.

Jiang et al. [6]: spectral clipping as Frank–Wolfe (and LLM viability). While we focus on polar-normalized directions (exact LMO for a spectral-norm geometry), Jiang et al. [6] study a different spectral primitive—post-spectral clipping—and connect it to (composite) Frank–Wolfe. This is best cited as complementary motivation: spectral geometry can be exploited through multiple algorithmic routes (polar/orthogonalization vs. clipping/projection), and our work helps clarify what is special about the polar route (e.g. iterate-dependent preconditioning effects and sensitivity to small singular values).

Appendix J. Matrix Smoothness: A Generalization of this Work

While this work focused on the matrix quadratic case, one can derive more general results through the lens of matrix smoothness. Bounds derived here become tight in the quadratic case, which we developed in Section 3. **Assumption (matrix curvature bounds).** There exist symmetric positive semidefinite matrices $M, L \in \mathbb{R}^{d\times d}$ with $M \preceq L$ such that for all $X \in \mathbb{R}^{d\times m}$ and all $D \in \mathbb{R}^{d\times m}$,

$$f(X) + \langle G, D\rangle + \frac{1}{2}\langle MD, D\rangle \;\le\; f(X+D) \;\le\; f(X) + \langle G, D\rangle + \frac{1}{2}\langle LD, D\rangle, \tag{54}$$

where $\langle LD, D\rangle = \mathrm{tr}(D^\top LD)$ (and similarly for M). Equivalently, the second-order remainder in direction D is bounded between $\frac{1}{2}\|M^{1/2}D\|_F^2$ and $\frac{1}{2}\|L^{1/2}D\|_F^2$.

Assumption (54) strictly generalizes the usual scalar bounds: if $L = \ell I_d$ and $M = \mu I_d$, then (54) is exactly ℓ-smoothness and μ-strong convexity with respect to the Frobenius norm. Allowing a general L captures anisotropic curvature (e.g., diagonal L yields coordinate/row-wise smoothness). For the quadratic $f(X) = \frac{1}{2}\mathrm{tr}(X^\top HX)$, the bounds hold with equality for $L = M = H$.

Model-based step size along a direction. Fix $D \in \mathbb{R}^{d\times m}$ and consider $X^+ = X - \gamma D$ with $\gamma \ge 0$. Plugging $-\gamma D$ into (54) gives, for all $\gamma \ge 0$,

$$f(X) - \gamma\,\langle G, D\rangle + \frac{\gamma^2}{2}\,\langle MD, D\rangle \;\le\; f(X-\gamma D) \;\le\; f(X) - \gamma\,\langle G, D\rangle + \frac{\gamma^2}{2}\,\langle LD, D\rangle. \tag{55}$$

Define the directional curvatures $c_L(D) := \langle LD, D\rangle$, $c_M(D) := \langle MD, D\rangle$ and the positive part $[a]_+ := \max\{a, 0\}$. Minimizing the upper quadratic model yields the closed-form step size

$$\gamma_L(D) := \frac{[\langle G, D\rangle]_+}{c_L(D)}, \qquad \text{(convention: } \gamma_L(D) = 0 \text{ if } c_L(D) = 0). \tag{56}$$

Proposition 38 (Model-based one-step decrease and exact line-search bracket) *Assume* (54)*. Fix* X *and a direction* D *with* $c_L(D) > 0$*. Let* $X^+ = X - \gamma_L(D)D$*, where* $\gamma_L(D)$ *is given by* (56)*. Then*

$$f(X^+) \le f(X) - \frac{[\langle G, D\rangle]_+^2}{2\, c_L(D)}. \tag{57}$$

Moreover, if $c_M(D) > 0$ *and* $\gamma^\star(D) \in \arg\min_{\gamma \ge 0} f(X - \gamma D)$ *is the exact line-search minimizer, then the exact step and exact decrease* $\Delta^\star(D) := f(X) - f(X - \gamma^\star(D)D)$ *satisfy*

$$\frac{[\langle G, D\rangle]_+}{c_L(D)} \le \gamma^\star(D) \le \frac{[\langle G, D\rangle]_+}{c_M(D)}, \qquad \frac{[\langle G, D\rangle]_+^2}{2\, c_L(D)} \le \Delta^\star(D) \le \frac{[\langle G, D\rangle]_+^2}{2\, c_M(D)}. \tag{58}$$

LMO-normalized and dual-adapted variants through the same formulas. After optimizing over γ, the update $X - \gamma^\star(D)D$ and the certified decreases (57)–(58) are invariant under rescaling $D \mapsto cD$ with $c > 0$. Hence all line-search comparisons below depend only on the *direction ray*. This lets us study LMO-normalized and dual-adapted variants through the same formulas.