Proceedings of Machine Learning Research vol 336:1–43, 2026 39th Annual Conference on Learning Theory

Adaptive Matrix Online Learning through Smoothing with Guarantees for Nonsmooth Nonconvex Optimization

Ruichen Jiang RJIANG@UTEXAS.EDU
The University of Texas at Austin

Zakaria Mhammedi MHAMMEDI@GOOGLE.COM
Google Research

Mehryar Mohri MOHRI@GOOGLE.COM
Google Research & Courant Institute of Mathematical Sciences

Aryan Mokhtari MOKHTARI@AUSTIN.UTEXAS.EDU
The University of Texas at Austin & Google Research

Editors: Steve Hanneke and Tor Lattimore

Abstract

We study online linear optimization with matrix variables constrained by the *operator norm*, a setting where the geometry renders designing *data-dependent* and *efficient* adaptive algorithms challenging. The best-known adaptive regret bounds are achieved by Shampoo-like methods, but they require solving a costly quadratic projection subproblem. To address this, we extend the gradient-based prediction scheme to adaptive matrix online learning and cast algorithm design as constructing a family of smoothed potentials for the nuclear norm. We define a notion of admissibility for such smoothings and prove any admissible smoothing yields a regret bound matching the best-known guarantees of one-sided Shampoo. We instantiate this framework with two efficient methods that avoid quadratic projections. The first is an adaptive Follow-the-Perturbed-Leader (FTPL) method using Gaussian stochastic smoothing. The second is Follow-the-Augmented-Matrix-Leader (FAML), which uses a deterministic hyperbolic smoothing in an augmented matrix space. By analyzing the admissibility of these smoothings, we show both methods admit closed-form updates and match one-sided Shampoo's regret up to a constant factor, while significantly reducing computational cost. Lastly, using the online-to-nonconvex conversion, we derive two matrix-based optimizers, *Pion* (from FTPL) and *Leon* (from FAML). We prove convergence guarantees for these methods in nonsmooth nonconvex settings, a guarantee that the popular Muon optimizer lacks.

Keywords: Matrix online learning, Adaptive regret, Operator-norm constraints

1. Introduction

While Online Linear Optimization (OLO) is well-established for vector spaces in $\mathbb{R}^d$, modern applications such as neural network training are often natively cast in terms of matrix variables $\mathbf{X} \in \mathbb{R}^{m\times n}$. In this context, relying on a naive reduction to the vector setting is problematic. Vectorization obscures intrinsic spectral structures, most notably operator-norm constraints, that are distinct from Euclidean geometry and essential for efficient optimization. To avoid the suboptimal regret bounds inherent in such reductions, we investigate OLO directly within the matrix domain. The problem is formally defined as follows:

The authors are listed in alphabetical order. The work of RJ was partially done while he was a Student Researcher at Google Research.

Matrix Online Linear Optimization
For $t = 1, \ldots, T$:

- Learner chooses a matrix $\mathbf{X}_t$ from a decision set $\mathcal{X}$;
- Environment selects a gradient matrix $\mathbf{G}_t \in \mathbb{R}^{m \times n}$ defining the linear loss $\ell_t(\mathbf{X}) = \langle \mathbf{G}_t, \mathbf{X} \rangle$;

Goal: Minimize $\mathrm{Reg}_T = \sum_{t=1}^{T} \langle \mathbf{G}_t, \mathbf{X}_t \rangle - \min_{\mathbf{X} \in \mathcal{X}} \sum_{t=1}^{T} \langle \mathbf{G}_t, \mathbf{X} \rangle$.

Matrix OLO has been extensively studied for problems like online variance minimization (Warmuth and Kuzmin, 2006), PCA (Warmuth and Kuzmin, 2008), and collaborative filtering (Hazan et al., 2012), typically under nuclear-norm constraints. To address these, matrix multiplicative weight updates (Arora and Kale, 2007; Arora et al., 2012) have been proposed as a natural generalization of the classical Hedge algorithm. Matrix OLO with Schatten-p constraints has also been applied to multi-task classification (Agarwal et al., 2008; Cavallanti et al., 2010), for which Kakade et al. (2012) developed mirror descent algorithms induced by matrix Bregman divergences.

In this paper, we focus on the case where the decision set $\mathcal{X}$ is an operator-norm ball, i.e., $\mathcal{X} = \{\mathbf{X} \in \mathbb{R}^{m \times n} : \|\mathbf{X}\|_{\mathrm{op}} \leq D\}$. As discussed in Section 1.1 and detailed further in Appendix A, this constraint naturally arises in classical problems such as learning rotations, as well as in modern optimization applications including the design of preconditioned gradient methods and neural network training. Our goal is to design adaptive online algorithms with data-dependent regret guarantees, in the spirit of AdaGrad (McMahan and Streeter, 2010; Duchi et al., 2011). In the matrix setting with operator-norm constraints, the strongest known regret guarantees are achieved by the one-sided Shampoo/ASGO methods proposed concurrently in (Xie et al., 2025; An et al., 2025). Building on the Shampoo preconditioner (Gupta et al., 2018), these methods attain a regret bound of $\mathcal{O}(D \operatorname{Tr}(\sqrt{\mathbf{M}_T}))$, where $\mathbf{M}_T = \sum_{t=1}^{T} \mathbf{G}_t \mathbf{G}_t^\top$, thereby offering better adaptivity over AdaGrad. However, enforcing feasibility requires solving a costly quadratic projection subproblem, which lacks a closed-form solution. This motivates the question: *can one retain the same adaptive regret guarantees with greater computational efficiency?*

Contributions We develop a unified framework for adaptive matrix online learning under operator-norm constraints that avoids quadratic projections while matching the best-known regret guarantees of one-sided Shampoo, answering the above question affirmatively. Our contributions are threefold.

First, we generalize the *Gradient-Based Prediction Algorithm* (GBPA) (Abernethy et al., 2016) to adaptive matrix online algorithms, where algorithm design is cast as constructing a family of smoothed potentials parameterized by a positive semidefinite (PSD) matrix $\mathbf{L}$ that captures problem geometry. We formalize the notion of (α, β)-*admissible smoothings* of the *nuclear norm* induced by the operator-norm constraint. We show that, when the parameter $\mathbf{L}$ is chosen adaptively, GBPA with any admissible potential family attains the regret bound $\mathrm{Reg}_T = \mathcal{O}\big(\sqrt{\alpha\beta}\, D \operatorname{Tr}(\sqrt{\mathbf{M}_T})\big)$, matching one-sided Shampoo up to constants (Theorem 4). Since the bound depends only on the product $\alpha\beta$, we characterize optimal admissibility by proving that any (α, β)-admissible smoothing must satisfy $\alpha\beta \geq \frac{1}{2}$, and by exhibiting a regularized smoothing that attains this lower bound (Proposition 5).

Second, to avoid the prohibitive cost of quadratic subproblems in one-sided Shampoo, we introduce two efficient online algorithms based on novel smoothed potentials. The first uses Gaussian stochastic smoothing, leading to an adaptive Follow-the-Perturbed-Leader (FTPL) algorithm that is parallelizable and relies on efficient matrix primitives; using noncentral Wishart theory, we show

that this smoothing is admissible up to a mild dimension-dependent factor (Theorem 6). Our main algorithm, Follow-the-Augmented-Matrix-Leader (FAML), is based on a *deterministic* and *explicit* hyperbolic smoothing tailored to the nuclear norm. By lifting to an augmented space, FAML admits closed-form updates, avoiding quadratic projections. It achieves near-optimal admissibility (Theorem 8) and matches one-sided Shampoo's regret up to a factor of two at substantially lower cost.

Finally, we leverage our adaptive matrix online learning framework and apply the *Online-to-Nonconvex Conversion* (O2NC) paradigm (Cutkosky et al., 2023) to obtain efficient matrix-based optimizers with provable guarantees for *nonsmooth nonconvex* optimization. We formally identify the popular Muon optimizer as an instance of spectrally constrained FTL, clarifying its lack of guarantees in nonsmooth settings. In contrast, we introduce Pion (derived from FTPL) and Leon (derived from FAML), and establish that both converge to (ρ, ε)-stationary points for general nonsmooth nonconvex objectives (Theorems 12 and 13).

1.1. Motivating Problems

Next, we focus on matrix-based deep learning optimization as the primary instance of the matrix OLO problem subject to operator-norm constraints. Additional applications, including learning rotations and online quasi-Newton updates, are deferred to Appendix A due to space constraints.

Matrix Optimization Algorithms Deep learning architectures are inherently matrix-valued. Spectral optimizers like Muon (Jordan et al., 2024) exploit this structure, demonstrating significant empirical advantages over element-wise baselines. By leveraging the Online-to-Nonconvex Conversion (O2NC) framework (Cutkosky et al., 2023; Ahn et al., 2025), these methods can be rigorously modeled as Matrix OLO instances, where minimizing regret guarantees convergence to Goldstein stationary points. As detailed in Section 5, Muon corresponds precisely to spectrally-constrained FTL. We use this connection to derive a new optimization method with rigorous convergence guarantees in the nonsmooth setting, addressing a key theoretical gap in the standard Muon algorithm.

2. Adaptive Matrix Online Learning

In the vector setting, it is known that achieving optimal regret requires aligning the adaptive preconditioner with the geometry of the constraint set; we defer a rigorous discussion of this alignment to Appendix B. Applying this principle to Matrix OLO via naive vectorization, however, encounters two fundamental barriers. First, treating an $m \times n$ matrix as a vector of dimension $d = mn$ implies a full preconditioner of size $d \times d$, incurring prohibitive storage and computational costs. Second, and more critically, vectorization obliterates the underlying spectral structure. Standard adaptive variants like Diagonal AdaGrad implicitly assume a hyper-rectangular constraint geometry, which is ill-suited for the spectral features of the operator-norm ball. Consequently, efficient optimization in this setting demands algorithms explicitly tailored to the matrix domain, which we review below.

A natural extension of adaptive Online Gradient Descent (OGD) to the matrix setting is characterized by the following update rule. Given a gradient matrix $\mathbf{G}_t \in \mathbb{R}^{m\times n}$, we compute:

$$\mathbf{X}_{t+1} = \arg\min_{\mathbf{X}\in\mathcal{X}} \Big\{ \langle \mathbf{G}_t, \mathbf{X} - \mathbf{X}_t \rangle + \frac{1}{2\eta} \operatorname{Tr}((\mathbf{X} - \mathbf{X}_t)^\top \mathbf{L}_t (\mathbf{X} - \mathbf{X}_t)\mathbf{R}_t) \Big\}, \tag{1}$$

where $\mathbf{L}_t \in \mathbb{S}_+^m$ and $\mathbf{R}_t \in \mathbb{S}_+^n$ denote the left and right preconditioners, respectively. This formulation admits a direct interpretation within the standard AdaGrad framework. Specifically, let

Algorithm	**Preconditioners $\mathbf{L}_t$ and $\mathbf{R}_t$**	**Regret Bound**
Shampoo† (Gupta et al., 2018)	$\mathbf{L}_t = \mathbf{M}_t^{1/4}, \mathbf{R}_t = \mathbf{N}_t^{1/4}$	$\|\mathcal{X}\|_{\mathrm{op}} \max_{\substack{(a,b)\in\{(F,F),\\(*,\mathrm{op}),(\mathrm{op},*)\}}} \|\mathbf{M}_T^{1/4}\|_a \, \|\mathbf{N}_T^{1/4}\|_b$
One-sided Shampoo (Xie et al., 2025; An et al., 2025)	$\mathbf{L}_t = \mathbf{M}_t^{1/2}, \mathbf{R}_t = \mathbf{I}$	$\|\mathcal{X}\|_{\mathrm{op}} \operatorname{Tr}(\mathbf{M}_T^{1/2})$

Table 1: Here, $\mathbf{M}_T = \sum_{t=1}^T \mathbf{G}_t\mathbf{G}_t^\top$ and $\mathbf{N}_T = \sum_{t=1}^T \mathbf{G}_t^\top\mathbf{G}_t$. Moreover, $\|\mathcal{X}\|_{\mathrm{op}} = \max_{\mathbf{X},\mathbf{Y}\in\mathcal{X}} \|\mathbf{X}-\mathbf{Y}\|_{\mathrm{op}}$. † The reported regret for Shampoo is proven in Theorem 16 (Appendix C) which improves the original bound.

$\boldsymbol{x} = \mathrm{vec}(\mathbf{X})$ and $\boldsymbol{g} = \mathrm{vec}(\mathbf{G})$ denote the vectorized decision variable and gradient obtained by stacking the columns of the matrix. The regularization term in (1) coincides with the quadratic form $\frac{1}{2\eta}(\boldsymbol{x}-\boldsymbol{x}_t)^\top\mathbf{H}_t(\boldsymbol{x}-\boldsymbol{x}_t)$ from vector AdaGrad, with the additional constraint that the preconditioner factorizes as a Kronecker product $\mathbf{H}_t = \mathbf{R}_t \otimes \mathbf{L}_t$. Under this structure, the penalty admits the efficient decomposition $(\boldsymbol{x}-\boldsymbol{x}_t)^\top(\mathbf{R}_t\otimes\mathbf{L}_t)(\boldsymbol{x}-\boldsymbol{x}_t) = \operatorname{Tr}\left((\mathbf{X}-\mathbf{X}_t)^\top\mathbf{L}_t(\mathbf{X}-\mathbf{X}_t)\mathbf{R}_t\right)$. Notably, Shampoo (Gupta et al., 2018) and its one-sided variants (Xie et al., 2025; An et al., 2025) correspond to different choices of $\mathbf{L}_t$ and $\mathbf{R}_t$; we summarize these selections and the resulting regret bounds in Table 1 and note that the latter achieves the strongest known guarantee.

Despite the success of these methods in mitigating vectorization overhead and preserving matrix structure, they face a fundamental obstruction when constrained to the operator-norm ball, i.e., $\mathcal{X} = \{\mathbf{X} \in \mathbb{R}^{m\times n} \colon \|\mathbf{X}\|_{\mathrm{op}} \le D\}$. In this regime, the update step in (1) necessitates minimizing a convex quadratic objective over the operator-norm constraint, a problem that admits no closed-form solution. While practitioners often skip this projection step for computational efficiency, such heuristics are known to invalidate worst-case regret guarantees (Orabona and Pál, 2018). Solving this subproblem at each round using iterative methods is also costly. Projection-based methods, such as accelerated gradient descent, achieve linear convergence but require a full Singular Value Decomposition (SVD) at every step. Moreover, the subproblem condition number depends on $\mathbf{H}_t = \mathbf{R}_t \otimes \mathbf{L}_t$ and in the worst case can scale as $\sqrt{T}$. As a result, reaching a desired accuracy may take $\tilde{O}(T^{1/4})$ projection steps per round, leading to a total of $\tilde{O}(T^{5/4})$ SVD computations over T rounds. Conversely, Frank–Wolfe methods replace projections with linear optimization oracles that admit efficient implementations via polar factorization, but only achieve sublinear $\mathcal{O}(1/K)$ convergence, requiring many oracle calls to attain high precision. We defer the detailed complexity analysis to Appendix F. The core challenge is therefore matching one-sided Shampoo's adaptive regret without incurring prohibitively expensive quadratic projections, which we address next.

Remark 1 *While FTRL with Kronecker regularizers is expected to match OGD-style Shampoo guarantees, a corresponding analysis appears to be missing in prior work. Our framework supplies such Shampoo-type bounds for adaptive matrix FTRL; however, it still requires solving the same expensive operator-norm-constrained quadratic subproblems, motivating the more efficient approach developed later in the paper.*

3. A Unified Framework for Adaptive Matrix Online Algorithms

Before addressing the bottleneck of quadratic projections, we introduce a broad class of algorithms for Matrix OLO. We establish their regret guarantees via a unified framework that isolates the struc-

ture essential for spectral adaptivity. In Section 4, we instantiate this framework to derive two concrete algorithms that effectively circumvent this computational barrier.

Our proposed framework can be considered as a generalization of the Gradient-Based Prediction Algorithm (GBPA) (Abernethy et al., 2016) to the matrix domain. By incorporating potentials that capture intrinsic spectral geometry, we specialize this approach to operator-norm constraints, enabling near-optimal adaptive guarantees that are unattainable via standard vector reductions.

Recalling the Matrix OLO formulation, a GBPA strategy generates the action $\mathbf{X}_{t+1}$ by evaluating the gradient of a convex potential $\widetilde{\Phi}_t : \mathbb{R}^{m\times n} \to \mathbb{R}$ at the cumulative gradient $\mathbf{S}_t = \sum_{s=1}^{t} \mathbf{G}_s$. To enforce the feasibility constraint $\|\mathbf{X}_{t+1}\|_{\mathrm{op}} \le D$, we restrict the potentials such that $\|\nabla\widetilde{\Phi}_t\|_{\mathrm{op}} \le 1$ and scale the output by D. This yields the explicit update:

$$\mathbf{X}_{t+1} = -D\nabla\widetilde{\Phi}_t(\mathbf{S}_t), \tag{2}$$

for $t \ge 1$, and the initial action is set as $\mathbf{X}_1 = 0$. Indeed, distinct choices of potential functions instantiate different algorithms for Matrix OLO. Now since our online learning problem is constrained to the operator-norm ball, we define the base potential function as $\Phi(\mathbf{S}) = \max_{\|\mathbf{X}\|_{\mathrm{op}}\le 1}\langle \mathbf{X}, \mathbf{S}\rangle$, which can be further simplified as $\Phi(\mathbf{S}) = \|\mathbf{S}\|_*$. Hence, the regret of the matrix OLO problem can be written as $\mathrm{Reg}_T = \sum_{t=1}^{T}\langle \mathbf{G}_t, \mathbf{X}_t\rangle + D\Phi(\mathbf{S}_T) = \sum_{t=1}^{T}\langle \mathbf{G}_t, \mathbf{X}_t\rangle + D\|\mathbf{S}_T\|_*$. The following lemma provides the central regret decomposition for this class of algorithms and is a direct corollary of (Abernethy et al., 2016, Lemma 1.2); the proof is deferred to Appendix D.1.

Lemma 2 *Define $\mathcal{B}_f(\mathbf{U} \,\|\, \mathbf{V}) := f(\mathbf{U}) - f(\mathbf{V}) - \langle \nabla f(\mathbf{V}), \mathbf{U} - \mathbf{V}\rangle$ as the Bregman divergence with respect to a function f. If $\{\mathbf{X}_t\}$ is generated by (2), then its regret can be decomposed as:*

$$\mathrm{Reg}_T = D\big(\Phi(\mathbf{S}_T) - \widetilde{\Phi}_T(\mathbf{S}_T)\big) + D\sum_{t=1}^{T-1} \mathcal{B}_{\widetilde{\Phi}_t}(\mathbf{S}_{t+1} \,\|\, \mathbf{S}_t) + D\sum_{t=1}^{T-1}\big(\widetilde{\Phi}_{t+1}(\mathbf{S}_{t+1}) - \widetilde{\Phi}_t(\mathbf{S}_{t+1})\big) + D\widetilde{\Phi}_1(\mathbf{G}_1).$$

This regret decomposition admits three interpretable terms. The first term, $D\big(\Phi(\mathbf{S}_T) - \widetilde{\Phi}_T(\mathbf{S}_T)\big)$, captures the underestimation error of the surrogate potential $\widetilde{\Phi}_T$ relative to the base potential Φ. The second term involves the Bregman divergence induced by $\widetilde{\Phi}_t$ and reflects the smoothness of the surrogate potential along the trajectory $\{\mathbf{S}_t\}$. The remaining term accounts for the temporal variation of the potential sequence $\{\widetilde{\Phi}_t\}$ and measures its stability across rounds.

To build intuition for choosing the surrogate potentials $\widetilde{\Phi}_t$, consider the simplest case where $\widetilde{\Phi}_t = \Phi$ for all $t \ge 1$. This choice recovers the classical Follow-the-Leader (FTL) algorithm, i.e.,

$$\mathbf{X}_{t+1} = \arg\min_{\|\mathbf{X}\|_{\mathrm{op}}\le D}\big\{\langle \textstyle\sum_{s=1}^{t} \mathbf{G}_s, \mathbf{X}\rangle\big\}.$$

In this case, both the underestimation error and the temporal variation term in (2) vanish, yielding $\mathrm{Reg}_T = D\sum_{t=1}^{T-1} \mathcal{B}_\Phi(\mathbf{S}_{t+1} \,\|\, \mathbf{S}_t) + D\Phi(\mathbf{G}_1)$. However, as the base potential Φ, i.e., the nuclear norm, is nonsmooth, this can be exploited by the adversary to incur a large Bregman divergence at each time step, leading to an $\Omega(T)$ regret bound.

This motivates the construction of a sequence of surrogate potentials $\{\widetilde{\Phi}_t\}$ that closely approximates the base potential Φ while enjoying favorable smoothness properties. This is our main point of departure from (Abernethy et al., 2016). There, the authors select $\{\widetilde{\Phi}_t\}$ from a scalar-parametrized family $\{\widetilde{\Phi}_\eta : \eta > 0\}$, achieving an adaptive regret bound similar to Scalar AdaGrad (in our setting, $\sqrt{\min\{m,n\}}D\sqrt{\sum_{t=1}^{T}\|\mathbf{G}_t\|_F^2}$). In contrast, to match the adaptive data-dependent

guarantees of Shampoo (Table 1), the smoothness of $\widetilde{\Phi}_t$ must depend on a preconditioner matrix, mirroring the adaptive regularizers in (1). Accordingly, we choose $\{\widetilde{\Phi}_t\}$ from a family of potentials $\{\widetilde{\Psi}(\cdot;\mathbf{L}) : \mathbf{L} \in \mathbb{S}_+^m\}$ parametrized by a PSD matrix $\mathbf{L}$, and formalize these desiderata in the following definition.

Definition 3 *Let $\{\widetilde{\Psi}(\cdot;\mathbf{L}) : \mathbf{L} \in \mathbb{S}_+^m\}$ be a family of potentials parametrized by a PSD matrix $\mathbf{L}$. We say that $\widetilde{\Psi}(\cdot;\cdot)$ is an (α,β)-admissible smoothing of $\|\cdot\|_*$ if the following conditions hold:*

(a) (Feasibility) *For all $\mathbf{L} \in \mathbb{S}_+^m$ and $\mathbf{X} \in \mathbb{R}^{m\times n}$, $\|\nabla_{\mathbf{X}}\widetilde{\Psi}(\mathbf{X};\mathbf{L})\|_{\mathrm{op}} \le 1$.*

(b) (Dominance) *For all $\mathbf{L} \in \mathbb{S}_+^m$ and $\mathbf{X} \in \mathbb{R}^{m\times n}$, $\widetilde{\Psi}(\mathbf{X};\mathbf{L}) \ge \|\mathbf{X}\|_*$ and $\widetilde{\Psi}(\mathbf{X};\mathbf{0}) = \|\mathbf{X}\|_*$.*

(c) (Upper stability) *For any $\mathbf{L}_1 \preceq \mathbf{L}_2$, $\sup_{\mathbf{X}}(\widetilde{\Psi}(\mathbf{X};\mathbf{L}_2) - \widetilde{\Psi}(\mathbf{X};\mathbf{L}_1)) \le \alpha(\mathrm{Tr}(\mathbf{L}_2) - \mathrm{Tr}(\mathbf{L}_1))$.*

(d) (Smoothness) *For any $\mathbf{L} \succ 0$, the function $\widetilde{\Psi}(\cdot;\mathbf{L})$ is continuously differentiable and satisfies $\mathcal{B}_{\widetilde{\Psi}(\cdot;\mathbf{L})}(\mathbf{Y} \,\|\, \mathbf{X}) \le \frac{\beta}{2}\mathrm{Tr}((\mathbf{X}-\mathbf{Y})^\top \mathbf{L}^{-1}(\mathbf{X}-\mathbf{Y}))$ for any $\mathbf{X}, \mathbf{Y} \in \mathbb{R}^{m\times n}$.*

The conditions in Definition 3 are directly motivated by the regret decomposition in Lemma 2 and are designed to control each term arising in (2). Specifically, feasibility ensures that the iterates $\mathbf{X}_{t+1}$ produced by (2) satisfy the operator-norm constraint; Dominance guarantees that the underestimation term is always non-positive; Upper stability controls the temporal variation term; and smoothness bounds the Bregman divergence term. Next, we show that (α,β)-admissible potentials $\widetilde{\Psi}$ control the terms in Lemma 2, yielding Shampoo-type regret for the presented GBPA class under a proper selection of $\mathbf{L}_t$. See Appendix D.2 for the proof.

Theorem 4 *Assume that $\|\mathbf{G}_t\|_{\mathrm{op}} \le G$ for all $t = 1,\dots,T$, and let $\widetilde{\Psi}$ be an (α,β)-admissible smoothing of the nuclear norm (Definition 3). Further, recall the definition $\mathbf{M}_t = \sum_{s=1}^t \mathbf{G}_s\mathbf{G}_s^\top$. Consider the GBPA update (2) with $\widetilde{\Phi}_t(\mathbf{S}) = \widetilde{\Psi}(\mathbf{S};\mathbf{L}_t/\eta)$, where*

$$\mathbf{L}_t \coloneqq \sqrt{G^2\mathbf{I} + \mathbf{M}_t}, \qquad \eta = \sqrt{\alpha/\beta}. \tag{3}$$

Then the regret of the algorithm satisfies

$$\mathrm{Reg}_T \le 2\sqrt{\alpha\beta}\, D\ \mathrm{Tr}\Big(\sqrt{G^2\mathbf{I} + \mathbf{M}_T}\Big) + (1 - \sqrt{\alpha\beta})\, D\|\mathbf{G}_1\|_*.$$

A couple of remarks follow. First, using the inequality $\mathrm{Tr}(\sqrt{G^2\mathbf{I} + \mathbf{M}_T}) \le mG + \mathrm{Tr}(\sqrt{\mathbf{M}_T})$, the above bound simplifies to $\mathcal{O}(\sqrt{\alpha\beta}D(\mathrm{Tr}(\sqrt{\mathbf{M}_T}) + mG))$. Neglecting the time-invariant term mG, this matches the one-sided Shampoo regret bound (Table 1) up to the factor $\sqrt{\alpha\beta}$. Second, provided the parameterization $\mathbf{L} = \mathbf{L}_t/\eta$ is fixed, the choice of potential $\widetilde{\Phi}_t(\mathbf{S})$ affects the regret solely through the admissibility constants α and β of its family.

Since the regret depends on the product $\alpha\beta$, two fundamental questions arise: what is the minimal achievable value for this product, and does there exist an (α,β)-admissible potential family that achieves it? We resolve both in the following result (proof in Appendix D.3).

Proposition 5 *For any (α,β)-admissible smoothing $\widetilde{\Psi}$, it holds that $\alpha\beta \ge \frac{1}{2}$. Moreover, the following smoothing $\widetilde{\Psi}^R$ is $(\frac{1}{2},1)$-admissible:*

$$\widetilde{\Psi}^R(\mathbf{S};\mathbf{L}) = \max_{\|\mathbf{X}\|_{\mathrm{op}}\le 1}\left\{\langle \mathbf{S},\mathbf{X}\rangle - \frac{1}{2}\mathrm{Tr}(\mathbf{X}^\top\mathbf{L}\mathbf{X})\right\} + \frac{1}{2}\mathrm{Tr}(\mathbf{L}). \tag{4}$$

By Danskin's theorem (Bertsekas, 1999), the gradient of $\widetilde{\Psi}^R(\mathbf{S};\mathbf{L})$ is given by $\nabla_{\mathbf{S}}\widetilde{\Psi}^R(\mathbf{S};\mathbf{L}) = \arg\max_{\|\mathbf{X}\|_{\mathrm{op}}\le 1}\left\{\langle \mathbf{S},\mathbf{X}\rangle - \frac{1}{2}\mathrm{Tr}(\mathbf{X}^\top\mathbf{L}\mathbf{X})\right\}$. Consequently, by setting $\widetilde{\Phi}_t(\mathbf{S}) = \widetilde{\Psi}^R(\mathbf{S};\mathbf{L}_t/\eta)$ with $\mathbf{L}_t$ defined in (3) and choosing $\eta = 1/\sqrt{2}$ as prescribed by Theorem 4, the update induced by (2) (up to a sign change) can be written as

$$\mathbf{X}_{t+1} = D\arg\min_{\|\mathbf{X}\|_{\mathrm{op}}\le 1}\left\{\langle \mathbf{S}_t,\mathbf{X}\rangle + \frac{1}{2\eta}\,\mathrm{Tr}(\mathbf{X}^\top\mathbf{L}_t\mathbf{X})\right\}, \tag{5}$$

which corresponds to a Follow-the-regularized-Leader (FTRL) method. As a direct corollary, the update in (5) attains the regret $\sqrt{2}D(\mathrm{Tr}(\sqrt{\mathbf{M}_T}) + mG) + (1-\frac{1}{\sqrt{2}})D\|\mathbf{G}_1\|_*$. Ignoring the lower-order terms DmG and $D\|\mathbf{G}_1\|_*$, this matches the regret of one-sided Shampoo in Table 1 up to a constant factor. Notably, this provides the first such guarantee for an adaptive matrix FTRL method, as a side result of our framework. However, like (1), the update (5) necessitates solving a costly, iterative quadratic projection over the operator-norm ball. To address this, in the next section, we introduce two alternative potentials with substantially cheaper gradient evaluations.

4. Proposed Algorithms

Next, we study alternative smoothings of the nuclear norm that yield more efficient algorithms. Section 4.1 introduces a stochastic smoothing that leads to a Follow-the-Perturbed-Leader (FTPL) method based on random perturbations, with parallelizable updates built from standard matrix primitives and an analysis relying on noncentral Wishart theory. Section 4.2 introduces the novel Follow-the-Augmented-Matrix-Leader (FAML) algorithm via a deterministic hyperbolic smoothing. We prove this smoothing attains near-optimal admissibility constants, matching those of regularized methods, while circumventing quadratic projections through efficient matrix primitives.

4.1. Stochastic Smoothing: Follow-the-Perturbed-Leader

As discussed, the central goal of GBPA is to construct a sequence of smooth, tractable surrogates for the nuclear norm that preserve its geometric properties while enabling efficient optimization. To this end, we use *stochastic smoothing*, a classical technique obtained by convolving the nonsmooth objective with a smooth probability density function (Glasserman, 1991; Yousefian et al., 2010; Duchi et al., 2012), where the key design choice is the perturbation distribution. Intuitively, to achieve an adaptive regret bound comparable to that of one-sided Shampoo, the perturbations themselves should adapt to the previously observed gradient sequence. Drawing a parallel with the regularized potential $\widetilde{\Psi}^R$ in (4), we consider the following family of stochastic smoothing potentials:

$$\widetilde{\Psi}^S(\mathbf{S};\mathbf{L}) = \mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathcal{MN}(0,\mathbf{I}_m,\mathbf{I}_n)}\|\mathbf{S}+\mathbf{L}\mathbf{Z}\|_*, \tag{6}$$

where $\mathcal{MN}(0,\mathbf{I}_m,\mathbf{I}_n)$ is the matrix normal distribution with independent standard Gaussian entries.

To derive the gradient of $\widetilde{\Psi}^S(\mathbf{S};\mathbf{L})$, we use the variational representation of the nuclear norm and rewrite $\widetilde{\Psi}^S$ in (6) as $\widetilde{\Psi}^S(\mathbf{S};\mathbf{L}) = \mathbb{E}_{\mathbf{Z}}\max_{\|\mathbf{X}\|_{\mathrm{op}}\le 1}\langle \mathbf{S}+\mathbf{L}\mathbf{Z},\mathbf{X}\rangle$. By (Bertsekas, 1973, Proposition 2.2), we can swap the order of expectation and differentiation to obtain $\nabla\widetilde{\Psi}^S(\mathbf{S};\mathbf{L}) = \mathbb{E}_{\mathbf{Z}}[\arg\max_{\|\mathbf{X}\|_{\mathrm{op}}\le 1}\langle \mathbf{S}+\mathbf{L}\mathbf{Z},\mathbf{X}\rangle]$. Following Theorem 4, we set $\widetilde{\Phi}_t(\mathbf{S}) = \widetilde{\Psi}(\mathbf{S};\mathbf{L}_t/\eta)$, where $\mathbf{L}_t$ is defined in (3). This choice leads to the following update:

$$\mathbf{X}_{t+1} = -D\mathbb{E}_{\mathbf{Z}}\left[\arg\max_{\|\mathbf{X}\|_{\mathrm{op}}\le 1}\left\{\langle \mathbf{S}_t + \frac{1}{\eta}\mathbf{L}_t\mathbf{Z},\mathbf{X}\rangle\right\}\right], \quad \text{where } \mathbf{S}_t := \sum_{s=1}^{t}\mathbf{G}_s. \tag{7}$$

The above update is indeed equivalent to an FTPL method with perturbation $(1/\eta)\mathbf{L}_t\mathbf{Z}$.

Implementation and computational efficiency In contrast to the OGD update in (1) or the FTRL update in (5), the update in (7) can be computed efficiently using standard linear algebra primitives such as Cholesky factorization and polar decomposition, as detailed below. To begin with, recall that $\mathbf{L}_t = \sqrt{G^2\mathbf{I} + \mathbf{M}_t}$, which may suggest explicitly computing a matrix square root. However, this is unnecessary; instead, we can compute the Cholesky factorization of $G^2\mathbf{I} + \mathbf{M}_t = \widetilde{\mathbf{L}}_t\widetilde{\mathbf{L}}_t^\top$, and observe that $\widetilde{\mathbf{L}}_t\mathbf{Z} \overset{d}{=} \mathbf{L}_t\mathbf{Z}$ when $\mathbf{Z}$ has independent standard Gaussian entries. The leading-order cost of Cholesky factorization is $\frac{1}{3}m^3$ (Golub and Van Loan, 2013).

Moreover, for a fixed perturbation matrix $\mathbf{Z}$, the maximization inside the expectation reduces to computing the *polar factor* of $\mathbf{S}_t + \frac{1}{\eta}\mathbf{L}_t\mathbf{Z}$. Specifically, for a matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$ with singular value decomposition $\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$, its polar factor is $\mathrm{polar}(\mathbf{X}) = \mathbf{U}\mathbf{V}^\top$ (Higham, 2008), which can be computed efficiently using well-established numerical iterative methods, including Newton–Schulz iteration (Higham, 2008), scaled Newton methods (Higham, 2008), and QDWH iterations (Nakatsukasa et al., 2010; Nakatsukasa and Higham, 2013); see also Amsel et al. (2025). In particular, Newton–Schulz iteration with K steps incurs a computational cost of $4Km^2n$ (see Appendix F.2).

Finally, the update in (7) is presented in its *deterministic* form based on the expected action. In practice, this expectation can be approximated via Monte Carlo sampling. Specifically, after computing the cumulative gradient $\mathbf{S}_t = \sum_{s=1}^t \mathbf{G}_s$ and the Cholesky factorization $\widetilde{\mathbf{L}}_t\widetilde{\mathbf{L}}_t^\top = G^2\mathbf{I} + \mathbf{M}_t$, where $\mathbf{M}_t = \sum_{s=1}^t \mathbf{G}_s\mathbf{G}_s^\top$, the update with k samples of the random matrix $\mathbf{Z}$ is

$$\mathbf{X}_{t+1} = -\frac{D}{k}\sum_{i=1}^{k} \mathrm{polar}\left(\mathbf{S}_t + \frac{1}{\eta}\widetilde{\mathbf{L}}_t\mathbf{Z}_t^{(i)}\right), \quad \mathbf{Z}_t^{(i)} \sim \mathcal{MN}(0, \mathbf{I}_m, \mathbf{I}_n) \quad \text{i.i.d.} \tag{8}$$

Moreover, the k polar factors in (8) can be computed in parallel. As a result, with sufficient parallel computing resources, the effective computational cost can be significantly reduced.

Regret guarantees By Theorem 4, the only remaining task is to determine the admissibility parameters (α, β) for the stochastic smoothing in (6), which will lead to a regret guarantee on the FTPL algorithm in (7). Using noncentral Wishart theory, we are able to establish the admissibility when $n \geq m + 2$. The proof is presented in Appendix E.1.

Theorem 6 *When $n \geq m + 2$, the stochastic smoothing $\widetilde{\Psi}^S(\mathbf{S}; \mathbf{L})$ is (α, β)-admissible with $\alpha = \sqrt{m} + \sqrt{n}$ and $\beta = \frac{1}{\sqrt{n-m-1}}$. As a corollary, $\mathrm{Reg}_T \leq 2\sqrt{2}D\left(\frac{n}{n-m-1}\right)^{1/4}\left(\mathrm{Tr}\left[\mathbf{M}_T^{1/2}\right] + mG\right)$.*

Comparing with the regret bound of one-sided Shampoo in Table 1, the bound in Theorem 6 incurs an additional factor of $(n/(n-m-1))^{1/4}$. That said, when $n \geq 2m$ and $m \geq 2$, this factor is at most $\sqrt{2}$ and the resulting guarantee matches one-sided Shampoo up to a constant factor.

Remark 7 *In OLO with an oblivious adversary, the sampled algorithm in (8) with $k = 1$ achieves the same expected regret as the deterministic version in (7). In our O2NC extension, however, the adversary is adaptive, so this equivalence no longer holds, and additional care is needed. In particular, the extra error from sampling can be controlled using a Khintchine-type inequality, which yields a deviation term of order $O(1/\sqrt{k})$, similar to the arguments in (Hazan and Minasyan, 2020). Therefore, by choosing $k = T$, we maintain the worst-case $\mathcal{O}(\sqrt{T})$ regret bound.*

4.2. Hyperbolic Smoothing: Follow-the-Augmented-Matrix-Leader

While the regularized smoothing $\widetilde{\Psi}^R$ achieves optimal admissibility (Proposition 5), its implicit definition via a quadratic program makes its gradient evaluation costly. In contrast, the randomized smoothing $\widetilde{\Psi}^S$ is cheaper to compute but incurs dimension-dependent admissibility factors (Theorem 6). This raises the question of whether one can obtain an explicit, inexpensive smoothing while retaining near-optimal admissibility. Exploiting the special structure of the nuclear norm, we answer this question in the affirmative. Specifically, we introduce the hyperbolic family of potentials

$$\widetilde{\Psi}^H(\mathbf{S};\mathbf{L}) := \mathrm{Tr}\Big(\sqrt{\mathbf{S}\mathbf{S}^\top + \mathbf{L}\mathbf{L}^\top}\Big). \tag{9}$$

Since $\|\mathbf{S}\|_* = \mathrm{Tr}(\sqrt{\mathbf{S}\mathbf{S}^\top})$, we have $\widetilde{\Psi}^H(\mathbf{S};\mathbf{0}) = \|\mathbf{S}\|_*$, and the additive term $\mathbf{L}\mathbf{L}^\top$ provides an explicit smoothing. Similar smoothing of the nuclear norm has been considered in low-rank optimization, e.g., (Mohan and Fazel, 2012). To the best of our knowledge, however, this particular smoothing has not been analyzed as a potential in an online learning framework.

To compute the gradient of $\widetilde{\Psi}^H$, we exploit the interesting fact that the hyperbolic smoothing in (9) can be interpreted as the nuclear norm of an augmented matrix. Specifically, define $\widehat{\mathbf{S}} := \begin{bmatrix}\mathbf{S} & \mathbf{L}\end{bmatrix} \in \mathbb{R}^{m\times(n+m)}$. Then $\widehat{\mathbf{S}}\widehat{\mathbf{S}}^\top = \mathbf{S}\mathbf{S}^\top + \mathbf{L}\mathbf{L}^\top$, and hence $\widetilde{\Psi}^H(\mathbf{S};\mathbf{L}) = \mathrm{Tr}(\sqrt{\widehat{\mathbf{S}}\widehat{\mathbf{S}}^\top}) = \|\widehat{\mathbf{S}}\|_*$. Consequently, the gradient with respect to $\mathbf{S}$ can be obtained by first differentiating the nuclear norm in the augmented space and then restricting to the leading block. Concretely, letting $\widehat{\mathbf{X}} = \nabla_{\widehat{\mathbf{S}}}\|\widehat{\mathbf{S}}\|_* = \arg\max_{\|\widehat{\mathbf{X}}'\|_{\mathrm{op}}\le 1}\langle\widehat{\mathbf{S}},\widehat{\mathbf{X}}'\rangle$, we obtain $\nabla\widetilde{\Psi}^H(\mathbf{S};\mathbf{L}) = \widehat{\mathbf{X}}[1:m,1:n]$. With the choice of $\widetilde{\Phi}_t(\mathbf{S}) = \widetilde{\Psi}^H(\mathbf{S};\mathbf{L}_t/\eta)$ as in Theorem 4, where $\mathbf{L}_t$ is defined in (3), the update (2) takes the form

$$\begin{aligned}
&\widehat{\mathbf{X}}_{t+1} = \arg\min_{\|\widehat{\mathbf{X}}\|_{\mathrm{op}}\le D}\langle\widehat{\mathbf{S}}_t,\widehat{\mathbf{X}}\rangle \quad \text{where} \quad \widehat{\mathbf{S}}_t = \Big[\mathbf{S}_t \;\; \tfrac{1}{\eta}\mathbf{L}_t\Big] \in \mathbb{R}^{m\times(n+m)} \text{ and } \mathbf{S}_t = \sum_{s=1}^{t}\mathbf{G}_s \\
&\mathbf{X}_{t+1} = \widehat{\mathbf{X}}_{t+1}[1:m,\,1:n].
\end{aligned} \tag{10}$$

The above update rule can be interpreted as an FTL-style update applied to the augmented matrix $\widehat{\mathbf{S}}_t$, which is the reason we refer to it as the *Follow-the-Augmented-Matrix-Leader (FAML)* algorithm.

Alternatively, using the identity $\nabla_{\widehat{\mathbf{S}}}\|\widehat{\mathbf{S}}\|_* = (\widehat{\mathbf{S}}\widehat{\mathbf{S}}^\top)^{-1/2}\widehat{\mathbf{S}}$, the gradient of (9) can also be written as $\nabla_{\mathbf{S}}\widetilde{\Psi}^H(\mathbf{S};\mathbf{L}) = \big(\mathbf{S}\mathbf{S}^\top + \mathbf{L}\mathbf{L}^\top\big)^{-1/2}\mathbf{S}$. This leads to an equivalent formulation of FAML:

$$\mathbf{X}_{t+1} = -\eta D\big(\eta^2\mathbf{S}_t\mathbf{S}_t^\top + (G^2\mathbf{I} + \mathbf{M}_t)\big)^{-1/2}\mathbf{S}_t, \text{ where } \mathbf{S}_t := \sum_{s=1}^{t}\mathbf{G}_s,\ \mathbf{M}_t := \sum_{s=1}^{t}\mathbf{G}_s\mathbf{G}_s^\top. \tag{11}$$

It is instructive to compare (11) with the FTRL update (5). If we ignore the operator-norm constraint in (5), the solution of the resulting unconstrained quadratic subproblem is $\mathbf{X}_{t+1} = -\eta D\mathbf{L}_t^{-1}\mathbf{S}_t = -\eta D(G^2\mathbf{I} + \mathbf{M}_t)^{-1/2}\mathbf{S}_t$. In contrast, the FAML update in (11) incorporates the additional term $\eta^2\mathbf{S}_t\mathbf{S}_t^\top$ inside the inverse square root. This guarantees that the iterate $\mathbf{X}_{t+1}$ automatically satisfies the operator-norm constraint, without requiring an explicit projection. Consequently, FAML achieves essentially the same computational cost as solving an unconstrained FTRL subproblem, which explains how FAML alleviates the main computational bottleneck of matrix FTRL.

Implementation In contrast to the updates of OGD in (1) and FTRL in (5), FAML can be implemented directly using (11), where the dominant cost comes from computing a matrix inverse square root. While a standard approach is to use SVD, this can be costly. A more efficient alternative is to use iterative methods such as coupled Newton–Schulz (NS) iteration (Higham, 2008; An et al., 2025). As detailed in Appendix F.3, the leading-order cost in terms of floating-point operations is $6m^2n + 6Km^3$, where K denotes the number of inner NS steps.

A drawback of this approach is that computing matrix inverse square roots is numerically unstable, particularly in low precision. In this regard, the formulation in (10) provides a more favorable alternative: similar to FTPL, it can be implemented by computing the polar factor of the augmented matrix $\widehat{\mathbf{S}}_t$. In fact, with a customized NS iteration tailored to this augmented formulation, we can avoid computing the matrix square root $\mathbf{L}_t$ and the leading-order computational cost becomes $(2m^2n + 4m^3)K + 4m^2n$, where again K is the number of inner NS steps (see Appendix F.3).

Regret guarantees The last step is to determine the admissibility constants α and β for the potential family defined in (9) and, together with Theorem 4, to derive the corresponding regret bound for FAML in (10). The proof is presented in Appendix E.2.

Theorem 8 *The hyperbolic smoothing $\widetilde{\Psi}^H(\mathbf{S}; \mathbf{L})$ defined in (9) is (α, β)-admissible with $\alpha = 1$ and $\beta = 1$. Hence, the regret of the FAML algorithm (10) satisfies $\mathrm{Reg}_T \leq 2D(\mathrm{Tr}(\mathbf{M}_T^{1/2}) + mG)$.*

Like FTRL (5), this bound matches the one-sided Shampoo regret up to constant factors, ignoring the lower-order term DmG. Crucially, FAML eliminates the need to solve a quadratic program and is instead implemented using computationally efficient matrix primitives.

5. Application: Nonsmooth Nonconvex Matrix Optimization

We apply our adaptive online methods to nonsmooth nonconvex optimization using the *Online-to-Nonconvex Conversion* (O2NC) framework. We start by reducing stochastic matrix optimization to Matrix OLO and identifying Muon (Jordan et al., 2024) as an FTL instance. We then introduce *Pion* and *Leon*, theoretically grounded optimizers derived from our adaptive FTPL and FAML algorithms.

5.1. Online-to-Nonconvex Conversion

The O2NC framework, proposed by Cutkosky et al. (2023) and refined in (Zhang and Cutkosky, 2024; Ahn and Cutkosky, 2024; Ahn et al., 2025), reduces finding a stationary point of a stochastic matrix optimization problem to an online matrix optimization task. To formalize this connection, we first define the stochastic matrix optimization problem:

$$\min_{\mathbf{W} \in \mathbb{R}^{m \times n}} L(\mathbf{W}) = \mathop{\mathbb{E}}_{\zeta \sim \mathcal{D}}[\ell(\mathbf{W}; \zeta)], \tag{12}$$

where $L : \mathbb{R}^{m \times n} \to \mathbb{R}$ is differentiable. As L may be nonsmooth and nonconvex, we adopt the notion of a (ρ, ε)-stationary point, a relaxation of the Goldstein stationary point (Goldstein, 1977).

Definition 9 ((ρ, ε)-stationary point) *Suppose L is differentiable and let $\|\cdot\|$ be a norm with dual norm $\|\cdot\|_\dagger$. Then $\mathbf{W}$ is a (ρ, ε)-stationary point if there exists a distribution p with finite support over $\mathbb{R}^{m \times n}$ with $\mathbb{E}_{\mathbf{Y} \sim p}[\mathbf{Y}] = \mathbf{W}$ such that $\|\mathbb{E}[\nabla L(\mathbf{Y})]\|_\dagger \leq \varepsilon$ and $\mathbb{E}\|\mathbf{Y} - \mathbf{W}\| \leq \rho$.*

Intuitively, a point $\mathbf{W}$ is (ρ, ε)-stationary if one can find finitely many nearby points whose weighted average is $\mathbf{W}$, such that the weighted average of their gradients is small. To simplify notation, for any $\mathbf{W} \in \mathbb{R}^{m\times n}$ and $\rho > 0$, define $\mathcal{P}(\mathbf{W}; \rho)$ as the set of all distributions p finitely supported on $\mathbb{R}^{m\times n}$ such that $\mathbb{E}_{\mathbf{Y}\sim p}[\mathbf{Y}] = \mathbf{W}$ and $\mathbb{E}\,\|\mathbf{Y} - \mathbf{W}\| \le \rho$. Then $\mathbf{W}$ is a (ρ, ε)-stationary point if and only if $\|\nabla L(\mathbf{W})\|_\dagger^{[\rho]} := \inf_{p\in\mathcal{P}(\mathbf{W};\rho)} \|\,\mathbb{E}_{\mathbf{Y}\sim p}[\nabla L(\mathbf{Y})]\|_\dagger \le \varepsilon$.

The Reduction Mechanism In the O2NC framework, for a given matrix norm $\|\cdot\|$, the optimizer queries an online learner at each step for a direction $\mathbf{X}_t \in \{\mathbf{X} \in \mathbb{R}^{m\times n} : \|\mathbf{X}\| \le D\}$ based on the history of observed gradients. The weight matrix is then updated via $\mathbf{W}_t = \mathbf{W}_{t-1} + \mathbf{X}_t$, and we evaluate a stochastic gradient at a random midpoint $\widetilde{\mathbf{W}}_t = \mathbf{W}_{t-1} + s_t\mathbf{X}_t$, where $s_t \sim \mathrm{Uniform}(0, 1)$ is a uniformly distributed scaling factor, yielding $\mathbf{G}_t = \nabla\ell(\widetilde{\mathbf{W}}_t; \zeta_t)$. As the gradients $\{\mathbf{G}_t\}$ depend on evolving parameters $\mathbf{W}_t$ through a nonconvex nonsmooth objective, the induced environment is highly non-stationary. To capture local geometry and decay stale information, we measure performance via *discounted regret* with $\beta \in (0, 1)$, defined as

$$\mathrm{Reg}_T^{[\beta]}(D) := \max_{\|\mathbf{X}\|\le D} [\sum_{t=1}^{T} \beta^{T-t}\langle \mathbf{G}_t, \mathbf{X}_t - \mathbf{X}\rangle]. \tag{13}$$

Discounted regret can be reduced to standard regret by defining the loss at time t as $\ell_t^{[\beta]}(\mathbf{X}) = \langle \beta^{-t}\mathbf{G}_t, \mathbf{X}\rangle$ for all t and then multiplying the resulting regret by β^T. The following proposition establishes the key guarantee: if the learner achieves a discounted regret that is sublinear in $\frac{1}{1-\beta}$, the sequence converges to a (ρ, ε)-stationary point. Detailed protocols and proofs are provided in Appendix G.

Proposition 10 (Informal O2NC bound) *Let $\{\bar{\mathbf{W}}_t\}$ denote the exponential moving average sequence of $\{\widetilde{\mathbf{W}}_t\}$ and let τ be a random index defined in Appendix G.1. If $D = \frac{(1-\beta)}{4\beta}\rho$, then the expected ρ-stationarity gap at $\bar{\mathbf{W}}_\tau$ satisfies*

$$\mathop{\mathbb{E}}_{\tau}\Big[\,\|\nabla L(\bar{\mathbf{W}}_\tau)\|_\dagger^{[\rho]}\Big] = \mathcal{O}\Big(\frac{L(\mathbf{W}_0) - L(\mathbf{W}^*)}{(1-\beta)\rho T} + (1-\beta)\,\mathbb{E}\left[\mathrm{Reg}_T^{[\beta]}(1)\right] + \textit{stochastic noise}\Big). \tag{14}$$

Note that random sampling $\bar{\mathbf{W}}_\tau$ is necessary as the last iterate lacks stationarity guarantees in nonconvex optimization. Since the stochastic noise term is algorithm-independent, minimizing the bound in (14) reduces to minimizing discounted regret of the online matrix learner. We instantiate this framework with our adaptive FTPL and FAML updates, yielding *Pion* and *Leon*, to establish rigorous convergence guarantees.

5.2. Matrix Optimization Algorithms: Muon, Pion, and Leon

In this section, we instantiate the O2NC framework with specific online learning algorithms. We first show that the Muon optimizer corresponds to the classic Follow-the-Leader (FTL) strategy, and then derive *Pion* and *Leon*, based on our adaptive FTPL and FAML frameworks, respectively. For consistency with optimizer implementations, we write updates as $\mathbf{W}_{t+1} = \mathbf{W}_t - \alpha_t\mathbf{P}_t$, where $\mathbf{P}_t = -\mathbf{X}_{t+1}/D$ is a spectrally normalized direction and D is absorbed into α_t. Throughout this section, we use the discounted statistics

$$\bar{\mathbf{G}}_t := \sum_{s=1}^{t} \beta^{t-s}\mathbf{G}_s, \qquad \bar{\mathbf{M}}_t := \sum_{s=1}^{t} \beta^{2(t-s)}\mathbf{G}_s\mathbf{G}_s^\top, \qquad \widetilde{\mathbf{L}}_t\widetilde{\mathbf{L}}_t^\top := G^2\beta^{-2}\mathbf{I} + \bar{\mathbf{M}}_t, \tag{15}$$

where $\{\mathbf{G}_t\}$ are the O2NC stochastic gradients, G satisfies $\|\mathbf{G}_t\|_{\mathrm{op}} \leq G$, $\bar{\mathbf{G}}_t$ is the discounted leader, and $\bar{\mathbf{M}}_t, \widetilde{\mathbf{L}}_t$ are its discounted second moment and regularized Cholesky factor.

Muon Consider applying the FTL algorithm to the underlying online learning problem. In the O2NC reduction, the learner receives discounted linear losses $\ell_s^{[\beta]}(\mathbf{X}) = \langle \beta^{-s}\mathbf{G}_s, \mathbf{X}\rangle$. Hence, at step t, the FTL update selects the direction $\mathbf{X}_{t+1}$ that minimizes the cumulative loss observed so far:

$$\mathbf{X}_{t+1} = \operatorname*{arg\,min}_{\|\mathbf{X}\|_{\mathrm{op}}\leq D}\Big\{\sum_{s=1}^{t}\ell_s^{[\beta]}(\mathbf{X})\Big\} = \operatorname*{arg\,min}_{\|\mathbf{X}\|_{\mathrm{op}}\leq D}\Big\{\Big\langle\sum_{s=1}^{t}\beta^{-s}\mathbf{G}_s, \mathbf{X}\Big\rangle\Big\} = -D\cdot\mathrm{polar}\Big(\sum_{s=1}^{t}\beta^{-s}\mathbf{G}_s\Big),$$

where the last equality follows from the fact that the solution to linear minimization over the operator-norm ball is given by the negative polar factor.

We now derive the optimization method for this update based on the reduction described in Section 5.1. Since the polar decomposition is scale-invariant—satisfying $\mathrm{polar}(c\mathbf{A}) = \mathrm{polar}(\mathbf{A})$ for any scalar $c > 0$—we may scale the argument by β^t without affecting the result. Using the EMA $\bar{\mathbf{G}}_t$ in (15), the update becomes

$$\mathbf{W}_{t+1} = \mathbf{W}_t - D\cdot\mathrm{polar}(\bar{\mathbf{G}}_t).$$

Equivalently, in optimizer notation, the update direction is $\mathbf{P}_t = \mathrm{polar}(\bar{\mathbf{G}}_t)$, which is precisely the direction used by the Muon optimizer (Jordan et al., 2024) (see Algorithm 1 in Appendix G). Thus, Muon can be viewed as a spectrally normalized optimization method: rather than applying coordinate-wise normalization, it follows the polar factor of the EMA of matrix gradients.

This derivation reveals that Muon is structurally equivalent to FTL under a spectral constraint. The only distinction is that, in the O2NC framework, $\mathbf{G}_t$ is evaluated at a random midpoint between $\mathbf{W}_{t-1}$ and $\mathbf{W}_t$ rather than at $\mathbf{W}_t$. However, because FTL does not generally achieve sublinear regret for nonsmooth losses, the existing O2NC framework alone is insufficient to establish a rigorous convergence guarantee for vanilla Muon in such settings.

We emphasize that this does not imply that Muon fails to converge or lacks a principled foundation. Rather, within the existing O2NC framework, vanilla Muon corresponds to an FTL-type method, and therefore the regret-to-stationarity reduction cannot certify its convergence in the nonsmooth nonconvex setting. Accordingly, we do not claim that Pion or Leon outperforms Muon in practice; our focus is primarily theoretical. Existing analyses of Muon rely on smoothness, whereas Pion and Leon, as smoothed variants of Muon, admit provable nonsmooth nonconvex convergence guarantees, as detailed below.

Pion We address the theoretical shortcomings of FTL by replacing it with the adaptive FTPL framework, which stabilizes the optimization process through stochastic perturbation of the cumulative gradients. In optimizer terms, Pion retains Muon's polar-update template, but replaces the single leader $\bar{\mathbf{G}}_t$ with randomly perturbed leaders. Applying adaptive FTPL within O2NC therefore gives one polar direction for each perturbation, and the final update is their Monte Carlo average. Extending (8) to the discounted case and using the scale invariance of the polar decomposition, the resulting update direction is

$$\mathbf{X}_{t+1} = -\frac{D}{k}\sum_{i=1}^{k}\Big[\mathrm{polar}\Big(\sum_{s=1}^{t}\beta^{t-s}\mathbf{G}_s + \frac{1}{\eta}\widetilde{\mathbf{L}}_t\mathbf{Z}_t^{(i)}\Big)\Big], \tag{16}$$

where $\mathbf{Z}_t^{(i)} \overset{\text{i.i.d.}}{\sim} \mathcal{MN}(0, \mathbf{I}_m, \mathbf{I}_n)$ are standard Gaussian matrices, and $\widetilde{\mathbf{L}}_t$ is the Cholesky factor defined in (15). The iterates then follow the standard O2NC protocol: $\mathbf{W}_{t+1} = \mathbf{W}_t + \mathbf{X}_{t+1}$.

We name this algorithm **Pion**, as it may be viewed as a perturbed variant of Muon. The perturbation is shaped by the discounted adaptive preconditioner $\widetilde{\mathbf{L}}_t$, which enables us to establish a valid regret bound and, consequently, convergence guarantees for nonsmooth nonconvex objectives. Computationally, as discussed in Section 4.1, Pion avoids matrix square roots by using $\widetilde{\mathbf{L}}_t$, and the k polar factors in (16) can be computed in parallel using Newton–Schulz iterations. The pseudocode is provided in Algorithm 2 in Appendix G.

Remark 11 *The formal O2NC analysis evaluates stochastic gradients at random midpoints, as described in Section 5.1 and Appendix G.1. For readability and consistency with standard optimizer implementations, the pseudocode for Pion and Leon in Algorithms 2 and 3 uses gradients evaluated at the current iterate. The convergence guarantees stated below correspond to the midpoint-gradient O2NC variant.*

We now establish the iteration complexity of Pion. We adopt standard assumptions on the stochastic gradients, consistent with prior work on adaptive methods (An et al., 2025).

Assumption 1 (Bounded Stochastic Gradient) *Recall that $\mathbf{G}_t = \nabla \ell(\widetilde{\mathbf{W}}_t; \zeta_t)$. We assume that $\|\mathbf{G}_t\|_{\mathrm{op}} \le G$ almost surely and there exists a positive semidefinite matrix $\mathbf{Q} \succeq 0$ such that $\mathbb{E}[\mathbf{G}_t \mathbf{G}_t^\top] \preceq \mathbf{Q}^2$ for all t.*

Combining the regret guarantee of adaptive FTPL (Theorem 6) with the O2NC reduction (Proposition 10), and denoting $\Delta L := L(\mathbf{W}_0) - L(\mathbf{W}^*)$, we obtain the following non-asymptotic complexity bound.

Theorem 12 (Convergence of Pion) *Suppose that $n \ge m + 2$. For target accuracy ε sufficiently small, Pion finds a (ρ, ε)-stationary point of Problem (12) with the norm $\|\cdot\| = \|\cdot\|_{\mathrm{op}}$ in T iterations, where $T = \mathcal{O}\big(\max\big\{\frac{C^2\|\mathbf{Q}\|_*^2 \Delta L}{\rho \varepsilon^3}, \frac{C^2\|\mathbf{Q}\|_*^2}{\varepsilon^2}, \frac{CmG}{\varepsilon}\big\}\big)$, and $C = (n/(n-m-1))^{1/4}$ is the dimensional factor from the FTPL analysis.*

Leon While Pion achieves robustness via stochastic perturbations, Monte Carlo sampling can introduce additional variance. To address this, we introduce Leon (Learning-Enabled Orthogonalization and Normalization), a deterministic alternative based on our FAML algorithm. The core idea of FAML is to replace the stochastic perturbation of FTPL with a deterministic augmentation in a higher-dimensional space. Instead of adding noise to the accumulated gradients, we construct an augmented matrix $\widehat{\mathbf{S}}_t$ by concatenating the accumulated gradient with a scaled preconditioner. Leon therefore smooths Muon's leader deterministically by computing one polar factor in a lifted space where the additional block $\eta^{-1}\widetilde{\mathbf{L}}_t$ acts as a stabilizer. Adapting the FAML update (10) to the discounted O2NC setting, we form the augmented matrix using the gradient EMA $\bar{\mathbf{G}}_t$ and the discounted preconditioner $\widetilde{\mathbf{L}}_t$ defined in (15). The update direction is then obtained by projecting the polar factor of this augmented matrix back onto the original space:

$$\mathbf{X}_{t+1} = \text{LeadingBlock}\left(-D \cdot \text{polar}\left(\begin{bmatrix} \bar{\mathbf{G}}_t & \frac{1}{\eta}\widetilde{\mathbf{L}}_t \end{bmatrix}\right)\right). \tag{17}$$

Method	Online learner	Direction $\mathbf{P}_t$	Smoothing
Muon	FTL	$\mathrm{polar}(\bar{\mathbf{G}}_t)$	None
Pion	Adaptive FTPL	$\frac{1}{k}\sum_{i=1}^{k}\mathrm{polar}\left(\bar{\mathbf{G}}_t+\eta^{-1}\widetilde{\mathbf{L}}_t\mathbf{Z}_t^{(i)}\right)$	Gaussian
Leon	FAML	$\left(\bar{\mathbf{G}}_t\bar{\mathbf{G}}_t^\top+\eta^{-2}\widetilde{\mathbf{L}}_t\widetilde{\mathbf{L}}_t^\top\right)^{-1/2}\bar{\mathbf{G}}_t$	Deterministic augmentation

Table 2: Comparison of Muon, Pion, and Leon in optimizer notation. Directions use the discounted statistics in (15).

Using the closed-form expression for the polar factor of a block matrix, this simplifies to:

$$\mathbf{X}_{t+1}=-D\left(\bar{\mathbf{G}}_t\bar{\mathbf{G}}_t^\top+\frac{1}{\eta^2}(G^2\beta^{-2}\mathbf{I}+\bar{\mathbf{M}}_t)\right)^{-1/2}\bar{\mathbf{G}}_t, \tag{18}$$

where $\bar{\mathbf{G}}_t$ and $\bar{\mathbf{M}}_t$ are defined in (15). The final weight update follows the standard O2NC protocol: $\mathbf{W}_{t+1}=\mathbf{W}_t+\mathbf{X}_{t+1}$. Leon can be interpreted as a deterministic, smoothed variant of Muon, or equivalently as a preconditioned gradient method in which the curvature correction is applied through a smoothed inverse square root. In contrast to existing analyses of Muon, which rely on smoothness, Leon inherits the adaptive regret guarantees of FAML and therefore admits provable convergence guarantees for nonsmooth nonconvex objectives, without requiring random sampling. As discussed in Section 4.2, a customized Newton–Schulz-type update applied to the augmented polar form (17) computes Leon's direction without explicitly forming a matrix square root. The pseudocode is provided in Algorithm 3 in Appendix G.

Theorem 13 (Convergence of Leon) *For target accuracy ε sufficiently small, Leon finds a (ρ,ε)-stationary point of Problem (12) with the norm $\|\cdot\|=\|\cdot\|_{\mathrm{op}}$ in T iterations, where $T=\mathcal{O}\left(\max\left\{\frac{\|\mathbf{Q}\|_*^2\Delta L}{\rho\varepsilon^3},\frac{\|\mathbf{Q}\|_*^2}{\varepsilon^2},\frac{mG}{\varepsilon}\right\}\right)$.*

Since FAML achieves tighter regret than FTPL, Leon improves Pion's convergence guarantee by a factor of $(n/(n-m-1))^{1/2}$ in dominant terms and $(n/(n-m-1))^{1/4}$ in the non-dominant term. Appendix H provides a synthetic nonsmooth optimization problem comparing the practical behavior of Pion and Leon with Muon.

6. Conclusion

In this work, we addressed the computational challenges of adaptive matrix online learning constrained by the operator norm. By casting algorithm design as a smoothing problem, we developed two efficient methods, adaptive FTPL and FAML, that matched the best-known regret guarantees without solving costly quadratic projections. Finally, we extended this framework to nonsmooth nonconvex optimization via O2NC, introducing the Pion and Leon optimizers and establishing convergence guarantees for these methods that the popular matrix-based optimizer Muon lacks.

Acknowledgments

This work was supported in part by NSF CAREER Award 2338846, the NSF AI Institute for Foundations of Machine Learning (IFML), NSF AI Institute for Future Edge Networks and Distributed Intelligence (AI-EDGE), the Machine Learning Lab (MLL) at UT Austin, and the Wireless Networking and Communications Group (WNCG) Industrial Affiliates Program. We are grateful for computing support on the Vista GPU Cluster through the Center for Generative AI (CGAI) and the Texas Advanced Computing Center (TACC) at UT Austin.

References

Jacob Abernethy, Chansoo Lee, and Ambuj Tewari. Perturbation techniques in online learning and optimization. *Perturbations, Optimization, and Statistics*, 233:17, 2016.

Alekh Agarwal, Alexander Rakhlin, and Peter Bartlett. Matrix regularization techniques for online multitask learning. *EECS Department, University of California, Berkeley, Tech. Rep. UCB/EECS-2008-138*, 2008.

Kwangjun Ahn and Ashok Cutkosky. Adam with model exponential moving average is effective for nonconvex optimization. *arXiv preprint arXiv:2405.18199*, 2024.

Kwangjun Ahn, Gagik Magakyan, and Ashok Cutkosky. General framework for online-to-nonconvex conversion: Schedule-free SGD is also effective for nonconvex optimization. In *Forty-second International Conference on Machine Learning*, 2025.

Noah Amsel, David Persson, Christopher Musco, and Robert M Gower. The polar express: Optimal matrix sign methods and their application to the Muon algorithm. *arXiv preprint arXiv:2505.16932*, 2025.

Kang An, Yuxing Liu, Rui Pan, Yi Ren, Shiqian Ma, Donald Goldfarb, and Tong Zhang. ASGO: Adaptive structured gradient optimization. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

Raman Arora. On learning rotations. *Advances in neural information processing systems*, 22, 2009.

Sanjeev Arora and Satyen Kale. A combinatorial, primal-dual approach to semidefinite programs. In *Proceedings of the thirty-ninth annual ACM symposium on Theory of computing*, pages 227–236, 2007.

Sanjeev Arora, Elad Hazan, and Satyen Kale. The multiplicative weights update method: a meta-algorithm and applications. *Theory of computing*, 8(1):121–164, 2012.

Dimitri P Bertsekas. Stochastic optimization problems with nondifferentiable cost functionals. *Journal of Optimization Theory and Applications*, 12(2):218–231, 1973.

Dimitri P Bertsekas. *Nonlinear Programming*. Athena Scientific, Belmont, MA, 2nd edition, 1999.

R. Bhatia. *Matrix Analysis*. Springer, New York, 1997.

Giovanni Cavallanti, Nicolo Cesa-Bianchi, and Claudio Gentile. Linear algorithms for online multitask classification. *The Journal of Machine Learning Research*, 11:2901–2934, 2010.

Ashok Cutkosky. Artificial constraints and hints for unbounded online learning. In *Conference on Learning Theory*, pages 874–894. PMLR, 2019.

Ashok Cutkosky. Better full-matrix regret via parameter-free online learning. *Advances in Neural Information Processing Systems*, 33:8836–8846, 2020.

Ashok Cutkosky, Harsh Mehta, and Francesco Orabona. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. In *International Conference on Machine Learning*, pages 6643–6670. PMLR, 2023.

John Duchi, Elad Hazan, and Yoram Singer. Adaptive subgradient methods for online learning and stochastic optimization. *Journal of machine learning research*, 12(7), 2011.

John C Duchi, Peter L Bartlett, and Martin J Wainwright. Randomized smoothing for stochastic optimization. *SIAM Journal on Optimization*, 22(2):674–701, 2012.

Paul Glasserman. *Gradient estimation via perturbation analysis*, volume 116. Springer Science & Business Media, 1991.

Allen A Goldstein. Optimization of Lipschitz continuous functions. *Mathematical Programming*, 13(1):14–22, 1977.

Gene H Golub and Charles F Van Loan. *Matrix computations*. JHU press, 2013.

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

Elad Hazan and Edgar Minasyan. Faster projection-free online learning. In *Conference on Learning Theory*, pages 1877–1893. PMLR, 2020.

Elad Hazan, Satyen Kale, and Manfred K. Warmuth. Corrigendum to "learning rotations with little regret". Self-published online correction, September 2010a.

Elad Hazan, Satyen Kale, and Manfred K Warmuth. Learning rotations with little regret. In *COLT*, pages 144–154, 2010b.

Elad Hazan, Satyen Kale, and Shai Shalev-Shwartz. Near-optimal algorithms for online matrix prediction. In *Conference on Learning Theory*, pages 38–1. JMLR Workshop and Conference Proceedings, 2012.

Elad Hazan, Satyen Kale, and Manfred K Warmuth. Learning rotations with little regret. *Machine Learning*, 104(1):129–148, 2016.

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

Grant Hillier and Raymond Kan. Properties of the inverse of a noncentral Wishart matrix. *Econometric Theory*, 38(6):1092–1116, 2022.

Ruichen Jiang and Aryan Mokhtari. Accelerated quasi-Newton proximal extragradient: Faster rate for smooth convex optimization. *Advances in Neural Information Processing Systems*, 36:8114–8151, 2023.

Ruichen Jiang and Aryan Mokhtari. Online learning guided quasi-Newton methods with global non-asymptotic convergence. *arXiv preprint arXiv:2410.02626*, 2024.

Ruichen Jiang, Qiujiang Jin, and Aryan Mokhtari. Online learning guided curvature approximation: A quasi-Newton method with global non-asymptotic superlinear convergence. In *The Thirty Sixth Annual Conference on Learning Theory*, pages 1962–1992. PMLR, 2023.

Ruichen Jiang, Aryan Mokhtari, and Francisco Patitucci. Improved complexity for smooth nonconvex optimization: a two-level online learning approach with quasi-Newton methods. In *Proceedings of the 57th Annual ACM Symposium on Theory of Computing*, pages 2225–2236, 2025.

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

Sham M Kakade, Shai Shalev-Shwartz, and Ambuj Tewari. Regularization techniques for learning with matrices. *The Journal of Machine Learning Research*, 13:1865–1890, 2012.

Adrian S Lewis and Hristo S Sendov. Twice differentiable spectral functions. *SIAM Journal on Matrix Analysis and Applications*, 23(2):368–386, 2001.

H Brendan McMahan. A survey of algorithms and analysis for adaptive online learning. *Journal of Machine Learning Research*, 18(90):1–50, 2017.

H Brendan McMahan and Matthew Streeter. Adaptive bound optimization for online convex optimization. *COLT 2010*, page 244, 2010.

Karthik Mohan and Maryam Fazel. Iterative reweighted algorithms for matrix rank minimization. *The Journal of Machine Learning Research*, 13(1):3441–3473, 2012.

Robb J. Muirhead. *Aspects of Multivariate Statistical Theory*. Wiley Series in Probability and Statistics. John Wiley & Sons, New York, 2005.

Yuji Nakatsukasa and Nicholas J Higham. Stable and efficient spectral divide and conquer algorithms for the symmetric eigenvalue decomposition and the SVD. *SIAM Journal on Scientific Computing*, 35(3):A1325–A1349, 2013.

Yuji Nakatsukasa, Zhaojun Bai, and François Gygi. Optimizing Halley's iteration for computing the matrix polar decomposition. *SIAM Journal on Matrix Analysis and Applications*, 31(5):2700–2720, 2010.

Jiazhong Nie. *Optimal online learning with matrix parameters*. University of California, Santa Cruz, 2015.

Francesco Orabona and Dávid Pál. Scale-free online learning. *Theoretical Computer Science*, 716: 50–69, 2018.

Matthew Streeter and H Brendan McMahan. Less regret via online conditioning. *arXiv preprint arXiv:1002.4862*, 2010.

Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. *arXiv preprint arXiv:1011.3027*, 2010.

Manfred K Warmuth and Dima Kuzmin. Online variance minimization. In *Proceedings of the 19th Annual Conference on Learning Theory (COLT '06)*, 2006.

Manfred K Warmuth and Dima Kuzmin. Randomized online PCA algorithms with regret bounds that are logarithmic in the dimension. *Journal of Machine Learning Research*, 9(10):2287–2320, 2008.

Shuo Xie, Tianhao Wang, Sashank J. Reddi, Sanjiv Kumar, and Zhiyuan Li. Structured preconditioners in adaptive optimization: A unified analysis. In *Forty-second International Conference on Machine Learning*, 2025.

Farzad Yousefian, Angelia Nedić, and Uday V Shanbhag. Convex nondifferentiable stochastic optimization: A local randomized smoothing technique. In *Proceedings of the 2010 American Control Conference*, pages 4875–4880. IEEE, 2010.

Qinzi Zhang and Ashok Cutkosky. Random scaling and momentum for non-smooth non-convex optimization. *arXiv preprint arXiv:2405.09742*, 2024.

Contents of Appendix

Appendix A. Additional Motivating Examples

Learning rotations The problem of learning rotations aims to recover an underlying rotation matrix from a sequence of examples (Arora, 2009; Hazan et al., 2010b,a; Nie, 2015; Hazan et al., 2016). This fundamental problem arises in a wide range of applications, including computer vision, face recognition, robotics, crystallography, and physics; we refer the reader to Arora (2009) for the references. Formally, at round t, the learner observes a unit vector $\boldsymbol{u}_t \in \mathbb{R}^n$ and chooses an orthogonal matrix $\mathbf{X}_t \in \mathbb{O}(n)$. Subsequently, a target unit vector $\boldsymbol{v}_t \in \mathbb{R}^n$ is revealed, and the learner suffers the squared prediction error $\frac{1}{2}\|\boldsymbol{v}_t - \mathbf{X}_t\boldsymbol{u}_t\|^2 = 1 - \boldsymbol{v}_t^\top \mathbf{X}_t \boldsymbol{u}_t$. As shown in Hazan et al. (2010a); Nie (2015), one natural approach is to convexify the decision set $\mathbb{O}(n)$ and use randomization to produce an orthogonal matrix. Since the convex hull of the orthogonal group is the operator-norm unit ball $\{\mathbf{X} \in \mathbb{R}^{n\times n} : \|\mathbf{X}\|_{\text{op}} \leq 1\}$, this problem can be formulated as an instance of matrix OLO with an operator-norm constraint, where the loss function takes the form $\ell_t(\mathbf{X}) = -\boldsymbol{v}_t^\top \mathbf{X} \boldsymbol{u}_t = -\langle \boldsymbol{v}_t \boldsymbol{u}_t^\top, \mathbf{X}\rangle$.

Online learning for quasi-Newton methods Quasi-Newton methods accelerate optimization by maintaining a Hessian approximation, thereby avoiding the prohibitive cost of exact curvature computation. Recent advances (Jiang et al., 2023; Jiang and Mokhtari, 2023, 2024; Jiang et al., 2025) have recast the update of this approximation as an online learning problem. Specifically, the optimizer sequentially selects a matrix $\mathbf{B}_t$ from the spectral set $\mathcal{K} = \{\mathbf{B} \in \mathbb{S}^n : \mu\mathbf{I} \preceq \mathbf{B} \preceq L\mathbf{I}\}$. Subsequently, the environment reveals the iterate difference $\boldsymbol{s}_t$ and gradient difference $\boldsymbol{y}_t$, and the learner suffers the loss $\ell_t(\mathbf{B}) = \frac{1}{2\|\boldsymbol{s}_t\|^2}\|\boldsymbol{y}_t - \mathbf{B}\boldsymbol{s}_t\|^2$. The regret bounds of this online sequence directly dictate the global convergence rate of the underlying optimization method (Jiang et al., 2023). The geometry of this problem is intrinsically spectral. The constraint $\mu\mathbf{I} \preceq \mathbf{B}_t \preceq L\mathbf{I}$ serves a dual purpose: it guarantees a baseline linear convergence rate comparable to gradient descent and ensures the condition number of subproblems remains bounded. Crucially, this constraint set is affinely equivalent to the unit operator-norm ball constraint $\|\hat{\mathbf{B}}\|_{\text{op}} \leq 1$ via the transformation $\hat{\mathbf{B}} = \frac{2}{L-\mu}(\mathbf{B} - \frac{\mu+L}{2}\mathbf{I})$. Since the loss ℓ_t is convex, standard linearization reduces the task to minimizing a sequence of linear losses over this spectral set. Consequently, designing efficient quasi-Newton updates reduces to solving a Matrix OLO problem over the operator-norm ball.

Appendix B. Adaptive Methods in the Vector Case

In this section, we review three canonical vector-based frameworks: (i) Online Gradient Descent (OGD), (ii) Follow-the-Regularized-Leader (FTRL), and (iii) Follow-the-Perturbed-Leader (FTPL). In the vector setting, these algorithms achieve adaptivity by explicitly tailoring the regularizer, local metric, or perturbation distribution to the structure of the constraint set. By clarifying how each framework exploits vector geometry, we motivate the need to adapt to spectral geometry in the matrix setting.

OGD A general formulation of adaptive OGD (Streeter and McMahan, 2010; Duchi et al., 2011) is given by

$$\boldsymbol{x}_{t+1} = \arg\min_{\boldsymbol{x}\in\mathcal{X}} \Big\{ \boldsymbol{g}_t^\top \boldsymbol{x} + \frac{1}{2\eta}(\boldsymbol{x} - \boldsymbol{x}_t)^\top \mathbf{H}_t (\boldsymbol{x} - \boldsymbol{x}_t) \Big\}, \tag{19}$$

where $\boldsymbol{g}_t$ is the gradient, $\eta > 0$ is a scaling parameter, and $\mathbf{H}_t$ is a preconditioner matrix constructed from the past gradients. It is the preconditioner $\mathbf{H}_t$ that enables data-dependent regret bounds: depending on its choice, one obtains different variants of AdaGrad, including scalar, diagonal, and

Algorithm	**Preconditioner $\mathbf{H}_t$**	**Regret Bound**
Scalar AdaGrad-OGD (Streeter and McMahan, 2010)	$\sqrt{\sum_{s=1}^t \|\boldsymbol{g}_s\|_2^2}\mathbf{I}$	$\|\mathcal{X}\|_2\sqrt{\sum_{t=1}^T \|\boldsymbol{g}_t\|_2^2}$
Diagonal AdaGrad-OGD (Streeter and McMahan, 2010; Duchi et al., 2011)	$\sqrt{\operatorname{diag}(\sum_{s=1}^t \boldsymbol{g}_s\boldsymbol{g}_s^\top)}$	$\|\mathcal{X}\|_\infty \operatorname{Tr}\left[\sqrt{\operatorname{diag}(\sum_{t=1}^T \boldsymbol{g}_t\boldsymbol{g}_t^\top)}\right]$
Full-matrix AdaGrad-OGD (Duchi et al., 2011)	$\sqrt{\sum_{s=1}^t \boldsymbol{g}_s\boldsymbol{g}_s^\top}$	$\|\mathcal{X}\|_2 \operatorname{Tr}\left[\sqrt{\sum_{t=1}^T \boldsymbol{g}_t\boldsymbol{g}_t^\top}\right]$
Scalar AdaGrad-FTRL (Duchi et al., 2011)	$\delta\mathbf{I} + \sqrt{\sum_{s=1}^t \|\boldsymbol{g}_s\|_2^2}\mathbf{I}$	$\|\mathcal{X}\|_2\sqrt{\sum_{t=1}^T \|\boldsymbol{g}_t\|_2^2} + \delta\|\mathcal{X}\|_2$
Diagonal AdaGrad-FTRL (Duchi et al., 2011)	$\delta\mathbf{I} + \sqrt{\operatorname{diag}(\sum_{s=1}^t \boldsymbol{g}_s\boldsymbol{g}_s^\top)}$	$\|\mathcal{X}\|_\infty \sum_{i=1}^d \sqrt{\sum_{t=1}^T g_{t,i}^2} + \delta\|\mathcal{X}\|_1$
Full-matrix AdaGrad-FTRL (Duchi et al., 2011)	$\delta\mathbf{I} + \sqrt{\sum_{s=1}^t \boldsymbol{g}_s\boldsymbol{g}_s^\top}$	$\|\mathcal{X}\|_2 \operatorname{Tr}\left[(\sum_{t=1}^T \boldsymbol{g}_t\boldsymbol{g}_t^\top)^{\frac{1}{2}}\right] + \delta\|\mathcal{X}\|_2$

Table 3: Regret guarantees for AdaGrad for OGD and FTRL-type updates. Here, we define $\|\mathcal{X}\|_p = \max_{\boldsymbol{x},\boldsymbol{y}\in\mathcal{X}} \|\boldsymbol{x}-\boldsymbol{y}\|_p$ for $p = 1, 2, \infty$. Moreover, $\delta \geq \max_t \|\boldsymbol{g}_t\|_2$ for the scalar and full-matrix cases, and $\delta \geq \max_t \|\boldsymbol{g}_t\|_\infty$ for the diagonal case.

full-matrix versions. In Table 3, we summarize the corresponding regret guarantees for these variants of AdaGrad. These choices naturally form a hierarchy of increasing expressive preconditioners, ranging from a scalar, to a diagonal matrix, to a full matrix. Somewhat surprisingly, however, this does not translate into a corresponding hierarchy of regret bounds. Instead, as we discuss further in Section B.1, their regret guarantees depend crucially on the geometry of the constraint set $\mathcal{X}$ and no single variant is uniformly superior. Consequently, obtaining the sharpest regret bound requires matching the choice of preconditioner to this geometry. Duchi et al. (2011) also proposed AdaGrad-style methods based on the FTRL update

$$\boldsymbol{x}_{t+1} = \arg\min_{\boldsymbol{x}\in\mathcal{X}} \left\{ \sum_{s=1}^t \boldsymbol{g}_s^\top \boldsymbol{x} + \frac{1}{2\eta}(\boldsymbol{x}-\boldsymbol{x}_1)^\top \mathbf{H}_t (\boldsymbol{x}-\boldsymbol{x}_1) \right\}.$$

The preconditioner matrix $\mathbf{H}_t$ is chosen similarly to the OGD case, except that an additional regularization term $\delta\mathbf{I}$ is typically needed to address the so-called "off-by-one" issue in the regret analysis (McMahan, 2017). The corresponding update rules and regret guarantees are also summarized in Table 3.

Remark 14 *Choosing the additional regularization term $\delta\mathbf{I}$ properly requires knowledge of the maximum gradient norm, and hence makes the algorithm not fully adaptive. Specifically, the scalar and full-matrix cases require $\delta \geq \max_t \|\boldsymbol{g}_t\|_2$, while the diagonal case requires $\delta \geq \max_t \|\boldsymbol{g}_t\|_\infty$. For the scalar and diagonal variants, this issue can be fixed by either setting $\delta = 0$ and using a more refined analysis (Orabona and Pál, 2018), or by applying the clipping technique introduced by Cutkosky (2019). Another alternative is to use Proximal FTRL as discussed in McMahan (2017).*

FTPL Adaptive algorithms for FTPL are less explored. Abernethy et al. (2016) proposed an FTPL with Gaussian perturbation for online learning with ℓ_2-Euclidean ball, given by

$$\boldsymbol{x}_{t+1} = \mathbb{E}_{\boldsymbol{r}\sim\mathcal{N}(0,\mathbf{I})}\left[\arg\min_{\|\boldsymbol{x}\|_2\leq D} \left\{ \left(\sum_{s=1}^t \boldsymbol{g}_s + \eta_t \boldsymbol{r}\right)^\top \boldsymbol{x} \right\}\right].$$

It is shown that when η_t is chosen adaptively, it recovers the same regret bound as Scalar AdaGrad-FTRL.

B.1. Comparison

From Table 3, we see that the geometry of the decision set $\mathcal{X}$ plays an important role in determining the regret of these AdaGrad variants, and hence which variant is preferable. In the following, we consider two special cases: (i) $\mathcal{X}$ is an ℓ_2-norm ball; (ii) $\mathcal{X}$ is an ℓ_∞-norm ball. Interestingly, both can be considered as a special case of the operator-norm ball in the matrix space. Specifically, when $m = 1$, the matrix decision variable $\mathbf{X} \in \mathbb{R}^{m\times n}$ can be identified by a vector $\boldsymbol{x} \in \mathbb{R}^n$, and it holds that $\|\mathbf{X}\|_{\text{op}} = \|\boldsymbol{x}\|_2$. Moreover, when $\mathbf{X}$ is restricted to be diagonal, i.e., $\mathbf{X} = \text{diag}(\boldsymbol{x})$, then $\|\mathbf{X}\|_{\text{op}} = \|\boldsymbol{x}\|_\infty$.

Remark 15 *It follows from Jensen's inequality that* $\text{Tr}\big(\sqrt{\sum_{t=1}^T \boldsymbol{g}_t\boldsymbol{g}_t^\top}\big) \geq \sqrt{\text{Tr}\big(\sum_{t=1}^T \boldsymbol{g}_t\boldsymbol{g}_t^\top\big)} = \sqrt{\sum_{t=1}^T \|\boldsymbol{g}_t\|_2^2}$*, and hence the regret bound of Full-matrix AdaGrad is always no better than that of Scalar AdaGrad; this phenomenon has been discussed in Cutkosky (2020). Thus, in the following, we focus on comparing the regret bound between Scalar AdaGrad and Diagonal AdaGrad.*

ℓ_2-norm ball When $\mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_2 \leq D\}$, we have $\|\mathcal{X}\|_2 = \|\mathcal{X}\|_\infty = 2D$. After optimizing the scaling parameter η, we obtain:

$$\text{Scalar AdaGrad: } D\sqrt{\sum_{t=1}^T \|\boldsymbol{g}_t\|_2^2} \quad \text{vs.} \quad \text{Diagonal AdaGrad: } D\sum_{i=1}^d \sqrt{\sum_{t=1}^T g_{t,i}^2}.$$

In this case, the regret bound of Scalar AdaGrad is always no worse than that of Diagonal AdaGrad. Moreover, Scalar AdaGrad is particularly favorable when the gradient sequence is *dense*, in the sense that $\sum_{t=1}^T g_{t,i}^2$ are of comparable magnitude across coordinates. In this regime, the scalar variant can outperform the diagonal one by a factor of $\sqrt{d}$.

ℓ_∞-norm ball When $\mathcal{X} = \{\boldsymbol{x} \in \mathbb{R}^d : \|\boldsymbol{x}\|_\infty \leq D\}$, we have $\|\mathcal{X}\|_2 = 2\sqrt{d}D$ and $\|\mathcal{X}\|_\infty = 2D$. Similarly, after optimizing the choice of η, we obtain

$$\text{Scalar AdaGrad: } \sqrt{d}D\sqrt{\sum_{t=1}^T \|\boldsymbol{g}_t\|_2^2} \quad \text{vs.} \quad \text{Diagonal AdaGrad: } D\sum_{i=1}^d \sqrt{\sum_{t=1}^T g_{t,i}^2}.$$

In this case, the comparison is reversed and Diagonal AdaGrad is always no worse than Scalar AdaGrad. As expected, Diagonal AdaGrad is particularly favorable when the gradient sequence is *sparse*, in the sense that only a small number of coordinates dominate. In this regime, the diagonal variant can outperform the scalar one by a factor of $\sqrt{d}$.

Appendix C. Improved Analysis of Shampoo

In the original Shampoo paper (Gupta et al., 2018), the authors establish the regret bound

$$\mathcal{O}\Big(\|\mathcal{X}\|_F\sqrt{r}\ \text{Tr}(\mathbf{M}_T^{1/4})\ \text{Tr}(\mathbf{N}_T^{1/4})\Big), \tag{20}$$

where $\mathbf{M}_T = \sum_{t=1}^T \mathbf{G}_t\mathbf{G}_t^\top$, $\mathbf{N}_T = \sum_{t=1}^T \mathbf{G}_t^\top\mathbf{G}_t$, $\|\mathcal{X}\|_F = \max_{\mathbf{X},\mathbf{Y}\in\mathcal{X}} \|\mathbf{X} - \mathbf{Y}\|_F$, and r is the maximum rank of $\{\mathbf{G}_t\}_{t=1}^T$. In this section, we show that a sharper regret bound for Shampoo can be obtained.

Theorem 16 *The Shampoo algorithm in (1) with $\mathbf{L}_t = \mathbf{M}_t^{1/4}$ and $\mathbf{R}_t = \mathbf{N}_t^{1/4}$ achieves*

$$\mathcal{O}\Big(\|\mathcal{X}\|_{\mathrm{op}} \max\Big\{\sqrt{\mathrm{Tr}(\mathbf{M}_T^{1/2})\ \mathrm{Tr}(\mathbf{N}_T^{1/2})},\ \|\mathbf{M}_T\|_{\mathrm{op}}^{1/4}\ \mathrm{Tr}(\mathbf{N}_T^{1/4}),\ \mathrm{Tr}(\mathbf{M}_T^{1/4})\,\|\mathbf{N}_T\|_{\mathrm{op}}^{1/4}\Big\}\Big). \quad (21)$$

Remark 17 *Using $\|\mathbf{M}_T^{1/4}\|_F = \sqrt{\mathrm{Tr}(\mathbf{M}_T^{1/2})}$ and $\|\mathbf{M}_T^{1/4}\|_{\mathrm{op}} = \|\mathbf{M}_T\|_{\mathrm{op}}^{1/4}$, the bound in (21) admits the symmetric expression*

$$\mathcal{O}\Big(\|\mathcal{X}\|_{\mathrm{op}} \max\Big\{\|\mathbf{M}_T^{1/4}\|_F\,\|\mathbf{N}_T^{1/4}\|_F,\ \|\mathbf{M}_T^{1/4}\|_{\mathrm{op}}\,\|\mathbf{N}_T^{1/4}\|_*,\ \|\mathbf{M}_T^{1/4}\|_*\,\|\mathbf{N}_T^{1/4}\|_{\mathrm{op}}\Big\}\Big).$$

Compared to (20), this bound removes the $\sqrt{r}$ factor and replaces the constraint-set dependence $\|\mathcal{X}\|_F$ by $\|\mathcal{X}\|_{\mathrm{op}}$. Moreover, since $\|\mathbf{A}\|_F \le \|\mathbf{A}\|_*$ and $\|\mathbf{A}\|_{\mathrm{op}} \le \|\mathbf{A}\|_*$, we have

$$\max\Big\{\|\mathbf{M}_T^{1/4}\|_F\,\|\mathbf{N}_T^{1/4}\|_F,\ \|\mathbf{M}_T^{1/4}\|_{\mathrm{op}}\,\|\mathbf{N}_T^{1/4}\|_*,\ \|\mathbf{M}_T^{1/4}\|_*\,\|\mathbf{N}_T^{1/4}\|_{\mathrm{op}}\Big\} \ \le\ \|\mathbf{M}_T^{1/4}\|_*\,\|\mathbf{N}_T^{1/4}\|_*.$$

In regimes where $\mathbf{M}_T^{1/4}$ and $\mathbf{N}_T^{1/4}$ are close to full-rank, the improvement over $\|\mathbf{M}_T^{1/4}\|_*\|\mathbf{N}_T^{1/4}\|_*$ can be as large as a factor on the order of $\min\{m,n\}$.

The following proposition is a standard regret bound for Online Mirror Descent.

Proposition 18 *Consider the update in (1). Then, we have:*

$$\begin{aligned}
&\sum_{t=1}^{T}\langle \mathbf{G}_t, \mathbf{X}_t - \mathbf{X}\rangle \\
&\le \frac{1}{2\eta}\sum_{t=2}^{T}\big[\mathrm{Tr}\big((\mathbf{X}_t-\mathbf{X})^\top \mathbf{L}_t(\mathbf{X}_t-\mathbf{X})\mathbf{R}_t\big) - \mathrm{Tr}\big((\mathbf{X}_t-\mathbf{X})^\top \mathbf{L}_{t-1}(\mathbf{X}_t-\mathbf{X})\mathbf{R}_{t-1}\big)\big] \\
&\quad + \frac{1}{2\eta}\mathrm{Tr}((\mathbf{X}_1-\mathbf{X})^\top \mathbf{L}_1(\mathbf{X}_1-\mathbf{X})\mathbf{R}_1) + \frac{\eta}{2}\sum_{t=1}^{T}\mathrm{Tr}\big(\mathbf{G}_t^\top \mathbf{L}_t^{-1}\mathbf{G}_t\mathbf{R}_t^{-1}\big).
\end{aligned}$$

Lemma 19 *For any sequence of matrices $\{\mathbf{G}_t\}_{t=1}^T$ with $\mathbf{G}_t \in \mathbb{R}^{m\times n}$, define $\mathbf{M}_t = \sum_{s=1}^t \mathbf{G}_s\mathbf{G}_s^\top$ and $\mathbf{N}_t = \sum_{s=1}^t \mathbf{G}_s^\top\mathbf{G}_s$. Then it holds that*

$$\sum_{t=1}^{T}\mathrm{Tr}(\mathbf{G}_t^\top \mathbf{M}_t^{-1/2}\mathbf{G}_t) \le 2\,\mathrm{Tr}(\sqrt{\mathbf{M}_T}) \quad \textit{and} \quad \sum_{t=1}^{T}\mathrm{Tr}(\mathbf{G}_t\mathbf{N}_t^{-1/2}\mathbf{G}_t^\top) \le 2\,\mathrm{Tr}(\sqrt{\mathbf{N}_T}).$$

Proof For ease of notation, define $\mathbf{M}_0 = 0$ and $\mathbf{N}_0 = 0$. Then we can write $\mathrm{Tr}(\mathbf{G}_t^\top \mathbf{M}_t^{-1/2}\mathbf{G}_t) = \mathrm{Tr}(\mathbf{M}_t^{-1/2}\mathbf{G}_t\mathbf{G}_t^\top) = \mathrm{Tr}(\mathbf{M}_t^{-1/2}(\mathbf{M}_t - \mathbf{M}_{t-1}))$ for any $t \ge 1$. Moreover, note that the matrix-valued function $\psi(\mathbf{M}) = 2\,\mathrm{Tr}(\mathbf{M}^{1/2})$ is concave and $\nabla\psi(\mathbf{M}) = \mathbf{M}^{-1/2}$. Thus, it holds that $\psi(\mathbf{M}_{t-1}) - \psi(\mathbf{M}_t) \le \langle \nabla\psi(\mathbf{M}_t), \mathbf{M}_{t-1} - \mathbf{M}_t\rangle$, which is equivalent to $\mathrm{Tr}(\mathbf{M}_t^{-1/2}(\mathbf{M}_t - \mathbf{M}_{t-1})) \le 2(\mathrm{Tr}(\mathbf{M}_t^{1/2}) - \mathrm{Tr}(\mathbf{M}_{t-1}^{1/2}))$. Thus, we get

$$\sum_{t=1}^{T}\mathrm{Tr}(\mathbf{G}_t^\top \mathbf{M}_t^{-1/2}\mathbf{G}_t) \le \sum_{t=1}^{T} 2(\mathrm{Tr}(\mathbf{M}_t^{1/2}) - \mathrm{Tr}(\mathbf{M}_{t-1}^{1/2})) = 2\,\mathrm{Tr}(\mathbf{M}_T^{1/2}).$$

This proves the first inequality, and the second one follows from similar arguments. ■

Lemma 20 *Recall that* $\mathbf{M}_t = \sum_{s=1}^t \mathbf{G}_s \mathbf{G}_s^\top$ *and* $\mathbf{N}_t = \sum_{s=1}^t \mathbf{G}_s^\top \mathbf{G}_s$. *With the choices* $\mathbf{L}_t = \mathbf{M}_t^{1/4}$ *and* $\mathbf{R}_t = \mathbf{N}_t^{1/4}$, *it holds that* $\sum_{t=1}^T \operatorname{Tr}\big(\mathbf{G}_t^\top \mathbf{L}_t^{-1} \mathbf{G}_t \mathbf{R}_t^{-1}\big) \leq 2\sqrt{\operatorname{Tr}(\mathbf{M}_T^{1/2}) \operatorname{Tr}(\mathbf{N}_T^{1/2})}$.

Proof By Cauchy-Schwarz inequality, we have $\operatorname{Tr}\big(\mathbf{G}_t^\top \mathbf{L}_t^{-1} \mathbf{G}_t \mathbf{R}_t^{-1}\big) \leq \|\mathbf{L}_t^{-1} \mathbf{G}_t\|_F \|\mathbf{G}_t \mathbf{R}_t^{-1}\|_F$ for any $t \in \{1, \ldots, T\}$. By applying Cauchy-Schwarz inequality again, we obtain that

$$\sum_{t=1}^T \operatorname{Tr}\big(\mathbf{G}_t^\top \mathbf{L}_t^{-1} \mathbf{G}_t \mathbf{R}_t^{-1}\big) \leq \sum_{t=1}^T \|\mathbf{L}_t^{-1} \mathbf{G}_t\|_F \|\mathbf{G}_t \mathbf{R}_t^{-1}\|_F \leq \sqrt{\sum_{t=1}^T \|\mathbf{L}_t^{-1} \mathbf{G}_t\|_F^2} \sqrt{\sum_{t=1}^T \|\mathbf{G}_t \mathbf{R}_t^{-1}\|_F^2}. \tag{22}$$

Since $\mathbf{L}_t = \mathbf{M}_t^{1/4}$ and $\mathbf{R}_t = \mathbf{N}_t^{1/4}$, we have $\|\mathbf{L}_t^{-1} \mathbf{G}_t\|_F^2 = \operatorname{Tr}\big(\mathbf{G}_t^\top \mathbf{L}_t^{-2} \mathbf{G}_t\big) = \operatorname{Tr}(\mathbf{G}_t^\top \mathbf{M}_t^{-1/2} \mathbf{G}_t)$ and similarly $\|\mathbf{G}_t \mathbf{R}_t^{-1}\|_F^2 = \operatorname{Tr}(\mathbf{G}_t \mathbf{N}_t^{-1/2} \mathbf{G}_t^\top)$. The statement now follows by using Lemma 19 in (22). ■

Lemma 21 *We have* $\operatorname{Tr}((\mathbf{X}_1 - \mathbf{X})^\top \mathbf{L}_1 (\mathbf{X}_1 - \mathbf{X}) \mathbf{R}_1) + \sum_{t=2}^T \big[\operatorname{Tr}\big((\mathbf{X}_t - \mathbf{X})^\top \mathbf{L}_t (\mathbf{X}_t - \mathbf{X}) \mathbf{R}_t\big) - \operatorname{Tr}\big((\mathbf{X}_t - \mathbf{X})^\top \mathbf{L}_{t-1} (\mathbf{X}_t - \mathbf{X}) \mathbf{R}_{t-1}\big)\big] \leq \|\mathcal{X}\|_{\mathrm{op}}^2 \|\mathbf{M}_T\|_{\mathrm{op}}^{1/4} \operatorname{Tr}(\mathbf{N}_T^{1/4}) + \|\mathcal{X}\|_{\mathrm{op}}^2 \operatorname{Tr}(\mathbf{M}_T^{1/4}) \|\mathbf{N}_T\|_{\mathrm{op}}^{1/4}$.

Proof We can write

$$\begin{aligned}
&\langle \mathbf{X}_t - \mathbf{X}, \mathbf{L}_t (\mathbf{X}_t - \mathbf{X}) \mathbf{R}_t \rangle - \langle \mathbf{X}_t - \mathbf{X}, \mathbf{L}_{t-1} (\mathbf{X}_t - \mathbf{X}) \mathbf{R}_{t-1} \rangle \\
&= \langle (\mathbf{X}_t - \mathbf{X}) \mathbf{R}_t (\mathbf{X}_t - \mathbf{X})^\top, \mathbf{L}_t - \mathbf{L}_{t-1} \rangle + \langle (\mathbf{X}_t - \mathbf{X})^\top \mathbf{L}_{t-1} (\mathbf{X}_t - \mathbf{X}), \mathbf{R}_t - \mathbf{R}_{t-1} \rangle.
\end{aligned}$$

Note that for two PSD matrices $\mathbf{A}$ and $\mathbf{B}$, $\langle \mathbf{A}, \mathbf{B} \rangle \leq \|\mathbf{A}\|_{\mathrm{op}} \operatorname{Tr}(\mathbf{B})$. Therefore, we have

$$\langle (\mathbf{X}_t - \mathbf{X}) \mathbf{R}_t (\mathbf{X}_t - \mathbf{X})^\top, \mathbf{L}_t - \mathbf{L}_{t-1} \rangle \leq \|(\mathbf{X}_t - \mathbf{X}) \mathbf{R}_t (\mathbf{X}_t - \mathbf{X})^\top\|_{\mathrm{op}} \operatorname{Tr}(\mathbf{L}_t - \mathbf{L}_{t-1}).$$

Moreover, $\|\mathbf{AB}\|_{\mathrm{op}} \leq \|\mathbf{A}\|_{\mathrm{op}} \|\mathbf{B}\|_{\mathrm{op}}$. Thus, we further have

$$\|(\mathbf{X}_t - \mathbf{X}) \mathbf{R}_t (\mathbf{X}_t - \mathbf{X})^\top\|_{\mathrm{op}} \leq \|\mathbf{X}_t - \mathbf{X}\|_{\mathrm{op}}^2 \|\mathbf{R}_t\|_{\mathrm{op}} \leq \|\mathcal{X}\|_{\mathrm{op}}^2 \|\mathbf{R}_T\|_{\mathrm{op}}.$$

Combining these two together leads to

$$\langle (\mathbf{X}_t - \mathbf{X}) \mathbf{R}_t (\mathbf{X}_t - \mathbf{X})^\top, \mathbf{L}_t - \mathbf{L}_{t-1} \rangle \leq \|\mathcal{X}\|_{\mathrm{op}}^2 \|\mathbf{R}_T\|_{\mathrm{op}} (\operatorname{Tr}(\mathbf{L}_t) - \operatorname{Tr}(\mathbf{L}_{t-1})).$$

Similarly, we have

$$\langle (\mathbf{X}_t - \mathbf{X})^\top \mathbf{L}_{t-1} (\mathbf{X}_t - \mathbf{X}), \mathbf{R}_t - \mathbf{R}_{t-1} \rangle \leq \|\mathcal{X}\|_{\mathrm{op}}^2 \|\mathbf{L}_T\|_{\mathrm{op}} (\operatorname{Tr}(\mathbf{R}_t) - \operatorname{Tr}(\mathbf{R}_{t-1})).$$

Thus, by summing the inequality from $t = 2$ to $t = T$, we conclude that

$$\begin{aligned}
&\sum_{t=2}^T [\langle \mathbf{X}_t - \mathbf{X}, \mathbf{L}_t (\mathbf{X}_t - \mathbf{X}) \mathbf{R}_t \rangle - \langle \mathbf{X}_t - \mathbf{X}, \mathbf{L}_{t-1} (\mathbf{X}_t - \mathbf{X}) \mathbf{R}_{t-1} \rangle] \\
&\leq \|\mathcal{X}\|_{\mathrm{op}}^2 \|\mathbf{R}_T\|_{\mathrm{op}} (\operatorname{Tr}(\mathbf{L}_T) - \operatorname{Tr}(\mathbf{L}_1)) + \|\mathcal{X}\|_{\mathrm{op}}^2 \|\mathbf{L}_T\|_{\mathrm{op}} (\operatorname{Tr}(\mathbf{R}_T) - \operatorname{Tr}(\mathbf{R}_1)) \\
&\leq \|\mathcal{X}\|_{\mathrm{op}}^2 \|\mathbf{R}_T\|_{\mathrm{op}} \operatorname{Tr}(\mathbf{L}_T) + \|\mathcal{X}\|_{\mathrm{op}}^2 \|\mathbf{L}_T\|_{\mathrm{op}} \operatorname{Tr}(\mathbf{R}_T),
\end{aligned}$$

where in the last inequality we used that $\mathrm{Tr}(\mathbf{L}_1) \geq 0$ and $\mathrm{Tr}(\mathbf{R}_1) \geq 0$. Finally, note that $\mathbf{L}_T = \mathbf{M}_T^{1/4}$ and $\mathbf{R}_T = \mathbf{N}_T^{1/4}$, and thus $\|\mathbf{L}_T\|_{\mathrm{op}} = \|\mathbf{M}_T^{1/4}\|_{\mathrm{op}} = \|\mathbf{M}_T\|_{\mathrm{op}}^{1/4}$ and similarly $\|\mathbf{R}_T\|_{\mathrm{op}} = \|\mathbf{N}_T\|_{\mathrm{op}}^{1/4}$. This concludes the proof. ■

Applying Lemmas 20 and 21 to Proposition 18 leads to

$$\sum_{t=1}^{T}\langle \mathbf{G}_t, \mathbf{X}_t - \mathbf{X}\rangle \leq \frac{\|\mathcal{X}\|_{\mathrm{op}}^2}{2\eta}\Big(\|\mathbf{M}_T\|_{\mathrm{op}}^{1/4}\,\mathrm{Tr}(\mathbf{N}_T^{1/4}) + \mathrm{Tr}(\mathbf{M}_T^{1/4})\|\mathbf{N}_T\|_{\mathrm{op}}^{1/4}\Big) + \eta\sqrt{\mathrm{Tr}(\mathbf{M}_T^{1/2})\,\mathrm{Tr}(\mathbf{N}_T^{1/2})}.$$

By setting $\eta = \|\mathcal{X}\|_{\mathrm{op}}/\sqrt{2}$, we obtain the bound in (21).

Appendix D. Gradient-based Prediction Algorithm

D.1. Proof of Lemma 2

It follows from the definition of regret that

$$\mathrm{Reg}_T = \sum_{t=1}^{T}\langle \mathbf{G}_t, \mathbf{X}_t\rangle - \min_{\|\mathbf{X}\|_{\mathrm{op}}\leq D}\sum_{t=1}^{T}\langle \mathbf{G}_t, \mathbf{X}\rangle = \sum_{t=1}^{T}\langle \mathbf{G}_t, \mathbf{X}_t\rangle + D\Big\|\sum_{t=1}^{T}\mathbf{G}_t\Big\|_*. \tag{23}$$

Recall that $\mathbf{S}_t = \sum_{s=1}^{t}\mathbf{G}_s$ and $\Phi(\mathbf{S}) = \|\mathbf{S}\|_*$, and hence the last term above is $D\Phi(\mathbf{S}_T)$. Moreover, for $t \geq 1$, using the update in (2), we write

$$\begin{aligned}
\langle \mathbf{G}_{t+1}, \mathbf{X}_{t+1}\rangle &= \langle \mathbf{S}_{t+1} - \mathbf{S}_t, -D\nabla\widetilde{\Phi}_t(\mathbf{S}_t)\rangle \\
&= D(\widetilde{\Phi}_t(\mathbf{S}_{t+1}) - \widetilde{\Phi}_t(\mathbf{S}_t) - \langle \nabla\widetilde{\Phi}_t(\mathbf{S}_t), \mathbf{S}_{t+1} - \mathbf{S}_t\rangle) - D\widetilde{\Phi}_t(\mathbf{S}_{t+1}) + D\widetilde{\Phi}_t(\mathbf{S}_t) \\
&= D\mathcal{B}_{\widetilde{\Phi}_t}(\mathbf{S}_{t+1}\,\|\,\mathbf{S}_t) - D\widetilde{\Phi}_t(\mathbf{S}_{t+1}) + D\widetilde{\Phi}_t(\mathbf{S}_t).
\end{aligned}$$

Adding and subtracting $D\widetilde{\Phi}_{t+1}(\mathbf{S}_{t+1})$, this can be written as

$$\langle \mathbf{G}_{t+1}, \mathbf{X}_{t+1}\rangle = D\mathcal{B}_{\widetilde{\Phi}_t}(\mathbf{S}_{t+1}\,\|\,\mathbf{S}_t) + D(\widetilde{\Phi}_{t+1}(\mathbf{S}_{t+1}) - \widetilde{\Phi}_t(\mathbf{S}_{t+1})) + D(\widetilde{\Phi}_t(\mathbf{S}_t) - \widetilde{\Phi}_{t+1}(\mathbf{S}_{t+1})).$$

Summing the above inequality from $t = 1$ to $t = T-1$, and noting that $\langle \mathbf{G}_1, \mathbf{X}_1\rangle = 0$, we obtain:

$$\sum_{t=1}^{T}\langle \mathbf{G}_t, \mathbf{X}_t\rangle = D\sum_{t=1}^{T-1}\mathcal{B}_{\widetilde{\Phi}_t}(\mathbf{S}_{t+1}\,\|\,\mathbf{S}_t) + D\sum_{t=1}^{T-1}(\widetilde{\Phi}_{t+1}(\mathbf{S}_{t+1}) - \widetilde{\Phi}_t(\mathbf{S}_{t+1})) + D(\widetilde{\Phi}_1(\mathbf{S}_1) - \widetilde{\Phi}_T(\mathbf{S}_T)). \tag{24}$$

Finally, combining (23) and (24) yields the desired result.

D.2. Proof of Theorem 4

We apply the regret decomposition in Lemma 2 and bound each term using properties from Definition 3.

Underestimation term By dominance (Definition 3, Property (b)), we have

$$\widetilde{\Phi}_T(\mathbf{S}_T) = \widetilde{\Psi}(\mathbf{S}_T; \mathbf{L}_T/\eta) \geq \|\mathbf{S}_T\|_* = \Phi(\mathbf{S}_T).$$

Hence, $\Phi(\mathbf{S}_T) - \widetilde{\Phi}_T(\mathbf{S}_T) \leq 0$.

Bregman divergence term Since $\widetilde{\Phi}_t = \widetilde{\Psi}(\cdot; \mathbf{L}_t/\eta)$, by smoothness (Property (d)) we have

$$\mathcal{B}_{\widetilde{\Phi}_t}(\mathbf{S}_{t+1} \,\|\, \mathbf{S}_t) \leq \frac{\beta\eta}{2} \operatorname{Tr}((\mathbf{S}_{t+1} - \mathbf{S}_t)^\top \mathbf{L}_t^{-1}(\mathbf{S}_{t+1} - \mathbf{S}_t)) = \frac{\beta\eta}{2} \operatorname{Tr}(\mathbf{G}_{t+1}^\top \mathbf{L}_t^{-1} \mathbf{G}_{t+1}).$$

Recall that $\mathbf{M}_t \coloneqq \sum_{s=1}^t \mathbf{G}_s \mathbf{G}_s^\top$. Since $\|\mathbf{G}_{t+1}\|_{\mathrm{op}} \leq G$, from (3) we have $\mathbf{L}_t = \sqrt{G^2\mathbf{I} + \mathbf{M}_t} \succeq \sqrt{\mathbf{M}_{t+1}}$. Applying Lemma 19, we obtain

$$\sum_{t=1}^{T-1} \mathcal{B}_{\widetilde{\Phi}_t}(\mathbf{S}_{t+1} \,\|\, \mathbf{S}_t) \leq \frac{\beta\eta}{2} \sum_{t=1}^{T-1} \operatorname{Tr}(\mathbf{G}_{t+1}^\top \mathbf{L}_t^{-1} \mathbf{G}_{t+1}) \leq \beta\eta\big(\operatorname{Tr}(\sqrt{\mathbf{M}_T}) - \|\mathbf{G}_1\|_*\big).$$

Stability term Since $\mathbf{L}_{t+1} \succeq \mathbf{L}_t$ from (3), we can use upper stability (Property (c)) to obtain

$$\widetilde{\Phi}_{t+1}(\mathbf{S}_{t+1}) - \widetilde{\Phi}_t(\mathbf{S}_{t+1}) = \widetilde{\Psi}(\mathbf{S}_{t+1}; \mathbf{L}_{t+1}/\eta) - \widetilde{\Psi}(\mathbf{S}_{t+1}; \mathbf{L}_t/\eta) \leq \alpha \left(\frac{1}{\eta} \operatorname{Tr}(\mathbf{L}_{t+1}) - \frac{1}{\eta} \operatorname{Tr}(\mathbf{L}_t) \right).$$

Summing the above inequality from $t = 1$ to $T - 1$, we get

$$\sum_{t=1}^{T-1} \big(\widetilde{\Phi}_{t+1}(\mathbf{S}_{t+1}) - \widetilde{\Phi}_t(\mathbf{S}_{t+1})\big) \leq \sum_{t=1}^{T-1} \frac{\alpha}{\eta} (\operatorname{Tr}(\mathbf{L}_{t+1}) - \operatorname{Tr}(\mathbf{L}_t)) = \frac{\alpha}{\eta} \big(\operatorname{Tr}(\mathbf{L}_T) - \operatorname{Tr}(\mathbf{L}_1)\big).$$

Moreover, using $\widetilde{\Psi}(\mathbf{X}; \mathbf{0}) = \|\mathbf{X}\|_*$ and upper stability once more, we have

$$\widetilde{\Phi}_1(\mathbf{G}_1) = \widetilde{\Psi}(\mathbf{G}_1; \mathbf{0}) + \widetilde{\Psi}(\mathbf{G}_1; \mathbf{L}_1/\eta) - \widetilde{\Psi}(\mathbf{G}_1; \mathbf{0}) \leq \|\mathbf{G}_1\|_* + \frac{\alpha}{\eta} \operatorname{Tr}(\mathbf{L}_1).$$

Combining the above,

$$\sum_{t=1}^{T-1} \big(\widetilde{\Phi}_{t+1}(\mathbf{S}_{t+1}) - \widetilde{\Phi}_t(\mathbf{S}_{t+1})\big) + \widetilde{\Phi}_1(\mathbf{G}_1) \leq \|\mathbf{G}_1\|_* + \frac{\alpha}{\eta} \operatorname{Tr}(\mathbf{L}_T).$$

Finally, Using $\sqrt{\mathbf{M}_T} \preceq \mathbf{L}_T$ and collecting all terms,

$$\operatorname{Reg}_T \leq D\Big((\beta\eta + \tfrac{\alpha}{\eta}) \operatorname{Tr}(\mathbf{L}_T) + (1 - \beta\eta)\|\mathbf{G}_1\|_*\Big).$$

Choosing $\eta = \sqrt{\alpha/\beta}$ yields the stated bound.

D.3. Proof of Proposition 5

We first show the lower bound that $\alpha\beta \geq \frac{1}{2}$. Let $\widetilde{\Psi}$ be any (α, β)-admissible smoothing of $\|\cdot\|_*$ according to Definition 3. By upper stability (Property (c)), for any $\mathbf{L} \succ 0$,

$$\widetilde{\Psi}(\mathbf{0}; \mathbf{L}) - \widetilde{\Psi}(\mathbf{0}; \mathbf{0}) \leq \alpha(\operatorname{Tr}(\mathbf{L}) - \operatorname{Tr}(\mathbf{0})) = \alpha \operatorname{Tr}(\mathbf{L}).$$

Moreover, from dominance in Property (b) we have $\widetilde{\Psi}(\mathbf{0}; \mathbf{0}) = \|\mathbf{0}\|_* = 0$, hence $\widetilde{\Psi}(\mathbf{0}; \mathbf{L}) \leq \alpha \operatorname{Tr}(\mathbf{L})$. Next, by smoothness (Property (d)), for any $\mathbf{X} \in \mathbb{R}^{m \times n}$, it holds that

$$\widetilde{\Psi}(\mathbf{X}; \mathbf{L}) \leq \widetilde{\Psi}(\mathbf{0}; \mathbf{L}) + \langle \nabla \widetilde{\Psi}(\mathbf{0}; \mathbf{L}), \mathbf{X} \rangle + \frac{\beta}{2} \operatorname{Tr}(\mathbf{X}^\top \mathbf{L}^{-1} \mathbf{X}),$$

$$\widetilde{\Psi}(-\mathbf{X}; \mathbf{L}) \leq \widetilde{\Psi}(\mathbf{0}; \mathbf{L}) - \langle \nabla \widetilde{\Psi}(\mathbf{0}; \mathbf{L}), \mathbf{X} \rangle + \frac{\beta}{2} \operatorname{Tr}(\mathbf{X}^\top \mathbf{L}^{-1} \mathbf{X}).$$

Averaging the two inequalities yields $\frac{1}{2}(\widetilde{\Psi}(\mathbf{X};\mathbf{L}) + \widetilde{\Psi}(-\mathbf{X};\mathbf{L})) \le \widetilde{\Psi}(\mathbf{0};\mathbf{L}) + \frac{\beta}{2}\operatorname{Tr}(\mathbf{X}^\top\mathbf{L}^{-1}\mathbf{X})$. Using dominance again, we have $\widetilde{\Psi}(\pm\mathbf{X};\mathbf{L}) \ge \|\mathbf{X}\|_*$, and combining with the bound on $\widetilde{\Psi}(\mathbf{0};\mathbf{L})$, we obtain:

$$\|\mathbf{X}\|_* \le \alpha \operatorname{Tr}(\mathbf{L}) + \frac{\beta}{2}\operatorname{Tr}(\mathbf{X}^\top\mathbf{L}^{-1}\mathbf{X}), \quad \forall\, \mathbf{L} \succ 0,\, \mathbf{X} \in \mathbb{R}^{m\times n}.$$

Now fix $\varepsilon > 0$ and choose $\mathbf{L} = \sqrt{\frac{\beta}{2\alpha}}\sqrt{\mathbf{X}\mathbf{X}^\top + \varepsilon\mathbf{I}} \succ 0$. With this choice, the right-hand side becomes

$$\begin{aligned}\alpha \operatorname{Tr}(\mathbf{L}) + \frac{\beta}{2}\operatorname{Tr}(\mathbf{X}^\top\mathbf{L}^{-1}\mathbf{X}) &= \sqrt{\frac{\alpha\beta}{2}}\operatorname{Tr}(\sqrt{\mathbf{X}\mathbf{X}^\top + \varepsilon\mathbf{I}}) + \sqrt{\frac{\alpha\beta}{2}}\operatorname{Tr}(\mathbf{X}^\top(\mathbf{X}\mathbf{X}^\top + \varepsilon\mathbf{I})^{-1/2}\mathbf{X}) \\ &\le \sqrt{2\alpha\beta}\operatorname{Tr}(\sqrt{\mathbf{X}\mathbf{X}^\top + \varepsilon\mathbf{I}}).\end{aligned}$$

Letting $\varepsilon \to 0$ and using $\operatorname{Tr}(\sqrt{\mathbf{X}\mathbf{X}^\top}) = \|\mathbf{X}\|_*$, we conclude that $\|\mathbf{X}\|_* \le \sqrt{2\alpha\beta}\|\mathbf{X}\|_*$ for all $\mathbf{X}$. This is possible only if $\alpha\beta \ge \frac{1}{2}$, completing the proof of the lower bound.

In the remaining, we show that the regularized smoothing defined in (4) is $(\frac{1}{2}, 1)$-admissible and verify the properties in Definition 3 one by one.

Feasibility By Danskin's theorem (Bertsekas, 1999), the gradient of $\widetilde{\Psi}^R(\mathbf{S};\mathbf{L})$ is given by solving the maximization problem

$$\nabla_{\mathbf{S}}\widetilde{\Psi}^R(\mathbf{S};\mathbf{L}) = \operatorname*{arg\,max}_{\|\mathbf{X}\|_{\mathrm{op}}\le 1}\left\{\langle \mathbf{S}, \mathbf{X}\rangle - \frac{1}{2}\operatorname{Tr}(\mathbf{X}^\top\mathbf{L}\mathbf{X})\right\}. \tag{25}$$

In particular, this implies that $\nabla_{\mathbf{S}}\widetilde{\Psi}^R(\mathbf{S};\mathbf{L})$ is a feasible point and thus $\|\nabla_{\mathbf{S}}\widetilde{\Psi}^R(\mathbf{S};\mathbf{L})\|_{\mathrm{op}} \le 1$.

Dominance It is easy to verify that $\widetilde{\Psi}^R(\mathbf{S};\mathbf{0}) = \max_{\|\mathbf{X}\|_{\mathrm{op}}\le 1}\langle \mathbf{S}, \mathbf{X}\rangle = \|\mathbf{S}\|_*$. Moreover, for any fixed $\mathbf{X}$ that satisfies $\|\mathbf{X}\|_{\mathrm{op}} \le 1$ and $\mathbf{L} \in \mathbb{S}_+^m$, it holds that $\operatorname{Tr}(\mathbf{X}^\top\mathbf{L}\mathbf{X}) \le \operatorname{Tr}(\mathbf{L})\|\mathbf{X}\|_{\mathrm{op}}^2 \le \operatorname{Tr}(\mathbf{L})$. Hence, we further have

$$\langle \mathbf{S}, \mathbf{X}\rangle - \frac{1}{2}\operatorname{Tr}(\mathbf{X}^\top\mathbf{L}\mathbf{X}) + \frac{1}{2}\operatorname{Tr}(\mathbf{L}) \ge \langle \mathbf{S}, \mathbf{X}\rangle.$$

Maximizing both sides over $\mathbf{X} \in \{\mathbf{X} : \|\mathbf{X}\|_{\mathrm{op}} \le 1\}$, we recognize the left-hand side as $\widetilde{\Psi}^R(\mathbf{S};\mathbf{L})$, while the right-hand side yields $\max_{\|\mathbf{X}\|_{\mathrm{op}}\le 1}\langle \mathbf{S}, \mathbf{X}\rangle = \|\mathbf{S}\|_*$. Hence, this proves that $\widetilde{\Psi}^R(\mathbf{S};\mathbf{L}) \ge \|\mathbf{S}\|_*$ for all $\mathbf{L} \in \mathbb{S}_+^m$ and $\mathbf{S}$.

Upper stability Consider any $\mathbf{L}_1 \preceq \mathbf{L}_2$ and fix $\mathbf{S} \in \mathbb{R}^{m\times n}$. For any $\mathbf{X} \in \mathbb{R}^{m\times n}$, we have $\operatorname{Tr}(\mathbf{X}^\top\mathbf{L}_1\mathbf{X}) \le \operatorname{Tr}(\mathbf{X}^\top\mathbf{L}_2\mathbf{X})$. This further implies that

$$\left\{\langle \mathbf{S}, \mathbf{X}\rangle - \frac{\operatorname{Tr}(\mathbf{X}^\top\mathbf{L}_2\mathbf{X})}{2} + \frac{\operatorname{Tr}(\mathbf{L}_2)}{2}\right\} \le \left\{\langle \mathbf{S}, \mathbf{X}\rangle - \frac{\operatorname{Tr}(\mathbf{X}^\top\mathbf{L}_1\mathbf{X})}{2} + \frac{\operatorname{Tr}(\mathbf{L}_1)}{2}\right\} + \frac{\operatorname{Tr}(\mathbf{L}_2) - \operatorname{Tr}(\mathbf{L}_1)}{2}.$$

Maximizing both sides over $\mathbf{X} \in \{\mathbf{X} : \|\mathbf{X}\|_{\mathrm{op}} \le 1\}$ yields

$$\widetilde{\Psi}^R(\mathbf{S};\mathbf{L}_2) \le \widetilde{\Psi}^R(\mathbf{S};\mathbf{L}_1) + \frac{1}{2}(\operatorname{Tr}(\mathbf{L}_2) - \operatorname{Tr}(\mathbf{L}_1)).$$

Since $\mathbf{S}$ is arbitrary, this proves Property (c) holds with $\alpha = \frac{1}{2}$.

Smoothness Fix $\mathbf{L} \succ 0$ and let $\iota_{\|\mathbf{X}\|_{\mathrm{op}} \leq 1}(\mathbf{X})$ denote the indicator function of the operator-norm unit ball. Define

$$g(\mathbf{X}) \;\coloneqq\; \iota_{\|\mathbf{X}\|_{\mathrm{op}} \leq 1}(\mathbf{X}) \;+\; \frac{1}{2}\operatorname{Tr}(\mathbf{X}^\top \mathbf{L}\mathbf{X}).$$

Then the regularized potential in (4) can be written as

$$\widetilde{\Psi}^R(\mathbf{S};\mathbf{L}) = \max_{\mathbf{X}} \big\{ \langle \mathbf{S}, \mathbf{X}\rangle - g(\mathbf{X}) \big\} \;+\; \frac{1}{2}\operatorname{Tr}(\mathbf{L}),$$

where $\langle \mathbf{S}, \mathbf{X}\rangle = \operatorname{Tr}(\mathbf{S}^\top \mathbf{X})$ denotes the Frobenius inner product. Ignoring the additive constant $\frac{1}{2}\operatorname{Tr}(\mathbf{L})$, which does not depend on $\mathbf{S}$, we may view $\widetilde{\Psi}^R(\cdot;\mathbf{L})$ as the Fenchel conjugate of g.

Since $\frac{1}{2}\operatorname{Tr}(\mathbf{X}^\top \mathbf{L}\mathbf{X})$ is 1-strongly convex with respect to the norm

$$\|\mathbf{X}\|_{\mathbf{L}} \coloneqq \sqrt{\operatorname{Tr}(\mathbf{X}^\top \mathbf{L}\mathbf{X})},$$

and the indicator $\iota_{\|\mathbf{X}\|_{\mathrm{op}} \leq 1}$ is convex, the function g is 1-strongly convex with respect to $\|\cdot\|_{\mathbf{L}}$. By Fenchel duality, its conjugate $\widetilde{\Psi}^R(\cdot;\mathbf{L})$ is therefore 1-smooth with respect to the dual norm

$$\|\mathbf{S}\|_{\mathbf{L}^{-1}} \coloneqq \sqrt{\operatorname{Tr}(\mathbf{S}^\top \mathbf{L}^{-1}\mathbf{S})}.$$

Consequently, for any $\mathbf{S}_1, \mathbf{S}_2$, the associated Bregman divergence satisfies

$$\begin{aligned}\mathcal{B}_{\widetilde{\Psi}^R(\cdot;\mathbf{L})}(\mathbf{S}_2 \,\|\, \mathbf{S}_1) &= \widetilde{\Psi}^R(\mathbf{S}_2;\mathbf{L}) - \widetilde{\Psi}^R(\mathbf{S}_1;\mathbf{L}) - \langle \nabla_{\mathbf{S}} \widetilde{\Psi}^R(\mathbf{S}_1;\mathbf{L}), \mathbf{S}_2 - \mathbf{S}_1\rangle \\ &\leq \frac{1}{2}\operatorname{Tr}\big((\mathbf{S}_2 - \mathbf{S}_1)^\top \mathbf{L}^{-1}(\mathbf{S}_2 - \mathbf{S}_1)\big).\end{aligned}$$

This verifies Property (d) with $\beta = 1$.

Appendix E. Proofs for Section 4

E.1. Proofs for FTPL (Theorem 6)

In this section, we verify that the stochastic smoothing potential family $\widetilde{\Psi}^S(\mathbf{S};\mathbf{L})$ satisfies all the properties in Definition 3 with $(\alpha, \beta) = (\sqrt{m}+\sqrt{n}, \frac{1}{\sqrt{n-m-1}})$, and the regret bound directly follows from Theorem 4. The first three are relatively straightforward while characterizing the smoothness property is the main challenge here; as we shall see, it requires tools from noncentral Wishart theory.

Feasibility Using the variational representation of the nuclear norm $\|\mathbf{S}\|_* = \max_{\|\mathbf{X}\|_{\mathrm{op}} \leq 1} \langle \mathbf{S}, \mathbf{X}\rangle$ and exchanging the order of expectation and differentiation via (Bertsekas, 1973, Proposition 2.2), we have

$$\nabla \widetilde{\Psi}^S(\mathbf{S};\mathbf{L}) = \mathop{\mathbb{E}}_{\mathbf{Z}}\Big[\operatorname*{arg\,max}_{\|\mathbf{X}\|_{\mathrm{op}} \leq 1} \langle \mathbf{S} + \mathbf{L}\mathbf{Z}, \mathbf{X}\rangle\Big].$$

Thus, since the operator norm is a convex function, by Jensen's inequality we have

$$\|\nabla \widetilde{\Psi}^S(\mathbf{S};\mathbf{L})\|_{\mathrm{op}} \leq \mathop{\mathbb{E}}_{\mathbf{Z}}\Big[\Big\|\operatorname*{arg\,max}_{\|\mathbf{X}\|_{\mathrm{op}} \leq 1} \langle \mathbf{S} + \mathbf{L}\mathbf{Z}, \mathbf{X}\rangle\Big\|_{\mathrm{op}}\Big] \leq 1.$$

Dominance It is easy to see that when $\mathbf{L} = 0$, the stochastic perturbation vanishes and hence $\widetilde{\Psi}^S(\mathbf{S}; \mathbf{0}) = \|\mathbf{S}\|_*$. Moreover, since the nuclear norm is convex and $\mathbb{E}[\mathbf{Z}] = \mathbf{0}$, Jensen's inequality implies that

$$\widetilde{\Psi}^S(\mathbf{S}; \mathbf{L}) = \mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathcal{MN}(0,\mathbf{I}_m,\mathbf{I}_n)} \|\mathbf{S} + \mathbf{L}\mathbf{Z}\|_* \geq \left\|\mathbf{S} + \mathbf{L} \mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathcal{MN}(0,\mathbf{I}_m,\mathbf{I}_n)} [\mathbf{Z}]\right\|_* = \|\mathbf{S}\|_*.$$

This proves Property (b).

Upper stability Consider any two matrices $\mathbf{L}_1 \preceq \mathbf{L}_2$ and fix $\mathbf{S} \in \mathbb{R}^{m\times n}$. Using the linearity of expectation and triangle inequality, we have

$$\begin{aligned}
\widetilde{\Psi}^S(\mathbf{S}; \mathbf{L}_2) - \widetilde{\Psi}^S(\mathbf{S}; \mathbf{L}_1) &= \mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathcal{MN}(0,\mathbf{I}_m,\mathbf{I}_n)} \|\mathbf{S} + \mathbf{L}_2\mathbf{Z}\|_* - \mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathcal{MN}(0,\mathbf{I}_m,\mathbf{I}_n)} \|\mathbf{S} + \mathbf{L}_1\mathbf{Z}\|_* \\
&= \mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathcal{MN}(0,\mathbf{I}_m,\mathbf{I}_n)} [\|\mathbf{S} + \mathbf{L}_2\mathbf{Z}\|_* - \|\mathbf{S} + \mathbf{L}_1\mathbf{Z}\|_*] \\
&\leq \mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathcal{MN}(0,\mathbf{I}_m,\mathbf{I}_n)} [\|(\mathbf{L}_2 - \mathbf{L}_1)\mathbf{Z}\|_*] .
\end{aligned}$$

Moreover, it holds that $\|(\mathbf{L}_2 - \mathbf{L}_1)\mathbf{Z}\|_* \leq \|\mathbf{Z}\|_{\mathrm{op}}\|\mathbf{L}_2 - \mathbf{L}_1\|_*$, and Gordon's inequality for Gaussian matrices (Vershynin, 2010, Theorem 5.32) establishes that $\mathbb{E}_{\mathbf{Z}\sim\mathcal{MN}(0,\mathbf{I}_m,\mathbf{I}_n)} \|\mathbf{Z}\|_{\mathrm{op}} \leq \sqrt{m} + \sqrt{n}$. Using the fact that $\|\mathbf{L}_2 - \mathbf{L}_1\|_* = \mathrm{Tr}(\mathbf{L}_2 - \mathbf{L}_1) = \mathrm{Tr}(\mathbf{L}_2) - \mathrm{Tr}(\mathbf{L}_1)$ since $\mathbf{L}_2 - \mathbf{L}_1 \succeq \mathbf{0}$, we obtain

$$\widetilde{\Psi}^S(\mathbf{S}; \mathbf{L}_2) - \widetilde{\Psi}^S(\mathbf{S}; \mathbf{L}_1) \leq \mathop{\mathbb{E}}_{\mathbf{Z}\sim\mathcal{MN}(0,\mathbf{I}_m,\mathbf{I}_n)} \|\mathbf{Z}\|_{\mathrm{op}}\|\mathbf{L}_2 - \mathbf{L}_1\|_* \leq (\sqrt{m} + \sqrt{n})(\mathrm{Tr}(\mathbf{L}_2) - \mathrm{Tr}(\mathbf{L}_1)).$$

This proves that Property (c) is satisfied with $\alpha = \sqrt{m} + \sqrt{n}$.

Smoothness To simplify the notation, in the following, we view the Hessian of a matrix function $F : \mathbb{R}^{m\times n} \to \mathbb{R}$ as a bilinear map on $\mathbb{R}^{m\times n}$. Equivalently, under the Frobenius inner product, we identify it with a self-adjoint linear operator and write

$$\nabla^2 F(\mathbf{X})[\mathbf{D}_1, \mathbf{D}_2] = \langle \nabla^2 F(\mathbf{X})[\mathbf{D}_1], \mathbf{D}_2\rangle, \quad \forall\, \mathbf{D}_1, \mathbf{D}_2 \in \mathbb{R}^{m\times n}.$$

Fix $\mathbf{L} \succ 0$. First, since the perturbation follows a matrix Gaussian distribution, by (Abernethy et al., 2016, Lemma 1.5), $\widetilde{\Psi}^S(\mathbf{S}; \mathbf{L})$ is twice differentiable. For any $\mathbf{S}_1, \mathbf{S}_2 \in \mathbb{R}^{m\times n}$, let $\Delta\mathbf{S} = \mathbf{S}_2 - \mathbf{S}_1$. By the fundamental theorem of calculus,

$$\widetilde{\Psi}^S(\mathbf{S}_2; \mathbf{L}) - \widetilde{\Psi}^S(\mathbf{S}_1; \mathbf{L}) = \int_0^1 \langle \nabla_{\mathbf{S}} \widetilde{\Psi}^S(\mathbf{S}_1 + t\Delta\mathbf{S}; \mathbf{L}), \Delta\mathbf{S}\rangle \, dt. \tag{26}$$

For any fixed $t \in [0, 1]$, we apply the fundamental theorem of calculus again to obtain

$$\begin{aligned}
\nabla_{\mathbf{S}} \widetilde{\Psi}^S(\mathbf{S}_1 + t\Delta\mathbf{S}; \mathbf{L}) - \nabla_{\mathbf{S}} \widetilde{\Psi}^S(\mathbf{S}_1; \mathbf{L}) &= \int_0^1 \nabla^2_{\mathbf{S}} \widetilde{\Psi}^S(\mathbf{S}_1 + st\Delta\mathbf{S}; \mathbf{L})[t\Delta\mathbf{S}] \, ds \\
&= \int_0^1 t\nabla^2_{\mathbf{S}} \widetilde{\Psi}^S(\mathbf{S}_1 + st\Delta\mathbf{S}; \mathbf{L})[\Delta\mathbf{S}] \, ds.
\end{aligned} \tag{27}$$

Combining (26) and (27) yields

$$\begin{aligned}
\mathcal{B}_{\widetilde{\Psi}^S(\cdot;\mathbf{L})}(\mathbf{S}_2 \,\|\, \mathbf{S}_1) &= \widetilde{\Psi}^S(\mathbf{S}_2; \mathbf{L}) - \widetilde{\Psi}^S(\mathbf{S}_1; \mathbf{L}) - \langle \nabla_{\mathbf{S}} \widetilde{\Psi}^S(\mathbf{S}_1; \mathbf{L}), \mathbf{S}_2 - \mathbf{S}_1\rangle \\
&= \int_0^1 \int_0^1 t \, \nabla^2_{\mathbf{S}} \widetilde{\Psi}^S(\mathbf{S}_1 + st\Delta\mathbf{S}; \mathbf{L})[\Delta\mathbf{S}, \Delta\mathbf{S}] \, ds \, dt.
\end{aligned} \tag{28}$$

Given the above expression of Bregman divergence, to prove Property (d), it suffices to show that

$$\nabla^2_{\mathbf{S}} \widetilde{\Psi}^S(\mathbf{S};\mathbf{L})[\mathbf{D},\mathbf{D}] \leq \beta \operatorname{Tr}(\mathbf{D}^\top \mathbf{L}^{-1} \mathbf{D}), \quad \forall\, \mathbf{D} \in \mathbb{R}^{m\times n}. \tag{29}$$

We now prove (29) in three steps; the proofs of several technical lemmas are deferred to Appendix E.1.1. Define the random matrices $\mathbf{G} := \mathbf{S} + \mathbf{L}\mathbf{Z}$ and $\mathbf{A} := \mathbf{G}\mathbf{G}^\top$. Since $n > m$ and $\mathbf{Z} \sim \mathcal{MN}(0, \mathbf{I}_m, \mathbf{I}_n)$, $\mathbf{G}$ is full row rank almost surely.

Step 1 (Hessian reduction to $\mathbb{E}[\mathbf{A}^{-1/2}]$) Let $\Phi(\mathbf{S}) = \|\mathbf{S}\|_*$. If $\mathbf{S}$ is full row rank, then Φ is twice differentiable at $\mathbf{S}$ and for all directions $\mathbf{D}$,

$$\nabla^2 \Phi(\mathbf{S})[\mathbf{D},\mathbf{D}] \;\leq\; \operatorname{Tr}\big(\mathbf{D}^\top (\mathbf{S}\mathbf{S}^\top)^{-1/2} \mathbf{D}\big) \qquad \text{(Lemma 22)}.$$

Using dominated convergence (or Leibniz' rule) to interchange expectation and differentiation, we obtain

$$\nabla^2_{\mathbf{S}} \widetilde{\Psi}^S(\mathbf{S};\mathbf{L})[\mathbf{D},\mathbf{D}] = \mathbb{E}\big[\nabla^2 \Phi(\mathbf{S}+\mathbf{L}\mathbf{Z})[\mathbf{D},\mathbf{D}]\big] \leq \operatorname{Tr}\Big(\mathbf{D}^\top\, \mathbb{E}[\mathbf{A}^{-1/2}]\, \mathbf{D}\Big). \tag{30}$$

Step 2 (bound $\mathbb{E}[\mathbf{A}^{-1}]$ via a Wishart argument) Let $\mathbf{Y} := \mathbf{L}^{-1}\mathbf{S}$ and $\tilde{\mathbf{A}} := (\mathbf{Y}+\mathbf{Z})(\mathbf{Y}+\mathbf{Z})^\top$. Then

$$\mathbf{A} = (\mathbf{S}+\mathbf{L}\mathbf{Z})(\mathbf{S}+\mathbf{L}\mathbf{Z})^\top = \mathbf{L}(\mathbf{L}^{-1}\mathbf{S}+\mathbf{Z})(\mathbf{L}^{-1}\mathbf{S}+\mathbf{Z})^\top \mathbf{L} = \mathbf{L}\,\tilde{\mathbf{A}}\,\mathbf{L},$$

and hence

$$\mathbf{A}^{-1} = \mathbf{L}^{-1}\tilde{\mathbf{A}}^{-1}\mathbf{L}^{-1} \quad \Rightarrow \quad \mathbb{E}[\mathbf{A}^{-1}] = \mathbf{L}^{-1}\,\mathbb{E}[\tilde{\mathbf{A}}^{-1}]\mathbf{L}^{-1}. \tag{31}$$

Moreover, $\tilde{\mathbf{A}}$ follows a (noncentral) Wishart distribution with n degrees of freedom and identity scale. In particular, it holds that

$$\mathbb{E}[\tilde{\mathbf{A}}^{-1}] \preceq \frac{1}{n-m-1}\mathbf{I} \qquad \text{(Lemma 24)}. \tag{32}$$

Combining (31) and (32) yields

$$\mathbb{E}[\mathbf{A}^{-1}] \preceq \frac{1}{n-m-1}\,\mathbf{L}^{-2}. \tag{33}$$

Step 3 (from $\mathbb{E}[\mathbf{A}^{-1}]$ to $\mathbb{E}[\mathbf{A}^{-1/2}]$) Since $t \mapsto \sqrt{t}$ is operator concave on $\mathbb{S}^m_+$ (Bhatia, 1997), Jensen's inequality gives

$$\mathbb{E}[\mathbf{A}^{-1/2}] = \mathbb{E}\big[(\mathbf{A}^{-1})^{1/2}\big] \preceq \big(\mathbb{E}[\mathbf{A}^{-1}]\big)^{1/2}.$$

Using (33) and the monotonicity of the matrix square root, we have

$$\big(\mathbb{E}[\mathbf{A}^{-1}]\big)^{1/2} \preceq \left(\frac{1}{n-m-1}\mathbf{L}^{-2}\right)^{1/2} = \frac{1}{\sqrt{n-m-1}}\mathbf{L}^{-1}.$$

Plugging this into (30) leads to

$$\nabla^2_{\mathbf{S}} \widetilde{\Psi}^S(\mathbf{S};\mathbf{L})[\mathbf{D},\mathbf{D}] \leq \frac{1}{\sqrt{n-m-1}} \operatorname{Tr}(\mathbf{D}^\top \mathbf{L}^{-1} \mathbf{D}).$$

This completes the proof.

E.1.1. TECHNICAL MATRIX LEMMAS

In the following lemma, we characterize the derivative of the nuclear norm.

Lemma 22 *Consider the nuclear norm function $\Phi(\mathbf{X}) = \|\mathbf{X}\|_*$, where $\mathbf{X} \in \mathbb{R}^{m\times n}$ with $m \le n$. If $\mathbf{X}$ is full row rank, then Φ is twice differentiable at $\mathbf{X}$, and for all $\mathbf{D} \in \mathbb{R}^{m\times n}$,*

$$\nabla^2\Phi(\mathbf{X})[\mathbf{D},\mathbf{D}] \le \operatorname{Tr}(\mathbf{D}^\top(\mathbf{X}\mathbf{X}^\top)^{-1/2}\mathbf{D}).$$

Proof We begin by rewriting the nuclear norm as $\Phi(\mathbf{X}) = \|\mathbf{X}\|_* = \operatorname{Tr}(\sqrt{\mathbf{X}\mathbf{X}^\top})$. Define $\mathbf{A}(\mathbf{X}) = \mathbf{X}\mathbf{X}^\top$ and $g(\mathbf{A}) = \operatorname{Tr}(\sqrt{\mathbf{A}})$, so that $\Phi = g \circ \mathbf{A}$. The mapping $\mathbf{A}(\mathbf{X})$ is a polynomial function and hence twice differentiable everywhere. Moreover, since $\mathbf{X}$ is full row rank, $\mathbf{A}(\mathbf{X})$ is positive definite. The function g is a spectral function induced by $t \mapsto t^{1/2}$, which is C^2 on $(0,\infty)$; therefore, g is twice Fréchet differentiable over $\mathbb{S}^m_+$ (Lewis and Sendov, 2001). Consequently, Φ is twice differentiable at $\mathbf{X}$.

Moreover, applying the chain rule for Fréchet derivatives, we obtain:

$$\nabla\Phi(\mathbf{X})[\mathbf{D}] = \nabla g(\mathbf{A})[\nabla\mathbf{A}(\mathbf{X})[\mathbf{D}]] \tag{34}$$

$$\nabla^2\Phi(\mathbf{X})[\mathbf{D}_1,\mathbf{D}_2] = \nabla^2 g(\mathbf{A})[\nabla\mathbf{A}(\mathbf{X})[\mathbf{D}_1],\nabla\mathbf{A}(\mathbf{X})[\mathbf{D}_2]] + \nabla g(\mathbf{A})[\nabla^2\mathbf{A}(\mathbf{X})[\mathbf{D}_1,\mathbf{D}_2]]. \tag{35}$$

We now compute the derivatives of $\mathbf{A}$ explicitly. For any $\mathbf{D} \in \mathbb{R}^{m\times n}$, we have $\nabla\mathbf{A}(\mathbf{X})[\mathbf{D}] = \mathbf{X}\mathbf{D}^\top + \mathbf{D}\mathbf{X}^\top$, and for any $\mathbf{D}_1, \mathbf{D}_2 \in \mathbb{R}^{m\times n}$, we have $\nabla^2\mathbf{A}(\mathbf{X})[\mathbf{D}_1,\mathbf{D}_2] = \mathbf{D}_1\mathbf{D}_2^\top + \mathbf{D}_2\mathbf{D}_1^\top$. Next, we consider the derivatives of g. For any symmetric matrix $\mathbf{D}$, $\nabla g(\mathbf{A})[\mathbf{D}] = \frac{1}{2}\operatorname{Tr}(\mathbf{A}^{-1/2}\mathbf{D})$. Furthermore, since the scalar function $t \mapsto t^{1/2}$ is operator concave (Bhatia, 1997), the induced spectral function g is concave on $\mathbb{S}^m_+$. As a result, $\nabla^2 g(\mathbf{A})[\mathbf{D},\mathbf{D}] \le 0$ for all symmetric $\mathbf{D}$. Specializing (35) to $\mathbf{D}_1 = \mathbf{D}_2 = \mathbf{D}$, we obtain

$$\nabla^2\Phi(\mathbf{X})[\mathbf{D},\mathbf{D}] = \nabla^2 g(\mathbf{A})[\nabla\mathbf{A}(\mathbf{X})[\mathbf{D}],\nabla\mathbf{A}(\mathbf{X})[\mathbf{D}]] + \nabla g(\mathbf{A})[\nabla^2\mathbf{A}(\mathbf{X})[\mathbf{D},\mathbf{D}]]. \tag{36}$$

The first term in (36) is nonpositive by concavity of g. For the second term, we compute

$$\nabla g(\mathbf{A})[\nabla^2\mathbf{A}(\mathbf{X})[\mathbf{D},\mathbf{D}]] = \nabla g(\mathbf{A})[2\mathbf{D}\mathbf{D}^\top] = \frac{1}{2}\operatorname{Tr}(\mathbf{A}^{-1/2}2\mathbf{D}\mathbf{D}^\top) = \operatorname{Tr}(\mathbf{D}^\top\mathbf{A}^{-1/2}\mathbf{D}).$$

Combining these bounds and substituting $\mathbf{A} = \mathbf{X}\mathbf{X}^\top$ yields the desired inequality. ■

Before presenting our key lemma on a noncentral Wishart matrix in Lemma 24, we first recall a basic property of the noncentral chi-square distribution (Muirhead, 2005). Recall that if $\boldsymbol{x} \in \mathbb{R}^k$ satisfies $\boldsymbol{x} \sim \mathcal{N}(\boldsymbol{\mu}, \mathbf{I}_k)$, then $\mathbf{X} := \|\boldsymbol{x}\|_2^2$ follows a noncentral chi-square distribution, denoted by $\chi^2_k(\lambda)$, with degrees of freedom k and noncentrality parameter $\lambda := \|\boldsymbol{\mu}\|_2^2$.

Lemma 23 *Suppose $\mathbf{X} \sim \chi^2_k(\lambda)$ with $k \ge 3$. Then*

$$\mathbb{E}\big[\mathbf{X}^{-1}\big] \;\le\; \frac{1}{k-2}.$$

Proof A noncentral chi-square admits a Poisson mixture representation (Muirhead, 2005, Corollary 1.3.5). Specifically, if $\mathbf{X} \sim \chi^2_k(\lambda)$ and $\mathbf{J} \sim \operatorname{Pois}(\lambda/2)$, then conditional on $\mathbf{J} = j$, we have $\mathbf{X} \,|\, (\mathbf{J} = j) \sim \chi^2_{k+2j}$. Hence, by the law of total expectation,

$$\mathbb{E}\big[\mathbf{X}^{-1}\big] = \mathbb{E}\big[\,\mathbb{E}\big[\mathbf{X}^{-1} \mid \mathbf{J}\big]\big] = \mathbb{E}\Big[\,\mathbb{E}\Big[\big(\chi^2_{k+2\mathbf{J}}\big)^{-1}\Big]\Big]. \tag{37}$$

For $\mathbf{U} \sim \chi^2_r$ with $r > 1$, a direct calculation gives

$$\mathbb{E}[\mathbf{U}^{-1}] = \frac{1}{2^{r/2}\Gamma(r/2)} \int_0^\infty x^{r/2-2} e^{-x/2}\, dx = \frac{\Gamma(\frac{r}{2}-1)}{2\,\Gamma(\frac{r}{2})} = \frac{1}{r-2}.$$

Applying this with $r = k + 2\mathbf{J}$ in (37) gives

$$\mathbb{E}\big[\mathbf{X}^{-1}\big] = \mathbb{E}\Big[\mathbb{E}\Big[\big(\chi^2_{k+2\mathbf{J}}\big)^{-1}\Big]\Big] \le \mathbb{E}\left[\frac{1}{k+2\mathbf{J}-2}\right] \le \frac{1}{k-2},$$

since $\mathbf{J} \ge 0$ almost surely and $t \mapsto 1/t$ is decreasing. ■

Using Lemma 23, we are ready to present our key result on the expected inverse of a Wishart matrix. Our proof is inspired by Hillier and Kan (2022).

Lemma 24 *Let $\mathbf{Z} \in \mathbb{R}^{m\times n}$ have i.i.d. $\mathcal{N}(0,1)$ entries and let $\mathbf{Y} \in \mathbb{R}^{m\times n}$ be deterministic. Assume $n \ge m+2$ and define $\mathbf{A}_\mathbf{Y} := (\mathbf{Z}+\mathbf{Y})(\mathbf{Z}+\mathbf{Y})^\top \in \mathbb{R}^{m\times m}$. Then, for any $\mathbf{Y}$,*

$$\mathbb{E}[\mathbf{A}_\mathbf{Y}^{-1}] \;\preceq\; \frac{1}{n-m-1}\mathbf{I}_m. \tag{38}$$

Proof **Step 1: reduce to diagonal $\mathbf{Y}$.** Let $\mathbf{Y} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top$ be a singular value decomposition, where $\mathbf{U} \in \mathbb{R}^{m\times m}$ and $\mathbf{V} \in \mathbb{R}^{n\times n}$ are orthogonal, and $\mathbf{\Sigma} = [\mathrm{diag}(\sigma_1,\dots,\sigma_m)\;\; 0] \in \mathbb{R}^{m\times n}$. Let $\widetilde{\mathbf{Z}} := \mathbf{U}^\top\mathbf{Z}\mathbf{V}$. By orthogonal invariance of the standard Gaussian, $\widetilde{\mathbf{Z}} \overset{d}{=} \mathbf{Z}$, and

$$\mathbf{A}_\mathbf{Y} = (\mathbf{Z}+\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top)(\mathbf{Z}+\mathbf{U}\mathbf{\Sigma}\mathbf{V}^\top)^\top = \mathbf{U}(\widetilde{\mathbf{Z}}+\mathbf{\Sigma})(\widetilde{\mathbf{Z}}+\mathbf{\Sigma})^\top\mathbf{U}^\top.$$

Hence $\mathbf{A}_\mathbf{Y}^{-1} = \mathbf{U}\,\mathbf{A}_\mathbf{\Sigma}^{-1}\mathbf{U}^\top$ with $\mathbf{A}_\mathbf{\Sigma} := (\widetilde{\mathbf{Z}}+\mathbf{\Sigma})(\widetilde{\mathbf{Z}}+\mathbf{\Sigma})^\top$, and taking expectations gives

$$\mathbb{E}[\mathbf{A}_\mathbf{Y}^{-1}] = \mathbf{U}\;\mathbb{E}[\mathbf{A}_\mathbf{\Sigma}^{-1}]\,\mathbf{U}^\top.$$

Since $\mathbf{U}$ is orthogonal, to prove (38) it suffices to bound $\mathbb{E}[\mathbf{A}_\mathbf{\Sigma}^{-1}]$.

Step 2: $\mathbb{E}[\mathbf{A}_\mathbf{\Sigma}^{-1}]$ is diagonal Let $\mathcal{T} := \{\mathbf{T} = \mathrm{diag}(\pm1,\dots,\pm1) \in \mathbb{R}^{m\times m}\}$ and define $\widetilde{\mathbf{T}} := \mathrm{diag}(\mathbf{T}, \mathbf{I}_{n-m}) \in \mathbb{R}^{n\times n}$. Since $\mathbf{\Sigma}$ has diagonal left block, we have $\mathbf{T}\mathbf{\Sigma}\widetilde{\mathbf{T}} = \mathbf{\Sigma}$. Moreover, $\mathbf{T}\widetilde{\mathbf{Z}}\widetilde{\mathbf{T}} \overset{d}{=} \widetilde{\mathbf{Z}}$, and therefore

$$\mathbf{A}_\mathbf{\Sigma} = (\widetilde{\mathbf{Z}}+\mathbf{\Sigma})(\widetilde{\mathbf{Z}}+\mathbf{\Sigma})^\top \overset{d}{=} (\mathbf{T}\widetilde{\mathbf{Z}}\widetilde{\mathbf{T}}+\mathbf{T}\mathbf{\Sigma}\widetilde{\mathbf{T}})(\mathbf{T}\widetilde{\mathbf{Z}}\widetilde{\mathbf{T}}+\mathbf{T}\mathbf{\Sigma}\widetilde{\mathbf{T}})^\top = \mathbf{T}\,\mathbf{A}_\mathbf{\Sigma}\,\mathbf{T}.$$

Inverting the matrices and taking expectations yields $\mathbb{E}[\mathbf{A}_\mathbf{\Sigma}^{-1}] = \mathbf{T}\;\mathbb{E}[\mathbf{A}_\mathbf{\Sigma}^{-1}]\,\mathbf{T}$ for all $\mathbf{T} \in \mathcal{T}$, which forces $\mathbb{E}[\mathbf{A}_\mathbf{\Sigma}^{-1}]$ to be diagonal.

Step 3: bound each diagonal entry Since $\mathbb{E}[\mathbf{A}_\mathbf{\Sigma}^{-1}]$ is diagonal, it is enough to show $\mathbb{E}[(\mathbf{A}_\mathbf{\Sigma}^{-1})_{ii}] \le \frac{1}{n-m-1}$ for each i. Fix $i = 1$ (the others are identical by relabeling rows). Write the first row of $\widetilde{\mathbf{Z}}+\mathbf{\Sigma}$ as $\tilde{z}_1^\top \in \mathbb{R}^n$ and the remaining rows as $\widetilde{\mathbf{Z}}_{-1} \in \mathbb{R}^{(m-1)\times n}$, so that $\widetilde{\mathbf{Z}}+\mathbf{\Sigma} = \begin{bmatrix}\tilde{z}_1^\top \\ \widetilde{\mathbf{Z}}_{-1}\end{bmatrix}$. A Schur complement computation gives

$$(\mathbf{A}_\mathbf{\Sigma}^{-1})_{11} = \frac{1}{\tilde{z}_1^\top\tilde{z}_1 - \tilde{z}_1^\top\widetilde{\mathbf{Z}}_{-1}^\top(\widetilde{\mathbf{Z}}_{-1}\widetilde{\mathbf{Z}}_{-1}^\top)^{-1}\widetilde{\mathbf{Z}}_{-1}\tilde{z}_1} = \frac{1}{\tilde{z}_1^\top\mathbf{P}\,\tilde{z}_1},$$

where $\mathbf{P} := \mathbf{I}_n - \widetilde{\mathbf{Z}}_{-1}^\top(\widetilde{\mathbf{Z}}_{-1}\widetilde{\mathbf{Z}}_{-1}^\top)^{-1}\widetilde{\mathbf{Z}}_{-1}$ is the orthogonal projector onto $\mathrm{Row}(\widetilde{\mathbf{Z}}_{-1})^\perp$. Almost surely $\mathrm{rank}(\widetilde{\mathbf{Z}}_{-1}) = m-1$, hence $\mathbf{P}$ has rank $n-m+1$.

Conditioned on $\widetilde{\mathbf{Z}}_{-1}$, the vector $\tilde{\boldsymbol{z}}_1 \sim \mathcal{N}(\sigma_1 \boldsymbol{e}_1, \mathbf{I}_n)$ is independent of $\widetilde{\mathbf{Z}}_{-1}$, and thus $\tilde{\boldsymbol{z}}_1^\top \mathbf{P}\tilde{\boldsymbol{z}}_1$ is a noncentral chi-square random variable with $k = n-m+1$ degrees of freedom (and some noncentrality parameter depending on $\mathbf{P}$ and σ_1). Since $n \geq m+2$, we have $k \geq 3$, and Lemma 23 implies

$$\mathbb{E}\Big[(\mathbf{A}_{\boldsymbol{\Sigma}}^{-1})_{11} \,\Big|\, \widetilde{\mathbf{Z}}_{-1}\Big] = \mathbb{E}\Big[(\tilde{\boldsymbol{z}}_1^\top \mathbf{P}\tilde{\boldsymbol{z}}_1)^{-1} \,\Big|\, \widetilde{\mathbf{Z}}_{-1}\Big] \leq \frac{1}{k-2} = \frac{1}{n-m-1}.$$

Taking expectations over $\widetilde{\mathbf{Z}}_{-1}$ yields $\mathbb{E}[(\mathbf{A}_{\boldsymbol{\Sigma}}^{-1})_{11}] \leq \frac{1}{n-m-1}$. Therefore $\mathbb{E}[\mathbf{A}_{\boldsymbol{\Sigma}}^{-1}] \preceq \frac{1}{n-m-1}\mathbf{I}_m$, and conjugating by $\mathbf{U}$ completes the proof. ■

E.2. Proofs for FAML (Theorem 8)

In this section, we verify that the hyperbolic smoothing potential family $\widetilde{\Psi}^H(\mathbf{S};\mathbf{L})$ satisfies all the properties in Definition 3 with $(\alpha,\beta) = (1,1)$.

Feasibility Recall from Section 4.2 that $\nabla_{\mathbf{S}}\widetilde{\Psi}^H(\mathbf{S};\mathbf{L}) = \left(\mathbf{S}\mathbf{S}^\top + \mathbf{L}\mathbf{L}^\top\right)^{-1/2}\mathbf{S}$. Since $\mathbf{L}\mathbf{L}^\top \succeq 0$, it holds that

$$\nabla_{\mathbf{S}}\widetilde{\Psi}^H(\mathbf{S};\mathbf{L})\nabla_{\mathbf{S}}\widetilde{\Psi}^H(\mathbf{S};\mathbf{L})^\top = \left(\mathbf{S}\mathbf{S}^\top + \mathbf{L}\mathbf{L}^\top\right)^{-1/2}\mathbf{S}\mathbf{S}^\top\left(\mathbf{S}\mathbf{S}^\top + \mathbf{L}\mathbf{L}^\top\right)^{-1/2} \preceq \mathbf{I}.$$

This implies $\|\nabla_{\mathbf{S}}\widetilde{\Psi}^H(\mathbf{S};\mathbf{L})\|_{\mathrm{op}} \leq 1$, verifying Property (a).

Dominance It is easy to see that $\widetilde{\Psi}^H(\mathbf{S};\mathbf{0}) = \mathrm{Tr}\big(\sqrt{\mathbf{S}\mathbf{S}^\top}\big) = \|\mathbf{S}\|_*$. Moreover, since $\mathbf{L}\mathbf{L}^\top \succeq 0$, we have $\sqrt{\mathbf{S}\mathbf{S}^\top + \mathbf{L}\mathbf{L}^\top} \succeq \sqrt{\mathbf{S}\mathbf{S}^\top}$, yielding $\widetilde{\Psi}^H(\mathbf{S};\mathbf{L}) \geq \|\mathbf{S}\|_*$.

Upper stability We rewrite the potential as $\widetilde{\Psi}^H(\mathbf{S};\mathbf{L}) = \|\begin{bmatrix}\mathbf{S} & \mathbf{L}\end{bmatrix}\|_*$. Then by triangle inequality, for any $\mathbf{L}_1 \preceq \mathbf{L}_2$ and $\mathbf{S} \in \mathbb{R}^{m\times n}$, we have

$$\widetilde{\Psi}^H(\mathbf{S};\mathbf{L}_2) - \widetilde{\Psi}^H(\mathbf{S};\mathbf{L}_1) = \big\|\begin{bmatrix}\mathbf{S} & \mathbf{L}_2\end{bmatrix}\big\|_* - \big\|\begin{bmatrix}\mathbf{S} & \mathbf{L}_1\end{bmatrix}\big\|_* \leq \big\|\begin{bmatrix}\mathbf{0} & \mathbf{L}_2 - \mathbf{L}_1\end{bmatrix}\big\|_* = \|\mathbf{L}_2 - \mathbf{L}_1\|_*.$$

Since $\mathbf{L}_2 - \mathbf{L}_1 \succeq 0$, we further have $\|\mathbf{L}_2 - \mathbf{L}_1\|_* = \mathrm{Tr}(\mathbf{L}_2) - \mathrm{Tr}(\mathbf{L}_1)$. This proves Property (c) with $\alpha = 1$.

Smoothness We rely on the characterization of the nuclear norm in Lemma 22. For any $\mathbf{L} \succ 0$ and $\mathbf{S} \in \mathbb{R}^{m\times n}$, the augmented matrix $\begin{bmatrix}\mathbf{S} & \mathbf{L}\end{bmatrix}$ is full row rank. Hence, $\widetilde{\Psi}^H(\mathbf{S};\mathbf{L}) = \|\begin{bmatrix}\mathbf{S} & \mathbf{L}\end{bmatrix}\|_*$ is twice differentiable and it follows from Lemma 22 that for all $\mathbf{D} \in \mathbb{R}^{m\times n}$,

$$\nabla_{\mathbf{S}}^2\widetilde{\Psi}^H(\mathbf{S};\mathbf{L})[\mathbf{D},\mathbf{D}] \leq \mathrm{Tr}(\mathbf{D}^\top(\mathbf{S}\mathbf{S}^\top + \mathbf{L}\mathbf{L}^\top)^{-1/2}\mathbf{D}).$$

Since $\mathbf{S}\mathbf{S}^\top \succeq 0$ and $\mathbf{L}$ is positive definite, this further implies that

$$\nabla_{\mathbf{S}}^2\widetilde{\Psi}^H(\mathbf{S};\mathbf{L})[\mathbf{D},\mathbf{D}] \leq \mathrm{Tr}(\mathbf{D}^\top(\mathbf{L}\mathbf{L}^\top)^{-1/2}\mathbf{D}) = \mathrm{Tr}(\mathbf{D}^\top\mathbf{L}^{-1}\mathbf{D}).$$

Hence, we follow the same argument as in Theorem 6 and use (28) to conclude that Property (d) is satisfied with $\beta = 1$.

Appendix F. The Cost of Solving Subproblems

F.1. Shampoo and One-Sided Shampoo

For Shampoo and one-sided Shampoo, each outer iteration requires solving the quadratic projection subproblem in (1) over the operator-norm ball. As discussed in Section 2, this typically necessitates an iterative inner solver, since the operator-norm geometry precludes a simple closed-form update.

For a clean comparison with FTPL and FAML, consider one-sided Shampoo and suppose SVD is the common primitive for matrix square roots, polar factors, and projections. Then solving the subproblem using accelerated gradient descent requires, beyond forming $\mathbf{L}_t$, one full SVD per inner iteration. Moreover, the subproblem condition number depends on $\mathbf{L}_t$ and in the worst case can scale as $\sqrt{T}$. As a result, reaching a desired accuracy may take $\tilde{O}(T^{1/4})$ iterations, leading to a total of $\tilde{O}(T^{5/4})$ SVD calls.

In contrast, as we shall establish in the next sections, our proposed methods, FTPL and FAML, avoid iterative projections and require fewer SVD-based primitives. FTPL requires only $k+1$ such calls per round: one matrix square root and k polar factors, the latter of which can be computed in parallel. FAML is even cheaper, requiring only two such calls per round: one matrix square root and one polar factor.

Practically, the gap can be even larger. FTPL and FAML rely mainly on polar computations, which can be implemented efficiently via Newton–Schulz using only matrix multiplications. By contrast, AGD requires projections that generally involve full SVDs, which are less GPU-friendly and often require higher numerical precision. This difference can lead to substantial empirical speedups for FTPL and FAML.

F.2. FTPL

As discussed in Section 4.1, each iteration of the update in (8) computes $\mathbf{G}_t\mathbf{G}_t^\top$, followed by a Cholesky factorization and k parallel computations of matrix polar factors. Below we describe a concrete implementation of the polar factor via the Newton–Schulz iteration. Given an input matrix $\mathbf{S} \in \mathbb{R}^{m\times n}$, initialize and iterate

$$\mathbf{X}^{(0)} = \frac{\mathbf{S}}{\|\mathbf{S}\|_F}, \qquad \mathbf{X}^{(i+1)} = \tfrac{1}{2}\big(3\mathbf{I} - \mathbf{X}^{(i)}(\mathbf{X}^{(i)})^\top\big)\mathbf{X}^{(i)}. \tag{39}$$

It is known that the iterates converge (locally) quadratically to $\mathrm{polar}(\mathbf{S})$, and in particular the iteration converges provided the singular values of $\mathbf{X}^{(0)}$ lie in $(0, \sqrt{3})$ (Higham, 2008). Each Newton–Schulz step is dominated by two matrix–matrix multiplications, for a leading cost $2 \times (2m^2 n) = 4m^2 n$ floating-point operations per step.

Let $s_{\text{par}} \geq 1$ denote the effective parallel speedup for the k polar-factor computations. Putting these components together, the leading-order *wall-clock* cost per iteration is

$$\underbrace{2m^2 n}_{\mathbf{G}_t\mathbf{G}_t^\top} + \underbrace{\tfrac{1}{3}m^3}_{\text{Cholesky}} + \underbrace{\frac{4kK\,m^2 n}{s_{\text{par}}}}_{\substack{K \text{ Newton–Schulz steps} \\ \text{for each of } k \text{ polar factors}}},$$

where K denotes the number of Newton–Schulz iterations used to compute each polar factor.

F.3. FAML

The first implementation based on (11) requires computing a matrix inverse square root. We describe a concrete realization using the coupled Newton–Schulz iteration. Given an input matrix $\mathbf{A} \in \mathbb{S}_+^m$, initialize

$$\mathbf{Y}^{(0)} = \frac{\mathbf{A}}{\sqrt{\|\mathbf{A}\|_F}}, \qquad \mathbf{Z}^{(0)} = \frac{\mathbf{I}}{\sqrt{\|\mathbf{A}\|_F}},$$

and iterate

$$\mathbf{T}^{(i)} = \mathbf{Z}^{(i)}\mathbf{Y}^{(i)}, \qquad \mathbf{Y}^{(i+1)} = \tfrac{1}{2}\mathbf{Y}^{(i)}(3\mathbf{I} - \mathbf{T}^{(i)}), \qquad \mathbf{Z}^{(i+1)} = \tfrac{1}{2}(3\mathbf{I} - \mathbf{T}^{(i)})\mathbf{Z}^{(i)}.$$

The coupled iteration converges provided $\|\mathbf{I} - \mathbf{A}/\|\mathbf{A}\|_F\|_{\mathrm{op}} \leq 1$. Each Newton–Schulz step is dominated by three $m \times m$ matrix–matrix multiplications, resulting in $6m^3$ floating-point operations per iteration.

In each outer iteration of the algorithm, the following operations are performed:

- Compute $\mathbf{G}_t\mathbf{G}_t^\top$: $2m^2n$ flops;
- Compute $\mathbf{S}_t\mathbf{S}_t^\top$: $2m^2n$ flops;
- Coupled Newton–Schulz iteration: $6m^3$ flops per step;
- Final preconditioning: $2m^2n$ flops.

Consequently, the leading-order per-iteration cost is

$$6m^2n + 6Km^3,$$

where K denotes the number of coupled Newton–Schulz iterations.

Alternatively, the update in (10) can be implemented by computing the polar factor of the augmented matrix $\widehat{\mathbf{S}}_t = \begin{bmatrix}\mathbf{S}_t & \frac{1}{\eta}\mathbf{L}_t\end{bmatrix}$. We describe a Newton–Schulz-based implementation that exploits the block structure of $\widehat{\mathbf{S}}_t$. Since the final update only uses the leading block of the polar factor, the procedure can be specialized to avoid explicitly forming $\mathbf{L}_t$. In particular, for the choice of $\mathbf{L}_t$ in (3), this avoids computing a matrix square root such as $(G^2\mathbf{I} + \mathbf{M}_t)^{1/2}$.

Let $\widehat{\mathbf{S}} = \begin{bmatrix}\mathbf{S} & \mathbf{L}\end{bmatrix}$ denote the input, and maintain iterates $\widehat{\mathbf{X}}^{(i)} = \begin{bmatrix}\mathbf{X}^{(i)} & \mathbf{Y}^{(i)}\end{bmatrix}$ at step i of the Newton–Schulz iteration. Applying the standard update (39) gives

$$\widehat{\mathbf{X}}^{(i+1)} = \tfrac{1}{2}\big(3\mathbf{I} - \widehat{\mathbf{X}}^{(i)}(\widehat{\mathbf{X}}^{(i)})^\top\big)\widehat{\mathbf{X}}^{(i)}.$$

Writing this update in block form yields

$$\mathbf{T}^{(i)} = \tfrac{1}{2}\Big(3\mathbf{I} - \mathbf{X}^{(i)}(\mathbf{X}^{(i)})^\top - \mathbf{Y}^{(i)}(\mathbf{Y}^{(i)})^\top\Big), \qquad \mathbf{X}^{(i+1)} = \mathbf{T}^{(i)}\mathbf{X}^{(i)}, \qquad \mathbf{Y}^{(i+1)} = \mathbf{T}^{(i)}\mathbf{Y}^{(i)}.$$

Crucially, the update of $\mathbf{X}^{(i)}$ depends on $\mathbf{Y}^{(i)}$ only through the Gram matrix $\mathbf{X}^{(i)}(\mathbf{X}^{(i)})^\top + \mathbf{Y}^{(i)}(\mathbf{Y}^{(i)})^\top$. We therefore eliminate $\mathbf{Y}^{(i)}$ by defining

$$\mathbf{B}^{(i)} := \mathbf{X}^{(i)}(\mathbf{X}^{(i)})^\top + \mathbf{Y}^{(i)}(\mathbf{Y}^{(i)})^\top.$$

Since

$$\mathbf{B}^{(i+1)} = \mathbf{X}^{(i+1)}(\mathbf{X}^{(i+1)})^\top + \mathbf{Y}^{(i+1)}(\mathbf{Y}^{(i+1)})^\top = \mathbf{T}^{(i)}\mathbf{B}^{(i)}\mathbf{T}^{(i)},$$

we obtain the equivalent recursion

$$\mathbf{T}^{(i)} = \tfrac{1}{2}\big(3\mathbf{I} - \mathbf{B}^{(i)}\big), \qquad \mathbf{X}^{(i+1)} = \mathbf{T}^{(i)}\mathbf{X}^{(i)}, \qquad \mathbf{B}^{(i+1)} = \mathbf{T}^{(i)}\mathbf{B}^{(i)}\mathbf{T}^{(i)}. \tag{40}$$

For initialization, note that

$$\|\widehat{\mathbf{S}}\|_F^2 \;=\; \|\mathbf{S}\|_F^2 + \|\mathbf{L}\|_F^2 \;=\; \|\mathbf{S}\|_F^2 + \mathrm{Tr}(\mathbf{L}\mathbf{L}^\top),$$

and set

$$\mathbf{X}^{(0)} = \frac{\mathbf{S}}{\|\widehat{\mathbf{S}}\|_F}, \qquad \mathbf{B}^{(0)} = \frac{\widehat{\mathbf{S}}\widehat{\mathbf{S}}^\top}{\|\widehat{\mathbf{S}}\|_F^2} = \frac{\mathbf{S}\mathbf{S}^\top + \mathbf{L}\mathbf{L}^\top}{\|\widehat{\mathbf{S}}\|_F^2}.$$

Moreover, with the choice of $\mathbf{L}_t$ in (3), we have $\mathbf{L}_t\mathbf{L}_t^\top = G^2\mathbf{I} + \mathbf{M}_t$, so $\mathbf{L}_t\mathbf{L}_t^\top$ can be formed directly from $\mathbf{M}_t \coloneqq \sum_{s=1}^t \mathbf{G}_s\mathbf{G}_s^\top$ without explicitly computing $(G^2\mathbf{I} + \mathbf{M}_t)^{1/2}$.

Each Newton–Schulz step in (40) requires one $m \times m$ by $m \times n$ multiplication (to update $\mathbf{X}^{(i+1)}$) and two $m \times m$ by $m \times m$ multiplications (to update $\mathbf{B}^{(i+1)}$), for a total of $2m^2n + 4m^3$ floating-point operations. Including the initialization cost of forming $\mathbf{S}\mathbf{S}^\top$ and $\mathbf{G}_t\mathbf{G}_t^\top$ (each $2m^2n$), the resulting leading-order per-iteration cost is

$$4m^2n \;+\; K\,(2m^2n + 4m^3),$$

where K denotes the number of Newton–Schulz iterations.

Appendix G. Proofs for Online-to-Nonconvex Conversion

We first describe the O2NC reduction protocol in full. At each iteration t, the environment draws a sample $\zeta_t \sim \mathcal{D}$ and reveals the corresponding stochastic gradient $\mathbf{G}_t$ at a random midpoint $\widetilde{\mathbf{W}}_t$. This gradient induces a (discounted) linear loss, which we feed to an online learner $\mathcal{A}$ to obtain the next update direction. Finally, we update the parameters using the direction selected by the online learner:

$$\widetilde{\mathbf{W}}_t = \mathbf{W}_{t-1} + s_t\mathbf{X}_t, \qquad s_t \sim \mathrm{Uniform}(0,1), \tag{41}$$

$$\mathbf{G}_t = \nabla_{\mathbf{W}}\ell(\widetilde{\mathbf{W}}_t;\zeta_t), \qquad \zeta_t \sim \mathcal{D}, \tag{42}$$

$$\ell_t^{[\beta]}(\mathbf{X}) \coloneqq \beta^{-t}\,\langle \mathbf{G}_t, \mathbf{X}\rangle, \qquad \text{(loss revealed)} \tag{43}$$

$$\mathbf{X}_{t+1} = \mathcal{A}\Big(\ell_t^{[\beta]}\Big) \in \mathcal{X}, \qquad \text{(learner selection)} \tag{44}$$

$$\mathbf{W}_{t+1} = \mathbf{W}_t + \mathbf{X}_{t+1}. \tag{45}$$

The choice of evaluating the stochastic gradient at the random midpoint is motivated by the following fact: by using the fundamental theorem of calculus, we have

$$L(\mathbf{W}_t) - L(\mathbf{W}_{t-1}) = \langle \boldsymbol{\nabla}_t, \mathbf{W}_t - \mathbf{W}_{t-1}\rangle = \langle \boldsymbol{\nabla}_t, \mathbf{X}_t\rangle,$$

where $\boldsymbol{\nabla}_t = \int_0^1 \nabla L(\mathbf{W}_{t-1} + s(\mathbf{W}_t - \mathbf{W}_{t-1}))\,ds$ is the average gradient along the line segment between $\mathbf{W}_{t-1}$ and $\mathbf{W}_t$. Moreover, $\mathbf{G}_t$ is an unbiased estimate of $\boldsymbol{\nabla}_t$, i.e., $\mathbb{E}_{s_t,\zeta_t}[\mathbf{G}_t] = \boldsymbol{\nabla}_t$. Hence, we have

$$\mathop{\mathbb{E}}_{s_t}\big[L(\mathbf{W}_t) - L(\mathbf{W}_{t-1})\big] = \mathop{\mathbb{E}}_{s_t,\zeta_t}\big[\langle \mathbf{G}_t, \mathbf{X}_t\rangle\big]. \tag{46}$$

Algorithm 1 Muon

Initialize: $\bar{\mathbf{G}}_0 = \mathbf{0}$
for iteration $t = 1, \dots, T$ **do**
  $\mathbf{G}_t \leftarrow \nabla_{\mathbf{W}} \ell(\mathbf{W}_t; \zeta_t) \in \mathbb{R}^{m \times n}$
  $\bar{\mathbf{G}}_t \leftarrow \beta \bar{\mathbf{G}}_{t-1} + \mathbf{G}_t$
  $\mathbf{P}_t \leftarrow \operatorname{polar}(\bar{\mathbf{G}}_t)$
  $\mathbf{W}_{t+1} \leftarrow \mathbf{W}_t - \alpha_t \mathbf{P}_t$
end for

Algorithm 2 Pion

Initialize: $\bar{\mathbf{G}}_0 = \mathbf{0}, \bar{\mathbf{M}}_0 = \mathbf{0}$
for iteration $t = 1, \dots, T$ **do**
  $\mathbf{G}_t \leftarrow \nabla_{\mathbf{W}} \ell(\mathbf{W}_t; \zeta_t) \in \mathbb{R}^{m \times n}$
  $\bar{\mathbf{G}}_t \leftarrow \beta_1 \bar{\mathbf{G}}_{t-1} + \mathbf{G}_t$
  $\bar{\mathbf{M}}_t \leftarrow \beta_2 \bar{\mathbf{M}}_{t-1} + \mathbf{G}_t \mathbf{G}_t^\top$
  Cholesky factorization $\bar{\mathbf{M}}_t = \widetilde{\mathbf{L}}_t \widetilde{\mathbf{L}}_t^\top$
  $\mathbf{P}_t \leftarrow \frac{1}{k} \sum_{i=1}^k \operatorname{polar}(\bar{\mathbf{G}}_t + \frac{1}{\eta} \widetilde{\mathbf{L}}_t \mathbf{Z}_t^{(i)})$, where $\mathbf{Z}_t^{(i)} \sim \mathcal{MN}(0, \mathbf{I}_m, \mathbf{I}_n)$
  $\mathbf{W}_{t+1} \leftarrow \mathbf{W}_t - \alpha_t \mathbf{P}_t$
end for

Algorithm 3 Leon

Initialize: $\bar{\mathbf{G}}_0 = \mathbf{0}, \bar{\mathbf{M}}_0 = \mathbf{0}$
for iteration $t = 1, \dots, T$ **do**
  $\mathbf{G}_t \leftarrow \nabla_{\mathbf{W}} \ell(\mathbf{W}_t; \zeta_t) \in \mathbb{R}^{m \times n}$
  $\bar{\mathbf{G}}_t \leftarrow \beta_1 \bar{\mathbf{G}}_{t-1} + \mathbf{G}_t$
  $\bar{\mathbf{M}}_t \leftarrow \beta_2 \bar{\mathbf{M}}_{t-1} + \mathbf{G}_t \mathbf{G}_t^\top$
  $\mathbf{P}_t \leftarrow \left(\bar{\mathbf{G}}_t \bar{\mathbf{G}}_t^\top + \frac{1}{\eta^2} \bar{\mathbf{M}}_t \right)^{-1/2} \bar{\mathbf{G}}_t$
  $\mathbf{W}_{t+1} \leftarrow \mathbf{W}_t - \alpha_t \mathbf{P}_t$
end for

For clarity and ease of exposition, we also provide the practical implementation version of Muon, Pion, and Leon in Algorithms 1, 2, and 3, respectively. For Pion and Leon, the theoretical setting corresponds to $\beta_1 = \beta$ and $\beta_2 = \beta^2$. Moreover, the displayed pseudocode omits the additive regularization term $\frac{G^2}{\beta_2} \mathbf{I}$ in the preconditioner; including this term recovers the theoretical updates in (16) and (18).

G.1. Proof of Proposition 10

Proposition 25 (Formal O2NC bound) *Define the exponentially weighted average (EWA) of the random midpoints*

$$\bar{\mathbf{W}}_t := \frac{1-\beta}{1-\beta^t} \sum_{s=1}^t \beta^{t-s} \widetilde{\mathbf{W}}_s, \qquad t = 1, \dots, T,$$

and the random time index $\tau \in \{1, \dots, T\}$ *with*

$$\Pr(\tau = t) = \begin{cases} \frac{1-\beta^t}{T}, & t = 1, \dots, T-1, \\ \frac{1-\beta^T}{(1-\beta)T}, & t = T. \end{cases}$$

Let $\mathbf{E}_t := \mathbf{G}_t - \nabla L(\widetilde{\mathbf{W}}_t)$ *denote the stochastic noise, and let*

$$\mathrm{Reg}_t^{[\beta]}(D) := \max_{\|\mathbf{X}\| \le D} \sum_{s=1}^{t} \beta^{t-s} \langle \mathbf{G}_s, \mathbf{X}_s - \mathbf{X} \rangle$$

be the discounted regret (with radius D*). If* $D = \frac{1-\beta}{4\beta}\rho$*, then the expected* ρ*-stationarity gap at* $\bar{\mathbf{W}}_\tau$ *satisfies*

$$\mathop{\mathbb{E}}_{\tau}\big[\|\nabla L(\bar{\mathbf{W}}_\tau)\|_\dagger^{[\rho]}\big] \le \frac{4(L(\mathbf{W}_0) - L(\mathbf{W}^*))}{(1-\beta)\rho T} + \frac{1}{T}\,\mathbb{E}\Big[\mathrm{Reg}_T^{[\beta]}(1) + (1-\beta)\sum_{t=1}^{T-1} \mathrm{Reg}_t^{[\beta]}(1)\Big] + \frac{1}{T}\,\mathbb{E}\Big[\Big\|\sum_{t=1}^{T} \beta^{T-t}\mathbf{E}_t\Big\|_\dagger\Big] + \frac{1-\beta}{T}\sum_{t=1}^{T-1}\mathbb{E}\Big[\Big\|\sum_{s=1}^{t}\beta^{t-s}\mathbf{E}_s\Big\|_\dagger\Big]. \tag{47}$$

Our proof for Proposition 25 is built on the following decomposition.

Lemma 26 (Ahn et al., 2025, Appendix A.1) *For any* $\beta \in (0,1)$*,*

$$L(\mathbf{W}_T) - L(\mathbf{W}_0) = \sum_{t=1}^{T} \beta^{T-t}\big(L(\mathbf{W}_t) - L(\mathbf{W}_{t-1})\big) + (1-\beta)\sum_{t=1}^{T-1}\sum_{s=1}^{t}\beta^{t-s}\big(L(\mathbf{W}_s) - L(\mathbf{W}_{s-1})\big).$$

Proof of Proposition 25 By combining (46) with Lemma 26 and taking total expectation, we obtain

$$\mathbb{E}[L(\mathbf{W}_T) - L(\mathbf{W}_0)] = \sum_{t=1}^{T}\beta^{T-t}\,\mathbb{E}[\langle \mathbf{G}_t, \mathbf{X}_t\rangle] + (1-\beta)\sum_{t=1}^{T-1}\sum_{s=1}^{t}\beta^{t-s}\,\mathbb{E}[\langle \mathbf{G}_s, \mathbf{X}_s\rangle].$$

We now relate the inner sums to regret. Specifically, for any comparator $\mathbf{X}$ with $\|\mathbf{X}\| \le D$, we have

$$\sum_{s=1}^{t}\beta^{t-s}\,\mathbb{E}[\langle \mathbf{G}_s, \mathbf{X}_s\rangle] = \mathbb{E}\Big[\sum_{s=1}^{t}\beta^{t-s}\langle \mathbf{G}_s, \mathbf{X}_s - \mathbf{X}\rangle\Big] + \mathbb{E}\Big[\Big\langle \sum_{s=1}^{t}\beta^{t-s}\mathbf{G}_s, \mathbf{X}\Big\rangle\Big].$$

Taking $\mathbf{X} = \arg\min_{\|\mathbf{X}\| \le D}\langle \sum_{s=1}^{t}\beta^{t-s}\mathbf{G}_s, \mathbf{X}\rangle$ and recalling the definition of the discounted regret in (13), we obtain

$$\sum_{s=1}^{t}\beta^{t-s}\,\mathbb{E}[\langle \mathbf{G}_s, \mathbf{X}_s\rangle] \le \mathbb{E}\Big[\mathrm{Reg}_t^{[\beta]}(D)\Big] - D\,\mathbb{E}\Big[\Big\|\sum_{s=1}^{t}\beta^{t-s}\mathbf{G}_s\Big\|_\dagger\Big].$$

Moreover, let $\mathbf{E}_t = \mathbf{G}_t - \nabla L(\widetilde{\mathbf{W}}_t)$ denote the stochastic noise. Then by the triangle inequality, we further have $-\big\|\sum_{s=1}^{t}\beta^{t-s}\mathbf{G}_s\big\|_\dagger \le -\big\|\sum_{s=1}^{t}\beta^{t-s}\nabla L(\widetilde{\mathbf{W}}_s)\big\|_\dagger + \big\|\sum_{s=1}^{t}\beta^{t-s}\mathbf{E}_s\big\|_\dagger$, yielding

$$\sum_{s=1}^{t}\beta^{t-s}\,\mathbb{E}[\langle \mathbf{G}_s, \mathbf{X}_s\rangle] \le \mathbb{E}\Big[\mathrm{Reg}_t^{[\beta]}(D)\Big] - D\,\mathbb{E}\Big[\Big\|\sum_{s=1}^{t}\beta^{t-s}\nabla L(\widetilde{\mathbf{W}}_s)\Big\|_\dagger\Big] + D\,\mathbb{E}\Big\|\sum_{s=1}^{t}\beta^{t-s}\mathbf{E}_s\Big\|_\dagger. \tag{48}$$

Apply (48) with $t = T$ and also with each $t = 1, \dots, T-1$, then combine them using Lemma 26. Rearranging yields

$$\begin{aligned} & D\ \mathbb{E}\Big\|\sum_{t=1}^{T} \beta^{T-t} \nabla L(\widetilde{\mathbf{W}}_t)\Big\|_\dagger + (1-\beta) D \sum_{t=1}^{T-1} \mathbb{E}\Big\|\sum_{s=1}^{t} \beta^{t-s} \nabla L(\widetilde{\mathbf{W}}_s)\Big\|_\dagger \\ & \le \mathbb{E}\big[L(\mathbf{W}_0) - L(\mathbf{W}_T)\big] + \mathbb{E}\big[\mathrm{Reg}_T^{[\beta]}(D)\big] + (1-\beta) \sum_{t=1}^{T-1} \mathbb{E}\big[\mathrm{Reg}_t^{[\beta]}(D)\big] \\ & \quad + D\ \mathbb{E}\Big\|\sum_{t=1}^{T} \beta^{T-t} \mathbf{E}_t\Big\|_\dagger + (1-\beta) D \sum_{t=1}^{T-1} \mathbb{E}\Big\|\sum_{s=1}^{t} \beta^{t-s} \mathbf{E}_s\Big\|_\dagger. \end{aligned} \tag{49}$$

Next we convert the discounted-gradient terms into a stationarity guarantee for an EWA iterate. For each $t = 1, \dots, T$, define a random iterate $\mathbf{Y}_t$ supported on $\{\widetilde{\mathbf{W}}_1, \dots, \widetilde{\mathbf{W}}_t\}$ by

$$\Pr(\mathbf{Y}_t = \widetilde{\mathbf{W}}_s) = \frac{1-\beta}{1-\beta^t} \beta^{t-s}, \qquad s = 1, \dots, t,$$

so that $\mathbb{E}[\mathbf{Y}_t] = \bar{\mathbf{W}}_t$. We will use the following concentration-of-the-mean bound.

Lemma 27 *For every $t = 1, \dots, T$, we have $\mathbb{E}\,\|\mathbf{Y}_t - \bar{\mathbf{W}}_t\| \le \frac{4\beta}{1-\beta} D$.*

Proof Fix t and define $q_{t,s} := \frac{1-\beta}{1-\beta^t} \beta^{t-s}$. Let $\widehat{\mathbf{Y}}_t$ be an independent copy of $\mathbf{Y}_t$. By Jensen's inequality,

$$\mathbb{E}\,\|\mathbf{Y}_t - \bar{\mathbf{W}}_t\| = \mathbb{E}\big\|\mathbf{Y}_t - \mathbb{E}[\widehat{\mathbf{Y}}_t]\big\| \le \mathbb{E}\,\|\mathbf{Y}_t - \widehat{\mathbf{Y}}_t\| = 2 \sum_{i=1}^{t} \sum_{j=1}^{i-1} q_{t,i} q_{t,j} \|\widetilde{\mathbf{W}}_i - \widetilde{\mathbf{W}}_j\|.$$

By the triangle inequality, $\|\widetilde{\mathbf{W}}_i - \widetilde{\mathbf{W}}_j\| \le \sum_{s=j+1}^{i} \|\widetilde{\mathbf{W}}_s - \widetilde{\mathbf{W}}_{s-1}\|$. Thus

$$\mathbb{E}\,\|\mathbf{Y}_t - \bar{\mathbf{W}}_t\| \le 2 \sum_{s=2}^{t} \Big(\sum_{i=s}^{t} \sum_{j=1}^{s-1} q_{t,i} q_{t,j} \Big) \|\widetilde{\mathbf{W}}_s - \widetilde{\mathbf{W}}_{s-1}\|.$$

Moreover,

$$\sum_{i=s}^{t} \sum_{j=1}^{s-1} q_{t,i} q_{t,j} = \Big(\frac{1-\beta}{1-\beta^t}\Big)^2 \Big(\sum_{i=s}^{t} \beta^{t-i}\Big)\Big(\sum_{j=1}^{s-1} \beta^{t-j}\Big) \le \frac{\beta^{t-s+1}}{1-\beta^t}.$$

Using $\|\widetilde{\mathbf{W}}_s - \widetilde{\mathbf{W}}_{s-1}\| \le \|\mathbf{X}_s\| + \|\mathbf{X}_{s-1}\| \le 2D$, we get

$$\mathbb{E}\,\|\mathbf{Y}_t - \bar{\mathbf{W}}_t\| \le 4 \sum_{s=2}^{t} \frac{\beta^{t-s+1}}{1-\beta^t} D \le \frac{4\beta}{1-\beta} D.$$

This completes the proof. ∎

Now take $D = \frac{1-\beta}{4\beta} \rho$ so Lemma 27 gives $\mathbb{E}\,\|\mathbf{Y}_t - \bar{\mathbf{W}}_t\| \le \rho$. By the definition of the ρ-stationarity gap, this implies

$$\|\nabla L(\bar{\mathbf{W}}_t)\|_\dagger^{[\rho]} \le \big\|\mathbb{E}[\nabla L(\mathbf{Y}_t)]\big\|_\dagger = \frac{1-\beta}{1-\beta^t} \Big\|\sum_{s=1}^{t} \beta^{t-s} \nabla L(\widetilde{\mathbf{W}}_s)\Big\|_\dagger.$$

Plugging this into (49) yields

$$\frac{1-\beta^T}{1-\beta}\,\mathbb{E}\big[\|\nabla L(\bar{\mathbf{W}}_T)\|_\dagger^{[\rho]}\big] + \sum_{t=1}^{T-1}(1-\beta^t)\,\mathbb{E}\big[\|\nabla L(\bar{\mathbf{W}}_t)\|_\dagger^{[\rho]}\big]$$
$$\le \frac{1}{D}\,\mathbb{E}\big[L(\mathbf{W}_0) - L(\mathbf{W}_T)\big] + \frac{1}{D}\,\mathbb{E}\big[\mathrm{Reg}_T^{[\beta]}(D)\big] + \frac{1-\beta}{D}\sum_{t=1}^{T-1}\mathbb{E}\big[\mathrm{Reg}_t^{[\beta]}(D)\big]$$
$$+ \mathbb{E}\Big\|\sum_{t=1}^{T}\beta^{T-t}\mathbf{E}_t\Big\|_\dagger + (1-\beta)\sum_{t=1}^{T-1}\mathbb{E}\Big\|\sum_{s=1}^{t}\beta^{t-s}\mathbf{E}_s\Big\|_\dagger.$$

Note that by a scaling argument, we have $\frac{1}{D}\mathrm{Reg}_t^{[\beta]}(D) = \mathrm{Reg}_t^{[\beta]}(1)$. Finally, by the definition of τ,

$$\mathop{\mathbb{E}}_{\tau}\big[\|\nabla L(\bar{\mathbf{W}}_\tau)\|_\dagger^{[\rho]}\big] = \frac{1}{T}\left(\sum_{t=1}^{T-1}(1-\beta^t)\,\mathbb{E}\big[\|\nabla L(\bar{\mathbf{W}}_t)\|_\dagger^{[\rho]}\big] + \frac{1-\beta^T}{1-\beta}\,\mathbb{E}\big[\|\nabla L(\bar{\mathbf{W}}_T)\|_\dagger^{[\rho]}\big]\right).$$

Divide the previous inequality by T, substitute $D = \frac{1-\beta}{4\beta}\rho$, use $\beta \le 1$, and use $\mathbb{E}[L(\mathbf{W}_0) - L(\mathbf{W}_T)] \le L(\mathbf{W}_0) - L(\mathbf{W}^*)$ to obtain (47). This completes the proof. ■

G.2. Proof of Theorem 12

For clarity, we assume the FTPL expectation is computed exactly; the finite-sample implementation follows similarly by a concentration argument as discussed in Remark 7.

Step 1: Bounding the discounted regret. By Theorem 6, for any $t = 1, \ldots, T$,

$$\mathrm{Reg}_t^{[\beta]}(1) \le 2\sqrt{2}\,C\left(\mathrm{Tr}\left[\sqrt{\sum_{s=1}^{t}\beta^{2(t-s)}\mathbf{G}_s\mathbf{G}_s^\top}\right] + \frac{mG}{\beta}\right), \qquad C := \left(\frac{n}{n-m-1}\right)^{1/4}, \tag{50}$$

where $G := \max_{s=1,\ldots,T}\|\mathbf{G}_s\|_{\mathrm{op}}$. Under Assumption 1, we have $\mathbb{E}[\mathbf{G}_s\mathbf{G}_s^\top] \preceq \mathbf{Q}^2$. Using Jensen's inequality and $\mathrm{Tr}(\sqrt{\mathbf{A}}) = \|\mathbf{A}^{1/2}\|_*$ for $\mathbf{A} \succeq 0$,

$$\mathbb{E}\,\mathrm{Tr}\left[\sqrt{\sum_{s=1}^{t}\beta^{2(t-s)}\mathbf{G}_s\mathbf{G}_s^\top}\right] \le \mathrm{Tr}\left[\sqrt{\sum_{s=1}^{t}\beta^{2(t-s)}\,\mathbb{E}[\mathbf{G}_s\mathbf{G}_s^\top]}\right]$$
$$\le \mathrm{Tr}\left[\sqrt{\sum_{s=1}^{t}\beta^{2(t-s)}\mathbf{Q}^2}\right] \le \frac{\|\mathbf{Q}\|_*}{\sqrt{1-\beta^2}}. \tag{51}$$

Combining (50)–(51) yields

$$\mathbb{E}\big[\mathrm{Reg}_t^{[\beta]}(1)\big] \le 2\sqrt{2}\,C\left(\frac{\|\mathbf{Q}\|_*}{\sqrt{1-\beta^2}} + \frac{mG}{\beta}\right). \tag{52}$$

Step 2: Bounding the discounted noise. Under Assumption 1, $\mathbb{E}[\mathbf{E}_s\mathbf{E}_s^\top] \preceq \mathbb{E}[\mathbf{G}_s\mathbf{G}_s^\top] \preceq \mathbf{Q}^2$. By Jensen's inequality and the same trace-sqrt manipulation,

$$\begin{aligned}\mathbb{E}\Big\|\sum_{s=1}^{t}\beta^{t-s}\mathbf{E}_s\Big\|_* &= \mathbb{E}\,\mathrm{Tr}\left(\sqrt{\Big(\sum_{s=1}^{t}\beta^{t-s}\mathbf{E}_s\Big)\Big(\sum_{s=1}^{t}\beta^{t-s}\mathbf{E}_s\Big)^\top}\right)\\ &\le \mathrm{Tr}\left(\sqrt{\mathbb{E}\Big[\sum_{s=1}^{t}\sum_{r=1}^{t}\beta^{t-s}\beta^{t-r}\mathbf{E}_s\mathbf{E}_r^\top\Big]}\right) \le \mathrm{Tr}\left(\sqrt{\sum_{s=1}^{t}\beta^{2(t-s)}\,\mathbb{E}[\mathbf{E}_s\mathbf{E}_s^\top]}\right)\\ &\le \mathrm{Tr}\left(\sqrt{\sum_{s=1}^{t}\beta^{2(t-s)}\mathbf{Q}^2}\right) \le \frac{\|\mathbf{Q}\|_*}{\sqrt{1-\beta^2}},\end{aligned} \tag{53}$$

where we used the conditional unbiasedness of the stochastic oracle, so cross terms vanish.

Step 3: Plug into the O2NC bound and choose parameters. Proposition 25 (with $D = \frac{1-\beta}{4\beta}\rho$) together with (52)–(53) implies

$$\begin{aligned}\mathop{\mathbb{E}}_{\tau}\Big[\,\|\nabla L(\bar{\mathbf{W}}_\tau)\|_*^{[\rho]}\Big] \le \frac{4\Delta L}{(1-\beta)\rho T} + \frac{1+(1-\beta)(T-1)}{T}\cdot 2\sqrt{2}\,C\left(\frac{\|\mathbf{Q}\|_*}{\sqrt{1-\beta^2}} + \frac{mG}{\beta}\right)\\ + \frac{1+(1-\beta)(T-1)}{T}\cdot\frac{\|\mathbf{Q}\|_*}{\sqrt{1-\beta^2}}.\end{aligned} \tag{54}$$

Using $1+(1-\beta)(T-1) \le \beta + (1-\beta)T$ and $\sqrt{1-\beta^2} \ge \sqrt{1-\beta}$, we obtain

$$\begin{aligned}\mathop{\mathbb{E}}_{\tau}\Big[\,\|\nabla L(\bar{\mathbf{W}}_\tau)\|_*^{[\rho]}\Big] \le \frac{4\Delta L}{(1-\beta)\rho T} + (2\sqrt{2}C+1)\Big(\frac{\beta}{T}+1-\beta\Big)\frac{\|\mathbf{Q}\|_*}{\sqrt{1-\beta}}\\ + 2\sqrt{2}C\Big(\frac{\beta}{T}+1-\beta\Big)\frac{mG}{\beta}.\end{aligned} \tag{55}$$

Since $2\sqrt{2}+1 \le 4$, we further simplify to

$$\mathop{\mathbb{E}}_{\tau}\Big[\,\|\nabla L(\bar{\mathbf{W}}_\tau)\|_*^{[\rho]}\Big] \le \frac{4\Delta L}{(1-\beta)\rho T} + 4C\|\mathbf{Q}\|_*\Big(\frac{1}{T\sqrt{1-\beta}} + \sqrt{1-\beta}\Big) + \frac{2\sqrt{2}CmG}{T} + 2\sqrt{2}C\frac{(1-\beta)mG}{\beta}. \tag{56}$$

We now choose β as a function of ε. Let

$$\sqrt{1-\beta} := \frac{\varepsilon}{20C\|\mathbf{Q}\|_*} \qquad\Longleftrightarrow\qquad \beta = 1 - \Big(\frac{\varepsilon}{20C\|\mathbf{Q}\|_*}\Big)^2,$$

and assume $\varepsilon \le 10\sqrt{2}\,C\|\mathbf{Q}\|_*$, so that $\beta \ge \frac{1}{2}$. Plugging these choices into (56) gives

$$\begin{aligned}\mathop{\mathbb{E}}_{\tau}\Big[\,\|\nabla L(\bar{\mathbf{W}}_\tau)\|_*^{[\rho]}\Big] \le \underbrace{\frac{4}{(1-\beta)\rho}\cdot\frac{\Delta L}{T}}_{\le \frac{1600C^2(\|\mathbf{Q}\|_*)^2\Delta L}{\rho\varepsilon^2 T}} + \underbrace{\frac{4C\|\mathbf{Q}\|_*}{T\sqrt{1-\beta}}}_{=\frac{80C^2(\|\mathbf{Q}\|_*)^2}{\varepsilon T}} + \underbrace{4C\|\mathbf{Q}\|_*\sqrt{1-\beta}}_{=\varepsilon/5}\\ + \frac{2\sqrt{2}CmG}{T} + 2\sqrt{2}C\frac{(1-\beta)mG}{\beta}.\end{aligned} \tag{57}$$

The last term satisfies

$$2\sqrt{2}C\frac{(1-\beta)mG}{\beta} \le 4\sqrt{2}C(1-\beta)mG = 4\sqrt{2}CmG\cdot\frac{\varepsilon^2}{400C^2(\|\mathbf{Q}\|_*)^2} = \frac{\sqrt{2}\,mG\,\varepsilon^2}{100\,C(\|\mathbf{Q}\|_*)^2},$$

which is $\le \varepsilon/5$ provided $\varepsilon \le \frac{10\sqrt{2}\,C(\|\mathbf{Q}\|_*)^2}{mG}$. Under this additional small-accuracy condition, (57) reduces to

$$\mathop{\mathbb{E}}_{\tau}\Big[\,\|\nabla L(\bar{\mathbf{W}}_\tau)\|_*^{[\rho]}\Big] \le \frac{1600C^2(\|\mathbf{Q}\|_*)^2\Delta L}{\rho\,\varepsilon^2 T} + \frac{80C^2(\|\mathbf{Q}\|_*)^2}{\varepsilon T} + \frac{2\sqrt{2}CmG}{T} + \frac{2\varepsilon}{5}.$$

Hence, if

$$T \;\ge\; \max\left\{\frac{8000\,C^2(\|\mathbf{Q}\|_*)^2\Delta L}{\rho\,\varepsilon^3},\; \frac{400\,C^2(\|\mathbf{Q}\|_*)^2}{\varepsilon^2},\; \frac{10\sqrt{2}\,CmG}{\varepsilon}\right\},$$

then $\mathbb{E}_\tau\Big[\,\|\nabla L(\bar{\mathbf{W}}_\tau)\|_*^{[\rho]}\Big] \le \varepsilon$, which means Pion outputs a (ρ,ε)-stationary point after T iterations. This proves the theorem.

G.3. Proof of Theorem 13

The proof follows the same O2NC reduction as in Theorem 12; the only change is the regret guarantee of the underlying online learner. Specifically, Leon (derived from FAML) achieves the same discounted-regret bound as Pion but without the dimensional factor, i.e., it corresponds to replacing $C = \left(\frac{n}{n-m-1}\right)^{1/4}$ by $C = 1$ in the proof of Theorem 12. Substituting $C = 1$ throughout yields the stated iteration complexity for Leon, with all other steps unchanged. For brevity, we do not repeat the argument.

Appendix H. Empirical Validation: Additional Details

To empirically validate the stability–convergence behavior suggested by our theory, we compare *Pion* and *Leon* against *Muon* on a synthetic Robust Matrix Sensing objective explicitly constructed to violate smooth-optimization assumptions. The purpose of this experiment is not to showcase best-case speed on benign losses, but to stress-test whether the *intrinsic (implicit) smoothing* built into the Matrix OLO formulation of Pion and Leon translates into visibly steadier descent dynamics when gradients are discontinuous and the landscape contains oscillatory nonconvex structure.

The key distinction is that Pion and Leon admit explicit stability–convergence guarantees in our framework, whereas Muon does not. This gap becomes most evident in nonsmooth regimes: without an intrinsic smoothing mechanism, Muon can react sharply to abrupt changes in the gradient matrix—exactly the behavior induced by objectives with kinked terms or rapidly varying components. In contrast, Pion and Leon inherit an *implicit smoothing effect* from the Matrix OLO geometry. Consequently, we expect Muon to exhibit more fluctuations on this stress test, while Pion and Leon follow smoother, more stable trajectories, with Leon typically enjoying a modest advantage consistent with our theoretical bounds.

We instantiate this stress test via a *Robust Matrix Sensing with Nonconvex Ripples* objective:

$$f(\mathbf{X}) = \frac{1}{N}\sum_{k=1}^{N}\left(|\langle\mathbf{A}_k,\mathbf{X}\rangle|\cdot\left(1-0.9\cos(3\langle\mathbf{A}_k,\mathbf{X}\rangle)\right)+0.5\right),$$

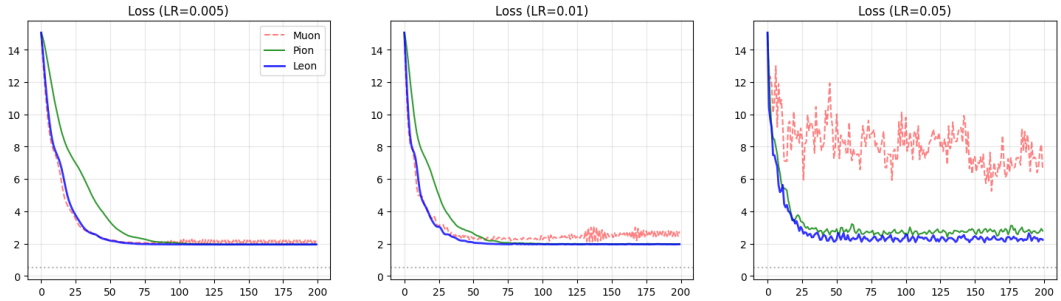

Figure 1: Convergence paths with different constant learning rates on the synthetic nonsmooth nonconvex stress test.

where $\mathbf{A}_k$ are random measurement matrices with i.i.d. Gaussian entries. In our experiments, $\mathbf{X} \in \mathbb{R}^{d\times d}$ with $d = 20$ and $N = 100$ measurements. For Leon, we set $\beta_1 = \beta_2 = 0.9$. The absolute-value term introduces ℓ_1-type nonsmoothness and hence gradient discontinuities near $\langle \mathbf{A}_k, \mathbf{X}\rangle = 0$, while the high-frequency cosine modulation creates persistent nonconvex "ripples" that repeatedly perturb the local geometry. Together, these components generate abruptly varying gradient signals designed to destabilize optimizers that track instantaneous gradients too closely.