FISMO: FISHER-STRUCTURED MOMENTUM-ORTHOGONALIZED OPTIMIZER

Anonymous authors
Paper under double-blind review

ABSTRACT

Training large-scale neural networks requires solving nonconvex optimization problems, where the choice of optimizer strongly affects both convergence behavior and computational efficiency. While adaptive methods like Adam have long dominated practice, the recently proposed Muon optimizer achieves superior performance through orthogonalized momentum updates that enforce isotropic geometry with uniform singular values. However, this strict isotropy discards potentially valuable curvature information encoded in gradient spectra, motivating optimization methods that balance geometric structure with adaptivity. We introduce **FISMO** (**FI**sher-**S**tructured **M**omentum-**O**rthogonalized), which generalizes isotropic updates to incorporate anisotropic curvature through Fisher information geometry. By reformulating the optimizer update as a trust-region problem constrained by a Kronecker-factored Fisher metric, FISMO achieves structured preconditioning that adapts to local loss landscape geometry while maintaining computational tractability. We establish convergence guarantees for FISMO in nonconvex settings, proving an $\mathcal{O}(T^{-1/2})$ deterministic rate and an $\mathcal{O}(T^{-1/4})$ stochastic rate for the expected gradient norm. Moreover, we theoretically show that FISMO can achieve sharper finite-time contraction than Muon-style isotropic orthogonalization through a pre-asymptotic analysis. Empirical evaluation on image classification and language modeling benchmarks demonstrates that FISMO achieves superior training efficiency and final performance compared to established baselines.

1 INTRODUCTION

The remarkable progress of large language models (LLMs) (Touvron et al., 2023; Achiam et al., 2023; Team et al., 2023) has intensified the challenge of training neural networks with millions to billions of parameters. This scale necessitates solving highly nonconvex optimization problems, where the choice of optimizer fundamentally determines both computational efficiency and model performance. While Adam (Kingma, 2014) and its variant AdamW (Loshchilov & Hutter, 2017) have served as the de facto standards for over a decade, the demands of modern deep learning continue to drive the search for more effective optimization algorithms.

A recent breakthrough in this pursuit is Muon (Jordan et al., 2024), a matrix-parameter optimizer that departs radically from element-wise adaptive methods. By orthogonalizing the momentum matrix to normalize all singular values to unity, Muon produces isotropic updates that prevent pathological amplification or suppression along specific directions. This geometric approach has demonstrated substantial improvements in both training stability and downstream accuracy across diverse applications (Liu et al., 2025; Wang et al., 2025; Tveit et al., 2025). In practice, Muon employs Newton–Schulz iterations (Higham, 2008) to efficiently approximate the orthogonalization operation.

Theoretically, Muon implements steepest descent under a spectral norm constraint (Bernstein & Newhouse, 2024; Chen et al., 2025). Following Bernstein (2025), Muon's update solves the constrained linear minimization oracle (LMO) problem (Lan, 2013):

$$\min_{\Delta W \in \mathbb{R}^{m\times n}} \langle \nabla_W \mathcal{L}, \Delta W \rangle \quad \text{s.t.} \quad \|\Delta W\|_2 \le \eta, \tag{1}$$

where $\mathcal{L}$ denotes the objective function, W represents the matrix parameters, and $\|\cdot\|_2$ denotes the spectral norm. This formulation seeks the update that maximally decreases the linearized objective

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

while constraining the worst-case amplification: $\|\Delta W\|_2 = \max_{\|x\|_2=1} \|\Delta W x\|_2$. The solution aligns with the gradient's singular directions and saturates the spectral constraint, yielding an isotropic update with uniform singular values (Chen et al., 2025; Bernstein, 2025).

However, the optimality of isotropic geometry for deep learning remains contested. The heterogeneous singular-value spectrum of gradient matrices encodes valuable information about curvature and scale variations across directions—information that uniform normalization may discard (Lau et al., 2025). Furthermore, theoretical analysis under isotropic-curvature models suggests that while Muon excels when curvature exhibits rapid transitions, such conditions may not characterize the loss landscapes of large neural networks (Su, 2025). Therefore, the isotropic update of Muon might not be the optimal choice in complicated nonconvex settings, especially in large-scale deep learning. These considerations raise a fundamental question:

Beyond isotropic updates, what additional structure should an optimizer leverage to produce more informative and effective update directions for training deep networks?

To address this question, we develop a structured-geometry framework for optimizer design. We reformulate optimizer updates through the LMO lens (Lan, 2013) and establish connections to natural-gradient methods (Martens, 2020) under Fisher information geometry (Costa et al., 2015). Recognizing that exact Fisher information is computationally intractable, we derive the optimal Kronecker-structured approximation (Martens & Grosse, 2015) and solve the resulting Fisher-structured LMO in closed form. This solution motivates a practical optimizer that balances computational efficiency with geometric informativeness, for which we establish convergence guarantees and demonstrate empirical effectiveness.

Our main contributions are summarized below.

- **Structured-Geometry Framework.** We develop a principled framework connecting second-order methods and LMO-based optimizers through Fisher information-induced trust regions. This framework reveals how local geometry shapes update directions, enabling adaptive, anisotropic updates that preserve curvature information lost in isotropic orthogonalization.
- **The FISMO Algorithm.** We introduce FISMO (**FI**sher-**S**tructured **M**omentum-**O**rthogonalized) optimizer, a practical instantiation of our theoretical framework. FISMO combines Kronecker-factored Fisher preconditioning with orthogonalized momentum, achieving computational efficiency while adapting to local curvature. This design synthesizes the stability of orthogonalization with the informativeness of second-order geometry.
- **Convergence and pre-asymptotic theory.** We prove nonconvex convergence rates for FISMO, with an $\mathcal{O}(T^{-1/2})$ deterministic rate and an $\mathcal{O}(T^{-1/4})$ stochastic rate. We further show that FISMO achieves sharper pre-asymptotic decay than Muon-style isotropic updates.
- **Empirical Validation.** Experiments on representative deep learning benchmarks show FISMO's improved training stability and performance relative to established optimizers.

Basic Notation. Lowercase and uppercase letters denote vectors and matrices, resp. Let I_d be the identity matrix and $\mathrm{tr}(\cdot)$ the trace. We use $\|\cdot\|_F$, $\|\cdot\|_2$, and $\|\cdot\|_*$ for the Frobenius, spectral, and nuclear norms, respectively. The operators $\mathrm{vec}(\cdot)$, $\otimes$, and $\mathrm{Orth}(\cdot)$ denote vectorization, Kronecker product, and orthogonal polar factor. Detailed notations can be found in Appendix A.

2 Related Work

Matrix-Parameter Optimizers. Matrix-parameter optimizers treat key network weights (e.g., attention and linear layers) as matrices and design update rules that explicitly exploit their row/column or block structure. Among these approaches, some works mainly aim to reduce Adam-style state and improve efficiency by sharing or factorizing second-moment statistics at the block or layer level, including Adafactor (Shazeer & Stern, 2018) and Adam-mini (Zhang et al., 2024). Other methods, such as Shampoo (Gupta et al., 2018) and its variant SOAP (Vyas et al., 2024), exploit tractable matrix factorizations (like Kronecker structure) to approximate large curvature preconditioners. More recently, orthogonalized matrix updates have emerged as a prominent design choice, with Muon

(Jordan et al., 2024) as a representative example and an expanding line of work that extends this matrix-geometry-driven update shaping. Closely related to Muon, ASGO (An et al., 2025) utilizes a single-sided structured preconditioner built from accumulated matrix-gradient second moments. PolarGrad (Lau et al., 2025) retains Muon's orthogonalized direction but adds a trace-based, curvature-dependent scaling, rather than enforcing a purely isotropic update. MuonBP (Khaled et al., 2025) adapts Muon's updates to large-scale parallel training via block-periodic orthogonalization.

Natural Gradient Methods. Natural gradient methods view learning as steepest descent under the Fisher information metric (Martens, 2020). In machine-learning settings, the Fisher matrix coincides with the generalized Gauss–Newton matrix (Shrestha, 2023), so natural gradient updates can be interpreted as estimates of Hessian. Since forming the full Fisher is infeasible at modern scales, practical algorithms therefore rely on structured approximations. For example, K-FAC (Martens & Grosse, 2015) uses a Kronecker-factored approximation of Fisher blocks, with extensions including recurrent-network variants and eigenbasis corrections such as EKFAC (George et al., 2018). Other approaches pursue lighter-weight or implicit approximations, such as quasi-diagonal Riemannian constructions like TONGA (Roux et al., 2007). Some works like TANGO (Ollivier, 2017) also asymptotically exact trajectories without explicit Fisher estimation.

3 PRELIMINARIES

Modern neural network training requires minimizing an expected loss over high-dimensional parameter spaces. Predominant optimizers, including Adam and other adaptive first-order methods, typically treat parameters as vectorized, unstructured arrays (Klein et al., 2009). However, a substantial portion of neural network parameters, especially in transformer architectures, are naturally matrix-valued weights $W \in \mathbb{R}^{m\times n}$. Such vectorization can discard important geometric structures, including low-rank gradient patterns, input-output coupling, and spectral characteristics that govern high-dimensional optimization dynamics. Recent evidence suggests that preserving matrix geometry can accelerate convergence, improve conditioning, and enhance training stability (Gupta et al., 2018).

The Muon Optimizer and Isotropic Geometry. The Muon optimizer (Jordan et al., 2024) represents a recent advance in this direction, preserving matrix geometry through polar decomposition. For matrix parameters $W \in \mathbb{R}^{m\times n}$, Muon maintains a momentum accumulator and orthogonalizes it as:

$$M_t = \beta M_{t-1} + (1-\beta)\nabla_W \mathcal{L}(W_{t-1}); \qquad W_t = W_{t-1} - \eta_t \,\mathrm{Orth}(M_t),$$

where $\beta \in [0, 1)$ is the momentum coefficient. This update mechanism has an elegant interpretation as the solution to a Linear Minimization Oracle (LMO) subproblem with a spectral norm constraint: $\min_{\|\Delta W\|_2 \le \eta} \langle M_t, \Delta W\rangle_F$. The solution, $\Delta W^* = -\eta \cdot \mathrm{Orth}(M_t)$, normalizes the singular values of the momentum matrix to unity while preserving the singular vectors. Consequently, it enforces a perfectly **isotropic** geometry: all update directions are scaled equally, irrespective of the curvature information encoded in the gradient's singular value spectrum. Computationally, Muon avoids the cost of a full SVD by using Newton-Schulz iterations, which rely only on efficient matrix multiplications. Within months of its release, Muon has been adopted across diverse training scenarios, from language model pretraining to vision transformers, consistently demonstrating 1.5-2× faster convergence than AdamW while maintaining or improving final performance (Jordan et al., 2024; Liu et al., 2025).

However, this isotropic constraint discards the heterogeneous curvature information encoded in the gradient's singular value spectrum—High-curvature directions (corresponding to large singular values) are treated identically to low-curvature ones (small singular values), potentially limiting adaptation to local landscape geometry.

Natural Gradient Descent and Anisotropic Geometry. Natural Gradient Descent (NGD) offers a framework for adapting to the local geometry of the loss landscape (Martens & Grosse, 2015). NGD preconditions the gradient with the Fisher Information Matrix (FIM), a canonical Riemannian metric on the statistical manifold of model parameters (Ly et al., 2017): $F(\theta) = \mathbb{E}_{z\sim p_\theta}\left[\nabla_\theta \log p_\theta(z) \nabla_\theta \log p_\theta(z)^\top\right]$. NGD then solves a FIM-regularized trust-region problem:

$$\min_{\Delta\theta} \langle \nabla\mathcal{L}(\theta_t), \Delta\theta\rangle \qquad \text{s.t. } \Delta\theta^\top F(\theta_t)\Delta\theta \le \eta^2, \tag{2}$$

which yields the update $\Delta\theta^* \propto -F(\theta_t)^{-1}\nabla\mathcal{L}(\theta_t)$. This **anisotropic** update adapts to local curvature by taking larger steps along flat directions and smaller steps along steep ones.

Inspired by FIM, deep learning practice often adopts the empirical Fisher as a surrogate for second-order information: $F_{\text{emp}}(\theta) = \mathbb{E}_{\xi\sim\mathcal{D}}\left[\nabla_\theta \ell(\theta;\xi)\nabla_\theta \ell(\theta;\xi)^\top\right]$ [1]. However, direct manipulation of empirical Fisher is computationally prohibitive for large-scale models. For a matrix parameter $W \in \mathbb{R}^{m\times n}$, vectorizing $\theta = \text{vec}(W)$ yields a Fisher-type matrix of size $mn \times mn$, making it costly to compute, store, and invert. This establishes a fundamental dichotomy in modern optimization: Muon achieves efficiency through isotropic orthogonalization but sacrifices curvature information, while NGD-style methods capture Fisher-type geometry only at prohibitive cost. Our work bridges this gap by incorporating empirical Fisher-derived geometry into an LMO framework, thereby achieving geometry-aware updates with lower computational efficiency.

4 Theoretical Framework

Building upon the NGD principles, we now develop our approach for optimizing matrix-valued parameters $W \in \mathbb{R}^{m\times n}$. Our goal is to derive a geometry-aware update rule that preserves the benefits of NGD while remaining computationally tractable for large-scale models.

4.1 Kronecker-Factored Approximation of FIM

We consider the empirical Fisher block $F_W = \mathbb{E}_{\xi\sim\mathcal{D}}[\text{vec}(\nabla_W\mathcal{L})\,\text{vec}(\nabla_W\mathcal{L})^\top]$, an $mn\times mn$ matrix. Direct manipulation of F_W is computationally prohibitive for large-scale neural networks. We therefore adopt the K-FAC strategy (Martens & Grosse, 2015) to approximate it through a Kronecker product. Specifically, we introduce two symmetric positive definite matrices, $P, Q \in \mathbb{S}^n_{++}$, such that $F_W \approx Q \otimes P$. Under this approximation, the trust region constraint in equation 2 becomes:

$$\text{vec}(\Delta W)^\top (Q\otimes P)\,\text{vec}(\Delta W) \le \eta^2,$$

where $\text{vec}(\Delta W)$ denotes the vectorization of the matrix update. Using the identity $\text{vec}(\Delta W)^\top(Q\otimes P)\,\text{vec}(\Delta W) = \text{tr}(\Delta W^\top P \Delta W Q)$, we reformulate this constraint as:

$$\text{tr}(\Delta W^\top P\Delta W Q) = \|P^{1/2}\Delta W Q^{1/2}\|_F^2 \le \eta^2.$$

While the Frobenius constraint arises naturally from the formulation, it controls only the aggregate update energy. To control the largest directional amplification, we instead impose a spectral-norm constraint. Since $\|X\|_2 \le \|X\|_F \le \sqrt{\text{rank}(X)}\|X\|_2$ (Horn & Johnson, 2012), this choice remains within the same whitened geometry while yielding updates that are less dominated by large singular directions. This leads to our central optimization problem, which seeks the steepest descent direction ΔW within a trust region defined by the Kronecker-factored geometry:

$$\boxed{\min_{\Delta W\in\mathbb{R}^{m\times n}} \langle G, \Delta W\rangle_F \qquad \text{s.t. } \|P^{1/2}\Delta W Q^{1/2}\|_2 \le \eta,} \tag{3}$$

where G denotes the gradient of the loss function. This formulation raises two fundamental questions:

[Q1] What choices of P and Q best approximate the empirical Fisher information matrix F_W?

[Q2] What is the optimal solution to equation 3 that determines our update rules?

4.2 The Best Kronecker Approximation

To answer question [Q1], we need to first define what constitutes the "best" Kronecker-product approximation. We measure the discrepancy between two positive definite matrices using the *log-det divergence* (Cichocki et al., 2015):

$$D_{\text{ld}}(A\,\|\,B) \coloneqq \text{tr}(B^{-1}A) - \log\det(B^{-1}A) - d,$$

[1] For notational simplicity, we use F to denote the empirical Fisher throughout the remainder of this paper.

which is equivalent (up to constants) to the Kullback–Leibler (KL) divergence between two centered multivariate Gaussian $\mathcal{N}(0, A)$ and $\mathcal{N}(0, B)$. We seek the preconditioners P and Q that minimize this divergence between the damped empirical FIM and its Kronecker approximation that:

$$\min_{P\in\mathbb{S}^m_{++},\, Q\in\mathbb{S}^n_{++}} \; D_{\mathrm{ld}}\left(F_W + \mu I_{mn} \,\middle\|\, Q\otimes P\right). \tag{4}$$

Here, $\mu > 0$ is a small damping constant that ensures the matrix is positive definite, a common practice in natural gradient methods. Using the properties $(Q\otimes P)^{-1} = Q^{-1}\otimes P^{-1}$ and $\log\det(Q\otimes P) = n\log\det P + m\log\det Q$, Problem equation 4 is equivalent to minimizing the following objective:

$$\begin{aligned}\mathcal{J}(P,Q) :=& \operatorname{tr}\left((Q^{-1}\otimes P^{-1})(F_W+\mu I_{mn})\right) \;+\; n\log\det P \;+\; m\log\det Q.\\ =&\mathbb{E}\big[\operatorname{tr}\big(P^{-1}GQ^{-1}G^\top\big)\big] + \mu\operatorname{tr}(P^{-1})\operatorname{tr}(Q^{-1}) \;+\; n\log\det P \;+\; m\log\det Q.\end{aligned} \tag{5}$$

The optimal preconditioners P and Q are characterized by the following theorem, which suggests an alternating minimization scheme for their computation. The proof is provided in Appendix E.1.

Theorem 4.1 (Optimal Kronecker Approximation)**.** *Suppose $\mathbb{E}\|G\|_F < \infty$. The objective $\mathcal{J}(P,Q)$ in equation 5 is convex in P for fixed Q, and convex in Q for fixed P. Moreover, the blockwise unique minimizers are coupled through the following fixed-point relations*

$$\begin{aligned}P^*(Q) &= \frac{1}{n}\mathbb{E}\big[GQ^{-1}G^\top\big] + \frac{\mu\operatorname{tr}(Q^{-1})}{n} I_m, \qquad Q\in\mathbb{S}^n_{++},\\ Q^*(P) &= \frac{1}{m}\mathbb{E}\big[G^\top P^{-1}G\big] + \frac{\mu\operatorname{tr}(P^{-1})}{m} I_n, \qquad P\in\mathbb{S}^m_{++}.\end{aligned}$$

4.3 The Best Update

With the best choice of P and Q, we now answer the question [Q2] to find the optimal solution for Problem equation 3. Applying a whitening change of variables to the objective, we discover a lower bound of equation 3 by von Neumann trace inequality D.1. By constructing a feasible update that attains this bound, we obtain the following theorem, whose proof is placed in Appendix E.2.

Theorem 4.2. *Suppose $G\in\mathbb{R}^{m\times n}$ is a nonzero matrix, $P\in\mathbb{S}^m_{++}, Q\in\mathbb{S}^n_{++}$ are defined as in Theorem 4.1. Let $\eta > 0$ be a given radius. Denote $\widetilde{G} := P^{-1/2}GQ^{-1/2}$, and suppose $\widetilde{G} = U\Sigma V^\top$ is the singular value decomposition of $\widetilde{G}$. Then the optimal solution to equation 3 is*

$$\Delta W^* = -\eta P^{-1/2}UV^\top Q^{-1/2} = -\eta P^{-1/2}\operatorname{Orth}(\widetilde{G})Q^{-1/2}.$$

Remark 4.3. *Theorem 4.2 gives a closed-form expression for the best update direction under the (P,Q)-preconditioned spectral-norm trust region. In particular, the optimal step is characterized by the polar factor $\operatorname{Orth}(\widetilde{G}) = UV^\top$ of the preconditioned gradient $\widetilde{G} = P^{-1/2}GQ^{-1/2}$. These characterizations will be used as the main building block for the algorithmic construction in Section 5.*

5 FISMO Algorithm

5.1 Algorithm Design

Based on the results of Theorem 4.1 and Theorem 4.2, we develop the FIsher Structured Momentum Orthogonalization (FISMO) algorithm, presented in Algorithm 1. The algorithm iteratively (i) updates the left and right preconditioners (P_t, Q_t), (ii) whitens the gradient, and (iii) computes the optimal update in the whitened space. Details are described as follows.

Gauss–Seidel Preconditioner Updates. Theorem 4.1 establishes that the optimal left preconditioner P depends on the right factor Q, and vice versa. Consequently, (P,Q) must be updated sequentially rather than simultaneously. We employ Gauss–Seidel iteration: as shown in Steps 4–5 of Algorithm 1, we compute P_t using Q_{t-1}, then compute Q_t using the newly obtained P_t.

This sequential updating offers advantages over Jacobi-style iteration, which computes both factors from (P_{t-1}, Q_{t-1}) simultaneously. First, the Gauss–Seidel method achieves superior conditioning by incorporating the updated left geometry when computing Q_t, thereby accelerating convergence of the whitening transform $P_t^{-1/2}(\cdot)Q_t^{-1/2}$. Second, it reduces memory requirements: after computing P_t, only the current value is retained, eliminating storage of P_{t-1}.

Algorithm 1 FISMO (FIsher-Structured Momentum-Orthogonalized Optimizer)

1: **Input:** $W_0 \in \mathbb{R}^{m\times n}$; learning rate η; momentum β; EMA γ; damping factor μ; $P_0 \leftarrow I_m$; $Q_0 \leftarrow I_n$; $M_0 = 0$.
2: **for** $t = 1, \dots, T$ **do**
3: Compute gradient $G_t = \frac{1}{B}\sum_{i=1}^{B} \nabla_W \ell(W_{t-1}; \xi_{t,i})$.
4: Update preconditioners:
5: Left: $\widetilde{P}_t \leftarrow \gamma P_{t-1} + (1-\gamma)\left(\frac{1}{n} G_t Q_{t-1}^{-1} G_t^\top + \mu \frac{\mathrm{tr}(Q_{t-1}^{-1})}{n} I_m\right)$; $P_t \leftarrow m\widetilde{P}_t / \mathrm{tr}(\widetilde{P}_t)$.
6: Right: $\widetilde{Q}_t \leftarrow \gamma Q_{t-1} + (1-\gamma)\left(\frac{1}{m} G_t^\top P_t^{-1} G_t + \mu \frac{\mathrm{tr}(P_t^{-1})}{m} I_n\right)$; $Q_t \leftarrow n\widetilde{Q}_t / \mathrm{tr}(\widetilde{Q}_t)$.
7: Whiten gradient $\widetilde{G}_t \leftarrow P_t^{-1/2} G_t Q_t^{-1/2}$.
8: Update momentum $M_t \leftarrow \beta M_{t-1} + (1-\beta)\widetilde{G}_t$.
9: Update parameters $W_t \leftarrow W_{t-1} - \eta P_t^{-1/2} \mathrm{Orth}(M_t) Q_t^{-1/2}$.
10: **end for**

Preconditioner Stabilization. The optimal preconditioner (P, Q) given in Theorem 4.1 are derived from instantaneous curvature estimates that may exhibit high variance in stochastic settings. To mitigate this, we employ three stabilization techniques in Algorithm 1.

1) Exponential Moving Average (EMA). Instead of directly applying the one-step update suggested by Theorem 4.1, we update the Kronecker factors through EMA smoothing, combining the current curvature estimate with the previous preconditioners P_{t-1} and Q_{t-1} (Step 5&6, Algorithm 1). This assigns larger weights to recent estimates while retaining historical curvature information. Specifically, the contribution of an observation from k steps ago is weighted by $\gamma^k(1-\gamma)$, ensuring smooth adaptation to changing curvature while suppressing high-frequency noise from mini-batch sampling.

2) Identity Regularization. The preconditioner updates include scaled-identity terms induced by the damping factor μ in Problem equation 4. As shown in Theorem 4.1, these identity components arise directly from the stabilized Kronecker approximation and keep the preconditioners positive definite even when empirical curvature estimates are rank-deficient. The trace-based scalings, such as $\mathrm{tr}(Q_{t-1}^{-1})/n$ and $\mathrm{tr}(P_t^{-1})/m$, adapt the regularization strength to the current inverse preconditioner magnitudes, preventing instability in poorly observed directions.

3) Trace Normalization. Algorithm 1 symmetrizes and normalizes the EMA estimates that: $P_t \leftarrow \mathrm{sym}\left(m\widetilde{P}_t / \mathrm{tr}(\widetilde{P}_t)\right)$, $Q_t \leftarrow \mathrm{sym}\left(n\widetilde{Q}_t / \mathrm{tr}(\widetilde{Q}_t)\right)$. This eliminates scale non-identifiability in the bilinear metric, since (P, Q) and $(cP, c^{-1}Q)$ induce identical preconditioned geometries for any $c > 0$. Fixing $\mathrm{tr}(P_t) = m$ and $\mathrm{tr}(Q_t) = n$ prevents scale drift between factors and ensures numerical stability. Combined with μ-damping from Theorem 4.1, this normalization yields uniform bounds on $\|P_t^{-1/2}\|_2$ and $\|Q_t^{-1/2}\|_2$, which is essential for convergence analysis.

Structured Orthogonalized Momentum. According to theorem 4.2, the optimal update for equation 3 requires an orthogonalization in preconditioned coordinates. Therefore, FISMO maintains momentum in whitened space and orthogonalizes it to determine the update. After whitening the stochastic gradient by current preconditioners P_t and Q_t (Step 7, Algorithm 1), we update the momentum in the whitened space (Step 8, Algorithm 1). Instead of using $\widetilde{G}_t$ directly, FISMO applies the orthogonalized momentum, $\mathrm{Orth}(M_t)$, which combines EMA with the structured direction in the (P, Q)-geometry. The resulting update direction is transformed back to the original coordinates (Step 9, Algorithm 1), as derived in Theorem 4.2.

Practical Efficiency. To reduce computational and memory overhead, FISMO updates the Kronecker preconditioners periodically rather than at every iteration. In our implementation, we refresh them every 10 iterations and reuse the latest factors in between, thereby amortizing the cost of curvature estimation, inverse square roots, and whitening.

Since exact polar decomposition via SVD is computationally prohibitive at scale, we follow Muon (Jordan et al., 2024) and approximate $\mathrm{Orth}(M_t)$ using Newton–Schulz iteration (Higham, 2008). This matrix-multiplication-only scheme enables efficient implementation on modern accelerators.

We also provide a simplified whitening-factor implementation that further reduces storage and computation; see Appendix B.

Remark 5.1. *In practice, exact gradient computation is often intractable, necessitating stochastic approximations. We therefore approximate the gradient matrix G in equation 3 by a mini-batch average in our algorithm implementation: $G_t = \frac{1}{B}\sum_{i=1}^{B} \nabla_W \ell(W_{t-1}; \xi_{t,i})$, where $\mathcal{B}_t = \{\xi_{t,1}, \ldots, \xi_{t,B}\}$ denotes a mini-batch of size B with samples drawn i.i.d. from the data distribution $\mathcal{D}$.*

5.2 Convergence Analysis

This subsection establishes convergence guarantees for Algorithm 1 in nonconvex optimization. Our analysis relies on the following standard assumptions.

Assumption 5.2 (Lower-Bounded Objective)**.** *The objective function $\mathcal{L}$ is lower-bounded with an inferior of $\mathcal{L}_* := \inf_W \mathcal{L}(W) > -\infty$.*

Assumption 5.3 (L-Smoothness)**.** *The funciton $\mathcal{L}(W)$ is Lipschitz smooth, i.e., there exist a constant L such that $\|\nabla\mathcal{L}(W_1) - \nabla\mathcal{L}(W_2)\|_F \le L\|W_1 - W_2\|_F$ for all $W_1, W_2 \in \mathbb{R}^{m\times n}$.*

Assumption 5.4 (Unbiased Stochastic Gradients)**.** *For all iteration t and any random sample ξ, the stochastic gradient is unbiased: $\mathbb{E}[\nabla\ell(W_{t-1};\xi)|\mathcal{F}_{t-1}] = \nabla\mathcal{L}(W_{t-1})$, where $\mathcal{F}_{t-1}$ is the σ-algebra generated by all random variables revealed up to iteration $t-1$.*

Assumption 5.5 (Bounded Variance)**.** *For all iteration t and any random sample ξ, there exists $\sigma^2 < \infty$ such that $\mathbb{E}\left[\|\nabla\ell(W_{t-1};\xi) - \nabla\mathcal{L}(W_{t-1})\|_F^2 \,\middle|\, \mathcal{F}_{t-1}\right] \le \sigma^2$.*

Assumption 5.6 (Bounded Gradient)**.** *The stochastic gradient is bounded at each iteration t, i.e., there is a constant G_* such that $\|G_t\|_* \le G_*$ holds for all t.*

Under these assumptions, we derive the following main convergence theorems. The detailed proofs of them are deferred to Appendix F.

Theorem 5.7 (Deterministic convergence theorem)**.** *Suppose the Assumptions 5.2–5.5 holds. Define $R := \mathcal{L}(W_0) - \mathcal{L}_\star$ with $\mathcal{L}_\star := \inf_W \mathcal{L}(W)$, and let $r := \min(m,n)$. Run Algorithm 1 for T iterations with stepsize $\eta = \frac{C}{\sqrt{T}}$ for some constant $C > 0$. Then the iterates $\{W_t\}_{t=0}^{T-1}$ satisfy:*

$$\frac{1}{T}\sum_{t=1}^{T} \|\nabla\mathcal{L}(W_{t-1})\|_* = \mathcal{O}\left(\frac{R + rL + G_*}{\sqrt{T}} + \frac{G_*}{T}\right).$$

Theorem 5.8 (Stochastic convergence theorem)**.** *Suppose Assumptions 5.2–5.5 hold. Then there exist a problem-dependent threshold $T_0 > 0$ and admissible parameter choices $\beta_T \in [0,1)$ and $\eta_T > 0$ such that, for all $T \ge T_0$, running Algorithm 1 for T iterations with $\beta = \beta_T$ and $\eta = \eta_T$ yields*

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\left[\|\nabla\mathcal{L}(W_{t-1})\|_*\right] = \mathcal{O}\left(\left(\frac{r^2 R\sigma^2}{B}\right)^{1/4} T^{-1/4} + \left(\frac{r^2 R^3 B}{\sigma^2}\right)^{1/4} T^{-3/4}\right).$$

Remark 5.9. *The detailed forms and proofs of Theorems 5.7 and 5.8 are provided in Appendix F. These two results establish that FISMO attains an $\mathcal{O}(T^{-1/2})$ convergence rate in the deterministic setting and an $\mathcal{O}(T^{-1/4})$ rate in the stochastic mini-batch setting, which is consistent with standard nonconvex first-order optimization theory (Shen et al., 2025; Li & Hong, 2025).*

5.3 Pre-asymptotic Analysis

We provide a local analysis showing that FISMO enjoys a pre-asymptotic advantage over isotropic update when Fisher-structured metric captures the dominant anisotropic curvature (Martens, 2020). In this subsection, we consider a simplified local quadratic model around a local minimizer W^*:

$$f(W) - f(W^\star) \approx \mathrm{vec}(W - W^\star)^\top H^\star \,\mathrm{vec}(W - W^\star)/2, \qquad H^\star \succ 0.$$

Let B_{pc} be a preconditioner of core update $W_{t+1} = W_t - \eta_t B_{pc}^{-1} G_t$. We have the following theorem:

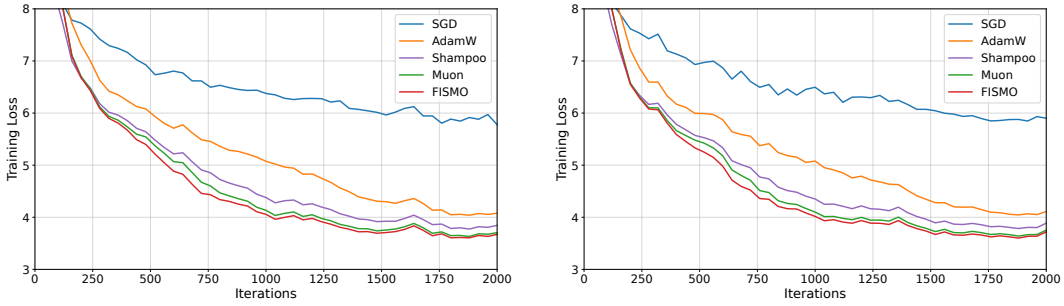
Theorem 5.10 (Local pre-asymptotic convergence)**.** *Assume the local quadratic model above and Assumptions 5.2–5.5 hold. Suppose B_{pc} is fixed in the local region. Suppose $b_{pc} := \lambda_{\min}(B_{pc}^{-1}H^\star) > \frac{1}{2}$, $\lambda_{\max}(B_{pc}^{-1}H^\star) \le a + 1$, and $\eta_t = (t + a + 1)^{-1}$, then after T iterations,*

$$\mathbb{E}[f(W_T)] - f(W^\star) = \mathcal{O}\big(T^{-1}\big) + \mathcal{O}\big(T^{-2b_{pc}}\big). \tag{6}$$

The two terms in equation 6 correspond to the noise-dependent error and the initialization-dependent error, respectively. For FISMO, the preconditioner is equivalently $B_{pc} = B_F = Q \otimes P$, whereas Muon's updates correspond to $B_{pc} = B_M = cI$. Since $Q \otimes P$ captures the anisotropic curvature structure, as characterized in Theorem 4.1, the effective curvature $B_F^{-1} H^\star$ can have a larger minimum eigenvalue than its isotropic counterpart. With the usual stability normalization, this yields $b_F > b_M$ in favorable anisotropic regimes, and hence $T^{-2b_F} = o(T^{-2b_M})$. Therefore, FISMO achieves a faster decay of the initialization-dependent term, explaining its pre-asymptotic advantage over isotropic updates. The detailed proof and the extension to orthogonalized updates are provided in Appendix G.

6 Experimental Results

To evaluate the performance of FISMO , we conduct experiments on two standard benchmarks: (i) the OpenWebText (Gokaslan & Cohen, 2019) corpus for language modeling, and (ii) the CIFAR-10 (Krizhevsky et al., 2009) dataset for image classification. For the language modeling task, we utilize the NanoGPT (Karpathy, 2022) framework to reproduce a 124M-parameter GPT-2 (Radford et al., 2019) model. For CIFAR-10, we employ the SimpleDLA (Yu et al., 2018) architecture. We compare FISMO with several established and state-of-the-art baseline optimizers, including widely adopted methods like SGD (Robbins & Monro, 1951) and AdamW (Loshchilov & Hutter, 2017), as well as advanced algorithms such as Shampoo (Gupta et al., 2018) and Muon (Jordan et al., 2024).



(a) Training loss (b) Validation loss

Figure 1: Training and validation loss versus training steps on the OpenWebText dataset (nanoGPT).

Figure 1 reports the training (1a) and validation (1b) loss versus training steps for the 124M-parameter GPT-2 model trained on OpenWebText under the NanoGPT framework, comparing FISMO with baselines SGD, AdamW, Shampoo and Muon. Across the whole training horizon, FISMO achieves the fastest loss reduction and the lowest final loss over the entire trajectory, surpassing even the strongest baseline Muon. The consistent ordering between training and validation suggests that the advantage of FISMO is not due to aggressive fitting but reflects more effective and stable optimization dynamics, aligning with FISMO's update mechanism that preconditions matrix gradients under a Fisher-inspired geometry while avoiding the overly isotropic spectrum imposed by strict orthogonalization.

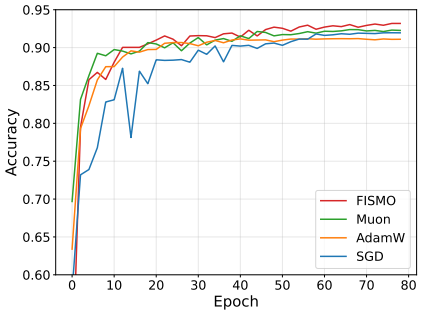


Figure 2: Validation accuracy versus the number of epochs on CIFAR-10.

Figure 2 illustrates the comparison results of FISMO optimizer and the baselines (SGD, AdamW, and Muon) on the CIFAR-10 image classification task. The results show that FISMO achieves the best final testing accuracy among the compared optimizers. Although Muon improves slightly faster in the very early stage, FISMO gradually catches up and surpasses it in the later phase of training, indicating that the Fisher-structured anisotropic update can provide more effective generalization-oriented progress. Compared with AdamW and SGD, FISMO also maintains a clear advantage in both initial training stages and final convergence. Detailed experimental results, including training loss, testing loss, training accuracy, and testing accuracy, are provided in Appendix C.

7 Further Discussion on FISMO: A Condition Number Perspective

To further understand the geometric efficacy of FISMO, we analyze the spectral properties of the update matrices, specifically focusing on the condition number. The condition number of a matrix (denoted as κ) is defined as the ratio of its greatest and lowest singular value, where a high condition indicates a highly anisotropic spectrum dominated by a few principal directions (Saad, 2003). Standard adaptive methods like Adam rely on element-wise scaling, often resulting in ill-conditioned updates with unconstrained global spectra. In contrast, Muon enforces strict orthogonality to achieve an isotropic update with an approximate condition number of 1, treating all spectral directions equally (Lau et al., 2025).

As illustrated in Figure 3, we track the average condition number of the update matrices throughout training, comparing FISMO against Adam and Muon with Newton-Schulz (NS) iteration numbers of 5 (NS=5) and 7 (NS=7). Note that increasing NS iterations in Muon will enforce stricter orthogonalization, driving the update towards the ideal isotropic limit ($\kappa \to 1$). Generally, we observe the following relation of condition number among them:

$$\kappa_{\text{Adam}} \gg \kappa_{\text{FISMO}} > \kappa_{\text{Muon Practical}} > \kappa_{\text{Muon Ideal}} = 1.$$

As observed, Adam exhibits pathologically high condition numbers ($\kappa > 10^8$). In contrast, increasing Muon's NS iterations progressively flattens the spectrum, with NS=7 approaching the ideal isotropic limit. Crucially, FISMO establishes a stable spectral profile in the range of 10^2–10^3. This represents a significant reduction in ill-conditioning compared to Adam (by several orders of magnitude) while maintaining a condition number generally higher than the orthogonalized updates from Muon.

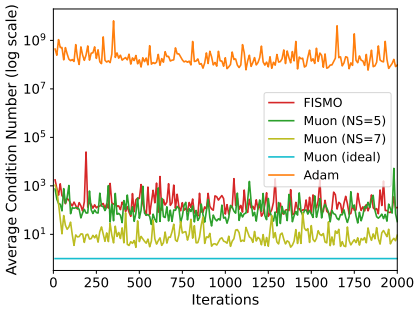


Figure 3: Average condition number of the update matrices over training iterations. FISMO is compared with Adam and Muon with 5 and 7 NS iterations, alongside the ideal isotropic Muon with strictly orthogonalized updates.

We characterize this regime as an *optimal conditioning trade-off*: FISMO effectively mitigates severe spectral skewness to ensure stability, yet retains a degree of anisotropy necessary to preserve relative curvature information that strict orthogonalization would eliminate.

These spectral properties perfectly align with the recently proposed *Isotropic Curvature Model* (Su, 2025), which challenges the optimality of strict gradient orthogonalization. Su (2025) demonstrates that under realistic super-quadratic curvature growth, the theoretically optimal update requires (partial) spectrum homogenization: the singular values should be compressed to be closer in magnitude to improve conditioning, but *not necessarily forced to be completely equal*. Specifically, the optimal spectral transformation is rarely perfectly uniform ($\kappa = 1$), because practical loss landscapes typically do not exhibit the extreme asymptotic behavior required to justify strict orthogonalization. This implies that there may exist update mechanisms superior to Muon that operate with a condition number theoretically larger than 1. By maintaining a condition number that is low but bounded away from unity, FISMO effectively implements partial spectrum homogenization, navigating the middle ground between the chaotic anisotropy of element-wise methods and the rigid isotropy of fully orthogonalized updates.

8 Conclusions

We proposed **FISMO**, a Fisher-structured momentum-orthogonalized optimizer for deep learning with matrix-parameter updates. By casting update as a trust-region problem under a Kronecker-factored Fisher metric, FISMO yields a structured preconditioning update rule that remains tractable while preserving informative spectral variation beyond strict isotropy. Our theoretical analysis establishes an $\mathcal{O}(T^{-1/4})$ stochastic nonconvex convergence rate and further explains FISMO's pre-asymptotic advantage over isotropic updates. Experiments validate the efficiency and stability of FISMO across representative benchmarks. A limitation of this work is that the empirical evaluation is constrained in scale. Future work will extend FISMO to larger models and broader training regimes.

AI USE STATEMENT

In this work, we used generative AI tools to assist with polishing the writing of the manuscript, including improving grammar, clarity, conciseness, and organization. We have not used generative AI tools for research ideation, methodology development, mathematical derivations, experimental design, data generation, or result interpretation, and the remaining required-disclosure tasks are not applicable to this work. All AI-assisted revisions were reviewed by the authors to ensure technical accuracy and consistency with the intended meaning. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

ETHICS STATEMENT

This work studies general-purpose optimization methods for training neural networks and does not involve human-subject experiments or the collection of new personal or sensitive data. Our empirical evaluation uses established public benchmark datasets and does not introduce or release new datasets. As a general-purpose optimization method, FISMO may be applied to a broad range of downstream models and applications; the ethical considerations of such uses depend primarily on the corresponding data, models, and application domains. We encourage users to follow applicable requirements regarding data privacy, licensing, safety, and responsible deployment.

REPRODUCIBILITY STATEMENT

We provide the algorithmic and theoretical details necessary to reproduce the main results of this work. The complete FISMO update is specified in Section 5 and Algorithm 1, with a simplified implementation described in Appendix B. The experimental settings and evaluations are described in Section 6, with additional results reported in Appendix C. Assumptions and complete derivations underlying the theoretical results are provided in Appendices D–G, including the derivation of the Fisher-structured update, the convergence analysis, and the pre-asymptotic analysis. We plan to release the implementation and experiment configurations upon publication.

REFERENCES

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Kang An, Yuxing Liu, Rui Pan, Yi Ren, Shiqian Ma, Donald Goldfarb, and Tong Zhang. Asgo: Adaptive structured gradient optimization. *arXiv preprint arXiv:2503.20762*, 2025.

Jeremy Bernstein. Deriving muon, 2025. URL `https://jeremybernste.in/writing/deriving-muon`.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. *arXiv preprint arXiv:2409.20325*, 2024.

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *arXiv preprint arXiv:2506.15054*, 2025.

Andrzej Cichocki, Sergio Cruces, and Shun-ichi Amari. Log-determinant divergences revisited: Alpha-beta and gamma log-det divergences. *Entropy*, 17(5):2988–3034, 2015.

Sueli IR Costa, Sandra A Santos, and Joao E Strapasson. Fisher information distance: A geometrical reading. *Discrete Applied Mathematics*, 197:59–69, 2015.

Thomas George, César Laurent, Xavier Bouthillier, Nicolas Ballas, and Pascal Vincent. Fast approximate natural gradient descent in a kronecker factored eigenbasis. *Advances in neural information processing systems*, 31, 2018.

Aaron Gokaslan and Vanya Cohen. Openwebtext corpus. `http://Skylion007.github.io/OpenWebTextCorpus`, 2019.

Vineet Gupta, Tomer Koren, and Yoram Singer. Shampoo: Preconditioned stochastic tensor optimization. In *International Conference on Machine Learning*, pp. 1842–1850. PMLR, 2018.

Nicholas J Higham. *Functions of matrices: theory and computation*. SIAM, 2008.

Roger A Horn and Charles R Johnson. *Topics in matrix analysis*. Cambridge university press, 1994.

Roger A Horn and Charles R Johnson. *Matrix analysis*. Cambridge university press, 2012.

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

Andrej Karpathy. nanoGPT. `https://github.com/karpathy/nanoGPT`, 2022.

Ahmed Khaled, Kaan Ozkara, Tao Yu, Mingyi Hong, and Youngsuk Park. Muonbp: Faster muon via block-periodic orthogonalization. *arXiv preprint arXiv:2510.16981*, 2025.

Diederik P Kingma. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Stefan Klein, Josien PW Pluim, Marius Staring, and Max A Viergever. Adaptive stochastic gradient descent optimisation for image registration. *International journal of computer vision*, 81(3): 227–239, 2009.

Alex Krizhevsky, Geoffrey Hinton, et al. Learning multiple layers of features from tiny images. 2009.

Guanghui Lan. The complexity of large-scale convex programming under a linear optimization oracle. *arXiv preprint arXiv:1309.5550*, 2013.

Tim Tsz-Kit Lau, Qi Long, and Weijie Su. Polargrad: A class of matrix-gradient optimizers from a unifying preconditioning perspective. *arXiv preprint arXiv:2505.21799*, 2025.

Jiaxiang Li and Mingyi Hong. A note on the convergence of muon. *arXiv preprint arXiv:2502.02900*, 2025.

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for llm training. *arXiv preprint arXiv:2502.16982*, 2025.

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

Alexander Ly, Maarten Marsman, Josine Verhagen, Raoul PPP Grasman, and Eric-Jan Wagenmakers. A tutorial on fisher information. *Journal of Mathematical Psychology*, 80:40–55, 2017.

James Martens. New insights and perspectives on the natural gradient method. *Journal of Machine Learning Research*, 21(146):1–76, 2020.

James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pp. 2408–2417. PMLR, 2015.

Yann Ollivier. True asymptotic natural gradient optimization. *arXiv preprint arXiv:1712.08449*, 2017.

Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

Herbert Robbins and Sutton Monro. A stochastic approximation method. *The annals of mathematical statistics*, pp. 400–407, 1951.

Nicolas Roux, Pierre-Antoine Manzagol, and Yoshua Bengio. Topmoumoute online natural gradient algorithm. *Advances in neural information processing systems*, 20, 2007.

Yousef Saad. *Iterative methods for sparse linear systems*. SIAM, 2003.

Noam Shazeer and Mitchell Stern. Adafactor: Adaptive learning rates with sublinear memory cost. In *International Conference on Machine Learning*, pp. 4596–4604. PMLR, 2018.

Wei Shen, Ruichuan Huang, Minhui Huang, Cong Shen, and Jiawei Zhang. On the convergence analysis of muon. *arXiv preprint arXiv:2505.23737*, 2025.

Rajesh Shrestha. Natural gradient methods: Perspectives, efficient-scalable approximations, and analysis. *arXiv preprint arXiv:2303.05473*, 2023.

Weijie Su. Isotropic curvature model for understanding deep learning optimization: Is gradient orthogonalization optimal? *arXiv preprint arXiv:2511.00674*, 2025.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

Amund Tveit, Bjørn Remseth, and Arve Skogvold. Muon optimizer accelerates grokking. *arXiv preprint arXiv:2504.16041*, 2025.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Mujin Kwun, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. Soap: Improving and stabilizing shampoo using adam. *arXiv preprint arXiv:2409.11321*, 2024.

Shuche Wang, Fengzhuo Zhang, Jiaxiang Li, Cunxiao Du, Chao Du, Tianyu Pang, Zhuoran Yang, Mingyi Hong, and Vincent YF Tan. Muon outperforms adam in tail-end associative memory learning. *arXiv preprint arXiv:2509.26030*, 2025.

Fisher Yu, Dequan Wang, Evan Shelhamer, and Trevor Darrell. Deep layer aggregation. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2403–2412, 2018.

Yushun Zhang, Congliang Chen, Ziniu Li, Tian Ding, Chenwei Wu, Diederik P Kingma, Yinyu Ye, Zhi-Quan Luo, and Ruoyu Sun. Adam-mini: Use fewer learning rates to gain more. *arXiv preprint arXiv:2406.16793*, 2024.

CONTENTS

1 Introduction **1**

2 Related Work **2**

3 Preliminaries **3**

4 Theoretical Framework **4**

4.1 Kronecker-Factored Approximation of FIM 4

4.2 The Best Kronecker Approximation 4

4.3 The Best Update 5

5 FISMO Algorithm **5**

5.1 Algorithm Design 5

5.2 Convergence Analysis 7

5.3 Pre-asymptotic Analysis 7

6 Experimental Results **8**

7 Further Discussion on FISMO: A Condition Number Perspective **9**

8 Conclusions **9**

A Notations **14**

B Simplified FISMO Algorithm **14**

C Additional Experimental Results **16**

D Preliminary Theorems on Matrix Analysis **16**

E Derivation of the Fisher-Structured Update **17**

E.1 Proof of Theorem 4.1 17

E.2 Proof of Theorem 4.2 17

F Convergence Analysis **18**

F.1 Auxiliary Lemmas for Convergence Analysis 18

F.2 Proof of the deterministic convergence theorem 26

F.3 Proof of the stochastic convergence theorem 30

G Pre-asymptotic Analysis of FISMO **36**

G.1 Without Orthogonalization 37

G.2 With Orthogonalization 41

A NOTATIONS

Table 1: Summary of notation used in the paper.

Notation	Meaning	Remarks / Conditions
x, g, v	Vectors	Lowercase letters are used for vectors.
A, B, G, W	Matrices	Uppercase letters are used for matrices.
I_d	Identity matrix	$d \times d$ identity matrix; the dimension is omitted when clear.
$\mathrm{tr}(A)$	Trace	Defined for square matrices A.
$\langle A, B\rangle_F$	Frobenius inner product	$\langle A, B\rangle_F := \mathrm{tr}(A^\top B)$ for $A, B \in \mathbb{R}^{m\times n}$.
$\|A\|_F$	Frobenius norm	$\|A\|_F := \sqrt{\langle A, A\rangle_F}$.
$\|A\|_2$, $\|A\|_{\mathrm{op}}$	Spectral/operator norm	Largest singular value of A.
$\|A\|_*$	Nuclear norm	Sum of singular values of A.
$\mathrm{vec}(A)$	Vectorization	Stacks the columns of A into a vector.
$A \otimes B$	Kronecker product	Defined for arbitrary matrices A and B with compatible dimensions.
$\mathrm{sym}(X)$	Symmetric part	$\mathrm{sym}(X) := \frac{1}{2}(X + X^\top)$ for square matrices X.
$\mathrm{Orth}(M)$	Orthogonal polar factor	For an SVD $M = U\Sigma V^\top$, $\mathrm{Orth}(M) := UV^\top$.
$A \succ B$	Positive definite order	For symmetric A, B, $A - B$ is positive definite.
$A \succeq B$	Positive semidefinite order	For symmetric A, B, $A-B$ is positive semidefinite.
$\mathbb{S}^d_{++}$	Positive definite cone	Set of $d \times d$ symmetric positive definite matrices.
$\mathbb{S}^d_{+}$	Positive semidefinite cone	Set of $d \times d$ symmetric positive semidefinite matrices.
$\lambda_{\max}(A)$	Largest eigenvalue	Used for symmetric matrices unless otherwise specified.
$\lambda_{\min}(A)$	Smallest eigenvalue	Used for symmetric matrices unless otherwise specified.
$\sigma_i(A)$	i-th singular value	Singular values are ordered nonincreasingly.
$\mathrm{rank}(A)$	Rank	Number of nonzero singular values of A.
$\mathbb{E}[\cdot]$	Expectation	Taken over the randomness specified by context.
$\mathcal{O}(\cdot)$	Big-O notation	Hides constants independent of the main asymptotic variable.

B SIMPLIFIED FISMO ALGORITHM

In this section, we present a simplified implementation form of FISMO by directly parameterizing the whitening factors, instead of explicitly maintaining the preconditioners themselves. The modified algorithm can be found in Algorithm 2.

Original formulation. In the original FISMO update (Algorithm 1), the left and right preconditioners are denoted by $P_t \in \mathbb{S}^m_{++}$ and $Q_t \in \mathbb{S}^n_{++}$, and the whitened gradient is $\widetilde{G}_t = P_t^{-1/2} G_t Q_t^{-1/2}$. The momentum is accumulated in the whitened space, $M_t = \beta M_{t-1} + (1-\beta)\widetilde{G}_t$, and the update direction is given by $\Delta W_t = P_t^{-1/2}\ \mathrm{Orth}(M_t)\, Q_t^{-1/2}$. Thus, the quantities that are actually used in whitening and in the final update are not P_t and Q_t themselves, but rather their inverse square roots $P_t^{-1/2}$ and $Q_t^{-1/2}$.

Whitening-factor reparameterization. Motivated by this observation, we introduce the whitening factors

$$X_t := P_t^{-1/2} \in \mathbb{S}^m_{++}, \qquad Y_t := Q_t^{-1/2} \in \mathbb{S}^n_{++}.$$

Equivalently, $P_t = X_t^{-2}, Q_t = Y_t^{-2}, P_t^{-1} = X_t^2, Q_t^{-1} = Y_t^2$. Under this change of variables, the whitened gradient becomes $\widetilde{G}_t = X_t G_t Y_t$, and the FISMO update can be written as

$$\Delta W_t = X_t\ \mathrm{Orth}(M_t)\, Y_t.$$

Algorithm 2 FISMO in whitening-factor form

Input: $W_0 \in \mathbb{R}^{m\times n}$; learning rate η; momentum β; EMA γ; damping factor μ; $X_0 \leftarrow I_m$; $Y_0 \leftarrow I_n$; $M_0 = 0$
for $t = 1, \cdots, T$ **do**
  Compute gradient $G_t = \frac{1}{B}\sum_{i=1}^{B} \nabla_W \ell(W_{t-1}; \xi_{t,i})$
  Update left whitening factor
    $\widehat{X}_t \leftarrow \sqrt{n}\,\left(G_t Y_{t-1}^2 G_t^\top + \mu I_m\right)^{-1/2}$
    $X_t \leftarrow \gamma X_{t-1} + (1-\gamma)\widehat{X}_t$
  Update right whitening factor
    $\widehat{Y}_t \leftarrow \sqrt{m}\,\left(G_t^\top X_t^2 G_t + \mu I_n\right)^{-1/2}$
    $Y_t \leftarrow \gamma Y_{t-1} + (1-\gamma)\widehat{Y}_t$
  Whiten gradient $\widetilde{G}_t \leftarrow X_t\, G_t\, Y_t$
  $M_t \leftarrow \beta M_{t-1} + (1-\beta)\widetilde{G}_t$
  $\Delta W_t \leftarrow X_t\, \mathrm{Orth}(M_t)\, Y_t$
  Update $W_t \leftarrow W_{t-1} - \eta\, \Delta W_t$
end for

Therefore, the full update takes the form

$$W_t = W_{t-1} - \eta\, X_t\ \mathrm{Orth}(M_t)\, Y_t.$$

Deriving the whitening-factor targets. Recall that, in the preconditioner form, the optimal Kronecker factors are motivated by the coupled relations

$$P^*(Q) = \frac{1}{n}\mathbb{E}[GQ^{-1}G^\top] \quad \text{and} \quad Q^*(P) = \frac{1}{m}\mathbb{E}[G^\top P^{-1} G],$$

up to the additional isotropic damping used in implementation. Substituting $Q^{-1} = Y^2$ and $P^{-1} = X^2$, we obtain the corresponding relations

$$P^*(Y) = \frac{1}{n}\mathbb{E}[GY^2G^\top], \qquad Q^*(X) = \frac{1}{m}\mathbb{E}[G^\top X^2 G].$$

Taking inverse square roots on both sides yields the corresponding whitening factors:

$$X^*(Y) = \sqrt{n}\left(\mathbb{E}[GY^2G^\top]\right)^{-1/2}, \qquad Y^*(X) = \sqrt{m}\left(\mathbb{E}[G^\top X^2 G]\right)^{-1/2}.$$

In practice, we replace the population expectations by the current mini-batch gradient and add a simple isotropic damping term μI for numerical stability, leading to the targets

$$\widehat{X}_t = \sqrt{n}\left(G_t Y_{t-1}^2 G_t^\top + \mu I_m\right)^{-1/2},$$

$$\widehat{Y}_t = \sqrt{m}\left(G_t^\top X_t^2 G_t + \mu I_n\right)^{-1/2}.$$

We then smooth these targets directly in whitening-factor space via EMA:

$$X_t = \gamma X_{t-1} + (1-\gamma)\widehat{X}_t, \qquad Y_t = \gamma Y_{t-1} + (1-\gamma)\widehat{Y}_t.$$

This gives rise to Algorithm 2.

Why this form is simpler. First, it directly parameterizes the quantities that are actually used in both whitening and the final update. In the original form, one maintains P_t and Q_t, but the algorithm ultimately requires $P_t^{-1/2}$ and $Q_t^{-1/2}$. The new parameterization removes this mismatch by treating $X_t = P_t^{-1/2}$ and $Y_t = Q_t^{-1/2}$ as the primary state variables.

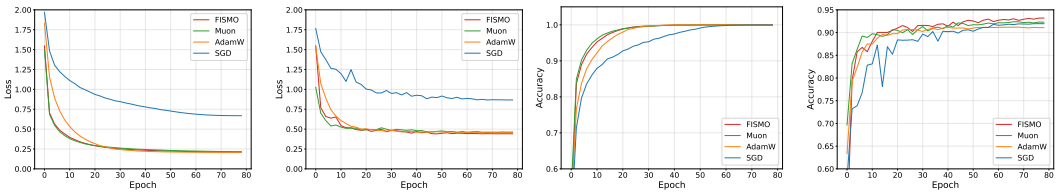
Second, it reduces optimizer-state storage. A direct implementation of the original form often needs to keep both the preconditioners and their inverse square roots (or inverse-related auxiliary quantities).

In contrast, the whitening-factor form stores only X_t and Y_t, which are already sufficient for both preconditioning and update construction.

Third, the damping term can be simplified. In the original derivation, one may introduce trace-dependent isotropic regularization terms to match the exact closed-form stationary equations. In the whitening-factor implementation, however, the role of the damping is mainly to ensure a positive spectral floor. Since the factors are subsequently smoothed by EMA and only used through their whitening effect, a constant damping term μI is often sufficient in practice.

C Additional Experimental Results

This section provides additional experimental results that complement the main text as in Figure 2. In particular, Figure 4 reports the complete training and validation trajectories on CIFAR-10, including both loss and accuracy curves. These results provide a more detailed view of the optimization dynamics and generalization behavior of different optimizers throughout training.



(a) Training loss (b) Validation loss (c) Training accuracy (d) Validation accuracy

Figure 4: Traning/validation loss/accuracy versus the number of steps on the CIFAR-10 dataset.

D Preliminary Theorems on Matrix Analysis

Lemma D.1 (Generalized von Neumann Trace Inequality (rectangular form)). *Let $A, B \in \mathbb{R}^{m\times n}$ be two given matrices, and let $p = \min\{m, n\}$. Suppose their singular values are $\sigma_1(A) \geq \sigma_2(A) \geq \cdots \geq \sigma_p(A)$ and $\sigma_1(B) \geq \sigma_2(B) \geq \cdots \geq \sigma_p(B)$. Then we have the following inequality*

$$|\operatorname{tr}(A^\top B)| \leq \sum_{i=1}^{p} \sigma_i(A)\sigma_i(B). \tag{7}$$

The above lemma can be found in the book *Topics in Matrix analysis* Horn & Johnson (1994).

Theorem D.2 (Hölder inequality for Schatten norm). *Let $A, B \in \mathbb{R}^{m\times n}$ and let $p, q \in [1, \infty]$ satisfy $\frac{1}{p} + \frac{1}{q} = 1$ (with the convention $\frac{1}{\infty} = 0$). Then*

$$\left|\langle A, B\rangle_F\right| = \left|\operatorname{tr}(A^\top B)\right| \;\leq\; \|A\|_{S_p}\, \|B\|_{S_q}, \tag{8}$$

where $\|\cdot\|_{S_p}$ denotes the Schatten-p norm. In particular, taking $(p, q) = (1, \infty)$ yields

$$\left|\operatorname{tr}(A^\top B)\right| \;\leq\; \|A\|_*\, \|B\|_{\mathrm{op}}, \tag{9}$$

where $\|A\|_* = \|A\|_{S_1}$ is the nuclear norm and $\|B\|_{\mathrm{op}} = \|B\|_{S_\infty}$ is the spectral (operator) norm.

Theorem D.3 (Generalized Hölder inequality for Schatten norm). *Let $A \in \mathbb{R}^{m\times m}$, $X \in \mathbb{R}^{m\times n}$, and $B \in \mathbb{R}^{n\times n}$. Let $p \in [1, \infty]$ and let $p_1, p_2, p_3 \in [1, \infty]$ satisfy*

$$\frac{1}{p} = \frac{1}{p_1} + \frac{1}{p_2} + \frac{1}{p_3}.$$

Then the following inequalities hold:

$$\|AXB\|_{S_p} \leq \|A\|_{S_{P_1}}\, \|X\|_{S_{p_2}}\, \|B\|_{S_{p_3}}.$$

In particular, taking $p = p_2 = 1$ and $p_1 = p_3 = \infty$ gives the commonly used nuclear-norm bound

$$\|AXB\|_* \leq \|A\|_{\mathrm{op}}\, \|X\|_*\, \|B\|_{\mathrm{op}}. \tag{10}$$

Lemma D.4 (Equivalence of Frobenius and Nuclear norm). *Let $A \in \mathbb{R}^{m\times n}$, suppose r is the rank of matrix A, then we have the following inequality:*

$$\|A\|_F \leq \|A\|_* \leq \sqrt{r}\|A\|_F. \tag{11}$$

E Derivation of the Fisher-Structured Update

E.1 Proof of Theorem 4.1

Proof. We prove the claim for P, then the proof of Q is analogous.

Fix $Q \in \mathbb{S}^n_{++}$. For simpler notation, we denote $A(Q) := \mathbb{E}[GQ^{-1}G^\top]$. Then the components of $\mathcal{J}(P, Q)$ depending on P is

$$J(P, Q) \;=\; \operatorname{tr}\left(P^{-1}A(Q)\right) + \mu \operatorname{tr}(Q^{-1}) \operatorname{tr}(P^{-1}) \;+\; n \log\det P.$$

By taking the gradient of $\mathcal{J}$ w.r.t. P, we have

$$\frac{\partial}{\partial P} J(P, Q) = -P^{-1}(A(Q) + \mu \operatorname{tr}(Q^{-1}) I_m) P^{-1} + nP^{-1}.$$

Setting $\frac{\partial}{\partial P} J(P, Q) = 0$, we have

$$(A(Q) + \mu \operatorname{tr}(Q^{-1}) I_m) P^{-1} = nI_m.$$

Therefore,

$$P = \frac{1}{n}(A(Q) + \mu \operatorname{tr}(Q^{-1}) I_m) = \frac{1}{n}\mathbb{E}[GQ^{-1}G^T] + \frac{\mu \operatorname{tr}(Q^{-1})}{n} I_m,$$

which is exactly the claim in Theorem 4.1.

It remains to prove the uniqueness. Notice that the function $P \mapsto \operatorname{tr}(P^{-1}S) + n \log\det P$ with $S \succ 0$ is strictly convex on $\mathbb{S}^m_{++}$. Since $\mu \operatorname{tr}(Q^{-1}) > 0$ when $\mu > 0$, here $S = \operatorname{tr}\left(P^{-1}A(Q)\right) + \mu \operatorname{tr}(Q^{-1})$ is positive definite. And hence $J(P, Q)$ is strictly convex w.r.t. P. Therefore, the stationary point for J is exactly the unique minimizer.

This complete the proof. □

E.2 Proof of Theorem 4.2

Proof of Theorem 4.2. The main idea of the proof is to transform the constraints into a standard spectral-norm ball. Let $\Phi := P^{1/2} \Delta W Q^{1/2}$ be the preconditioned update. Since P and Q are positive definite, the mapping from ΔW to Φ is bijective, and therefore $\Delta W = P^{-1/2} \Phi Q^{-1/2}$. Then the constraint of equation 3 becomes $\|\Phi\|_2 \le \eta$.

By the cyclic property of trace, the objective of problem equation 3 becomes

$$\begin{aligned} \langle G, \Delta W \rangle_F &= \operatorname{tr}\left(G^\top P^{-1/2} \Phi Q^{-1/2}\right) = \operatorname{tr}\left(Q^{-1/2} G^\top P^{-1/2} \Phi\right) \\ &= \operatorname{tr}\left(\left(P^{-1/2} G Q^{-1/2}\right)^\top \Phi\right) = \langle \widetilde{G}, \Phi \rangle_F. \end{aligned} \tag{12}$$

Therefore, the problem equation 3 is equivalent to

$$\min_{\Phi \in \mathbb{R}^{m \times n}} \langle \widetilde{G}, \Phi \rangle_F \quad \text{s.t.} \quad \|\Phi\|_2 \le \eta. \tag{13}$$

Lower Bound. Let $p = \min\{m, n\}$. Apply lemma D.1 with $A = \widetilde{G}$ and $B = \Phi$, we have

$$\langle \widetilde{G}, \Phi \rangle_F = \operatorname{tr}(\widetilde{G}^\top \Phi) \ge -\sum_{i=1}^{p} \sigma_i(\widetilde{G}) \sigma_i(\Phi),$$

where the singular values σ_i of both $\widetilde{G}$ and Φ are arranged in descending order as in lemma D.1. Notice that the spectral norm (i.e. $\|\cdot\|_2$) of a matrix equals to its largest singular value, therefore,

$$\sigma_i(\Phi) \le \sigma_1(\Phi) = \|\Phi\|_2 \le \eta, \quad \forall i = 1, \cdots, p.$$

Therefore,

$$\langle \widetilde{G}, \Phi \rangle_F \ge -\sum_{i=1}^{p} \sigma_i(\widetilde{G}) \cdot \eta = -\eta \sum_{i=1}^{p} \sigma_i(\widetilde{G}) = -\eta \|\widetilde{G}\|_*,$$

which means that $-\eta\|\widetilde{G}\|_*$ is a lower bound on the minimum of the problem 13.

Feasibility and Achievability. We consider the SVD $\widetilde{G} = U\Sigma V^\top$, where $\Sigma = \mathrm{diag}(\sigma_1(\widetilde{G}), \cdots, \sigma_r(\widetilde{G}))$ with $r = \mathrm{rank}(\widetilde{G}) \le p$. Define

$$\Phi^* := -\eta UV^\top. \tag{14}$$

Since the matrix $UV^\top$ has singular values equal to 1, so $\|UV^\top\|_2 = 1$. So $\|\Phi^*\|_2 = \eta\|UV^\top\|_2 = \eta$, and thus Φ^* is feasible.

Now we consider the objective value of problem 13 given $\Phi = \Phi^*$:

$$\begin{aligned}\langle \widetilde{G}, \Phi^\star\rangle_F &= -\eta\langle U\Sigma V^\top, UV^\top\rangle_F = -\eta\,\mathrm{tr}((U\Sigma V^\top)^\top UV^\top)\\ &= -\eta\,\mathrm{tr}(V\Sigma U^\top UV^\top) = -\eta\,\mathrm{tr}(V\Sigma V^\top)\\ &= -\eta\sum_{i=1}^{r}\sigma_i(\widetilde{G}) = -\eta\|\widetilde{G}\|_*.\end{aligned}$$

Therefore, $\|\Phi^*\| = UV^\top$ attains the global lower bound and is optimal for problem 13. Thus the solution for problem 3 is:

$$\Delta W^\star = P^{-1/2}\Phi^\star Q^{-1/2} = -\eta P^{-1/2}UV^\top Q^{-1/2} = -\eta P^{-1/2}\,\mathrm{Orth}(P^{-1/2}GQ^{-1/2})Q^{-1/2}.$$

□

F Convergence Analysis

F.1 Auxiliary Lemmas for Convergence Analysis

Lemma F.1 (Positive definiteness of P and Q)**.** *For each t, the P_t and Q_t derived from Algorithm 1 are positive definite.*

Proof of Lemma F.1. We prove the lemma by mathematical induction. By initialization, clearly $P_0 = I_m \succ 0$ and $Q_0 = I_n \succ 0$.

Assume $P_{t-1} \succ 0$ and $Q_{t-1} \succ 0$ for some $t \ge 1$. We need to show that $P_t \succ 0$ and $Q_t \succ 0$. By the algorithm we have

$$L_t \;=\; \frac{1}{n}G_tQ_{t-1}^{-1}G_t^\top \;+\; \mu\frac{\mathrm{tr}(P_{t-1})}{m}I_m.$$

Since $Q_{t-1} \succ 0$, we have $Q_{t-1}^{-1} \succ 0$. Hence $\frac{1}{n}G_tQ_{t-1}^{-1}G_t^\top \succeq 0$.

Also, $P_{t-1} \succ 0$ implies $\mathrm{tr}(P_{t-1}) > 0$, and with $\mu > 0$ the term $\mu\frac{\mathrm{tr}(P_{t-1})}{m}I_m$ is strictly positive definite. Therefore, for any $x \neq 0$,

$$x^\top L_t x = \frac{1}{n}(G_t^\top x)^\top Q_{t-1}^{-1}(G_t^\top x) + \mu\frac{\mathrm{tr}(P_{t-1})}{m}\|x\|_2^2 \ge \mu\frac{\mathrm{tr}(P_{t-1})}{m}\|x\|_2^2 > 0,$$

which shows $L_t \succ 0$.

Since $P_{t-1} \succ 0$ and $L_t \succ 0$ and $\gamma \in [0,1)$, therefore, $\widetilde{P}_t \;=\; \gamma P_{t-1} + (1-\gamma)L_t$ is positive definite.

Finally, Algorithm 1 normalizes $P_t \;=\; \mathrm{sym}\Big(\frac{m}{\mathrm{tr}(\widetilde{P}_t)}\,\widetilde{P}_t\Big)$. Because $\widetilde{P}_t$ is symmetric positive definite and $\frac{m}{\mathrm{tr}(\widetilde{P}_t)} > 0$, the scaled matrix $\frac{m}{\mathrm{tr}(\widetilde{P}_t)}\widetilde{P}_t$ is symmetric positive definite. Moreover, $\mathrm{sym}(\cdot)$ preserves symmetry and does not affect positive definiteness here. Thus P_t is positive definite.

Similarly, we can show that Q_t is positive definite as well, and by induction, the positive-definiteness of P and Q extend to all iteration t's.

□

Lemma F.2 (Upper bound of $K_{PQ}(t)$)**.** *In Algorithm 1, suppose there exists $G > 0$ such that for all t, $\|G_t\|_F \le G$. Moreover, assume $\mu > G^2/(mn)$. Then there exist constant $\overline{r}, \overline{s}$ such that for all t,*

$$\|P_t^{-1}\|_2 \le \overline{r}, \quad \|Q_t^{-1}\|_2 \le \overline{s},$$

and thus

$$K_{PQ}(t) = \left\|P_t^{-1/2}\right\|_2 \left\|Q_t^{-1/2}\right\|_2 \le \sqrt{\overline{r}\,\overline{s}} =: \overline{K},$$

which means $K_{PQ}(t)$ is uniformly upper-bounded.

Proof of Lemma F.2. According to the update rules in Algorithm 1,

$$P_t = \mathrm{sym}\left(\frac{m}{\mathrm{tr}(\widetilde{P}_t)}\widetilde{P}_t\right), \qquad Q_t = \mathrm{sym}\left(\frac{n}{\mathrm{tr}(\widetilde{Q}_t)}\widetilde{Q}_t\right).$$

Since $\mathrm{sym}(\cdot)$ preserves trace, we have, for all $t \ge 0$,

$$\mathrm{tr}(P_t) = \mathrm{tr}\left(\frac{m}{\mathrm{tr}(\widetilde{P}_t)}\widetilde{P}_t\right) = m, \qquad \mathrm{tr}(Q_t) = \mathrm{tr}\left(\frac{m}{\mathrm{tr}(\widetilde{Q}_t)}\widetilde{Q}_t\right) = n.$$

Hence

$$\mu\frac{\mathrm{tr}(P_{t-1})}{m} I_m = \mu I_m, \qquad \mu\frac{\mathrm{tr}(Q_{t-1})}{n} I_n = \mu I_n.$$

Moreover, by Lemma F.1, $Q_{t-1}^{-1} \succ 0$, we have $\frac{1}{n} G_t Q_{t-1}^{-1} G_t^\top \succeq 0$, hence

$$L_t = \frac{1}{n} G_t Q_{t-1}^{-1} G_t^\top + \mu I_m \succeq \mu I_m. \tag{15}$$

Similarly, we have

$$R_t = \frac{1}{m} G_t^\top P_t^{-1} G_t + \mu I_n \succeq \mu I_n. \tag{16}$$

Denote $\lambda_{\min}(A)$ as the smallest eigenvalue of matrix A. From $\widetilde{P}_t = \gamma P_{t-1} + (1-\gamma)L_t$ and equation 15, by Weyl's eigenvalue inequality,

$$\lambda_{\min}(\widetilde{P}_t) \ge \gamma\,\lambda_{\min}(P_{t-1}) + (1-\gamma)\lambda_{\min}(L_t) \ge (1-\gamma)\mu. \tag{17}$$

Since $P_t = \frac{m}{\mathrm{tr}(\widetilde{P}_t)}\widetilde{P}_t$, we have

$$P_t^{-1} = \frac{\mathrm{tr}(\widetilde{P}_t)}{m}\,\widetilde{P}_t^{-1}, \quad \text{hence} \quad \|P_t^{-1}\|_2 = \frac{\mathrm{tr}(\widetilde{P}_t)}{m} \cdot \frac{1}{\lambda_{\min}(\widetilde{P}_t)}.$$

It remains to upper bound $\mathrm{tr}(\widetilde{P}_t)$. By equation F.1,

$$\mathrm{tr}(\widetilde{P}_t) = \gamma\,\mathrm{tr}(P_{t-1}) + (1-\gamma)\mathrm{tr}(L_t) = \gamma m + (1-\gamma)\mathrm{tr}(L_t). \tag{18}$$

We also have,

$$\mathrm{tr}(L_t) = \frac{1}{n}\mathrm{tr}(G_t Q_{t-1}^{-1} G_t^\top) + \mu m = \frac{1}{n}\mathrm{tr}(Q_{t-1}^{-1} G_t^\top G_t) + \mu m.$$

By von Neumann trace inequality, for positive definite matrix $A, B \in \mathbb{S}^n$, we have

$$\mathrm{tr}(AB) \le \sum_{i=1}^n \lambda_i(A)\lambda_i(B) \le \sum_{i=1}^n \lambda_{\max}(A)\lambda_i(B) = \lambda_{\max}(A)\sum_{i=1}^n \lambda_i(B) = \|A\|_2\,\mathrm{tr}(B).$$

Since $\mathrm{tr}(G_t^\top G_t) = \|G_t\|_F^2$, we obtain

$$\mathrm{tr}(Q_{t-1}^{-1} G_t^\top G_t) \le \|Q_{t-1}^{-1}\|_2\,\|G_t\|_F^2.$$

By assuming $\|G_t\|_F \le G$ for all t,

$$\mathrm{tr}(L_t) \le \mu m + \frac{G^2}{n}\|Q_{t-1}^{-1}\|_2.$$

Substituting this bound into equation 18 and combining with equation 17 yields

$$\|P_t^{-1}\|_2 \le \frac{\gamma m + (1-\gamma)\left(\mu m + \frac{G^2}{n}\|Q_{t-1}^{-1}\|_2\right)}{m(1-\gamma)\mu} = a + b\,\|Q_{t-1}^{-1}\|_2, \tag{19}$$

where

$$a := \frac{\gamma + (1-\gamma)\mu}{(1-\gamma)\mu}, \qquad b := \frac{G^2}{mn\,\mu}. \tag{20}$$

Similarly, from $\widetilde{Q}_t = \gamma Q_{t-1} + (1-\gamma)R_t$ and equation 16,

$$\lambda_{\min}(\widetilde{Q}_t) \geq (1-\gamma)\mu, \qquad \mathrm{tr}(\widetilde{Q}_t) = \gamma n + (1-\gamma)\mathrm{tr}(R_t). \tag{21}$$

Moreover,

$$\mathrm{tr}(R_t) = \frac{1}{m}\mathrm{tr}(G_t^\top P_t^{-1} G_t) + \mu n = \frac{1}{m}\mathrm{tr}(P_t^{-1} G_t G_t^\top) + \mu n \leq \frac{1}{m}\|P_t^{-1}\|_2 \, \|G_t\|_F^2 + \mu n \leq \mu n + \frac{G^2}{m}\|P_t^{-1}\|_2.$$

Therefore, by the same normalization argument as for P_t and equation 21, we have

$$\|Q_t^{-1}\|_2 \leq \frac{\gamma n + (1-\gamma)\left(\mu n + \frac{G^2}{m}\|P_t^{-1}\|_2\right)}{n(1-\gamma)\mu} = a + b\,\|P_t^{-1}\|_2, \tag{22}$$

with the same a, b as in equation 20.

For simplicity, we denote $r_t := \|P_t^{-1}\|_2$ and $s_t := \|Q_t^{-1}\|_2$. Equations equation 19–equation 22 give

$$r_t \leq a + b\,s_{t-1}, \qquad s_t \leq a + b\,r_t.$$

Combining them yields

$$s_t \leq a + b(a + b\,s_{t-1}) = a(1+b) + b^2 s_{t-1}.$$

If $b < 1$, then $b^2 < 1$ and iterating the above recursion gives

$$s_t \leq a(1+b)\sum_{k=0}^{t-1} b^{2k} + b^{2t} s_0 \leq \frac{a(1+b)}{1-b^2} + s_0.$$

Since $Q_0 = I_n$, we have $s_0 = \|Q_0^{-1}\|_2 = 1$. By defining

$$\overline{s} := \frac{a(1+b)}{1-b^2} + 1,$$

we obtain $s_t \leq \overline{s}$ for all t. Plugging into $r_t \leq a + b s_{t-1}$ yields

$$r_t \leq a + b\,\overline{s} =: \overline{r}, \qquad \forall t.$$

Finally,

$$K_{PQ}(t) = \|P_t^{-1/2}\|_2 \|Q_t^{-1/2}\|_2 = \sqrt{\|P_t^{-1}\|_2}\,\sqrt{\|Q_t^{-1}\|_2} = \sqrt{r_t s_t} \leq \sqrt{\overline{r}\,\overline{s}} =: \overline{K},$$

which proves that $K_{PQ}(t)$ is uniformly upper-bounded.

□

Lemma F.3 (Lower bound of K_{PQ})**.** *Under the same condition as in Lemma F.2, there exist a constant $\underline{K}$ such that*

$$K_{PQ}(t) \geq \underline{K}. \tag{23}$$

for all iteration t.

Proof of Lemma F.3. Since $P_t \succ 0$ by Lemma F.1, following the similar analysis on P_t, we have

$$\|P_t^{-1/2}\|_2 \;=\; \sqrt{\lambda_{\max}(P_t^{-1})} \;=\; \frac{1}{\sqrt{\lambda_{\min}(P_t)}}.$$

Moreover, since $P_t \succ 0$ is positive semidefinite, its largest eigenvalue is bounded by its trace:

$$\lambda_{\max}(P_t) \;\leq\; \mathrm{tr}(P_t) \;=\; m.$$

It follows that

$$\lambda_{\min}(P_t) \;\le\; m \qquad \Longrightarrow \qquad \|P_t^{-1/2}\|_2 \;=\; \frac{1}{\sqrt{\lambda_{\min}(P_t)}} \;\ge\; \frac{1}{\sqrt{m}}.$$

Similarly, since $Q_t \succ 0$ and using $\mathrm{tr}(Q_t) = n$, we obtain

$$\|Q_t^{-1/2}\|_2 \;\ge\; \frac{1}{\sqrt{n}}.$$

Combining the two inequalities yields

$$K_{PQ}(t) \;=\; \|P_t^{-1/2}\|_2 \, \|Q_t^{-1/2}\|_2 \;\ge\; \frac{1}{\sqrt{m}} \cdot \frac{1}{\sqrt{n}} \;=\; \frac{1}{\sqrt{mn}} =: \underline{K}.$$

This completes the proof.

□

Lemma F.4 (Lipschitz bound for inverse square root)**.** *Let $A, B \in \mathbb{R}^{m\times m}$ be positive definite matrices. Assume that $A \succeq \underline{c}\, I_m$ and $B \succeq \underline{c}\, I_m$ for some constant $\underline{c} > 0$. Then*

$$\left\|A^{-1/2} - B^{-1/2}\right\|_2 \;\le\; \frac{1}{2\,\underline{c}^{3/2}}\, \|A - B\|_2.$$

Proof of Lemma F.4. For any positive definite matrix X, the following identity holds:

$$X^{-1/2} = \frac{1}{\pi} \int_0^\infty t^{-1/2}\, (X + tI_m)^{-1}\, \mathrm{d}t. \tag{24}$$

This follows from the identity $x^{-1/2} = \frac{1}{\pi}\int_0^\infty t^{-1/2}(x+t)^{-1}\,\mathrm{d}t$.

Therefore, we have

$$A^{-1/2} - B^{-1/2} = \frac{1}{\pi} \int_0^\infty t^{-1/2} \Big[(A + tI_m)^{-1} - (B + tI_m)^{-1}\Big] \mathrm{d}t. \tag{25}$$

Moreover, the resolvent identity gives

$$(A + tI_m)^{-1} - (B + tI_m)^{-1} = (A + tI_m)^{-1}(B - A)(B + tI_m)^{-1}. \tag{26}$$

Substituting equation 26 into equation 25 yields

$$A^{-1/2} - B^{-1/2} = \frac{1}{\pi} \int_0^\infty t^{-1/2} (A + tI_m)^{-1}(B - A)(B + tI_m)^{-1}\, \mathrm{d}t. \tag{27}$$

Taking spectral norm of both sides and applying submultiplicativity, we get

$$\left\|A^{-1/2} - B^{-1/2}\right\|_2 \le \frac{1}{\pi} \int_0^\infty t^{-1/2} \left\|(A + tI_m)^{-1}\right\|_2 \|A - B\|_2 \left\|(B + tI_m)^{-1}\right\|_2 \mathrm{d}t. \tag{28}$$

Since $A \succeq \underline{c} I_m$, we have $A + tI_m \succeq (\underline{c} + t) I_m$, hence

$$\left\|(A + tI_m)^{-1}\right\|_2 \le \frac{1}{\underline{c} + t}, \qquad \left\|(B + tI_m)^{-1}\right\|_2 \le \frac{1}{\underline{c} + t}. \tag{29}$$

Plugging equation 29 into equation 28 gives

$$\left\|A^{-1/2} - B^{-1/2}\right\|_2 \le \frac{\|A - B\|_2}{\pi} \int_0^\infty t^{-1/2}\, \frac{1}{(\underline{c} + t)^2}\, \mathrm{d}t. \tag{30}$$

For simplicity, we change the variable by $t = \underline{c}u$. Then $\mathrm{d}t = \underline{c}\,\mathrm{d}u$, $t^{-1/2} = \underline{c}^{-1/2}u^{-1/2}$, and $(\underline{c} + t)^2 = \underline{c}^2(1 + u)^2$. Therefore,

$$\int_0^\infty t^{-1/2}\, \frac{1}{(\underline{c} + t)^2}\, \mathrm{d}t = \underline{c}^{-3/2} \int_0^\infty \frac{u^{-1/2}}{(1 + u)^2}\, \mathrm{d}u. \tag{31}$$

Clearly the integral is in form of Beta-function. By the Beta-function identity

$$\int_0^\infty \frac{u^{a-1}}{(1+u)^{a+b}}\,\mathrm{d}u = \mathrm{B}(a,b) = \frac{\Gamma(a)\Gamma(b)}{\Gamma(a+b)},$$

with $a = \frac{1}{2}$ and $b = \frac{3}{2}$, we obtain

$$\int_0^\infty \frac{u^{-1/2}}{(1+u)^2}\,\mathrm{d}u = \mathrm{B}\big(\tfrac{1}{2},\tfrac{3}{2}\big) = \frac{\Gamma(\frac{1}{2})\Gamma(\frac{3}{2})}{\Gamma(2)} = \frac{\sqrt{\pi}\cdot\frac{1}{2}\sqrt{\pi}}{1} = \frac{\pi}{2}. \tag{32}$$

Combining equation 31 and equation 32 yields

$$\int_0^\infty t^{-1/2}\,\frac{1}{(\underline{c}+t)^2}\,\mathrm{d}t = \underline{c}^{-3/2}\cdot\frac{\pi}{2}. \tag{33}$$

Plug equation 33 into equation 30, we obtain

$$\big\|A^{-1/2} - B^{-1/2}\big\|_2 \le \frac{\|A-B\|_2}{\pi}\cdot\underline{c}^{-3/2}\cdot\frac{\pi}{2} = \frac{1}{2\,\underline{c}^{3/2}}\|A-B\|_2,$$

which completes the proof.

□

Lemma F.5 (Preconditioner drift bound by step size)**.** *Suppose the EMA hyperparameter γ and the learning rate η in Algorithm 1 satisfies $1-\gamma \le c_\gamma\eta$, for some constant $c_\gamma > 0$. Under the same condition as in LemmaF.2, there exist constants C_P and C_Q such that*

$$\big\|P_t^{-1/2} - P_{t-1}^{-1/2}\big\|_2 \le C_P\eta, \qquad \big\|Q_t^{-1/2} - Q_{t-1}^{-1/2}\big\|_2 \le C_Q\eta. \tag{34}$$

Proof of Lemma F.5. We first prove that there exist positive $\underline{p}$ and $\overline{p}$ such that $\underline{p}I \preceq P_t \preceq \overline{p}I$ for all t.

From Lemma F.2 and Lemma F.3 we know there exist positive a_P and b_P such that $a_P \le \big\|P_t^{-1/2}\big\|_2 \le b_P$ for all t. In particular, $a_P \le 1/\sqrt{m}$ and $b_P \ge \sqrt{r}$. Also we have

$$\big\|P_t^{-1/2}\big\|_2 = \lambda_{\max}(P_t^{-1/2}) = \frac{1}{\sqrt{\lambda_{\min}(P_t)}} \le b_P.$$

Therefore,

$$\lambda_{\min}(P_t) \ge \frac{1}{b_P^2}.$$

Take $\underline{p} = 1/b_P^2$. For the upper bound, we can simply take $\overline{p} = m$, since $\lambda_{\max}(P_t) \le \mathrm{tr}(P_t) = m$. Therefore, we have

$$\lambda_{\min}(P_t) \ge \underline{p}, \qquad \lambda_{\max}(P_t) \le \overline{p}. \tag{35}$$

Hence $\underline{p}I \preceq P_t \preceq \overline{p}I$.

Similarly, we conclude that there exist positive $\underline{q}$ and $\overline{q}$ such that $\underline{q}I \preceq Q_t \preceq \overline{q}I$ for all t.

According to Lemma F.4, we have

$$\big\|P_t^{-1/2} - P_{t-1}^{-1/2}\big\|_2 \le \frac{1}{2\underline{p}^{3/2}}\|P_t - P_{t-1}\|_2.$$

Therefore, it suffice to prove

$$\|P_t - P_{t-1}\|_2 \le c_P\eta \tag{36}$$

for some constant c_P, after which setting $C_P = c_p/(2\underline{p}^{3/2})$ will give the result.

Recall that from Algorithm 1, the update rules are

$$\widetilde{P}_t = \gamma P_{t-1} + (1-\gamma)L_t, \qquad P_t = \frac{m}{\mathrm{tr}(\widetilde{P}_t)}\,\widetilde{P}_t.$$

Define $\alpha_t := \frac{m}{\mathrm{tr}(\widetilde{P}_t)} > 0$, then $P_t = \alpha_t\widetilde{P}_t$. Then we have

$$P_t - P_{t-1} = \alpha_t\widetilde{P}_t - P_{t-1} = (\alpha_t - 1)P_{t-1} + \alpha_t(\widetilde{P}_t - P_{t-1}).$$

Taking operator norms on both sides, we get

$$\|P_t - P_{t-1}\|_2 \le |\alpha_t - 1|\,\|P_{t-1}\|_2 + \alpha_t\,\|\widetilde{P}_t - P_{t-1}\|_2. \tag{37}$$

By the definition of EMA in Algorithm 1,

$$\widetilde{P}_t - P_{t-1} = \gamma P_{t-1} + (1-\gamma)L_t - P_{t-1} = (1-\gamma)(L_t - P_{t-1}). \tag{38}$$

Hence

$$\|\widetilde{P}_t - P_{t-1}\|_2 \le (1-\gamma)\big(\|L_t\|_2 + \|P_{t-1}\|_2\big). \tag{39}$$

Using $\|P_{t-1}\|_2 \le \overline{p}$, it remains to bound $\|L_t\|_2$. Recall

$$L_t = \frac{1}{n} G_t Q_{t-1}^{-1} G_t^\top + \mu \frac{\mathrm{tr}(P_{t-1})}{m} I_m = \frac{1}{n} G_t Q_{t-1}^{-1} G_t^\top + \mu I_m.$$

Therefore,

$$\|L_t\|_2 \le \frac{1}{n}\|G_t\|_2^2\,\|Q_{t-1}^{-1}\|_2 + \mu \le \frac{1}{n}\,G_2^2 \cdot \frac{1}{\underline{q}} + \mu =: \overline{L}_P. \tag{40}$$

Plugging equation 40 into equation 39 yields

$$\|\widetilde{P}_t - P_{t-1}\|_2 \le (1-\gamma)\big(\overline{L}_P + \overline{p}\big). \tag{41}$$

Under the timescale condition $1-\gamma \le c_\gamma \eta$, we obtain

$$\|\widetilde{P}_t - P_{t-1}\|_2 \le c_\gamma\big(\overline{L}_P + \overline{p}\big)\eta. \tag{42}$$

Now we consider α_t and $|\alpha_t - 1|$. Since $\widetilde{P}_t = \gamma P_{t-1} + (1-\gamma)L_t \succeq \gamma P_{t-1} \succeq \gamma \underline{p} I_m$, therefore, $\mathrm{tr}(\widetilde{P}_t) \ge \gamma \underline{p}\, m$. Hence we have

$$\alpha_t = \frac{m}{\mathrm{tr}(\widetilde{P}_t)} \le \frac{1}{\gamma \underline{p}}. \tag{43}$$

For $|\alpha_t - 1|$, we have

$$|\alpha_t - 1| = \left|\frac{m - \mathrm{tr}(\widetilde{P}_t)}{\mathrm{tr}(\widetilde{P}_t)}\right| \le \frac{|m - \mathrm{tr}(\widetilde{P}_t)|}{\gamma \underline{p}\, m}. \tag{44}$$

Taking trace of $\widetilde{P}_t = \gamma P_{t-1} + (1-\gamma)L_t$ gives

$$\mathrm{tr}(\widetilde{P}_t) = \gamma\,\mathrm{tr}(P_{t-1}) + (1-\gamma)\,\mathrm{tr}(L_t) = \gamma m + (1-\gamma)\,\mathrm{tr}(L_t),$$

hence

$$m - \mathrm{tr}(\widetilde{P}_t) = (1-\gamma)\big(m - \mathrm{tr}(L_t)\big). \tag{45}$$

Combining equation 44–equation 45 yields

$$|\alpha_t - 1| \le \frac{1-\gamma}{\gamma \underline{p}\, m}\,|m - \mathrm{tr}(L_t)|. \tag{46}$$

It remains to find an upper bound for $|m - \mathrm{tr}(L_t)|$. Using $L_t = \frac{1}{n} G_t Q_{t-1}^{-1} G_t^\top + \mu I_m$,

$$\mathrm{tr}(L_t) = \frac{1}{n}\,\mathrm{tr}(Q_{t-1}^{-1} G_t^\top G_t) + \mu m \le \frac{1}{n}\|Q_{t-1}^{-1}\|_2\,\mathrm{tr}(G_t^\top G_t) + \mu m.$$

Moreover, $\|Q_{t-1}^{-1}\|_2 \le 1/\underline{q}$ and $\mathrm{tr}(G_t^\top G_t) = \|G_t\|_F^2 \le \min(m,n)\|G_t\|_2^2 \le \min(m,n) G_2^2$. Thus there exists a constant $c_L > 0$ such that $\mathrm{tr}(L_t) \le c_L m$, and consequently

$$|m - \mathrm{tr}(L_t)| \le m + \mathrm{tr}(L_t) \le (1 + c_L) m. \tag{47}$$

Plugging equation 47 into equation 46 gives

$$|\alpha_t - 1| \le \frac{1 + c_L}{\gamma \underline{p}}\,(1-\gamma) \le \frac{1 + c_L}{\gamma \underline{p}}\,c_\gamma\,\eta. \tag{48}$$

Substitute equation 42, equation 43, and equation 48 into equation 37, and use $\|P_{t-1}\|_2 \leq \overline{p}$:

$$\begin{aligned}\|P_t - P_{t-1}\|_2 &\leq |\alpha_t - 1| \, \|P_{t-1}\|_2 + \alpha_t \, \|\widetilde{P}_t - P_{t-1}\|_2 \\ &\leq \left(\frac{1 + c_L}{\gamma \underline{p}} c_\gamma \, \eta\right) \overline{p} \; + \; \left(\frac{1}{\gamma \underline{p}}\right) \left(c_\gamma (\overline{L}_P + \overline{p}) \eta\right) \\ &= \left[\frac{(1 + c_L) c_\gamma \, \overline{p}}{\gamma \underline{p}} + \frac{c_\gamma (\overline{L}_P + \overline{p})}{\gamma \underline{p}}\right] \eta =: c_P \, \eta,\end{aligned}$$

which proves equation 36. Therefore, we finish the boundedness of the drift for preconditioner P_t. i.e.

$$\left\|P_t^{-1/2} - P_{t-1}^{-1/2}\right\|_2 \leq \frac{c_P}{2\underline{p}^{3/2}} \eta = C_P \eta,$$

as claimed.

Similarly, replacing $(m, P, L, \underline{p}, \overline{p})$ by $(n, Q, R, \underline{q}, \overline{q})$, we can prove that there exist a constant C_Q such that

$$\left\|Q_t^{-1/2} - Q_{t-1}^{-1/2}\right\|_2 \leq C_Q \eta,$$

which finishs the whole proof.

□

Lemma F.6. *Given the update rules in Algorithm 1, the following inequality holds:*

$$\sum_{t=1}^{T} \|\widetilde{G}_t - M_t\|_* \leq \frac{\beta}{1-\beta} \sum_{t=2}^{T} \|\widetilde{G}_t - \widetilde{G}_{t-1}\|_* + \frac{\beta}{1-\beta} \sup_{1 \leq t \leq T} \|\widetilde{G}_t\|_*. \tag{49}$$

Proof of Lemma F.6. We first claim that the closed form for M_t is given by

$$M_t = (1-\beta) \sum_{s=1}^{t} \beta^{t-s} \widetilde{G}_s. \tag{50}$$

We prove the claim by induction as follows. Clear equation equation 50 holds for $t = 1$. Suppose equation 50 also holds for $t - 1$, then

$$\begin{aligned}M_t &= \beta M_{t-1} + (1-\beta)\widetilde{G}_t \\ &= \beta \left((1-\beta) \sum_{s=1}^{t-1} \beta^{t-1-s} \widetilde{G}_s\right) + (1-\beta)\widetilde{G}_t \\ &= (1-\beta) \sum_{s=1}^{t-1} \beta^{t-s} \widetilde{G}_s + (1-\beta)\beta^0 \widetilde{G}_t \\ &= (1-\beta) \sum_{s=1}^{t} \beta^{t-s} \widetilde{G}_s,\end{aligned}$$

which proves the claim.

Therefore, we have

$$\begin{aligned}\widetilde{G}_t - M_t &= \widetilde{G}_t - (1-\beta) \sum_{s=1}^{t} \beta^{t-s} \widetilde{G}_s \\ &= \beta \widetilde{G}_t - (1-\beta) \sum_{s=1}^{t-1} \beta^{t-s} \widetilde{G}_s.\end{aligned} \tag{51}$$

Notice that by geometric-sum,

$$(1-\beta) \sum_{s=1}^{t-1} \beta^{t-s} = (1-\beta) \sum_{\hat{s}=1}^{t-1} \beta^{\hat{s}} = \beta(1 - \beta^{t-1}) = \beta - \beta^t.$$

Therefore,

$$\beta\widetilde{G}_t = (1-\beta)\sum_{s=1}^{t-1}\beta^{t-s}\widetilde{G}_t + \beta^t\widetilde{G}_t.$$

Hence

$$\widetilde{G}_t - M_t = (1-\beta)\sum_{s=1}^{t-1}\beta^{t-s}(\widetilde{G}_t - \widetilde{G}_s) + \beta^t\widetilde{G}_t. \tag{52}$$

Take nuclear norm on both sides and apply triangle inequality, we get

$$\|\widetilde{G}_t - M_t\|_* \le (1-\beta)\sum_{s=1}^{t-1}\beta^{t-s}\|\widetilde{G}_t - \widetilde{G}_s\|_* + \beta^t\|\widetilde{G}_t\|_*. \tag{53}$$

For each $1 \le s \le t-1$, telescoping yields

$$\|\widetilde{G}_t - \widetilde{G}_s\|_* \le \sum_{j=s+1}^{t}\|\widetilde{G}_j - \widetilde{G}_{j-1}\|_*. \tag{54}$$

Plugging equation 54 into equation 53 gives

$$\begin{aligned}
\|\widetilde{G}_t - M_t\|_* &\le (1-\beta)\sum_{s=1}^{t-1}\beta^{t-s}\sum_{j=s+1}^{t}\|\widetilde{G}_j - \widetilde{G}_{j-1}\|_* + \beta^t\|\widetilde{G}_t\|_* \\
&= \sum_{j=2}^{t}\Big((1-\beta)\sum_{s=1}^{j-1}\beta^{t-s}\Big)\|\widetilde{G}_j - \widetilde{G}_{j-1}\|_* + \beta^t\|\widetilde{G}_t\|_* \\
&\le \sum_{j=2}^{t}\beta^{t-j+1}\|\widetilde{G}_j - \widetilde{G}_{j-1}\|_* \;+\; \beta^t\|\widetilde{G}_t\|_*,
\end{aligned} \tag{55}$$

where the third line use the fact that, for each $j \le t$,

$$(1-\beta)\sum_{s=1}^{j-1}\beta^{t-s} = (1-\beta)\beta^{t-j+1}\sum_{r=0}^{j-2}\beta^r \le \beta^{t-j+1}.$$

Summing equation 55 over $t = 1, \dots, T$,

$$\begin{aligned}
\sum_{t=1}^{T}\|\widetilde{G}_t - M_t\|_* &\le \sum_{t=2}^{T}\sum_{j=2}^{t}\beta^{t-j+1}\|\widetilde{G}_j - \widetilde{G}_{j-1}\|_* + \sum_{t=1}^{T}\beta^t\|\widetilde{G}_t\|_* \\
&= \sum_{j=2}^{T}\Big(\sum_{t=j}^{T}\beta^{t-j+1}\Big)\|\widetilde{G}_j - \widetilde{G}_{j-1}\|_* + \sum_{t=1}^{T}\beta^t\|\widetilde{G}_t\|_*.
\end{aligned} \tag{56}$$

Notice that, for the geometric tail,

$$\sum_{t=j}^{T}\beta^{t-j+1} \le \sum_{r=1}^{\infty}\beta^r = \frac{\beta}{1-\beta}.$$

Thus the first term in equation 56 is bounded by

$$\frac{\beta}{1-\beta}\sum_{j=2}^{T}\|\widetilde{G}_j - \widetilde{G}_{j-1}\|_*.$$

For the residual term, since $\|\widetilde{G}_t\|_* \le \sup_{1\le s\le T}\|\widetilde{G}_s\|_*$, we obtain

$$\sum_{t=1}^{T}\beta^t\|\widetilde{G}_t\|_* \le \Big(\sum_{t=1}^{\infty}\beta^t\Big)\sup_{1\le s\le T}\|\widetilde{G}_s\|_* = \frac{\beta}{1-\beta}\sup_{1\le s\le T}\|\widetilde{G}_s\|_*.$$

This complete the proof.

□

Lemma F.7 (Variance reduction w.r.t. batch size)**.** *Define the gradient sampling noise*

$$\begin{aligned} N_t :=& G_t - \nabla\mathcal{L}(W_{t-1}) \\ =& \frac{1}{B}\sum_{i=1}^{B}\left(\nabla\ell(W_{t-1};\xi_{t,i}) - \nabla\mathcal{L}(W_{t-1})\right). \end{aligned}$$

Under Assumption 5.4–5.5, we have

$$\mathbb{E}[\|N_t\|_F^2 \mid \mathcal{F}_{t-1}] \le \frac{\sigma^2}{B}. \tag{57}$$

Proof. Let $Z_{t,i} := \nabla\ell(W_{t-1};\xi_{t,i}) - \nabla\mathcal{L}(W_{t-1})$. Then $N_t = \frac{1}{B}\sum_{i=1}^{B} Z_{t,i}$ and $\mathbb{E}[Z_{t,i} \mid \mathcal{F}_{t-1}] = 0$. Using independence across i and expanding the squared Frobenius norm:

$$\begin{aligned} \mathbb{E}[\|N_t\|_F^2 \mid \mathcal{F}_{t-1}] &= \mathbb{E}\left[\left\|\frac{1}{B}\sum_{i=1}^{B} Z_{t,i}\right\|_F^2 \middle| \mathcal{F}_{t-1}\right] \\ &= \frac{1}{B^2}\mathbb{E}\left[\sum_{i=1}^{B}\|Z_{t,i}\|_F^2 + \sum_{i\neq j}\langle Z_{t,i}, Z_{t,j}\rangle_F \middle| \mathcal{F}_{t-1}\right]. \end{aligned}$$

The cross terms vanish:

$$\mathbb{E}[\langle Z_{t,i}, Z_{t,j}\rangle_F \mid \mathcal{F}_{t-1}] = \Big\langle \mathbb{E}[Z_{t,i} \mid \mathcal{F}_{t-1}], \mathbb{E}[Z_{t,j} \mid \mathcal{F}_{t-1}] \Big\rangle_F = 0 \quad (i \neq j),$$

so

$$\mathbb{E}[\|N_t\|_F^2 \mid \mathcal{F}_{t-1}] = \frac{1}{B^2}\sum_{i=1}^{B}\mathbb{E}[\|Z_{t,i}\|_F^2 \mid \mathcal{F}_{t-1}] \le \frac{1}{B^2}\cdot B\sigma^2 = \frac{\sigma^2}{B}.$$

□

F.2 Proof of the deterministic convergence theorem

Lemma F.8 (One-step descent)**.** *Suppose that Assumptions 5.2–5.5 hold. Denote* $K_{PQ}(t) = \|P_t^{-1/2}\|_2 \cdot \|Q_t^{-1/2}\|_2$. *Then for each iteration* $t \ge 1$ *in Algorithm 1, the following one-step upper bound for* $\mathcal{L}(W_t)$ *holds:*

$$\begin{aligned} \mathcal{L}(W_t) \le \mathcal{L}(W_{t-1}) &- \eta K_{PQ}(t)\|\nabla\mathcal{L}(W_{t-1})\|_* + 2\eta K_{PQ}(t)\|\nabla\mathcal{L}(W_{t-1}) - G_t\|_* \\ &+ 2\eta\|\widetilde{G}_t - M_t\|_* + \frac{L\eta^2}{2}\min(m,n)K_{PQ}^2(t). \end{aligned}$$

Proof of Lemma F.8. We summarize the update rules for Algorithm 1 as follows:

$$G_t \leftarrow \frac{1}{B}\sum_{i=1}^{B}\nabla_W\ell(W_{t-1};\xi_{t,i}), \quad \widetilde{G}_t := P_t^{-1/2}G_tQ_t^{-1/2}, \quad M_t := \beta M_{t-1} + (1-\beta)\widetilde{G}_t,$$

$$\Delta W_t := P_t^{-1/2}\,\mathrm{Orth}(M_t)\,Q_t^{-1/2}, \quad W_t := W_{t-1} - \eta\,\Delta W_t.$$

Denote the whitened true gradient

$$\widetilde{\nabla}\mathcal{L}(W_{t-1}) := P_t^{-1/2}\,\nabla\mathcal{L}(W_{t-1})\,Q_t^{-1/2}. \tag{58}$$

Apply the Lipschitz smoothness condition with respect to W_t and W_{t-1}, we have

$$\begin{aligned} \mathcal{L}(W_t) &\le \mathcal{L}(W_{t-1}) + \langle\nabla\mathcal{L}(W_{t-1}),\, W_t - W_{t-1}\rangle_F + \frac{L}{2}\|W_t - W_{t-1}\|_F^2 \\ &= \mathcal{L}(W_{t-1}) - \eta\langle\nabla\mathcal{L}(W_{t-1}),\, \Delta W_t\rangle_F + \frac{L\eta^2}{2}\|\Delta W_t\|_F^2. \end{aligned} \tag{59}$$

We first consider the term $\langle \nabla \mathcal{L}(W_{t-1}), \Delta W_t \rangle_F$. Denote $O_t = \mathrm{Orth}(M_t)$, so $\Delta W_t = P_t^{-1/2} O_t Q_t^{-1/2}$. By the cyclic property of trace (as in equation equation 12), we have

$$\begin{aligned}
\langle \nabla \mathcal{L}(W_{t-1}),\, \Delta W_t \rangle_F &= \left\langle \nabla \mathcal{L}(W_{t-1}),\, P_t^{-1/2} O_t Q_t^{-1/2} \right\rangle_F \\
&= \left\langle P_t^{-1/2} \nabla \mathcal{L}(W_{t-1}) Q_t^{-1/2},\, O_t \right\rangle_F \\
&= \left\langle \widetilde{\nabla} \mathcal{L}(W_{t-1}),\, O_t \right\rangle_F \\
&= \langle M_t, O_t \rangle_F + \left\langle \widetilde{\nabla} \mathcal{L}(W_{t-1}) - M_t,\, O_t \right\rangle_F . \qquad (60)
\end{aligned}$$

Since $O_t = \mathrm{Orth}(M_t)$, we have $\|O_t\|_{\mathrm{op}} = \|O_t\|_2 = 1$ and $\langle M_t, O_t \rangle_F = \|M_t\|_*$. By Hölder inequality D.2, we have

$$\left\langle \widetilde{\nabla} \mathcal{L}(W_{t-1}) - M_t,\, O_t \right\rangle_F \geq -\left\| \widetilde{\nabla} \mathcal{L}(W_{t-1}) - M_t \right\|_* \cdot \|O_t\|_{\mathrm{op}} = -\left\| \widetilde{\nabla} \mathcal{L}(W_{t-1}) - M_t \right\|_* . \qquad (61)$$

Moreover, by the triangle inequality,

$$\|M_t\|_* \geq \left\| \widetilde{\nabla} \mathcal{L}(W_{t-1}) \right\|_* - \left\| \widetilde{\nabla} \mathcal{L}(W_{t-1}) - M_t \right\|_* . \qquad (62)$$

Combining equation 60–equation 62 yields

$$\begin{aligned}
\left\langle \widetilde{\nabla} \mathcal{L}(W_{t-1}), O_t \right\rangle_F &\geq \|M_t\|_* - \left\| \widetilde{\nabla} \mathcal{L}(W_{t-1}) - M_t \right\|_* \\
&\geq \left\| \widetilde{\nabla} \mathcal{L}(W_{t-1}) \right\|_* - 2 \left\| \widetilde{\nabla} \mathcal{L}(W_{t-1}) - M_t \right\|_* . \qquad (63)
\end{aligned}$$

Plugging equation 63 into equation 59, we have

$$\begin{aligned}
\mathcal{L}(W_t) &\leq \mathcal{L}(W_{t-1}) - \eta \left\langle \widetilde{\nabla} \mathcal{L}(W_{t-1}), O_t \right\rangle_F + \frac{L\eta^2}{2} \|\Delta W_t\|_F^2 \\
&\leq \mathcal{L}(W_{t-1}) - \eta \left\| \widetilde{\nabla} \mathcal{L}(W_{t-1}) \right\|_* + 2\eta \left\| \widetilde{\nabla} \mathcal{L}(W_{t-1}) - M_t \right\|_* + \frac{L\eta^2}{2} \|\Delta W_t\|_F^2 . \qquad (64)
\end{aligned}$$

For the second term in the RHS of equation 64,

$$\left\| \widetilde{\nabla} \mathcal{L}(W_{t-1}) \right\|_* = \left\| P_t^{-1/2} \nabla \mathcal{L}(W_{t-1}) Q_t^{-1/2} \right\|_* \leq \left\| P_t^{-1/2} \right\|_2 \cdot \left\| \nabla \mathcal{L}(W_{t-1}) \right\|_* \cdot \left\| P_t^{-1/2} \right\|_2 \qquad (65)$$

For the third term in the RHS of equation 64, subtract and add a $\widetilde{G}_t$ and apply triangle inequality, we have

$$\left\| \widetilde{\nabla} \mathcal{L}(W_{t-1}) - M_t \right\|_* \leq \left\| \widetilde{\nabla} \mathcal{L}(W_{t-1}) - \widetilde{G}_t \right\|_* + \left\| \widetilde{G}_t - M_t \right\|_* ,$$

where

$$\begin{aligned}
\left\| \widetilde{\nabla} \mathcal{L}(W_{t-1}) - \widetilde{G}_t \right\|_* &= \left\| P_t^{-1/2} \left(\nabla \mathcal{L}(W_{t-1}) - G_t \right) Q_t^{-1/2} \right\|_* \\
&\leq \left\| P_t^{-1/2} \right\|_2 \cdot \left\| \nabla \mathcal{L}(W_{t-1}) - G_t \right\|_* \cdot \left\| P_t^{-1/2} \right\|_2 , \qquad (66)
\end{aligned}$$

according to the Generalized Hölder inequality for Schatten norm in Theorem D.3.

For the fourth term in the RHS of equation 64, we have

$$\|\Delta W_t\|_F = \|P_t^{-1/2} O_t Q_t^{-1/2}\|_F \leq \|P_t^{-1/2}\|_2 \|O_t\|_F \|Q_t^{-1/2}\|_2 . \qquad (67)$$

Since $O_t = \mathrm{Orth}(M_t)$ is a partial isometry, we have $\|O_t\|_F^2 = \mathrm{rank}(O_t) = \mathrm{rank}(M_t) \leq \min(m, n)$.

Plug equation 65–equation 67 into equation 64, we derive the final formula in Lemma F.8.

□

Proof of Theorem 5.7. In the deterministic setting, we use the full gradient $G_t = \nabla \mathcal{L}(W_{t-1})$ for further analysis. By Lemma F.8, for each $t = 1, \ldots, T$, we have

$$\begin{aligned}
\mathcal{L}(W_t) &\leq \mathcal{L}(W_{t-1}) - \eta K_{PQ}(t) \|\nabla \mathcal{L}(W_{t-1})\|_* \qquad (68) \\
&\quad + 2\eta \|\widetilde{G}_t - M_t\|_* + \frac{L\eta^2}{2} \min(m, n) K_{PQ}^2(t) . \qquad (69)
\end{aligned}$$

Here the stochastic gradient noise term vanishes because

$$\|\nabla\mathcal{L}(W_{t-1}) - G_t\|_* = 0.$$

Summing equation 69 from $t = 1$ to T yields

$$\mathcal{L}(W_T) \le \mathcal{L}(W_0) - \eta \sum_{t=1}^{T} K_{PQ}(t)\|\nabla\mathcal{L}(W_{t-1})\|_* + 2\eta \sum_{t=1}^{T} \|\widetilde{G}_t - M_t\|_* + \frac{L\eta^2}{2}\min(m,n)\sum_{t=1}^{T} K_{PQ}^2(t), \tag{70}$$

where the objective terms telescope as $\sum_{t=1}^{T}(\mathcal{L}(W_t) - \mathcal{L}(W_{t-1})) = \mathcal{L}(W_T) - \mathcal{L}(W_0)$. By Assumption 5.2, the objective function is lower-bounded by $\mathcal{L}_\star$, so $\mathcal{L}(W_T) \ge \mathcal{L}_\star$. Rearranging equation 70 and dividing both sides by ηT, we obtain

$$\frac{1}{T}\sum_{t=1}^{T} K_{PQ}(t)\|\nabla\mathcal{L}(W_{t-1})\|_* \le \frac{\mathcal{L}(W_0) - \mathcal{L}_\star}{\eta T} + \underbrace{\frac{2}{T}\sum_{t=1}^{T}\|\widetilde{G}_t - M_t\|_*}_{\text{Momentum tracking error}} + \underbrace{\frac{L\eta}{2}\min(m,n)\cdot\frac{1}{T}\sum_{t=1}^{T} K_{PQ}^2(t)}_{\text{Quadratic term (smoothness penalty)}}. \tag{71}$$

Now we consider the terms on the right-hand side of equation 71 separately.

Quadratic term. According to Lemma F.2, the term $K_{PQ}(t)$ is upper-bounded by $\overline{K}$. Therefore,

$$\frac{1}{T}\sum_{t=1}^{T} K_{PQ}^2(t) \le \overline{K}^2. \tag{72}$$

Momentum tracking term. By Lemma F.6, we have

$$\sum_{t=1}^{T}\|\widetilde{G}_t - M_t\|_* \le \frac{\beta}{1-\beta}\sum_{t=2}^{T}\|\widetilde{G}_t - \widetilde{G}_{t-1}\|_* + \frac{\beta}{1-\beta}\sup_{1\le t\le T}\|\widetilde{G}_t\|_*. \tag{73}$$

It remains to bound the increment $\|\widetilde{G}_t - \widetilde{G}_{t-1}\|_*$. By adding and subtracting intermediate terms, we have

$$\begin{aligned}\widetilde{G}_t - \widetilde{G}_{t-1} &= (P_t^{-1/2} - P_{t-1}^{-1/2})G_tQ_t^{-1/2} \\ &\quad + P_{t-1}^{-1/2}(G_t - G_{t-1})Q_t^{-1/2} \\ &\quad + P_{t-1}^{-1/2}G_{t-1}(Q_t^{-1/2} - Q_{t-1}^{-1/2}).\end{aligned}$$

Using Lemma D.3 and the triangle inequality, we obtain

$$\|\widetilde{G}_t - \widetilde{G}_{t-1}\|_* \le \|P_t^{-1/2} - P_{t-1}^{-1/2}\|_2\|G_t\|_*\|Q_t^{-1/2}\|_2 \tag{74}$$

$$+ \|P_{t-1}^{-1/2}\|_2\|G_t - G_{t-1}\|_*\|Q_t^{-1/2}\|_2 \tag{75}$$

$$+ \|P_{t-1}^{-1/2}\|_2\|G_{t-1}\|_*\|Q_t^{-1/2} - Q_{t-1}^{-1/2}\|_2. \tag{76}$$

By Lemma F.2 and Lemma F.5, there exist constants $\overline{K}, C_P, C_Q > 0$ such that, for all t,

$$K_{PQ}(t) := \|P_t^{-1/2}\|_2\|Q_t^{-1/2}\|_2 \le \overline{K},$$

and

$$\|P_t^{-1/2} - P_{t-1}^{-1/2}\|_2 \le C_P\eta, \qquad \|Q_t^{-1/2} - Q_{t-1}^{-1/2}\|_2 \le C_Q\eta.$$

Since we assume a uniform bound on the full gradient magnitude in nuclear norm,

$$\|G_t\|_* = \|\nabla \mathcal{L}(W_{t-1})\|_* \le G_*, \qquad t = 1, \dots, T, \tag{77}$$

the first and third terms in equation 74–equation 76 are bounded as

$$\|P_t^{-1/2} - P_{t-1}^{-1/2}\|_2 \|G_t\|_* \|Q_t^{-1/2}\|_2 \le C_P \eta \, \overline{K} G_*,$$

and

$$\|P_{t-1}^{-1/2}\|_2 \|G_{t-1}\|_* \|Q_t^{-1/2} - Q_{t-1}^{-1/2}\|_2 \le C_Q \eta \, \overline{K} G_*.$$

Moreover, the middle term satisfies

$$\|P_{t-1}^{-1/2}\|_2 \|G_t - G_{t-1}\|_* \|Q_t^{-1/2}\|_2 \le \overline{K} \|G_t - G_{t-1}\|_*.$$

Hence,

$$\|\widetilde{G}_t - \widetilde{G}_{t-1}\|_* \le (C_P + C_Q)\eta \, \overline{K} G_* + \overline{K} \|G_t - G_{t-1}\|_*. \tag{78}$$

Since $\mathcal{L}$ is L -smooth in Frobenius norm by Assumption 5.3, and $\|A\|_* \le \sqrt{\min(m,n)}\|A\|_F$, we have

$$\begin{aligned} \|G_t - G_{t-1}\|_* &\le \sqrt{\min(m,n)} \|G_t - G_{t-1}\|_F \\ &\le L\sqrt{\min(m,n)} \|W_{t-1} - W_{t-2}\|_F. \end{aligned} \tag{79}$$

It remains to bound $\|W_{t-1} - W_{t-2}\|_F$. Since

$$W_{t-1} - W_{t-2} = -\eta \Delta W_{t-1},$$

we have

$$\|W_{t-1} - W_{t-2}\|_F = \eta \|\Delta W_{t-1}\|_F.$$

Recall that

$$\Delta W_{t-1} = P_{t-1}^{-1/2} \, \mathrm{Orth}(M_{t-1}) \, Q_{t-1}^{-1/2}.$$

By Theorem D.3 and $\|\mathrm{Orth}(M_{t-1})\|_F \le \sqrt{\min(m,n)}$, we get

$$\|\Delta W_{t-1}\|_F \le \|P_{t-1}^{-1/2}\|_2 \|\mathrm{Orth}(M_{t-1})\|_F \|Q_{t-1}^{-1/2}\|_2 \le \sqrt{\min(m,n)} \, K_{PQ}(t-1) \le \sqrt{\min(m,n)} \, \overline{K}. \tag{80}$$

Plugging equation 80 into equation 79, we obtain

$$\|G_t - G_{t-1}\|_* \le L\sqrt{\min(m,n)} \cdot \eta \sqrt{\min(m,n)} \, \overline{K} = L\eta \, \min(m,n) \, \overline{K}. \tag{81}$$

Combining equation 78 and equation 81 yields

$$\|\widetilde{G}_t - \widetilde{G}_{t-1}\|_* \le \eta \Big[(C_P + C_Q)\overline{K} G_* + L \min(m,n) \overline{K}^2 \Big]. \tag{82}$$

Furthermore, by Lemma F.2 and $\|G_t\|_* \le G_*$,

$$\|\widetilde{G}_t\|_* = \|P_t^{-1/2} G_t Q_t^{-1/2}\|_* \le K_{PQ}(t) \|G_t\|_* \le \overline{K} G_*.$$

Plugging equation 82 and $\sup_t \|\widetilde{G}_t\|_* \le \overline{K} G_*$ into equation 73, we obtain

$$\sum_{t=1}^{T} \|\widetilde{G}_t - M_t\|_* \le \frac{\beta}{1-\beta}(T-1)\eta \Big[(C_P + C_Q)\overline{K} G_* + L \min(m,n) \overline{K}^2 \Big] + \frac{\beta}{1-\beta} \overline{K} G_*.$$

Dividing by T, we have

$$\frac{1}{T} \sum_{t=1}^{T} \|\widetilde{G}_t - M_t\|_* \le \frac{\beta}{1-\beta} \eta \Big[(C_P + C_Q)\overline{K} G_* + L \min(m,n) \overline{K}^2 \Big] + \frac{\beta}{1-\beta} \frac{\overline{K} G_*}{T}. \tag{83}$$

Final upper bound. Plugging equation 72 and equation 83 into equation 71, we obtain

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T} K_{PQ}(t)\|\nabla\mathcal{L}(W_{t-1})\|_* \le\ & \frac{\mathcal{L}(W_0)-\mathcal{L}_\star}{\eta T} + \frac{2\beta}{1-\beta}\frac{\overline{K}G_*}{T} \\ & + \eta\left[\frac{L}{2}\min(m,n)\overline{K}^2 + \frac{2\beta}{1-\beta}\Big((C_P+C_Q)\overline{K}G_* + L\min(m,n)\overline{K}^2\Big)\right].\end{aligned} \tag{84}$$

For simplicity, denote

$$R := \mathcal{L}(W_0)-\mathcal{L}_\star, \qquad C_{dr} := C_P + C_Q, \qquad r := \min(m,n).$$

By Lemma F.3, there exists $\underline{K} > 0$ such that $K_{PQ}(t) \ge \underline{K}$ for all t. Then, for each t,

$$K_{PQ}(t)\|\nabla\mathcal{L}(W_{t-1})\|_* \ge \underline{K}\|\nabla\mathcal{L}(W_{t-1})\|_*.$$

Averaging over $t = 1,\dots,T$ gives

$$\frac{1}{T}\sum_{t=1}^{T} K_{PQ}(t)\|\nabla\mathcal{L}(W_{t-1})\|_* \ge \underline{K}\cdot\frac{1}{T}\sum_{t=1}^{T}\|\nabla\mathcal{L}(W_{t-1})\|_*. \tag{85}$$

Combining equation 85 with equation 84, we obtain

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\|\nabla\mathcal{L}(W_{t-1})\|_* &\le \frac{1}{\underline{K}}\left[\frac{R}{\eta T} + \frac{2\beta}{1-\beta}\frac{\overline{K}G_*}{T} + \eta\left(\frac{L}{2}r\overline{K}^2 + \frac{2\beta}{1-\beta}\Big(C_{dr}\overline{K}G_* + Lr\overline{K}^2\Big)\right)\right] \\ &= \frac{R}{\underline{K}\eta T} + \frac{2\beta}{1-\beta}\frac{\overline{K}G_*}{\underline{K}T} + \frac{2\beta}{1-\beta}\frac{C_{dr}\overline{K}G_*}{\underline{K}}\eta + \frac{1+3\beta}{2-2\beta}\frac{L\overline{K}^2}{\underline{K}}r\eta.\end{aligned} \tag{86}$$

In particular, if we take $\eta = C/\sqrt{T}$ for some constant $C > 0$, then

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\|\nabla\mathcal{L}(W_{t-1})\|_* \le\ & \frac{R}{\underline{K}C\sqrt{T}} + \frac{2\beta}{1-\beta}\frac{\overline{K}G_*}{\underline{K}T} \\ & + \frac{2\beta}{1-\beta}\frac{C_{dr}\overline{K}G_*}{\underline{K}}\frac{C}{\sqrt{T}} + \frac{1+3\beta}{2-2\beta}\frac{L\overline{K}^2}{\underline{K}}r\frac{C}{\sqrt{T}}.\end{aligned} \tag{87}$$

Therefore,

$$\frac{1}{T}\sum_{t=1}^{T}\|\nabla\mathcal{L}(W_{t-1})\|_* = \mathcal{O}\left(\frac{R + rL + G_*}{\sqrt{T}} + \frac{G_*}{T}\right). \tag{88}$$

This finishes the proof. □

F.3 Proof of the stochastic convergence theorem

Notation. Recall that FISMO updates the matrix parameter by

$$W_t = W_{t-1} - \eta\, P_t^{-1/2}\ \mathrm{Orth}(M_t)\, Q_t^{-1/2},$$

where

$$M_t = \beta M_{t-1} + (1-\beta)\widetilde{G}_t, \qquad \widetilde{G}_t := P_t^{-1/2} G_t Q_t^{-1/2}.$$

We further define the whitened true gradient

$$\widetilde{H}_t := P_t^{-1/2}\nabla L(W_{t-1})Q_t^{-1/2},$$

and the EMA of the whitened true gradients

$$\widetilde{C}_t := \beta\widetilde{C}_{t-1} + (1-\beta)\widetilde{H}_t.$$

Lemma F.9 (One-step descent)**.** *Under Assumptions 5.3 and Lemma F.5, the FISMO iterate satisfies*

$$L(W_t) \le L(W_{t-1}) - \eta\, K_{PQ}(t)\, \|\nabla L(W_{t-1})\|_* + 2\eta\, \|\widetilde{H}_t - M_t\|_* + \frac{L\eta^2}{2}\, r\, K_{PQ}(t)^2. \tag{89}$$

Proof. By L-smoothness,

$$L(W_t) \le L(W_{t-1}) + \langle \nabla L(W_{t-1}), W_t - W_{t-1}\rangle_F + \frac{L}{2}\|W_t - W_{t-1}\|_F^2.$$

Using

$$W_t - W_{t-1} = -\eta\, P_t^{-1/2}\, \mathrm{Orth}(M_t)\, Q_t^{-1/2},$$

we obtain

$$L(W_t) \le L(W_{t-1}) - \eta\, \langle \widetilde{H}_t, \mathrm{Orth}(M_t)\rangle_F + \frac{L\eta^2}{2}\, \|P_t^{-1/2}\, \mathrm{Orth}(M_t) Q_t^{-1/2}\|_F^2.$$

Since $\mathrm{Orth}(M_t)$ is partial orthogonal, its nonzero singular values are all equal to 1, hence

$$\|\,\mathrm{Orth}(M_t)\|_F \le \sqrt{r}.$$

Therefore

$$\|P_t^{-1/2}\, \mathrm{Orth}(M_t) Q_t^{-1/2}\|_F \le \|P_t^{-1/2}\|_2\, \|\,\mathrm{Orth}(M_t)\|_F\, \|Q_t^{-1/2}\|_2 \le K_{PQ}(t)\sqrt{r},$$

and so

$$L(W_t) \le L(W_{t-1}) - \eta\, \langle \widetilde{H}_t, \mathrm{Orth}(M_t)\rangle_F + \frac{L\eta^2}{2}\, r\, K_{PQ}(t)^2.$$

Next, by the standard polar inequality, for all X, Y,

$$\langle X, \mathrm{Orth}(Y)\rangle_F \ge \|X\|_* - 2\|X - Y\|_*, \tag{90}$$

hence

$$\langle \widetilde{H}_t, \mathrm{Orth}(M_t)\rangle_F \ge \|\widetilde{H}_t\|_* - 2\|\widetilde{H}_t - M_t\|_*.$$

Finally,

$$\|\widetilde{H}_t\|_* = \|P_t^{-1/2} \nabla L(W_{t-1}) Q_t^{-1/2}\|_* \ge K_{PQ}(t)\, \|\nabla L(W_{t-1})\|_*.$$

Substituting these bounds yields the claim. □

Lemma F.10 (Variance reduction in the whitened space)**.** *Under Assumption 5.5, the whitened stochastic noise satisfies*

$$\mathbb{E}\Big[\|\widetilde{G}_t - \widetilde{C}_t\|_F^2 \,\Big|\, W_{t-1}\Big] \le \frac{\widetilde{\sigma}^2}{B}, \qquad \textit{with } \widetilde{\sigma} := \overline{K}\sigma.$$

Define $D_t := \widetilde{C}_t - M_t$. Then, for all $t \ge 1$,

$$\mathbb{E}\|D_t\|_F^2 \le \beta^{2t}\mathbb{E}\|D_0\|_F^2 + \frac{1-\beta}{1+\beta}\big(1 - \beta^{2t}\big)\frac{\widetilde{\sigma}^2}{B}.$$

Consequently,

$$\mathbb{E}\|D_t\|_* \le \sqrt{r}\, \beta^t \Big(\mathbb{E}\|D_0\|_F^2\Big)^{1/2} + \sqrt{\frac{r(1-\beta)}{1+\beta}}\, \frac{\widetilde{\sigma}}{\sqrt{B}}. \tag{91}$$

Proof. Using

$$\widetilde{C}_t = \beta \widetilde{C}_{t-1} + (1-\beta)\widetilde{H}_t, \qquad M_t = \beta M_{t-1} + (1-\beta)\widetilde{G}_t,$$

we get $D_t = \beta D_{t-1} - (1-\beta)\widetilde{N}_t$. Hence

$$\|D_t\|_F^2 = \beta^2 \|D_{t-1}\|_F^2 + (1-\beta)^2 \|\widetilde{N}_t\|_F^2 - 2\beta(1-\beta)\langle D_{t-1}, \widetilde{N}_t\rangle_F.$$

Taking conditional expectation with respect to $\mathcal{G}_{t-1}$ and using $\mathbb{E}[\widetilde{N}_t \mid \mathcal{G}_{t-1}] = 0$ gives

$$\mathbb{E}[\|D_t\|_F^2 \mid \mathcal{G}_{t-1}] \le \beta^2 \|D_{t-1}\|_F^2 + (1-\beta)^2 \frac{\widetilde{\sigma}^2}{B}. \tag{92}$$

Taking total expectation and unrolling the recursion yields

$$\mathbb{E}\|D_t\|_F^2 \le \beta^{2t}\mathbb{E}\|D_0\|_F^2 + (1-\beta)^2\frac{\widetilde{\sigma}^2}{B}\sum_{j=0}^{t-1}\beta^{2j}.$$

Since

$$\sum_{j=0}^{t-1}\beta^{2j} = \frac{1-\beta^{2t}}{1-\beta^2} = \frac{1-\beta^{2t}}{(1-\beta)(1+\beta)},$$

we obtain

$$\mathbb{E}\|D_t\|_F^2 \le \beta^{2t}\mathbb{E}\|D_0\|_F^2 + \frac{1-\beta}{1+\beta}\big(1-\beta^{2t}\big)\frac{\widetilde{\sigma}^2}{B}.$$

Now Jensen's inequality gives

$$\mathbb{E}\|D_t\|_F \le \sqrt{\mathbb{E}\|D_t\|_F^2},$$

hence

$$\mathbb{E}\|D_t\|_F \le \beta^t\Big(\mathbb{E}\|D_0\|_F^2\Big)^{1/2} + \sqrt{\frac{1-\beta}{1+\beta}}\frac{\widetilde{\sigma}}{\sqrt{B}}.$$

Finally,

$$\|D_t\|_* \le \sqrt{r}\,\|D_t\|_F,$$

which implies

$$\mathbb{E}\|D_t\|_* \le \sqrt{r}\,\beta^t\Big(\mathbb{E}\|D_0\|_F^2\Big)^{1/2} + \sqrt{\frac{r(1-\beta)}{1+\beta}}\frac{\widetilde{\sigma}}{\sqrt{B}}.$$

□

Lemma F.11 (One-step drift of the whitened true gradient)**.** *Under Assumptions 5.3–5.6 and Lemma F.5, for all $t \ge 2$,*

$$\|\widetilde{H}_t - \widetilde{H}_{t-1}\|_* \le A\,\eta,$$

where

$$A := \sqrt{r}\,L_P G K_Q + rL(K_P K_Q)^2 + \sqrt{r}\,K_P G L_Q,$$

with

$$K_P := \sup_t \|P_t^{-1/2}\|_2, \qquad K_Q := \sup_t \|Q_t^{-1/2}\|_2.$$

Proof. We write $\widetilde{H}_t - \widetilde{H}_{t-1} = A_t + B_t + C_t$, where

$$A_t := (P_t^{-1/2} - P_{t-1}^{-1/2})\nabla L(W_{t-1})Q_t^{-1/2},$$

$$B_t := P_{t-1}^{-1/2}\big(\nabla L(W_{t-1}) - \nabla L(W_{t-2})\big)Q_t^{-1/2},$$

$$C_t := P_{t-1}^{-1/2}\nabla L(W_{t-2})(Q_t^{-1/2} - Q_{t-1}^{-1/2}).$$

Thus

$$\|\widetilde{H}_t - \widetilde{H}_{t-1}\|_* \le \|A_t\|_* + \|B_t\|_* + \|C_t\|_*.$$

For the first term,

$$\|A_t\|_* \le \sqrt{r}\,\|P_t^{-1/2} - P_{t-1}^{-1/2}\|_2\,\|\nabla L(W_{t-1})\|_F\,\|Q_t^{-1/2}\|_2 \le \sqrt{r}\,L_P G K_Q\,\eta.$$

For the third term,

$$\|C_t\|_* \le \sqrt{r}\,\|P_{t-1}^{-1/2}\|_2\,\|\nabla L(W_{t-2})\|_F\,\|Q_t^{-1/2} - Q_{t-1}^{-1/2}\|_2 \le \sqrt{r}\,K_P G L_Q\,\eta.$$

For the middle term, by L-smoothness,

$$\|\nabla L(W_{t-1}) - \nabla L(W_{t-2})\|_F \le L\|W_{t-1} - W_{t-2}\|_F.$$

Moreover,

$$W_{t-1} - W_{t-2} = -\eta\, P_{t-1}^{-1/2}\, \mathrm{Orth}(M_{t-1}) Q_{t-1}^{-1/2},$$

so

$$\|W_{t-1} - W_{t-2}\|_F \le \eta\, K_P K_Q \sqrt{r}.$$

Hence

$$\|B_t\|_* \le \sqrt{r}\, K_P K_Q\, L\, \|W_{t-1} - W_{t-2}\|_F \le rL(K_P K_Q)^2\, \eta.$$

Adding the three bounds gives

$$\|\widetilde{H}_t - \widetilde{H}_{t-1}\|_* \le \Big(\sqrt{r}\, L_P G K_Q + rL(K_P K_Q)^2 + \sqrt{r}\, K_P G L_Q\Big)\eta = A\eta.$$

□

Lemma F.12 (Tracking bound for the whitened true-gradient EMA)**.** *Under the assumptions of Lemma F.11,*

$$\|\widetilde{H}_t - \widetilde{C}_t\|_* \le \beta^t \|\widetilde{H}_0 - \widetilde{C}_0\|_* + \frac{A\beta}{1-\beta}\eta \qquad \textit{for all } t \ge 1.$$

In particular, if $\widetilde{C}_0 = \widetilde{H}_0$*, then*

$$\|\widetilde{H}_t - \widetilde{C}_t\|_* \le \frac{A\beta}{1-\beta}\eta, \qquad \mathbb{E}\|\widetilde{H}_t - \widetilde{C}_t\|_* \le \frac{A\beta}{1-\beta}\eta.$$

Proof. Let $E_t := \widetilde{H}_t - \widetilde{C}_t$. Using $\widetilde{C}_t = \beta\widetilde{C}_{t-1} + (1-\beta)\widetilde{H}_t$, we obtain

$$E_t = \beta(\widetilde{H}_t - \widetilde{C}_{t-1}) = \beta(\widetilde{H}_t - \widetilde{H}_{t-1}) + \beta E_{t-1}.$$

Therefore

$$\|E_t\|_* \le \beta\|\widetilde{H}_t - \widetilde{H}_{t-1}\|_* + \beta\|E_{t-1}\|_*.$$

By Lemma F.11, $\|E_t\|_* \le \beta A\eta + \beta\|E_{t-1}\|_*$. Unrolling the recursion gives

$$\|E_t\|_* \le \beta^t\|E_0\|_* + \beta A\eta \sum_{j=0}^{t-1}\beta^j \le \beta^t\|E_0\|_* + \frac{A\beta}{1-\beta}\eta.$$

Since $E_0 = \widetilde{H}_0 - \widetilde{C}_0$, the claim follows. The expectation bound is immediate. □

Lemma F.13. *Suppose Assumptions 5.2–5.6 hold. Then, for any constant stepsize* $\eta > 0$ *and any* $\beta \in [0, 1)$*,*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla L(W_{t-1})\|_* \le \frac{\Delta}{K\eta T} + \frac{Lr\overline{K}^2}{2K}\eta + \frac{2}{K}\sqrt{\frac{r(1-\beta)}{1+\beta}}\frac{\widetilde{\sigma}}{\sqrt{B}} + \frac{2\sqrt{r}}{K}\frac{\beta}{(1-\beta)T}\Gamma_0 + \frac{2A}{K}\frac{\beta}{1-\beta}\eta,$$

where

$$\Gamma_0 := \Big(\mathbb{E}\|\widetilde{C}_0 - M_0\|_F^2\Big)^{1/2}.$$

In particular, if $\widetilde{C}_0 = M_0$ *and* $\widetilde{C}_0 = \widetilde{H}_0$*, then*

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla L(W_{t-1})\|_* \le \frac{\Delta}{K\eta T} + \frac{Lr\overline{K}^2}{2K}\eta + \frac{2}{K}\sqrt{\frac{r(1-\beta)}{1+\beta}}\frac{\widetilde{\sigma}}{\sqrt{B}} + \frac{2A}{K}\frac{\beta}{1-\beta}\eta.$$

Proof. By Lemma F.9,

$$L(W_t) \le L(W_{t-1}) - \eta\, K_{PQ}(t)\, \|\nabla L(W_{t-1})\|_* + 2\eta\, \|\widetilde{H}_t - M_t\|_* + \frac{L\eta^2}{2}\, r\, K_{PQ}(t)^2.$$

Using

$$\|\widetilde{H}_t - M_t\|_* \le \|\widetilde{H}_t - \widetilde{C}_t\|_* + \|\widetilde{C}_t - M_t\|_*,$$

together with

$$K_{PQ}(t) \ge K, \qquad K_{PQ}(t) \le \overline{K},$$

we get

$$L(W_t) \le L(W_{t-1}) - \eta K \|\nabla L(W_{t-1})\|_* + 2\eta \|\widetilde{H}_t - \widetilde{C}_t\|_* + 2\eta \|\widetilde{C}_t - M_t\|_* + \frac{L\eta^2}{2} r\overline{K}^2.$$

Taking expectations and summing from $t = 1$ to T yields

$$\mathbb{E}[L(W_T)] \le L(W_0) - \eta K \sum_{t=1}^{T} \mathbb{E}\|\nabla L(W_{t-1})\|_* + 2\eta \sum_{t=1}^{T} \mathbb{E}\|\widetilde{H}_t - \widetilde{C}_t\|_*$$

$$+2\eta \sum_{t=1}^{T} \mathbb{E}\|\widetilde{C}_t - M_t\|_* + \frac{L\eta^2}{2} r\overline{K}^2 T.$$

Since $\mathbb{E}[L(W_T)] \ge L^\star$, we have

$$\eta K \sum_{t=1}^{T} \mathbb{E}\|\nabla L(W_{t-1})\|_* \le R + 2\eta \sum_{t=1}^{T} \mathbb{E}\|\widetilde{H}_t - \widetilde{C}_t\|_* + 2\eta \sum_{t=1}^{T} \mathbb{E}\|\widetilde{C}_t - M_t\|_* + \frac{L\eta^2}{2} r\overline{K}^2 T.$$

Divide both sides by ηKT:

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\|\nabla L(W_{t-1})\|_* \le \frac{R}{K\eta T} + \frac{2}{KT} \sum_{t=1}^{T} \mathbb{E}\|\widetilde{H}_t - \widetilde{C}_t\|_*$$

$$+\frac{2}{KT} \sum_{t=1}^{T} \mathbb{E}\|\widetilde{C}_t - M_t\|_* + \frac{Lr\overline{K}^2}{2K} \eta.$$

Now apply Lemma F.12:

$$\mathbb{E}\|\widetilde{H}_t - \widetilde{C}_t\|_* \le \beta^t \|\widetilde{H}_0 - \widetilde{C}_0\|_* + \frac{A\beta}{1-\beta}\eta.$$

If $\widetilde{C}_0 = \widetilde{H}_0$, then

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\|\widetilde{H}_t - \widetilde{C}_t\|_* \le \frac{A\beta}{1-\beta}\eta.$$

Next, by Lemma F.10,

$$\mathbb{E}\|\widetilde{C}_t - M_t\|_* \le \sqrt{r}\, \beta^t \Gamma_0 + \sqrt{\frac{r(1-\beta)}{1+\beta}} \frac{\widetilde{\sigma}}{\sqrt{B}}.$$

Hence

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\|\widetilde{C}_t - M_t\|_* \le \frac{\sqrt{r}\,\Gamma_0}{T} \sum_{t=1}^{T} \beta^t + \sqrt{\frac{r(1-\beta)}{1+\beta}} \frac{\widetilde{\sigma}}{\sqrt{B}}.$$

Since

$$\sum_{t=1}^{T} \beta^t \le \frac{\beta}{1-\beta},$$

we obtain

$$\frac{1}{T} \sum_{t=1}^{T} \mathbb{E}\|\widetilde{C}_t - M_t\|_* \le \frac{\sqrt{r}\,\Gamma_0}{T} \frac{\beta}{1-\beta} + \sqrt{\frac{r(1-\beta)}{1+\beta}} \frac{\widetilde{\sigma}}{\sqrt{B}}.$$

Substituting these bounds into the averaged descent inequality proves the theorem. □

The following is a more detailed version of Theorem 5.8.

Theorem F.14 (Stochastic convergence of FISMO)**.** *Under the assumptions of Theorem F.13, define* $\Lambda := \max\left\{\overline{K}^2 L,\ \frac{2A}{r}\right\}$. *Suppose* $\widetilde{C}_0 = M_0 = \widetilde{H}_0$, *and choose*

$$1-\beta = \min\left\{\frac{\sqrt{\Lambda R B}}{\widetilde{\sigma}\sqrt{T}},\ 1\right\}, \qquad \eta = \sqrt{\frac{(1-\beta)R}{rT\Lambda}}.$$

Then, whenever $T \geq \frac{\Lambda R B}{\widetilde{\sigma}^2}$, *so that* $1-\beta = \frac{\sqrt{\Lambda R B}}{\widetilde{\sigma}\sqrt{T}}$, *the iterates satisfy*

$$\begin{aligned}\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla L(W_{t-1})\|_* \leq{}& \frac{1}{K}(\Lambda R)^{1/4} r^{1/2}\left(\frac{\widetilde{\sigma}^2}{B}\right)^{1/4} T^{-1/4} + \frac{1}{2K} r^{1/2}(\Lambda R)^{3/4}\left(\frac{B}{\widetilde{\sigma}^2}\right)^{1/4} T^{-3/4} \\ &+ \frac{2}{K} r^{1/2}(\Lambda R)^{1/4}\left(\frac{\widetilde{\sigma}^2}{B}\right)^{1/4}\frac{1}{\sqrt{1+\beta}} T^{-1/4} \\ &+ \frac{2A}{K\Lambda}(\Lambda R)^{1/4} r^{1/2}\left(\frac{\widetilde{\sigma}^2}{B}\right)^{1/4} T^{-1/4}.\end{aligned} \tag{93}$$

Consequently,

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla L(W_{t-1})\|_* = \mathcal{O}(T^{-1/4}),$$

and therefore FISMO reaches an ε*-stationary point in nuclear norm within* $T = \mathcal{O}(\varepsilon^{-4})$ *iterations.*

Proof. Under the stated initialization, Theorem F.13 gives

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla L(W_{t-1})\|_* \leq \frac{R}{K\eta T} + \frac{Lr\overline{K}^2}{2K}\eta + \frac{2}{K}\sqrt{\frac{r(1-\beta)}{1+\beta}}\frac{\widetilde{\sigma}}{\sqrt{B}} + \frac{2A}{K}\frac{\beta}{1-\beta}\eta.$$

Substituting $\eta = \sqrt{\frac{(1-\beta)R}{rT\Lambda}}$, we obtain

$$\frac{R}{K\eta T} = \frac{1}{K}\sqrt{\frac{r\Lambda R}{(1-\beta)T}},$$

$$\frac{Lr\overline{K}^2}{2K}\eta \leq \frac{r\Lambda}{2K}\eta = \frac{1}{2K}\sqrt{\frac{r\Lambda R(1-\beta)}{T}},$$

and, using $\beta \leq 1$,

$$\frac{2A}{K}\frac{\beta}{1-\beta}\eta \leq \frac{2A}{K}\frac{1}{1-\beta}\eta = \frac{2A}{K\Lambda}\sqrt{\frac{r\Lambda R}{(1-\beta)T}}.$$

Hence

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla L(W_{t-1})\|_* \leq \frac{1}{K}\sqrt{\frac{r\Lambda R}{(1-\beta)T}} + \frac{1}{2K}\sqrt{\frac{r\Lambda R(1-\beta)}{T}} + \frac{2}{K}\sqrt{\frac{r(1-\beta)}{1+\beta}}\frac{\widetilde{\sigma}}{\sqrt{B}} + \frac{2A}{K\Lambda}\sqrt{\frac{r\Lambda R}{(1-\beta)T}}.$$

Now assume $T \geq \frac{\Lambda R B}{\widetilde{\sigma}^2}$, so that the minimum in the definition of β is attained by the first branch:

$$1-\beta = \frac{\sqrt{\Lambda R B}}{\widetilde{\sigma}\sqrt{T}}.$$

We substitute this identity into each term.

For the first term,

$$\frac{1}{K}\sqrt{\frac{r\Lambda R}{(1-\beta)T}} = \frac{1}{K}\sqrt{\frac{r\Lambda R}{\left(\frac{\sqrt{\Lambda R B}}{\widetilde{\sigma}\sqrt{T}}\right)T}}$$

$$= \frac{1}{K}\sqrt{r\Lambda R \cdot \frac{\widetilde{\sigma}}{\sqrt{\Lambda R\, B}} \cdot \frac{1}{\sqrt{T}}} = \frac{1}{K}\,(\Lambda R)^{1/4}\, r^{1/2} \left(\frac{\widetilde{\sigma}^2}{B}\right)^{1/4} T^{-1/4}.$$

For the second term,

$$\frac{1}{2K}\sqrt{\frac{r\Lambda R(1-\beta)}{T}} = \frac{1}{2K}\sqrt{\frac{r\Lambda R}{T} \cdot \frac{\sqrt{\Lambda R\, B}}{\widetilde{\sigma}\sqrt{T}}} = \frac{1}{2K}\, r^{1/2}\, (\Lambda R)^{3/4} \left(\frac{B}{\widetilde{\sigma}^2}\right)^{1/4} T^{-3/4}.$$

For the stochastic term,

$$\frac{2}{K}\sqrt{\frac{r(1-\beta)}{1+\beta}}\frac{\widetilde{\sigma}}{\sqrt{B}} = \frac{2}{K}\sqrt{\frac{r}{1+\beta} \cdot \frac{\sqrt{\Lambda R\, B}}{\widetilde{\sigma}\sqrt{T}}}\frac{\widetilde{\sigma}}{\sqrt{B}} = \frac{2}{K}\, r^{1/2}\, (\Lambda R)^{1/4} \left(\frac{\widetilde{\sigma}^2}{B}\right)^{1/4} \frac{1}{\sqrt{1+\beta}}\, T^{-1/4}.$$

For the fourth term,

$$\frac{2A}{K\Lambda}\sqrt{\frac{r\Lambda R}{(1-\beta)T}} = \frac{2A}{K\Lambda}\,(\Lambda R)^{1/4}\, r^{1/2} \left(\frac{\widetilde{\sigma}^2}{B}\right)^{1/4} T^{-1/4}.$$

Combining these four expressions proves the explicit bound as in Equation equation 93. Finally, since $\beta \in [0, 1)$, we have $1 + \beta \geq 1$, hence the third term is also of order $T^{-1/4}$. Therefore the dominant terms are of order $T^{-1/4}$, while the second term is of smaller order $T^{-3/4}$. It follows that

$$\frac{1}{T}\sum_{t=1}^{T} \mathbb{E}\|\nabla L(W_{t-1})\|_* = \mathcal{O}(T^{-1/4})$$

. Solving $T^{-1/4} \leq \varepsilon$ gives $T = \mathcal{O}(\varepsilon^{-4})$, which proves the complexity claim. □

G Pre-asymptotic Analysis of FISMO

In this section, we provide a theoretical justification of the pre-asymptotic advantages of FISMO against Muon as shown in Theorem 5.10. The proof is followed by the schemes from Martens (2020).

In particular, we consider the following simplified local quadratic model:

$$f(W) = f(W^\star) + \frac{1}{2}(w - w^\star)^\top H^\star (w - w^\star), \tag{94}$$

where $w = \mathrm{vec}(W)$, $w^\star = \mathrm{vec}(W^\star)$, and $H^\star \in \mathbb{S}^{mn}_{++}$. Hence $\nabla f(w) = H^\star(w - w^\star)$.

We consider the main structure of FISMO, ignoring momentum and the slow drift of the EMA preconditioners. For fixed positive definite matrices $P \in \mathbb{S}^m_{++}$ and $Q \in \mathbb{S}^n_{++}$, define $B_F := Q \otimes P$. The core update is $W_{t+1} = W_t - \eta_t P^{-1} G_t Q^{-1}$, where G_t is a stochastic gradient matrix. By vectorization, this update can be written as

$$w_{t+1} = w_t - \eta_t B_F^{-1} g_t, \qquad g_t := \mathrm{vec}(G_t).$$

Thus, $B_F = Q \otimes P$ plays the role of a Fisher-structured curvature matrix.

We also made the following assumption more clearly:

Assumption G.1 (Unbiased stochastic gradient)**.** *The stochastic gradient satisfies*

$$g_t = \nabla f(w_t) + \xi_t, \qquad \mathbb{E}[\xi_t \mid w_t] = 0, \qquad \mathbb{E}[\xi_t \xi_t^\top \mid w_t] = \Sigma_g,$$

where $\Sigma_g \succeq 0$ is independent of w_t within the local region.

Assumption G.2 (Stable Kronecker Fisher approximation)**.** *The Kronecker Fisher preconditioner $B_F = Q \otimes P$ is fixed in the local region and spectrally approximates the local curvature: there exist constants $0 < \rho_- \leq \rho_+$ such that*

$$\rho_- B_F \preceq H^\star \preceq \rho_+ B_F.$$

Equivalently,

$$\lambda\big(B_F^{-1} H^\star\big) \subseteq [\rho_-, \rho_+].$$

Remark G.3. *Assumption G.2 does not require the Kronecker factorization to exactly match the full curvature. It only requires spectral equivalence. When $\rho_+/\rho_- \ll \kappa(H^\star)$, the Fisher-structured geometry significantly improves the conditioning of the local problem.*

G.1 WITHOUT ORTHOGONALIZATION

Lemma G.4 (Vectorized form)**.** *Let $B_F = Q \otimes P$. The matrix update*

$$W_{t+1} = W_t - \eta_t P^{-1} G_t Q^{-1}$$

is equivalent to

$$w_{t+1} = w_t - \eta_t B_F^{-1} g_t, \qquad g_t = \mathrm{vec}(G_t).$$

Proof. Using the identity

$$\mathrm{vec}(AGB) = (B^\top \otimes A)\,\mathrm{vec}(G),$$

with $A = P^{-1}$ and $B = Q^{-1}$, and using the symmetry of Q, we obtain

$$\mathrm{vec}(P^{-1} G_t Q^{-1}) = (Q^{-1} \otimes P^{-1})\,\mathrm{vec}(G_t).$$

Since

$$(Q \otimes P)^{-1} = Q^{-1} \otimes P^{-1},$$

the claim follows. □

Theorem G.5 (Local stochastic convergence of the FISMO core)**.** *Suppose Assumptions G.1–G.2 hold. Consider the FISMO core iteration*

$$w_{t+1} = w_t - \eta_t B_F^{-1} g_t, \qquad B_F = Q \otimes P.$$

Define

$$U_F := B_F^{-1} \Sigma_g B_F^{-1}.$$

(i) Constant learning rate. *If $\eta_t = \eta$ and*

$$\eta \lambda_{\max}(B_F^{-1} H^\star) \le 1,$$

then

$$\ell_F(t) \le \mathbb{E}[f(W_t)] - f(W^\star) \le u_F(t),$$

where

$$u_F(t) = \left[1 - (1-\varepsilon_1)^{2t}\right] \frac{\eta}{4} \,\mathrm{tr}\left[\left(B_F - \frac{\eta}{2} H^\star\right)^{-1} \Sigma_g\right] + (1-\varepsilon_2)^{2t}\big(f(W_0) - f(W^\star)\big),$$

and

$$\ell_F(t) = \left[1 - (1-\varepsilon_2)^{2t}\right] \frac{\eta}{4} \,\mathrm{tr}\left[\left(B_F - \frac{\eta}{2} H^\star\right)^{-1} \Sigma_g\right] + (1-\varepsilon_1)^{2t}\big(f(W_0) - f(W^\star)\big),$$

with

$$\varepsilon_1 = \eta \lambda_{\max}(B_F^{-1} H^\star), \qquad \varepsilon_2 = \eta \lambda_{\min}(B_F^{-1} H^\star).$$

(ii) Decaying learning rate. *If the learning rate is set to be $\eta_t = \frac{1}{t+a+1}$, and we assume that $b_F := \lambda_{\min}(B_F^{-1} H^\star) > \frac{1}{2}$, and $\lambda_{\max}(B_F^{-1} H^\star) \le a+1$, then*

$$\mathbb{E}[f(W_t)] - f(W^\star) \le \frac{1}{4(t+a)} \,\mathrm{tr}\left[\left(B_F^{-1} - \frac{1}{2}(H^\star)^{-1}\right)^{-1} U_F\right] + \mathcal{O}\left(\frac{1}{(t+a)^2}\right) + \big(f(W_0) - f(W^\star)\big)\left(\frac{a+1}{t+a}\right)^{2b_F}.$$

In particular,

$$\mathbb{E}[f(W_t)] - f(W^\star) = \mathcal{O}\left(\frac{1}{t}\right) + \mathcal{O}\big(t^{-2b_F}\big),$$

where the $\mathcal{O}(1/t)$ term is noise-dependent and the $\mathcal{O}(t^{-2b_F})$ term is initialization-dependent.

Proof. Denote $e_t := w_t - w^\star$. We have $\nabla f(w_t) = H^\star e_t$. Using Assumption G.1, we have

$$g_t = H^\star e_t + \xi_t.$$

Therefore,

$$\begin{aligned} e_{t+1} &= e_t - \eta_t B_F^{-1}(H^\star e_t + \xi_t) \\ &= (I - \eta_t B_F^{-1} H^\star) e_t - \eta_t B_F^{-1}\xi_t. \end{aligned}$$

Taking conditional expectation and using $\mathbb{E}[\xi_t \mid w_t] = 0$, we obtain the mean recursion

$$\mathbb{E}[e_{t+1}] = (I - \eta_t B_F^{-1} H^\star)\mathbb{E}[e_t].$$

For the simplicity of notation, we denote

$$A_t := I - \eta_t B_F^{-1} H^*.$$

Since the gradient is stochastic, we also need to evaluate the variance formulas, denote $\mu_t = \mathbb{E}[e_t]$, and

$$V_t := \mathrm{Var}(e_t) = \mathbb{E}[(e_t - \mu_t)(e_t - \mu_t)^\top].$$

Notice that

$$\begin{aligned} e_{t+1} - \mu_{t+1} &= A_t e_t - \eta_t B_F^{-1}\xi_t - A_t\mu_t, \\ &= A_t(e_t - \mu_t) - \eta_t B_F^{-1}\xi_t, \end{aligned}$$

we have

$$\begin{aligned} V_{t+1} &= \mathbb{E}\left[[A_t(e_t - \mu_t) - \eta_t B_F^{-1}\xi_t][A_t(e_t - \mu_t) - \eta_t B_F^{-1}\xi_t]^\top\right] \\ &= A_t V_t A_t^\top + \eta_t^2 B_F^{-1}\mathbb{E}[\xi_t\xi_t^\top]B_F^{-1}. \end{aligned}$$

Note that the two middle terms vanishes since

$$\mathbb{E}\left[(e_t - \mu_t)\xi_t^\top\right] = \mathbb{E}\left[\mathbb{E}\left[(e_t - \mu_t)\xi_t^\top | w_t\right]\right],$$

where e_t is determined for fixed w_t.

Therefore, we obtain the variance recursion formula

$$V_{t+1} = A_t V_t A_t^\top + \eta_t^2 U_F, \qquad A_t := I - \eta_t B_F^{-1} H^\star,$$

where

$$U_F = B_F^{-1}\Sigma_g B_F^{-1}.$$

Moreover, for the locally quadratic model, we have

$$f(W_t) - f(W^\star) = \frac{1}{2} e_t^\top H^\star e_t.$$

Since

$$\mathbb{E}[e_t^\top H^* e_t] = \mathrm{tr}(H^*\mathbb{E}[e_t e_t^T]) = \mathrm{tr}(H^*(V_t + \mu_t\mu_t^\top)),$$

we have the decomposition

$$\mathbb{E}[f(W_t)] - f(W^\star) = \frac{1}{2}\,\mathrm{tr}(H^\star V_t) + \frac{1}{2}\mathbb{E}[e_t]^\top H^\star \mathbb{E}[e_t].$$

The first term is the noise-dependent term, while the second term is the initialization-dependent term.

Constant learning rate. Let $\eta_t = \eta$ and define

$$C_F := \eta(H^\star)^{1/2} B_F^{-1}(H^\star)^{1/2}.$$

Then the eigenvalues of C_F are those of $\eta B_F^{-1} H^\star$. Hence

$$\lambda_{\max}(C_F) = \varepsilon_1, \qquad \lambda_{\min}(C_F) = \varepsilon_2.$$

The stability assumption implies $0 \preceq C_F \preceq I$.

For the mean term,

$$\mathbb{E}[e_t] = (I - \eta B_F^{-1} H^\star)^t e_0.$$

Equivalently,

$$(H^\star)^{1/2}\mathbb{E}[e_t] = (I - C_F)^t (H^\star)^{1/2} e_0.$$

Therefore,

$$\begin{aligned}(1-\varepsilon_1)^{2t}\big(f(W_0) - f(W^\star)\big) &\le \frac{1}{2}\mathbb{E}[e_t]^\top H^\star \mathbb{E}[e_t] \\ &\le (1-\varepsilon_2)^{2t}\big(f(W_0) - f(W^\star)\big).\end{aligned}$$

Next we consider the variance term. Remind that $A := I - \eta B_F^{-1} H^\star$. Denote $\Lambda(X) := AXA^\top$. The variance recursion becomes

$$V_{t+1} = \Lambda(V_t) + \eta^2 U_F.$$

Its stationary solution is

$$V_\infty = \eta^2 (I - \Lambda)^{-1}(U_F).$$

Since $V_0 = 0$, by the recursive formula we have

$$V_t = \eta^2 \sum_{t=0}^{t-1} \Lambda^i (U_F).$$

Notice that $(1-\Lambda)^{-1} = \sum_{i=0}^{\infty} \Lambda^i$, we have

$$V_t = (I - \Lambda^t)(V_\infty).$$

Using the standard Lyapunov trace identity,

$$\operatorname{tr}\big(H^\star (I-\Lambda)^{-1}(X)\big) = \frac{1}{2\eta} \operatorname{tr}\left[\left(B_F - \frac{\eta}{2} H^\star\right)^{-1} B_F X B_F\right],$$

with $X = U_F = B_F^{-1} \Sigma_g B_F^{-1}$, we get

$$\frac{1}{2}\operatorname{tr}(H^\star V_\infty) = \frac{\eta}{4} \operatorname{tr}\left[\left(B_F - \frac{\eta}{2} H^\star\right)^{-1} \Sigma_g\right].$$

Using the eigenvalue bounds of $I - C_F$, the variance contribution is bounded between

$$\left[1 - (1-\varepsilon_2)^{2t}\right] \frac{\eta}{4} \operatorname{tr}\left[\left(B_F - \frac{\eta}{2} H^\star\right)^{-1} \Sigma_g\right]$$

and

$$\left[1 - (1-\varepsilon_1)^{2t}\right] \frac{\eta}{4} \operatorname{tr}\left[\left(B_F - \frac{\eta}{2} H^\star\right)^{-1} \Sigma_g\right].$$

Combining the bounds for the mean term and the variance term gives the constant learning-rate result.

Decaying learning rate. Now let

$$\eta_t = \frac{1}{t + a + 1}.$$

The mean recursion gives

$$\mathbb{E}[e_t] = \prod_{j=0}^{t-1} \left(I - \frac{1}{j+a+1} B_F^{-1} H^\star\right) e_0.$$

Denote

$$b_F = \lambda_{\min}(B_F^{-1} H^\star).$$

Using the spectral assumption

$$\lambda_{\max}(B_F^{-1} H^\star) \le a + 1,$$

and the standard product bound

$$\prod_{j=0}^{t-1} \left(1 - \frac{b_F}{j+a+1}\right) \le \left(\frac{a+1}{t+a}\right)^{b_F},$$

we obtain

$$\frac{1}{2}\mathbb{E}[e_t]^\top H^\star \mathbb{E}[e_t] \le \big(f(W_0) - f(W^\star)\big)\left(\frac{a+1}{t+a}\right)^{2b_F}.$$

For the variance term, define the linear operator

$$\Xi_F(X) := B_F^{-1}H^\star X + X H^\star B_F^{-1}.$$

Since $b_F > 1/2$, the operator $\Xi_F - I$ is invertible on symmetric matrices. The variance recursion implies the asymptotic expansion

$$V_t = \frac{1}{t+a}(\Xi_F - I)^{-1}(U_F) + \mathcal{O}\left(\frac{1}{(t+a)^2}\right).$$

Therefore,

$$\frac{1}{2}\operatorname{tr}(H^\star V_t) = \frac{1}{2(t+a)}\operatorname{tr}\left[H^\star(\Xi_F - I)^{-1}(U_F)\right] + \mathcal{O}\left(\frac{1}{(t+a)^2}\right).$$

Using the trace identity

$$\operatorname{tr}\left[H^\star(\Xi_F - I)^{-1}(X)\right] = \frac{1}{2}\operatorname{tr}\left[\left(B_F^{-1} - \frac{1}{2}(H^\star)^{-1}\right)^{-1} X\right],$$

with $X = U_F$, we get

$$\frac{1}{2}\operatorname{tr}(H^\star V_t) = \frac{1}{4(t+a)}\operatorname{tr}\left[\left(B_F^{-1} - \frac{1}{2}(H^\star)^{-1}\right)^{-1} U_F\right] + \mathcal{O}\left(\frac{1}{(t+a)^2}\right).$$

Combining this with the mean-term bound proves the decaying learning-rate result. □

Corollary G.6 (Pre-asymptotic advantage over isotropic geometry)**.** *Under Assumption G.2,*

$$\kappa(B_F^{-1}H^\star) \le \frac{\rho_+}{\rho_-}.$$

Hence the initialization-dependent term in Theorem G.5 decays with exponent

$$2b_F = 2\lambda_{\min}(B_F^{-1}H^\star) \ge 2\rho_-.$$

In contrast, an isotropic method (like Muon) with $B_M = cI$ has

$$b_M = \lambda_{\min}(B_M^{-1}H^\star) = \frac{\lambda_{\min}(H^\star)}{c}.$$

With the stability-motivated choice $c = \lambda_{\max}(H^\star)$, this becomes

$$b_M = \frac{1}{\kappa(H^\star)}.$$

Therefore, when the Kronecker Fisher structure captures the anisotropic curvature so that $\rho_- > \frac{1}{\kappa(H^\star)}$, FISMO has a sharper initialization-dependent finite-time bound than isotropic preconditioned methods.

Remark G.7 (Connection to Muon)**.** *Muon performs orthogonalization in the original parameter geometry, whereas FISMO first whitens the gradient using the Fisher-structured metric $Q \otimes P$, applies an orthogonalized update in the whitened coordinates, and maps the update back to the original coordinates. The above corollary formalizes the advantage of this anisotropic geometry: if $Q \otimes P$ spectrally captures the local curvature, then the effective curvature $B_F^{-1}H^\star$ is better conditioned than the raw curvature $H^\star$, leading to faster decay of the starting-point-dependent error.*

Remark G.8 (From the core update to the full FISMO algorithm)**.** *Theorem G.5 analyzes the frozen-preconditioner linear core of FISMO. The full algorithm contains three additional components: EMA updates of P_t, Q_t, momentum in whitened gradient and polar projection. These components can be viewed as perturbations of the core recursion under suitable tracking assumptions. For example, if*

$$\|B_t^{-1} - B_F^{-1}\|_{H^\star \to H^\star} \le \delta_t, \qquad B_t := Q_t \otimes P_t,$$

and $\sum_{t=0}^{T-1}\eta_t\delta_t$ is small, the same decomposition remains valid up to an additional perturbation term.

G.2 With Orthogonalization

We provide a local quadratic analysis to explain why performing the polar update in a Fisher-whitened geometry can be more favorable than performing it in the isotropic parameter geometry, such as Muon.

Let $W \in \mathbb{R}^{m\times n}$, $w = \mathrm{vec}(W)$, and consider the local quadratic model around a local minimizer $W^\star$:

$$f(W) - f(W^\star) = \frac{1}{2}(w - w^\star)^\top H^\star (w - w^\star), \qquad H^\star \succ 0.$$

For any symmetric positive definite matrix $B \succ 0$, define the whitened coordinate

$$z_B := B^{1/2}(w - w^\star),$$

and the corresponding effective curvature

$$\bar{H}_B := B^{-1/2} H^\star B^{-1/2}.$$

Then the local objective can be written as

$$\phi_B(z_B) := f(W) - f(W^\star) = \frac{1}{2} z_B^\top \bar{H}_B z_B.$$

Its gradient in the whitened coordinate is

$$s_B := \nabla_{z_B} \phi_B(z_B) = \bar{H}_B z_B.$$

Let

$$S_B := \mathrm{mat}(s_B) \in \mathbb{R}^{m\times n},$$

where $\mathrm{mat}(\cdot)$ reshapes a vectorized matrix back to its matrix form. We also define

$$L_B := \lambda_{\max}(\bar{H}_B), \qquad \mu_B := \lambda_{\min}(\bar{H}_B), \qquad \kappa_B := \frac{L_B}{\mu_B},$$

and $r := \min\{m, n\}$.

For this whitened geometry, consider the Orth-B update

$$z_B^+ = z_B - \eta u_B,$$

where $u_B := \mathrm{vec}(\mathrm{Orth}(S_B))$. Equivalently, in the original parameter coordinate,

$$w^+ = w - \eta B^{-1/2} u_B.$$

Theorem G.9 (Local contraction of Orth updates under a preconditioned geometry)**.** *Under the local quadratic model above, the following statements hold.*

1. General Orth-B descent formula. *For any $B \succ 0$, the Orth-B update satisfies*

$$\phi_B(z_B^+) \le \phi_B(z_B) - \eta \|S_B\|_* + \frac{\eta^2}{2} L_B r.$$

Moreover, if we choose the local normalized step size $\eta_B^\star := \frac{\|S_B\|_*}{L_B r}$, then

$$\phi_B(z_B^+) \le \left(1 - \frac{1}{r\kappa_B}\right) \phi_B(z_B).$$

Consequently, after t steps with the same local contraction condition,

$$\phi_B(z_{B,t}) \le \left(1 - \frac{1}{r\kappa_B}\right)^t \phi_B(z_{B,0}).$$

2. Specialization to FISMO and isotropic orthogonalization. *For FISMO, the Fisher-structured preconditioner is $B_F := Q \otimes P$, and therefore*

$$\bar{H}_F := B_F^{-1/2} H^\star B_F^{-1/2} = (Q \otimes P)^{-1/2} H^\star (Q \otimes P)^{-1/2}.$$

The corresponding local contraction factor is

$$1 - \frac{1}{r\kappa_F}, \qquad \kappa_F := \kappa\Big((Q \otimes P)^{-1/2} H^\star (Q \otimes P)^{-1/2}\Big).$$

For isotropic update methods such as Muon, the update is performed in the original parameter geometry. This corresponds to $B = I$, and hence $\bar{H}_{\mathrm{iso}} = H^\star, \kappa_{\mathrm{iso}} = \kappa(H^\star)$. Thus, the isotropic local contraction factor is $1 - \frac{1}{r\kappa(H^\star)}$.

Therefore, if the Fisher-structured Kronecker preconditioner improves the effective conditioning, namely

$$\kappa_F = \kappa\Big((Q \otimes P)^{-1/2} H^\star (Q \otimes P)^{-1/2}\Big) < \kappa(H^\star),$$

then FISMO has a strictly sharper local contraction factor than isotropic orthogonalization:

$$1 - \frac{1}{r\kappa_F} < 1 - \frac{1}{r\kappa(H^\star)}.$$

In particular,

$$f(W_t^{\mathrm{FISMO}}) - f(W^\star) \le \left(1 - \frac{1}{r\kappa_F}\right)^t \big(f(W_0) - f(W^\star)\big),$$

whereas the corresponding isotropic bound is

$$f(W_t^{\mathrm{iso}}) - f(W^\star) \le \left(1 - \frac{1}{r\kappa(H^\star)}\right)^t \big(f(W_0) - f(W^\star)\big).$$

Hence, whenever $Q \otimes P$ reduces the effective condition number of the local curvature, Orth-FISMO enjoys a strictly better local pre-asymptotic contraction bound than Muon-style isotropic orthogonalization.

Proof. We first prove the general Orth-B formula. Since $\phi_B(z_B) = \frac{1}{2} z_B^\top \bar{H}_B z_B$, we have the exact quadratic expansion

$$\phi_B(z_B^+) = \phi_B(z_B - \eta u_B) = \phi_B(z_B) - \eta \langle \nabla_{z_B} \phi_B(z_B), u_B \rangle + \frac{\eta^2}{2} u_B^\top \bar{H}_B u_B.$$

By definition,

$$\nabla_{z_B} \phi_B(z_B) = s_B, \qquad u_B = \mathrm{vec}(\mathrm{Orth}(S_B)).$$

Using the standard polar identity $\langle S, \mathrm{Orth}(S) \rangle = \|S\|_*$, we obtain

$$\langle \nabla_{z_B} \phi_B(z_B), u_B \rangle = \langle S_B, \mathrm{Orth}(S_B) \rangle = \|S_B\|_*.$$

Moreover, $u_B^\top \bar{H}_B u_B \le L_B \|u_B\|_2^2$. Since

$$\|u_B\|_2^2 = \|\,\mathrm{Orth}(S_B)\|_F^2 \le r,$$

we have $u_B^\top \bar{H}_B u_B \le L_B r$. Therefore,

$$\phi_B(z_B^+) \le \phi_B(z_B) - \eta \|S_B\|_* + \frac{\eta^2}{2} L_B r.$$

Now choose $\eta = \eta_B^\star = \frac{\|S_B\|_*}{L_B r}$. Then

$$\phi_B(z_B^+) \le \phi_B(z_B) - \frac{\|S_B\|_*^2}{2 L_B r}.$$

Next, since $\|S_B\|_* \ge \|S_B\|_F = \|s_B\|_2$, we have

$$\|S_B\|_*^2 \ge \|s_B\|_2^2.$$

Because $s_B = \bar{H}_B z_B$, it follows that

$$\|s_B\|_2^2 = z_B^\top \bar{H}_B^2 z_B \geq \mu_B z_B^\top \bar{H}_B z_B.$$

Using $\phi_B(z_B) = \frac{1}{2} z_B^\top \bar{H}_B z_B$, we obtain

$$\|S_B\|_*^2 \geq 2\mu_B \phi_B(z_B).$$

Therefore,

$$\phi_B(z_B^+) \leq \phi_B(z_B) - \frac{2\mu_B \phi_B(z_B)}{2L_B r} = \left(1 - \frac{\mu_B}{L_B r}\right) \phi_B(z_B) = \left(1 - \frac{1}{r\kappa_B}\right) \phi_B(z_B).$$

Iterating this inequality gives the multi-step contraction bound.

We now specialize the result. For Orth-FISMO, $B = B_F = Q \otimes P$, so the effective curvature is

$$\bar{H}_F = (Q \otimes P)^{-1/2} H^\star (Q \otimes P)^{-1/2},$$

and the contraction factor is

$$1 - \frac{1}{r\kappa_F}.$$

For isotropic orthogonalization ($B = I$), $\bar{H}_{\text{iso}} = H^\star$, and $\kappa_{\text{iso}} = \kappa(H^\star)$, and contraction factor is

$$1 - \frac{1}{r\kappa(H^\star)}.$$

If $\kappa_F < \kappa(H^\star)$, then $1 - \frac{1}{r\kappa_F} < 1 - \frac{1}{r\kappa(H^\star)}$. Thus Orth-FISMO has a strictly sharper local contraction factor than isotropic orthogonalization whenever the Fisher-structured metric reduces the effective curvature condition number. □