MUON WITH FINITE NEWTON–SCHULZ: THE SMOOTHING BENEFIT IN NONSMOOTH NONCONVEX OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Muon has emerged as a strong optimizer for the matrix-valued parameters in large language model pretraining, approximately orthogonalizing its momentum with a few Newton–Schulz iterations. Existing theory either replaces this iteration with the exact polar factor it approximates, or treats it as a source of approximation error, and thus the iteration Muon actually runs can only hurt the guarantees. We show that finite Newton–Schulz can instead be beneficial for nonsmooth nonconvex optimization. To this end, we analyze Muon through the online-to-nonconvex conversion, which views the update rule of an optimizer as an online learner and converts its regret bound into a stationarity guarantee. We observe that the finite Newton–Schulz iteration smooths the discontinuous polar map into a Lipschitz map of the singular values, and Muon with finite Newton–Schulz can be regarded as an online learner with a smoothed spectral potential. Building on this observation, we prove that a Newton–Schulz depth growing only logarithmically in the target accuracy suffices for convergence to stationary points in nonsmooth nonconvex optimization, whereas Muon with the exact-polar update may fail to converge. The resulting sample complexity bounds match the best-known guarantees for nonsmooth nonconvex optimization and are optimal for smooth nonconvex optimization up to problem-dependent factors. The argument extends beyond Newton–Schulz to general spectral maps that are Lipschitz smoothings of the polar map.

1 INTRODUCTION

In large language model (LLM) pretraining, where most trainable parameters are matrices, Muon (Jordan et al., 2024) has emerged as a strong alternative to the de facto standard AdamW (Loshchilov & Hutter, 2019). Scaling studies report substantial efficiency gains (Liu et al., 2025), and a systematic benchmark places Muon among the strongest pretraining optimizers (Wen et al., 2026). Muon-based optimizers have also been used to train frontier models such as Kimi K2 and GLM-4.5 (Kimi Team, 2025; GLM-4.5 Team, 2025). This success depends on an efficient implementation of the operation at the core of Muon, the orthogonalization of the update direction.

At each round, Muon maintains an exponential moving average M of the stochastic gradients and updates the parameters along an orthogonalized version of M. For a singular value decomposition $M = U\Sigma V^\top$, the ideal direction is the polar factor $UV^\top$, which replaces every positive singular value of M with one (Jordan et al., 2024; Bernstein & Newhouse, 2024). Instead of computing $UV^\top$ through a singular value decomposition, Muon approximates it with a few iterations of Newton–Schulz, a classical scheme that repeatedly applies a fixed odd matrix polynomial using only matrix–matrix multiplications (Kovarik, 1970; Björck & Bowie, 1971). We refer to the iteration truncated at a finite depth as finite Newton–Schulz. This inexpensive orthogonalization is a key reason why Muon is practical at the scale of LLM pretraining.

Muon's empirical success has prompted a rapidly growing convergence theory, so far mostly for smooth objectives. Most of it, however, replaces the Newton–Schulz loop with the exact polar factor, or, more generally, with an exact linear minimization oracle over a norm ball (*e.g.,* Li & Hong, 2025; Kovalev, 2025; Pethick et al., 2025; Shen et al., 2026; Chen et al., 2026; Riabinin et al., 2026; Sfyraki & Wang, 2026), and thus the finite Newton–Schulz iteration never enters the analysis. More

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

recent analyses account for the finite Newton–Schulz iteration, but treat its finite-depth effect as an approximation error relative to the exact-polar update (Kim & Oh, 2026; Shulgin et al., 2026a; Choudhury et al., 2026). In these analyses, finite Newton–Schulz can only hurt because the guarantees improve when the iteration more closely approximates the exact polar factor.

A separate line of theoretical work removes the smoothness assumption on the objective and builds on online learning. The online-to-nonconvex conversion (O2NC) of Cutkosky et al. (2023) lets an online learner choose the update increments of an optimizer. The learner's regret then translates into a bound on a relaxed, Goldstein-type stationarity measure (Goldstein, 1977), the standard target in nonsmooth nonconvex optimization. Applying this framework to matrix optimizers, Jiang et al. (2026a) observe that Muon with the exact-polar update coincides with follow-the-leader over the operator-norm ball, an online algorithm that can suffer linear regret and thus cannot guarantee convergence to stationary points. They therefore replace follow-the-leader, injecting stochastic perturbations or augmenting the momentum matrix, and derive the Pion and Leon algorithms, which come with stationarity guarantees for nonsmooth objectives. These methods, however, depart from Muon as implemented. They add ingredients that Muon does not use, and their analyses do not involve the finite Newton–Schulz iteration.

Thus one line of work admits finite Newton–Schulz only as an error to be controlled and requires smoothness, while the other handles nonsmooth objectives but replaces the update that Muon actually performs. This leaves the basic question open:

Can the finite Newton–Schulz iteration be a benefit rather than an error, enabling Muon to find stationary points of nonsmooth objectives?

1.1 Contributions of this paper

We answer this question affirmatively. With momentum and finite Newton–Schulz, Muon finds stationary points of nonsmooth objectives, whereas with the exact-polar update it may fail to converge (Parshakova et al., 2026). The finite iteration replaces the discontinuous polar map with a Lipschitz map of the singular values, and this smoothing enables the online-to-nonconvex conversion to provide a stationarity guarantee.

The following informal statement summarizes the resulting guarantee.

Theorem 1.1 (Informal version of Theorem 5.4)**.** *Consider a Lipschitz objective that may be neither smooth nor convex, accessed through an unbiased stochastic gradient oracle with bounded noise. Then, for any radius $\rho > 0$ and accuracy $\varepsilon > 0$, Muon with momentum and $q = O(\log(1/\varepsilon))$ Newton–Schulz steps per round finds a (ρ, ε)-stationary point in expectation within $O(\rho^{-1}\varepsilon^{-3} + \varepsilon^{-2})$ stochastic gradient evaluations.*

The stationarity criterion is the (ρ, ε)-stationarity of Jiang et al. (2026a), formally defined in Definition 2.2, and the leading $\rho^{-1}\varepsilon^{-3}$ dependence matches the guarantees established for Pion and Leon under the same criterion (see Appendix A.2). To our knowledge, this is the first stationarity guarantee for nonsmooth nonconvex objectives in which finite Newton–Schulz acts as the smoothing mechanism that enables convergence rather than as an approximation error to be controlled. The required depth grows only logarithmically in the target accuracy ε, which is consistent with the empirical observation that a few Newton–Schulz iterations suffice in practice (Jordan et al., 2024). The precise algorithm we analyze and its remaining differences from deployed Muon are specified in Section 2 and discussed in Section 6.

Technical contributions. By the online-to-nonconvex conversion, it suffices to bound the discounted regret that Muon's update rule incurs as an online learner over the operator-norm ball (Cutkosky et al., 2023; Jiang et al., 2026a). We show that this update is the gradient of a smoothed spectral potential evaluated at the momentum, which makes the learner a gradient-based prediction algorithm (Abernethy et al., 2014; 2016). The discounted regret of such an algorithm decomposes into a penalty term, which shrinks as the potential approaches the nuclear norm, and a stability term, which grows with the Lipschitz constant of the induced spectral map. Unlike the standard decomposition in terms of the cumulative gradients (Jiang et al., 2026a), ours is carried out in terms of the momentum, which yields the optimal $O(1/\varepsilon^2)$ dependence for deterministic smooth objectives (Section 5.3). Our key lemma (Lemma 5.2) quantifies both terms for Muon with finite

Newton–Schulz: the penalty term decays and the stability term grows, both geometrically in the Newton–Schulz depth q. The depth therefore governs a penalty–stability tradeoff, and balancing the two terms yields sublinear discounted regret at $q = O(\log(1/\varepsilon))$ and, through the conversion, the guarantee above (Section 5). As $q \to \infty$, the update approaches the exact polar factor, and the stability term grows without bound, reflecting the linear regret that follow-the-leader can suffer.

The update of Muon with finite Newton–Schulz can also be regarded as follow-the-regularized-leader (FTRL) on the discounted linear losses, with a spectral regularizer given by the Fenchel conjugate of the smoothed potential (Theorem 5.3). The depth controls the amount of regularization: at $q = 0$ the regularizer becomes the squared Frobenius regularizer restricted to the operator-norm ball, and as $q \to \infty$ it decays to zero on the ball and the update approaches follow-the-leader.

Our analysis is not specific to the Newton–Schulz iteration. The regret bound holds for more general spectral maps of the singular values (Theorem 4.2), which gives sufficient conditions for such maps to yield stationarity guarantees on nonsmooth objectives. In particular, recent work designs a smooth relaxation of the polar map and derives its associated convex regularizer (Mustafi et al., 2026; Feoktistov et al., 2026), and our general analysis covers such relaxations (Appendix G). In contrast to these studies, we show that the finite Newton–Schulz iteration itself smooths the polar map and induces an FTRL regularizer, without introducing a separate relaxation. A detailed discussion of related work is provided in Appendix A.

2 Optimization setting and Muon with finite Newton–Schulz

Notation. We work with matrices in $\mathbb{R}^{m\times n}$ with $1 \le m \le n$ and set $r := \min\{m,n\} = m$. The case $m > n$ is reduced to this one by transposing every matrix. We write $\sigma_1(X) \ge \cdots \ge \sigma_r(X) \ge 0$ for the singular values of $X \in \mathbb{R}^{m\times n}$ and set $\sigma(X) := (\sigma_1(X), \dots, \sigma_r(X))$. We write $\operatorname{diag}(x_1, \dots, x_r) \in \mathbb{R}^{r\times r}$ for the diagonal matrix with diagonal entries $x_1, \dots, x_r$. We write $\langle X, Y\rangle := \operatorname{tr}(X^\top Y)$ for the Frobenius inner product and define the operator, Frobenius, and nuclear norms by $\|X\|_{\mathrm{op}} := \sigma_1(X)$, $\|X\|_{\mathrm{F}} := \sqrt{\sum_{i=1}^r \sigma_i(X)^2}$, and $\|X\|_* := \sum_{i=1}^r \sigma_i(X)$. These norms satisfy $\|X\|_{\mathrm{op}} \le \|X\|_{\mathrm{F}} \le \sqrt{r}\,\|X\|_{\mathrm{op}}$ and $\|X\|_{\mathrm{F}} \le \|X\|_* \le \sqrt{r}\,\|X\|_{\mathrm{F}}$. For a positive integer k, let $[k] := \{1, \dots, k\}$. For a scalar function h, we use $\mathrm{Lip}(h)$ to denote its Lipschitz constant. For a differentiable function Φ, we use $B_\Phi(M'\|M) := \Phi(M') - \Phi(M) - \langle \nabla\Phi(M), M' - M\rangle$ to denote the Bregman divergence from M to M' induced by Φ.

Polar factor and singular-value maps. Let $S = U \operatorname{diag}(\sigma_1(S), \dots, \sigma_r(S))V^\top$ be a thin singular value decomposition. Define its polar factor by $\operatorname{polar}(S) := U \operatorname{diag}(\mathbf{1}\{\sigma_i(S) > 0\})V^\top$, where $\mathbf{1}\{\cdot\}$ denotes the indicator function. The operator and nuclear norms are dual, so for every $S \in \mathbb{R}^{m\times n}$ and radius $D > 0$, $\sup_{\|X\|_{\mathrm{op}} \le D} \langle S, X\rangle = D\|S\|_*$, where the supremum is attained by the matrix $D \operatorname{polar}(S)$. For a scalar function $h\colon [0,\infty) \to \mathbb{R}$ with $h(0) = 0$, define the singular-value map

$$\mathcal{H}_h(S) := U \operatorname{diag}\left(h(\sigma_1(S)), \dots, h(\sigma_r(S))\right) V^\top \tag{1}$$

for $S = U \operatorname{diag}(\sigma_1(S), \dots, \sigma_r(S))V^\top$. The value of $\mathcal{H}_h(S)$ is independent of the chosen thin singular value decomposition. Indeed, a thin singular value decomposition is unique up to a simultaneous orthogonal change of basis in U and V within each group of repeated positive singular values, and this change cancels in (1). The singular vectors associated with zero singular values can be chosen independently in U and V, but the corresponding terms vanish since $h(0) = 0$.

2.1 Objective and stationarity

Let $\mathcal{D}$ be a probability distribution on a sample space $\mathcal{Z}$, and let $\ell\colon \mathbb{R}^{m\times n} \times \mathcal{Z} \to \mathbb{R}$ be a loss function. We consider the nonconvex matrix optimization problem

$$\min_{W\in\mathbb{R}^{m\times n}} \mathcal{L}(W) := \mathbb{E}_{\zeta\sim\mathcal{D}}[\ell(W;\zeta)].$$

We impose the following conditions on the objective $\mathcal{L}$ and the stochastic gradient oracle.

Assumption 2.1. *The function $\mathcal{L}$ is differentiable*[1] *and bounded below. There are constants $\Gamma > 0$ and $\sigma \ge 0$ such that, at every query point $W \in \mathbb{R}^{m\times n}$, the oracle uses a random seed independent*

[1] As discussed in Cutkosky et al. (2023, Proposition 2 and Corollary 6), for a locally Lipschitz objective that is not differentiable everywhere, we can construct a differentiable surrogate objective $\widehat{\mathcal{L}}_\delta(W) := \mathbb{E}_U[\mathcal{L}(W + \delta U)]$,

Algorithm 1: Muon online learner with momentum and finite Newton–Schulz

Require : momentum $\beta \in (0,1)$, radius $D > 0$, gradient bound $G_{\mathrm{op}} > 0$, depth $q \in \{0, 1, \dots\}$

1 Set $M_0 \leftarrow 0$ and $X_1 \leftarrow 0$;
2 **for** $t = 1, 2, \dots$ **do**
3 Receive gradient feedback G_t after playing X_t;
4 Update momentum $M_t \leftarrow \beta M_{t-1} + (1-\beta) G_t$;
5 Normalize momentum by $Y \leftarrow M_t / G_{\mathrm{op}}$;
6 **for** $j = 1, \dots, q$ **do**
7 $B \leftarrow YY^\top, \quad P \leftarrow BY, \quad Q \leftarrow BP; \quad Y \leftarrow \frac{15}{8} Y - \frac{5}{4} P + \frac{3}{8} Q;$
8 Set $X_{t+1} \leftarrow -DY$;

of the preceding history and returns a stochastic gradient G satisfying $\mathbb{E}[G \mid W] = \nabla \mathcal{L}(W)$, $\mathbb{E}[\|G\|_{\mathrm{F}}^2 \mid W] \le \Gamma^2$, and $\mathbb{E}[\|G - \nabla \mathcal{L}(W)\|_{\mathrm{F}}^2 \mid W] \le \sigma^2$.

For the initial point $W_0 \in \mathbb{R}^{m \times n}$, let $\Delta_{\mathcal{L}} := \mathcal{L}(W_0) - \inf_{W \in \mathbb{R}^{m \times n}} \mathcal{L}(W) < \infty$. Assumption 2.1 does not require smoothness of $\mathcal{L}$, and our guarantees are therefore stated in terms of the following notion of (ρ, ε)-stationarity.

Definition 2.2 (Jiang et al., 2026a, Definition 9)**.** For $W \in \mathbb{R}^{m \times n}$ and $\rho > 0$, let $\mathcal{P}(W; \rho)$ be the set of finitely supported probability distributions p on $\mathbb{R}^{m \times n}$ such that $\mathbb{E}_{Y \sim p}[Y] = W$ and $\mathbb{E}_{Y \sim p} \|Y - W\|_{\mathrm{op}} \le \rho$. For $\varepsilon \ge 0$, a point W is said to be a (ρ, ε)-*stationary point* if $\|\nabla \mathcal{L}(W)\|_*^{[\rho]} := \inf_{p \in \mathcal{P}(W;\rho)} \|\mathbb{E}_{Y \sim p} \nabla \mathcal{L}(Y)\|_* \le \varepsilon$.

The point mass at W belongs to $\mathcal{P}(W; \rho)$, so $\|\nabla \mathcal{L}(W)\|_*^{[\rho]} \le \|\nabla \mathcal{L}(W)\|_*$. Other distributions in $\mathcal{P}(W; \rho)$ allow gradients at different points to cancel on average. Under Assumption 2.1, Jensen's inequality gives $\|\nabla \mathcal{L}(W)\|_{\mathrm{F}} \le \Gamma$ for every W, so $\mathcal{L}$ is Γ-Lipschitz with respect to the Frobenius norm, and for general nonsmooth Lipschitz objectives such relaxations are unavoidable (Zhang et al., 2020; Kornowski & Shamir, 2022).

2.2 Muon with momentum and finite Newton–Schulz

For matrix parameters, Muon maintains gradient momentum and approximately orthogonalizes it with a few Newton–Schulz iterations. The specific Muon update investigated in this paper is given in Algorithm 1. The learner in Algorithm 1 maintains the momentum sequence

$$M_0 := 0, \qquad M_t := \beta M_{t-1} + (1-\beta) G_t, \tag{2}$$

where $\beta \in (0,1)$ is the momentum parameter, as in Line 4. To normalize the momentum, we make the following assumption.

Assumption 2.3. *There is a constant $G_{\mathrm{op}} > 0$ such that, at every query point $W \in \mathbb{R}^{m \times n}$, the oracle returns a stochastic gradient G satisfying $\|G\|_{\mathrm{op}} \le G_{\mathrm{op}}$ almost surely.*

Under Assumption 2.3, we assume without loss of generality that $\Gamma \le \sqrt{r}\, G_{\mathrm{op}}$, since $\mathbb{E}[\|G\|_{\mathrm{F}}^2 \mid W] \le r G_{\mathrm{op}}^2$. The normalization in Line 5 then satisfies $\|M_t / G_{\mathrm{op}}\|_{\mathrm{op}} \le 1$. Then the learner applies q Newton–Schulz steps in Line 7 (Kovarik, 1970; Björck & Bowie, 1971). Finally, the learner sets the next update to $X_{t+1} = -DY$ in Line 8.

For the Newton–Schulz polynomial (Line 7), we take the degree-five polynomial, which is commonly used in Muon. On the singular values, one step of Line 7 then acts as

$$f(x) := \frac{15}{8} x - \frac{5}{4} x^3 + \frac{3}{8} x^5, \qquad A := f'(0) = \frac{15}{8}.$$

For Newton–Schulz polynomials of general degree, truncating the Taylor expansion of $\lambda^{-1/2}$ around $\lambda = 1$ at order $\kappa \ge 1$ gives

$$p_\kappa(\lambda) := \sum_{s=0}^{\kappa} c_s (1-\lambda)^s, \qquad c_s := \frac{(2s)!}{4^s (s!)^2}, \tag{3}$$

where $\delta > 0$ and U is uniform on the Frobenius unit ball, and this does not worsen the resulting stationarity guarantees.

and $f(x) = x\,p_2(x^2)$ is the case $\kappa = 2$. We defer the analysis at a general order κ to Appendix E.1. For $q \in \{0, 1, 2, \ldots\}$, let $f^{\circ q}$ denote the q-fold composition of f, with $f^{\circ 0}$ the identity, and define

$$h_q := f^{\circ q} \quad \text{on } [0,1], \qquad h_q(x) := 1 \quad \text{for } x \geq 1. \tag{4}$$

Under Assumption 2.3, the procedure in Lines 5–8 can be written as

$$X_{t+1} = -D\mathcal{H}_{h_q}(M_t/G_{\mathrm{op}}). \tag{5}$$

In the discounted online-to-nonconvex conversion framework introduced in Section 3, we use the learner in Algorithm 1 as the online learner $\mathcal{A}$ to determine the update direction in the optimizer.

The learner in Algorithm 1 retains momentum and finitely many Newton–Schulz steps, the two features that our analysis concerns. It differs from practical Muon implementations in other respects, including the fixed normalization discussed in Section 6.

3 Discounted Online-to-Nonconvex Conversion

Online-to-nonconvex conversion (O2NC) uses an online learner to select update directions and converts a bound on regret into a stationarity guarantee (Cutkosky et al., 2023). In particular, we use the discounted O2NC framework of Jiang et al. (2026a), since it uses the same discount factor β as the momentum update in Algorithm 1.

Conversion mechanism. Let W_t denote the iterate at time t, with initial point W_0. At each round t, an online learner $\mathcal{A}$ chooses an increment X_t satisfying $\|X_t\|_{\mathrm{op}} \leq D$ from the preceding gradient feedback, and sets $W_t = W_{t-1} + X_t$. The learner then queries the oracle at $\widetilde{W}_t := W_{t-1} + u_t X_t$ with $u_t \sim \mathrm{Unif}([0,1])$ to obtain a stochastic gradient G_t satisfying Assumption 2.1 with $W = \widetilde{W}_t$, and feeds the linear loss $\ell_t^{[\beta]}(X) := \beta^{-t}\langle G_t, X\rangle$ to $\mathcal{A}$. The fundamental theorem of calculus along the segment from W_{t-1} to W_t gives $\mathcal{L}(W_t) - \mathcal{L}(W_{t-1}) = \int_0^1 \langle \nabla\mathcal{L}(W_{t-1} + uX_t), X_t\rangle\,\mathrm{d}u$, and averaging over the uniform u_t and using $\mathbb{E}[G_t \mid \widetilde{W}_t] = \nabla\mathcal{L}(\widetilde{W}_t)$ identifies this integral with $\mathbb{E}[\langle G_t, X_t\rangle \mid W_{t-1}, X_t]$. By this identity, controlling the decrease of the objective reduces to online linear optimization with gradient feedback G_t, without any smoothness of $\mathcal{L}$. The full protocol is given in Algorithm 2 in Appendix C.

After T rounds, the learner returns $\bar{W}_\tau$, where $\bar{W}_t$ is the exponentially weighted average (EWA) of the query points and $\tau \in [T]$ is drawn independently according to

$$\bar{W}_t := \frac{1-\beta}{1-\beta^t}\sum_{s=1}^{t}\beta^{t-s}\widetilde{W}_s, \qquad \mathbb{P}(\tau = t) := \begin{cases} \dfrac{1-\beta^t}{T}, & 1 \leq t < T, \\ \dfrac{1-\beta^T}{(1-\beta)T}, & t = T. \end{cases} \tag{6}$$

Conversion guarantee. The discounted O2NC framework transforms a discounted regret bound of an online learner into a guarantee for (ρ, ε)-stationarity in Definition 2.2. Define the discounted regret over the feasible set $\{X \in \mathbb{R}^{m\times n} : \|X\|_{\mathrm{op}} \leq D\}$ by

$$\mathrm{Reg}_t^{[\beta]}(D) := \max_{\|X\|_{\mathrm{op}} \leq D} \sum_{s=1}^{t}\beta^{t-s}\langle G_s, X_s - X\rangle. \tag{7}$$

The following lemma gives this conversion, with proof in Appendix C.

Lemma 3.1 (Corollary of Jiang et al., 2026a, Proposition 25)**.** *Under Assumption 2.1, fix $\rho > 0$ and run Algorithm 2 with $D = (1-\beta)\rho/(4\beta)$. Then*

$$\mathbb{E}\Big[\|\nabla\mathcal{L}(\bar{W}_\tau)\|_*^{[\rho]}\Big] \leq \frac{4\Delta_{\mathcal{L}}}{(1-\beta)\rho T} + \frac{1}{T}\mathbb{E}\left[\mathrm{Reg}_T^{[\beta]}(1) + (1-\beta)\sum_{t=1}^{T-1}\mathrm{Reg}_t^{[\beta]}(1)\right] + \left(1 - \beta + \frac{\beta}{T}\right)\frac{\sqrt{r}\,\sigma}{\sqrt{1-\beta^2}},$$

where $\mathrm{Reg}_t^{[\beta]}(1) := \max_{\|U\|_{\mathrm{op}}\leq 1}\sum_{s=1}^{t}\beta^{t-s}\langle G_s, X_s/D - U\rangle$ *so that* $\mathrm{Reg}_t^{[\beta]}(D) = D\,\mathrm{Reg}_t^{[\beta]}(1)$.

In the following sections, we bound $\mathbb{E}[\mathrm{Reg}_t^{[\beta]}(1)]$ for the online learner.

4 DISCOUNTED REGRET FOR A MATRIX ONLINE LEARNER WITH A GENERAL SPECTRAL MAP

The conversion of Section 3 reduces (ρ, ε)-stationarity to bounding the discounted regret of the online learner over the operator-norm ball. This section bounds the discounted regret of the learner that applies a general spectral map to the momentum. The analysis is based on the gradient-based prediction algorithm (GBPA) framework (Abernethy et al., 2014; 2016), in which the learner plays the gradient of a differentiable potential function of the aggregated gradient feedback. In our setting, this aggregate corresponds to the momentum M_{t-1}, so the learner plays $X_t = -D\nabla\widetilde{\Phi}_t(M_{t-1})$ for some potential function $\widetilde{\Phi}_t$, and we consider the regret decomposition in terms of the momentum sequence. Section 4.1 states the bound and Section 4.2 sketches its proof, while Section 5 specializes the result to Muon with finite Newton–Schulz in Algorithm 1. The proofs of this section are deferred to Appendix D.

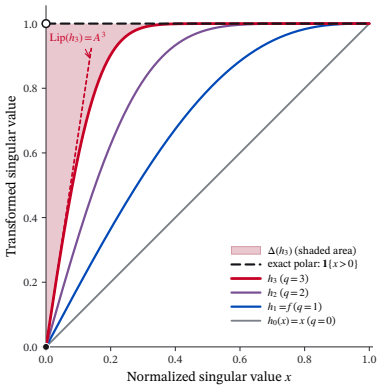


Figure 1: Finite Newton–Schulz smooths the exact polar map. Increasing q decreases $\Delta(h_q)$ but increases $\mathrm{Lip}(h_q) = A^q$. The shaded area equals $\Delta(h_3)$.

We consider a spectral map $h\colon [0,\infty) \to [0,1]$ satisfying the following conditions.

Assumption 4.1. *The function $h\colon [0,\infty) \to [0,1]$ is continuous, nondecreasing, and satisfies $h(0) = 0$, $\mathrm{Lip}(h) < \infty$, and $\Delta(h) := \int_0^\infty (1 - h(u))\,\mathrm{d}u < \infty$.*

Here the approximation error $\Delta(h)$ measures the gap between h and the exact polar map, which sends every positive singular value to one, and a smaller $\Delta(h)$ forces a larger $\mathrm{Lip}(h)$. The function h_q in (4) satisfies Assumption 4.1, as verified in Section 5.1, and Figure 1 illustrates this tradeoff for $h = h_q$.

4.1 ONLINE LEARNER AND DISCOUNTED-REGRET BOUND

Using the momentum (2) and a normalization scale $C_t > 0$, we consider the online learner that takes the update

$$X_t^h := -D\mathcal{H}_h\left(\frac{M_{t-1}}{C_t}\right), \tag{8}$$

which satisfies $\|X_t^h\|_{\mathrm{op}} \le D$ since h takes values in $[0,1]$. For $h = h_q$ and $C_t = G_{\mathrm{op}}$, the update (8) coincides with the finite Newton–Schulz update in (5) provided that $\|M_t\|_{\mathrm{op}} \le G_{\mathrm{op}}$ for all t. Replacing $\mathcal{H}_h(\cdot)$ with $\mathrm{polar}(\cdot)$ recovers Muon with the exact-polar computation. To regard the update (8) as a GBPA update, we need a potential whose gradient is the singular-value map $\mathcal{H}_h(\cdot/C_t)$. The potential corresponding to exact-polar Muon is the base potential $\Phi(M) := \|M\|_*$, one of whose subgradients is the exact polar factor. The update (8) replaces this base potential Φ by the smoothed potential given by

$$\widetilde{\Phi}_{h,C}(M) := C\sum_{i=1}^{r}\phi_h\left(\frac{\sigma_i(M)}{C}\right) \qquad \text{for} \quad \phi_h(x) := \int_0^x h(u)\,\mathrm{d}u. \tag{9}$$

The regret bound under the update (8) depends on h through two scalar quantities: the Lipschitz constant $\mathrm{Lip}(h)$ and the approximation error $\Delta(h)$.

Theorem 4.2. *Suppose Assumption 4.1 holds. For any $t \ge 1$, let $(C_s)_{s=1}^t$ be positive reals satisfying $C_{s+1} \ge \beta C_s$ for $s \in [t-1]$. Then the online learner (8) achieves*

$$\mathrm{Reg}_t^{[\beta]}(D) \le D\left(\frac{rC_t\Delta(h)}{1-\beta} + \frac{1}{1-\beta}\sum_{s=1}^{t}\beta^{t-s}B_{\widetilde{\Phi}_{h,C_s}}(M_s\|M_{s-1})\right). \tag{10}$$

If, in addition, Assumption 2.1 holds and $C_s = C$ for all $s \in [t]$, then

$$\mathbb{E}[\mathrm{Reg}_t^{[\beta]}(D)] \le D\left(\frac{rC\Delta(h)}{1-\beta} + \frac{2\Gamma^2\,\mathrm{Lip}(h)}{C}\right). \tag{11}$$

The coarser bound (11) is sufficient for the guarantee for nonsmooth nonconvex objectives in Theorem 5.4, while the guarantee for smooth objectives in Theorem 5.7 relies on the bound depending on the Bregman divergence (10) to obtain its optimal $O(1/\varepsilon^2)$ dependence in the deterministic case. Notably, the bound (11) has the same structure as the well-known penalty–stability tradeoff in the analysis of FTRL and online mirror descent (*e.g.,* Shalev-Shwartz, 2012; McMahan, 2017; Zimmert & Seldin, 2021; Tsuchiya et al., 2023), with a penalty term proportional to $\Delta(h)$ and a stability term proportional to $\mathrm{Lip}(h)$.

4.2 Proof sketch of Theorem 4.2

We first use the following lemma to show that (8) is a GBPA update.

Lemma 4.3. *Under Assumption 4.1, for every $C > 0$ the function $\widetilde{\Phi}_{h,C}$ is convex and continuously differentiable on all of $\mathbb{R}^{m\times n}$, and its gradient is $\nabla\widetilde{\Phi}_{h,C}(M) = \mathcal{H}_h(M/C)$.*

By Lemma 4.3, the update (8) can be written as $X_t^h = -D\nabla\widetilde{\Phi}_{h,C_t}(M_{t-1})$, which is GBPA with the time-varying potential $\widetilde{\Phi}_t = \widetilde{\Phi}_{h,C_t}$ evaluated at the momentum M_{t-1}. Unlike the standard GBPA decomposition in terms of the cumulative gradients (Abernethy et al., 2014; 2016), employed by Jiang et al. (2026a), the following lemma decomposes the discounted regret directly in terms of the momentum sequence.

Lemma 4.4. *Under Assumption 4.1, the discounted regret for the online learner (8) is decomposed as*

$$\mathrm{Reg}_t^{[\beta]}(D) = \frac{D}{1-\beta}\left(\Phi(M_t) - \widetilde{\Phi}_{h,C_t}(M_t) + \sum_{s=1}^{t}\beta^{t-s}B_{\widetilde{\Phi}_{h,C_s}}(M_s\|M_{s-1}) + \sum_{s=1}^{t-1}\beta^{t-s-1}V_s\right), \quad (12)$$

where $V_s \coloneqq \widetilde{\Phi}_{h,C_{s+1}}(M_s) - \beta\widetilde{\Phi}_{h,C_s}(M_s) - (1-\beta)\langle M_s, \mathcal{H}_h(M_s/C_{s+1})\rangle$.

On the right-hand side of (12), the first two terms $\Phi(M_t) - \widetilde{\Phi}_{h,C_t}(M_t)$ form a penalty term, the gap between the base potential Φ and the smoothed potential $\widetilde{\Phi}_{h,C_t}$, the third term is a stability term, the accumulated Bregman divergences along the momentum trajectory, and the last term is due to the variation of the potentials across rounds. The following three lemmas bound these terms: Lemma 4.5 bounds the penalty term by $rC_t\Delta(h)$, Lemma 4.6 bounds each Bregman divergence in the stability term by $\frac{\mathrm{Lip}(h)}{2C_s}\|M_s - M_{s-1}\|_{\mathrm{F}}^2$, and Lemma 4.7 shows that $V_s \le 0$ under the scale condition $C_{s+1} \ge \beta C_s$.

Lemma 4.5. *Under Assumption 4.1, for every $M \in \mathbb{R}^{m\times n}$ and $C > 0$, $0 \le \Phi(M) - \widetilde{\Phi}_{h,C}(M) \le rC\Delta(h)$.*

Lemma 4.6. *Under Assumption 4.1, for every $C > 0$ the gradient of $\widetilde{\Phi}_{h,C}$ is $(\mathrm{Lip}(h)/C)$-Lipschitz in the Frobenius norm. Consequently, $B_{\widetilde{\Phi}_{h,C}}(M'\|M) \le \frac{\mathrm{Lip}(h)}{2C}\|M' - M\|_{\mathrm{F}}^2$.*

Lemma 4.7. *Let $\beta \in (0,1)$ and $C, C' > 0$ with $C' \ge \beta C$. Under Assumption 4.1, for every $M \in \mathbb{R}^{m\times n}$, $\widetilde{\Phi}_{h,C'}(M) - \beta\widetilde{\Phi}_{h,C}(M) - (1-\beta)\langle M, \nabla\widetilde{\Phi}_{h,C'}(M)\rangle \le 0$. Consequently, $V_s \le 0$ whenever $C_{s+1} \ge \beta C_s$.*

Combining Lemmas 4.5 and 4.7 with the decomposition in Lemma 4.4 gives (10). For $C_s = C$, Assumption 2.1 gives $\mathbb{E}[\|M_s - M_{s-1}\|_{\mathrm{F}}^2] \le 4(1-\beta)^2\Gamma^2$. Applying Lemma 4.6 and using $(1-\beta)\sum_{s=1}^{t}\beta^{t-s} \le 1$ then yields (11) of Theorem 4.2.

5 Regret and stationarity guarantees for Muon with finite Newton–Schulz

This section specializes the general theory of Section 4 to the finite Newton–Schulz map and to Algorithm 1. The proofs of this section are deferred to Appendices E and F.

5.1 Discounted regret for Muon with finite Newton–Schulz

We bound the discounted regret of Algorithm 1 by specializing Theorem 4.2 to $h = h_q$ from (4). Throughout this section, we use the constant normalization $C_t = G_{\mathrm{op}}$ under Assumption 2.3. First,

we verify that h_q satisfies Assumption 4.1. By (4), h_q is the q-fold composition of f on $[0,1]$, extended by the constant 1 on $[1,\infty)$. Since f is a polynomial that is nondecreasing on $[0,1]$ with $f(0)=0$ and $f(1)=1$, $h_q\colon[0,\infty)\to[0,1]$ is continuous and nondecreasing, satisfies $h_q(0)=0$ and $h_q(x)=1$ for $x\geq 1$, and has a finite Lipschitz constant, as illustrated in Figure 1.

Since the momentum update (2) and Assumption 2.3 give $\|M_t\|_{\mathrm{op}}\leq(1-\beta^t)G_{\mathrm{op}}\leq G_{\mathrm{op}}$ for every t, all singular values of M_t/G_{op} lie in $[0,1]$, and the finite Newton–Schulz iteration coincides with the singular-value map $\mathcal{H}_{h_q}(M_t/G_{\mathrm{op}})$. Thus Algorithm 1 is exactly the online learner of (8) with normalization $C_t=G_{\mathrm{op}}$, which satisfies the scale condition $C_{t+1}=G_{\mathrm{op}}\geq\beta C_t$ and keeps the update feasible, $\|X_{t+1}\|_{\mathrm{op}}\leq D$. Specializing Theorem 4.2 to $h=h_q$ and choosing the depth q via our key lemma (Lemma 5.2 below), we obtain the following discounted-regret bound.

Theorem 5.1. *Suppose Assumptions 2.1 and 2.3 hold and $\Gamma>0$. For every $t\geq 1$, the learner of Algorithm 1 with depth*

$$q=\left\lceil \log_A \frac{G_{\mathrm{op}}\sqrt{r}}{\Gamma\sqrt{1-\beta}}\right\rceil \quad \textit{achieves} \quad \mathbb{E}[\mathrm{Reg}_t^{[\beta]}(D)]\leq(1+2A)D\Gamma\sqrt{\frac{r}{1-\beta}}. \tag{13}$$

The bound of Theorem 5.1 is obtained from (11) in Theorem 4.2 with $h=h_q$, by balancing a penalty term controlled by the approximation error $\Delta(h_q)$ against a stability term controlled by the Lipschitz constant $\mathrm{Lip}(h_q)$. The following lemma shows that $\mathrm{Lip}(h_q)$ grows exactly as A^q, while $\Delta(h_q)$ remains within constant factors of A^{-q}.

Lemma 5.2. *For every integer $q\geq 0$, the function h_q in (4) satisfies*

$$\mathrm{Lip}(h_q)=A^q, \qquad \frac{1}{2A^q}\leq\Delta(h_q)\leq\frac{1}{A^q+1}\leq\frac{1}{A^q}, \qquad A=\frac{15}{8}. \tag{14}$$

Proof sketch of Theorem 5.1. Substituting $\mathrm{Lip}(h_q)=A^q$ and $\Delta(h_q)\leq A^{-q}$ from Lemma 5.2 into (11) in Theorem 4.2 with $C=G_{\mathrm{op}}$ gives $\mathbb{E}[\mathrm{Reg}_t^{[\beta]}(D)]\leq D\left(\frac{rG_{\mathrm{op}}}{(1-\beta)A^q}+\frac{2A^q\Gamma^2}{G_{\mathrm{op}}}\right)$. Balancing the two terms over the integer depth q gives (13). □

This tradeoff explains the benefit of a finite Newton–Schulz iteration over exact orthogonalization. As $q\to\infty$, h_q approaches the exact polar map, and the update approaches the exact polar factor, the ideal Muon update. In this limit, $\Delta(h_q)\to 0$, but $\mathrm{Lip}(h_q)\to\infty$, which reflects the discontinuity of the polar map, and the stability term is no longer controlled. Lemma E.1 shows that the same tradeoff holds at any finite Taylor order and that our analysis therefore extends beyond the degree-five Newton–Schulz polynomial investigated here. Moreover, the same penalty–stability tradeoff is unavoidable for general spectral maps, as every h satisfying Assumption 4.1 obeys $\Delta(h)\geq 1/(2\,\mathrm{Lip}(h))$. By Lemma 5.2, the finite Newton–Schulz map satisfies $\Delta(h_q)\leq 1/\mathrm{Lip}(h_q)$ and hence attains this lower bound up to a factor of two. Appendix G proves this lower bound and instantiates our guarantees for a recently proposed smooth relaxation of the polar map (Mustafi et al., 2026; Feoktistov et al., 2026). Appendix H illustrates this depth tradeoff numerically on a synthetic nonsmooth objective.

Interpretation as FTRL. By Fenchel duality, the online learner (8) is follow-the-regularized-leader (FTRL) on the discounted linear losses, with a spectral regularizer given by the Fenchel conjugate of the smoothed potential. Define the Fenchel conjugates $\phi_h^*(a):=\sup_{x\geq 0}\{ax-\phi_h(x)\}$ and $\widetilde{\Phi}_{h,C}^*(W):=\sup_{M\in\mathbb{R}^{m\times n}}\{\langle W,M\rangle-\widetilde{\Phi}_{h,C}(M)\}$.

Theorem 5.3. *Suppose Assumption 4.1 holds and let $(C_t)_{t\geq 1}$ be a positive sequence. For any $t\geq 1$, the update (8) is the FTRL update*

$$X_t^h=-D\nabla\widetilde{\Phi}_{h,C_t}(M_{t-1})\in\operatorname*{arg\,min}_{\|X\|_{\mathrm{op}}\leq D}\left\{\left\langle\sum_{s=1}^{t-1}\beta^{-s}G_s,X\right\rangle+\frac{R_{h,C_t}(X)}{(1-\beta)\beta^{t-1}}\right\}, \tag{15}$$

where the regularizer is $R_{h,C}(X):=D\widetilde{\Phi}_{h,C}^*(X/D)=DC\sum_{i=1}^r\phi_h^*(\sigma_i(X)/D)$.

By Theorem 5.3 with $h=h_q$, the finite Newton–Schulz update of Algorithm 1 corresponds to FTRL with the regularizer R_{h_q,C_t}. At $q=0$ the regularizer is the Euclidean quadratic $(C_t/2D)\|X\|_{\mathrm{F}}^2$ on the ball, and as $q\to\infty$ it decays to zero and X_t approaches $-D\,\mathrm{polar}(M_{t-1})$, the follow-the-leader (FTL) update of exact-polar Muon. We discuss this in more detail in Appendix F.

5.2 Sample complexity in nonsmooth nonconvex optimization

Substituting the discounted-regret bound in Theorem 5.1 into the O2NC conversion in Lemma 3.1, we can obtain the following stationarity guarantee for nonsmooth nonconvex optimization.

Theorem 5.4. *Suppose Assumptions 2.1 and 2.3 hold and $\Gamma > 0$, fix $\rho > 0$ and $\varepsilon > 0$, and let $B_{\mathrm{NS}} := \sqrt{r}((1+2A)\Gamma + \sigma)$. Running Algorithm 2 using Algorithm 1 as its online learner $\mathcal{A}$ with $1-\beta = \min\{1/9, (\varepsilon/3B_{\mathrm{NS}})^2\}$, $D = (1-\beta)\rho/(4\beta)$, and depth $q = \lceil \log_A \max\{1,\ 3B_{\mathrm{NS}}G_{\mathrm{op}}\sqrt{r}/(\Gamma\varepsilon)\}\rceil$ returns a (ρ,ε)-stationary point $\bar{W}_\tau$ in expectation with gradient-oracle complexity $T = O\left(\max\left\{\frac{\Delta_{\mathcal{L}}\, r(\Gamma+\sigma)^2}{\rho\varepsilon^3},\ \frac{r(\Gamma+\sigma)^2}{\varepsilon^2}\right\}\right)$.*

The leading $\rho^{-1}\varepsilon^{-3}$ term matches the bounds for Pion and Leon (Jiang et al., 2026a), and Appendix A.2 gives a detailed comparison.

5.3 Sample complexity in smooth nonconvex optimization

Using the same discounted O2NC reduction, we also obtain a guarantee for smooth objectives. The proofs of this subsection are deferred to Appendix E.4. In addition to Assumptions 2.1 and 2.3, we assume the following smoothness of $\mathcal{L}$ with respect to the operator norm.

Assumption 5.5. *There is a constant $L_{\mathrm{op}} > 0$ such that $\|\nabla\mathcal{L}(W) - \nabla\mathcal{L}(W')\|_* \le L_{\mathrm{op}}\|W - W'\|_{\mathrm{op}}$ for all $W, W' \in \mathbb{R}^{m\times n}$.*

For smooth objectives we use the bound (10) of Theorem 4.2, stated in terms of the Bregman divergences. Under Assumption 5.5, and for $h = h_q$ and $C_s = G_{\mathrm{op}}$, each Bregman divergence $B_{\widetilde{\Phi}_{h_q,G_{\mathrm{op}}}}(M_s\|M_{s-1})$ is bounded more sharply, as shown in Lemma E.2. The resulting bound on $\|\nabla\mathcal{L}(W)\|_*^{[\rho]}$ is converted to a bound on $\|\nabla\mathcal{L}(W)\|_*$ by the following lemma.

Lemma 5.6. *Under Assumption 5.5, for every $W \in \mathbb{R}^{m\times n}$, $\|\nabla\mathcal{L}(W)\|_* \le \|\nabla\mathcal{L}(W)\|_*^{[\rho]} + L_{\mathrm{op}}\rho$.*

Thus, the O2NC conversion of Lemma 3.1 gives the following sample complexity guarantee.

Theorem 5.7. *Suppose Assumptions 2.1, 2.3 and 5.5 hold and fix $\varepsilon > 0$. Running Algorithm 2 using Algorithm 1 as its online learner $\mathcal{A}$ with $1-\beta = \min\{1/2, \varepsilon^2/(144r\sigma^2)\}$, $D = (1-\beta)\varepsilon/(48\beta L_{\mathrm{op}})$, and depth $q = \lceil \log_A \max\{1, 12rG_{\mathrm{op}}/\varepsilon\}\rceil$ returns $\bar{W}_\tau$ satisfying $\mathbb{E}\left[\|\nabla\mathcal{L}(\bar{W}_\tau)\|_*\right] \le \varepsilon$ with gradient-oracle complexity $T = O\left(\max\left\{\frac{L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon^2},\ \frac{r\sigma^2}{\varepsilon^2},\ \frac{r\sigma^2 L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon^4}\right\}\right)$.*

The ε^{-2} dependence in the deterministic case matches the classical lower bound for smooth nonconvex optimization (Carmon et al., 2020). Under the standard unbiased bounded-variance stochastic first-order oracle, the ε^{-4} dependence matches the lower bound of Arjevani et al. (2023). The leading terms coincide with those established for exact-polar Muon (Kovalev, 2025; Shen et al., 2026), as summarized in Appendix A.2. The resulting complexity has the same dependence as that of Kim & Oh (2026), except that the middle term improves from $r^2\sigma^2/\varepsilon^2$ to $r\sigma^2/\varepsilon^2$, though the algorithms under analysis are slightly different.

6 Conclusion and future work

This paper analyzed Muon with momentum and finite Newton–Schulz through the discounted online-to-nonconvex conversion, viewing it as an online learner with a smoothed spectral potential. In this view, the finite Newton–Schulz iteration is a smoothing mechanism rather than an approximation error, and its depth controls a penalty–stability tradeoff between the polar-approximation error and the Lipschitz constant of the update. Balancing the two terms at a depth that grows only logarithmically in the target accuracy yields sublinear discounted regret, and hence convergence to stationary points of nonsmooth nonconvex objectives, whereas Muon with the exact-polar update may fail to converge.

The analysis leaves a gap to practical Muon. We use a fixed normalization constant known in advance and the classical Newton–Schulz coefficients obtained by truncating the Taylor expansion. In practice, Muon uses data-dependent normalization and empirically tuned coefficients (Jordan et al., 2024). The conversion also evaluates gradients at random points between consecutive iterates, whereas practical Muon evaluates them at the current iterate. Extending the analysis to these practical choices is left for future work.

AI USE STATEMENT

The central research idea and overall approach originated with the authors, who also prepared the initial manuscript primarily themselves. We used generative AI tools (GPT-5.6-Sol, GPT-6-Astra, and Fable-5) for examining and refining mathematical arguments, assisting with the writing and checking of proofs, and implementing the methods used in the numerical experiments. We also used these tools to identify related literature, improve the clarity and readability of the manuscript, and create the experimental figures. We reviewed all AI-assisted work, including the experimental code, figures, and edited text. In particular, we checked the mathematical arguments against the stated assumptions and our own calculations, and verified the cited results against the original sources. We take responsibility for the final content of this work.

ETHICS STATEMENT

This work is theoretical and does not involve human subjects, personal data, or applications that raise ethical concerns.

REPRODUCIBILITY STATEMENT

The optimization setting and oracle assumptions are specified in Section 2, and the Muon update and discounted online-to-nonconvex conversion are given in Algorithms 1 and 2. A proof sketch of the general discounted-regret bound is provided in Section 4, with auxiliary results and deferred proofs in Appendices B–G. The synthetic objective, the distribution used to generate the problem instance, the compared methods, and their parameter settings are described in Appendix H.

REFERENCES

Jacob Abernethy, Chansoo Lee, Abhinav Sinha, and Ambuj Tewari. Online linear optimization via smoothing. In *Proceedings of the 27th Conference on Learning Theory*, volume 35, pp. 807–823, 2014. 2, 6, 7

Jacob Abernethy, Chansoo Lee, and Ambuj Tewari. Perturbation techniques in online learning and optimization. In *Perturbations, Optimization, and Statistics*, chapter 8, pp. 233–264. MIT Press, 2016. 2, 6, 7

Kwangjun Ahn and Ashok Cutkosky. Adam with model exponential moving average is effective for nonconvex optimization. In *Advances in Neural Information Processing Systems*, volume 37, pp. 94909–94933, 2024. 14

Kwangjun Ahn, Zhiyu Zhang, Yunbum Kook, and Yan Dai. Understanding Adam optimizer via online learning of updates: Adam is FTRL in disguise. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 619–640, 2024. 14

Kwangjun Ahn, Gagik Magakyan, and Ashok Cutkosky. General framework for online-to-nonconvex conversion: Schedule-free SGD is also effective for nonconvex optimization. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 772–795, 2025. 14

Noah Amsel, David Persson, Christopher Musco, and Robert M. Gower. The Polar Express: Optimal matrix sign methods and their application to the Muon algorithm. In *International Conference on Learning Representations*, 2026. 14

Fredrik Andersson, Marcus Carlsson, and Karl-Mikael Perfekt. Operator-Lipschitz estimates for the singular value functional calculus. *Proceedings of the American Mathematical Society*, 144(5): 1867–1875, 2016. 16

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199(1–2): 165–214, 2023. 9

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. In *OPT 2024: 16th Annual Workshop on Optimization for Machine Learning*, 2024. 1, 14

Jeremy Bernstein and Laker Newhouse. Modular duality in deep learning. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 3920–3930, 2025. 14

Åke Björck and C. Bowie. An iterative algorithm for computing the best estimate of an orthogonal matrix. *SIAM Journal on Numerical Analysis*, 8(2):358–364, 1971. 1, 4, 14

Guillaume Braun, Han Bao, Wei Huang, and Masaaki Imaizumi. Spectral gradient descent mitigates anisotropy-driven misalignment: A case study in phase retrieval. In *International Conference on Machine Learning*, 2026. 14

David E. Carlson, Edo Collins, Ya-Ping Hsieh, Lawrence Carin, and Volkan Cevher. Preconditioned spectral descent for deep learning. In *Advances in Neural Information Processing Systems*, volume 28, pp. 2971–2979, 2015. 14

Yair Carmon, John C. Duchi, Oliver Hinder, and Aaron Sidford. Lower bounds for finding stationary points I. *Mathematical Programming*, 184:71–120, 2020. 9

Lizhang Chen, Jonathan Li, and Qiang Liu. Muon optimizes under spectral norm constraints. *Transactions on Machine Learning Research*, 2026. 1, 14

Sayantan Choudhury, Xiaoran Cheng, Martin Takáč, Sen Na, and Mladen Kolar. Muon with Nesterov momentum: Heavy-tailed noise and (randomized) inexact polar decomposition. *arXiv preprint arXiv:2605.06884*, 2026. 2, 14

Ashok Cutkosky, Harsh Mehta, and Francesco Orabona. Optimal stochastic non-smooth non-convex optimization through online-to-non-convex conversion. In *Proceedings of the 40th International Conference on Machine Learning*, volume 202, pp. 6643–6670, 2023. 2, 3, 5, 14

Damek Davis and Dmitriy Drusvyatskiy. When do spectral gradient updates help in deep learning? *arXiv preprint arXiv:2512.04299*, 2025. 14

Yihe Dong and Will Sawin. Muon^p: Muon with fractional spectral powers. *arXiv preprint arXiv:2606.13867*, 2026. 14

Chen Fan, Mark Schmidt, and Christos Thrampoulidis. Implicit bias of spectral descent and Muon on multiclass separable data. In *Advances in Neural Information Processing Systems*, volume 38, pp. 39622–39669, 2025. 14

Dmitrii Feoktistov, Timofey Belinsky, Andrey Veprikov, Amir Zainullin, and Aleksandr Beznosikov. Softsign: Smooth sign in your optimizer for better parameter heterogeneity handling. In *International Conference on Machine Learning*, 2026. 3, 8, 14, 28

GLM-4.5 Team. GLM-4.5: Agentic, reasoning, and coding (ARC) foundation models. *arXiv preprint arXiv:2508.06471*, 2025. 1

A. A. Goldstein. Optimization of Lipschitz continuous functions. *Mathematical Programming*, 13 (1):14–22, 1977. 2

Antoine Gonon, Andreea-Alexandra Muşat, and Nicolas Boumal. Insights on Muon from simple quadratics. *arXiv preprint arXiv:2602.11948*, 2026. 14

Fanfan Ji and Xiaotong Yuan. Derandomized online-to-non-convex conversion for stochastic weakly convex optimization. In *International Conference on Learning Representations*, 2026. 14

Ruichen Jiang, Zakaria Mhammedi, Mehryar Mohri, and Aryan Mokhtari. Adaptive matrix online learning through smoothing with guarantees for nonsmooth nonconvex optimization. In *Proceedings of Thirty Ninth Conference on Learning Theory*, volume 336, pp. 3782–3824, 2026a. 2, 4, 5, 7, 9, 14, 15, 17, 25, 26, 29

Xiaowen Jiang, Andrei Semenov, and Sebastian U. Stich. Enhancing LLM training via spectral clipping. In *International Conference on Machine Learning*, 2026b. 14

Keller Jordan, Yuchen Jin, Vlado Boza, Jiacheng You, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. URL https://kellerjordan.github.io/posts/muon/. 1, 2, 9, 14

Gyu Yeol Kim and Min-hwan Oh. Convergence of Muon with Newton–Schulz. In *International Conference on Learning Representations*, 2026. 2, 9, 14, 15

Kimi Team. Kimi K2: Open agentic intelligence. *arXiv preprint arXiv:2507.20534*, 2025. 1

Guy Kornowski and Ohad Shamir. Oracle complexity in nonsmooth nonconvex optimization. *Journal of Machine Learning Research*, 23(314):1–44, 2022. 4

Dmitry Kovalev. Understanding gradient orthogonalization for deep learning via non-Euclidean trust-region optimization. *arXiv preprint arXiv:2503.12645*, 2025. 1, 9, 14, 15

Zdislav Kovarik. Some iterative methods for improving orthonormality. *SIAM Journal on Numerical Analysis*, 7(3):386–389, 1970. 1, 4, 14

Adrian S. Lewis. The convex analysis of unitarily invariant matrix functions. *Journal of Convex Analysis*, 2(1–2):173–183, 1995. 16

Jiaxiang Li and Mingyi Hong. A note on the convergence of Muon. *arXiv preprint arXiv:2502.02900*, 2025. 1, 14

Xianliang Li, Zihan Zhang, Weiyang Liu, and Han Bao. Denoise first, orthogonalize later: Understanding momentum in Muon via spectral filtering. *arXiv preprint arXiv:2606.03899*, 2026. 14

Jingyuan Liu, Jianlin Su, Xingcheng Yao, Zhejun Jiang, Guokun Lai, Yulun Du, Yidao Qin, Weixin Xu, Enzhe Lu, Junjie Yan, et al. Muon is scalable for LLM training. *arXiv preprint arXiv:2502.16982*, 2025. 1

Langqi Liu, Yibo Wang, and Lijun Zhang. High-probability bound for non-smooth non-convex stochastic optimization with heavy tails. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 32122–32138, 2024. 14

Zijian Liu. Online convex optimization with heavy tails: Old algorithms, new regrets, and applications. In *Proceedings of The 37th International Conference on Algorithmic Learning Theory*, volume 313, pp. 1–47, 2026. 14

Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *International Conference on Learning Representations*, 2019. 1

Jianhao Ma, Yu Huang, Yuejie Chi, and Yuxin Chen. Preconditioning benefits of spectral orthogonalization in Muon. *arXiv preprint arXiv:2601.13474*, 2026. 14

H. Brendan McMahan. A survey of algorithms and analysis for adaptive online learning. *Journal of Machine Learning Research*, 18(90):1–50, 2017. 7

Aratrika Mustafi, Soumya Mukherjee, and Bharath K. Sriperumbudur. Move on Muon: A Hamiltonian probability gradient flow perspective of Muon optimizer. *arXiv preprint arXiv:2605.23871*, 2026. 3, 8, 14, 28

Tetiana Parshakova, Ahmed Khaled, Michael Crawshaw, Guillaume Garrigos, and Robert M. Gower. Muon does not converge on convex Lipschitz functions. *arXiv preprint arXiv:2605.08980*, 2026. 2, 14, 26

Francisco Patitucci, Ruichen Jiang, and Aryan Mokhtari. Improving online-to-nonconvex conversion for smooth optimization via double optimism. In *International Conference on Learning Representations*, 2026. 14

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267, pp. 49069–49104, 2025. 1, 14

Xianbiao Qi, Marco Chen, Jiaquan Ye, Yelin He, and Rong Xiao. Delving into Muon and beyond: Deep analysis and extensions. In *International Conference on Machine Learning*, 2026. 14

Artem Riabinin, Egor Shulgin, Kaja Gruntkowska, and Peter Richtárik. From Muon to Gluon: Bridging theory and practice of LMO-based optimizers for LLMs. In *International Conference on Machine Learning*, 2026. 1, 14

Maria-Eleni Sfyraki and Jun-Kun Wang. Lions and Muons: Optimization via stochastic Frank–Wolfe under heavy-tailed noise. In *International Conference on Machine Learning*, 2026. 1, 14

Shai Shalev-Shwartz. Online learning and online convex optimization. *Foundations and Trends in Machine Learning*, 4(2):107–194, 2012. 7

Wei Shen, Ruichuan Huang, Minhui Huang, Cong Shen, and Jiawei Zhang. On the convergence analysis of Muon. *Transactions on Machine Learning Research*, 2026. 1, 9, 14, 15

Egor Shulgin, Sultan AlRashed, Peter Richtárik, and Francesco Orabona. Beyond the ideal: Analyzing the inexact Muon update. In *International Conference on Artificial Intelligence and Statistics*, volume 300, pp. 5167–5175, 2026a. 2, 14, 15

Egor Shulgin, Sam Laing, Antonio Orvieto, and Peter Richtárik. A quadratic lens on Muon: Orthogonalization, invariance, and implicit preconditioning. In *HiLD 2026: 4th Workshop on High-dimensional Learning Dynamics*, 2026b. 14

Taira Tsuchiya, Shinji Ito, and Junya Honda. Stability-penalty-adaptive follow-the-regularized-leader: Sparsity, game-dependency, and best-of-both-worlds. In *Advances in Neural Information Processing Systems*, volume 36, pp. 47406–47437, 2023. 7

Kaiyue Wen, David Hall, Tengyu Ma, and Percy Liang. Fantastic pretraining optimizers and where to find them. In *International Conference on Learning Representations*, 2026. 1

Fangzhou Wu, Rikhav Shah, Sandeep Silwal, and Qiuyi Zhang. DynMuon: A dynamic spectral shaping view of Muon. In *HiLD 2026: 4th Workshop on High-dimensional Learning Dynamics*, 2026. 14

Jingzhao Zhang, Hongzhou Lin, Stefanie Jegelka, Suvrit Sra, and Ali Jadbabaie. Complexity of finding stationary points of nonconvex nonsmooth functions. In *Proceedings of the 37th International Conference on Machine Learning*, volume 119, pp. 11173–11182, 2020. 4

Qinzi Zhang and Ashok Cutkosky. Random scaling and momentum for non-smooth non-convex optimization. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235, pp. 58780–58799, 2024. 14

Julian Zimmert and Yevgeny Seldin. Tsallis-INF: An optimal algorithm for stochastic and adversarial bandits. *Journal of Machine Learning Research*, 22(28):1–49, 2021. 7

A Relation to prior work

This section supplements the discussion in Section 1 with further related work and compare our convergence guarantees with those of prior work.

A.1 Additional related work

Muon optimizer. Muon was proposed by Jordan et al. (2024) as an optimizer for the matrix-shaped hidden layers of neural networks. Spectral update directions were used earlier in preconditioned spectral descent (Carlson et al., 2015), and the steepest-descent and duality interpretations of Muon are developed by Bernstein & Newhouse (2024; 2025). Convergence guarantees for Muon with the exact-polar update or an exact linear minimization oracle have been established under smoothness and generalized smoothness (Li & Hong, 2025; Kovalev, 2025; Pethick et al., 2025; Shen et al., 2026; Chen et al., 2026; Riabinin et al., 2026; Sfyraki & Wang, 2026). Its implicit bias, the denoising role of momentum, and the regimes in which spectral updates outperform Euclidean ones have also been studied (Fan et al., 2025; Li et al., 2026; Davis & Drusvyatskiy, 2025; Braun et al., 2026). For Muon with the exact-polar update, Parshakova et al. (2026) construct convex Lipschitz objectives on which the iterates fail to converge, which sharpens the question of what changes under finite Newton–Schulz on nonsmooth problems. Our work takes up this question, analyzing the update that retains both momentum and finite Newton–Schulz on nonsmooth nonconvex objectives.

Newton–Schulz and general spectral maps. Iterative orthogonalization by matrix polynomials goes back to Kovarik (1970) and Björck & Bowie (1971), and the iterations used in Muon range from empirically tuned polynomials (Jordan et al., 2024) to optimal polynomial schemes designed for this purpose (Amsel et al., 2026). Analyses of Muon that include the finite iteration treat the effect of the finite iteration as an error: for smooth nonconvex objectives (Kim & Oh, 2026), through inexact linear minimization oracles (Shulgin et al., 2026a), and for Nesterov momentum under heavy-tailed noise (Choudhury et al., 2026). A few recent works identify benefits of orthogonalization and of its finite approximations in structured settings: on matrix quadratics, finite Newton–Schulz damps directions associated with small singular values near rank deficiency (Shulgin et al., 2026b), inexact polar updates can improve reachability on simple strongly convex quadratics (Gonon et al., 2026), and spectral orthogonalization acts as a preconditioner in matrix factorization and in-context learning models (Ma et al., 2026). These benefits are confined to quadratic or otherwise structured objectives. In contrast, we show that finite Newton–Schulz enables stationarity guarantees for general nonsmooth nonconvex objectives and characterize how the depth trades off polar approximation against the stability of the online updates. Beyond the exact polar factor, recent work has considered more general spectral maps (Qi et al., 2026; Dong & Sawin, 2026; Jiang et al., 2026b; Wu et al., 2026), including smooth relaxations of spectral normalization and orthogonalization (Feoktistov et al., 2026; Mustafi et al., 2026). Our theory in Section 4 gives sufficient conditions under which such maps enjoy stationarity guarantees for nonsmooth nonconvex objectives.

Online-to-nonconvex conversion. The online-to-nonconvex conversion (O2NC) framework was introduced by Cutkosky et al. (2023), who convert online regret guarantees into optimal stationarity guarantees for stochastic nonsmooth nonconvex optimization. The framework can also exploit additional properties of an objective: it attains the optimal first-order complexity for deterministic smooth objectives (Cutkosky et al., 2023), and the best-known complexity when both the gradient and Hessian are Lipschitz (Patitucci et al., 2026). The discounted O2NC framework, which is central to our analysis of Muon's momentum, grew out of O2NC variants based on random scaling and model exponential moving averages (Zhang & Cutkosky, 2024; Ahn & Cutkosky, 2024). The broader online-learning perspective has also been used to understand practical optimizers such as Adam (Ahn et al., 2024) and schedule-free SGD (Ahn et al., 2025), and to handle heavy-tailed gradient noise (Liu et al., 2024; Liu, 2026) and weakly convex optimization (Ji & Yuan, 2026). Most closely related to our matrix setting, Jiang et al. (2026a) construct a family of smoothed potentials for the nuclear norm and use the resulting adaptive matrix online learners to derive Pion and Leon, which have a convergence guarantee in nonsmooth optimization. In contrast, we analyze the finite Newton–Schulz transformation underlying Muon and show that finite depth smooths the exact-polar update by quantifying how the depth trades off polar approximation against the stability of successive updates from the online-to-nonconvex conversion perspective.

A.2 Comparison with prior convergence guarantees

Here we compare the guarantees of this work with those of prior work on Muon and related methods.

Table 1: Gradient-oracle complexity for nonsmooth nonconvex objectives under the (ρ, ε)-stationarity criterion of Definition 2.2. Here $r = m$, $\nu = (n/(n-r-1))^{1/4}$, and NS abbreviates Newton–Schulz. The bounds for Pion and Leon are stated for sufficiently small ε.

Reference	Smoothing	Polar computation	Gradient-oracle complexity
Exact-polar Muon	None	Exact polar	Not covered by discounted O2NC
Pion (Jiang et al., 2026a)	Gaussian perturbation	Exact polar	$O\left(\max\left\{\frac{\nu^2\|Q\|_*^2\Delta_{\mathcal{L}}}{\rho\varepsilon^3}, \frac{\nu^2\|Q\|_*^2}{\varepsilon^2}, \frac{\nu r G}{\varepsilon}\right\}\right)$
Leon (Jiang et al., 2026a)	Hyperbolic smoothing	Exact polar	$O\left(\max\left\{\frac{\|Q\|_*^2\Delta_{\mathcal{L}}}{\rho\varepsilon^3}, \frac{\|Q\|_*^2}{\varepsilon^2}, \frac{rG}{\varepsilon}\right\}\right)$
This work (Theorem 5.4)	Finite NS	$q = O(\log(1/\varepsilon))$ NS steps	$O\left(\max\left\{\frac{r(\Gamma+\sigma)^2\Delta_{\mathcal{L}}}{\rho\varepsilon^3}, \frac{r(\Gamma+\sigma)^2}{\varepsilon^2}\right\}\right)$

Table 2: Gradient-oracle complexity for finding ε-stationary points of smooth nonconvex objectives. The results are specialized to smoothness in the operator norm and stationarity in the nuclear norm. The "Deterministic" and "Stochastic" columns give the complexity with exact gradients and with unbiased stochastic gradients of variance at most σ^2, respectively, and the bounds for the stochastic setting assume one sample per oracle query. For Shulgin et al. (2026a), $\delta < 1$ is a uniform additive error of the inexact linear minimization oracle. For Kim & Oh (2026), $\chi_q = (1-\varepsilon_q)^{-1}$, where ε_q is the uniform operator-norm error of the q-step Newton–Schulz output relative to the polar factor, and under their uniform condition on the initial residual, χ_q converges to one doubly exponentially in q. NS abbreviates Newton–Schulz.

Reference	Deterministic	Stochastic	Polar computation
Kovalev (2025)	$O\left(\frac{L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon^2}\right)$	$O\left(\max\left\{\frac{L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon^2}, \frac{\sqrt{r}\sigma}{\varepsilon}, \frac{r^{3/2}\sigma^3}{\varepsilon^3}, \frac{r\sigma^2 L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon^4}\right\}\right)$	Exact polar
Shen et al. (2026)	$O\left(\frac{L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon^2}\right)$	$O\left(\max\left\{\frac{L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon^2}, \frac{r^2\sigma^4}{L_{\mathrm{op}}\Delta_{\mathcal{L}}\varepsilon^2}, \frac{r\sigma^2 L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon^4}\right\}\right)$	Exact polar
Shulgin et al. (2026a)	$O\left(\frac{(1+\delta)^2 L_{\mathrm{op}}\Delta_{\mathcal{L}}}{(1-\delta)^2\varepsilon^2}\right)$	$O\left(\frac{(1+\delta) r\sigma^2 L_{\mathrm{op}}\Delta_{\mathcal{L}}}{(1-\delta)^4\varepsilon^4}\right)$	Inexact LMO
Kim & Oh (2026)	$O\left(\frac{\chi_q^2 L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon^2}\right)$	$O\left(\max\left\{\frac{\chi_q^2 L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon^2}, \frac{\chi_q^2 r^2\sigma^2}{\varepsilon^2}, \frac{\chi_q^4 r\sigma^2 L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon^4}\right\}\right)$	Finite NS
This work (Theorem 5.7)	$O\left(\frac{L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon^2}\right)$	$O\left(\max\left\{\frac{L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon^2}, \frac{r\sigma^2}{\varepsilon^2}, \frac{r\sigma^2 L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon^4}\right\}\right)$	Finite NS

In Table 1, we summarize the gradient-oracle complexity and the treatment of the polar computation for nonsmooth nonconvex objectives. The leading $\rho^{-1}\varepsilon^{-3}$ term in the bound of this work matches the bounds for Pion and Leon. The bound for Pion assumes an exact expectation of polar factors, and the Monte Carlo implementation of Pion uses $k = T$ samples per round to preserve this bound. For exact-polar Muon, the discounted O2NC framework alone does not provide a general stationarity guarantee for nonsmooth objectives, as discussed in Appendix F. For Pion and Leon, the matrix $Q \succeq 0$ and the constant $G > 0$ satisfy $\mathbb{E}[G_t G_t^\top] \preceq Q^2$ and $\|G_t\|_{\mathrm{op}} \le G$ almost surely. The quantities $\|Q\|_*^2$ and $r(\Gamma+\sigma)^2$ reflect different oracle assumptions and are not directly comparable.

In Table 2, we summarize the gradient-oracle complexity and the treatment of the polar computation for smooth nonconvex objectives. The leading terms of this work coincide with those established for exact-polar Muon (Kovalev, 2025; Shen et al., 2026). Compared with Kim & Oh (2026), the bounds of this work involve no factor χ_q, and the middle term improves from $r^2\sigma^2/\varepsilon^2$ to $r\sigma^2/\varepsilon^2$, though the algorithms under analysis are slightly different. The output criteria follow the respective references: an expected minimum for Kovalev (2025), an average expected gradient norm for the other prior works, and the randomized output in Theorem 5.7 for this work. The bound for Shulgin et al. (2026a) in the stochastic setting specializes their norm-compatibility factor to the nuclear norm, whose square is r. Their guarantee for the inexact LMO requires a uniform error $\delta < 1$, which is

not automatically available for a fixed-depth polar approximation near rank deficiency. This work additionally assumes an almost-sure operator-norm bound and a Frobenius second-moment bound for the stochastic gradients.

B AUXILIARY LEMMAS ON SINGULAR-VALUE FUNCTIONS

For ease of reference, this appendix collects the standard singular-value lemmas used in this paper. Some of them are stated in the special forms in which we use them.

The following three lemmas specialize results of Lewis (1995) to separable functions of the singular values.

Lemma B.1 (Convexity of separable singular-value functions)**.** *Let $\phi\colon [0,\infty) \to \mathbb{R}$ be convex and nondecreasing. Then $S \mapsto \sum_{i=1}^{r} \phi(\sigma_i(S))$ is convex on $\mathbb{R}^{m\times n}$.*

Proof. Define $f\colon \mathbb{R}^r \to \mathbb{R}$ by $f(\gamma) := \sum_{i=1}^{r} \phi(|\gamma_i|)$. Since ϕ is convex and nondecreasing, the map $t \mapsto \phi(|t|)$ is convex. Thus, f is convex, lower semicontinuous, and absolutely symmetric. Applying Lewis (1995, Corollary 2.6) to f shows that the function $S \mapsto (f \circ \sigma)(S) = \sum_{i=1}^{r} \phi(\sigma_i(S))$ is convex, which completes the proof. □

The next lemma gives the corresponding gradient formula.

Lemma B.2 (Singular-value gradient formula)**.** *Let $\phi\colon [0,\infty) \to \mathbb{R}$ be convex and differentiable with $\phi'(0) = 0$, and define $F(S) := \sum_{i=1}^{r} \phi(\sigma_i(S))$. Then F is differentiable on all of $\mathbb{R}^{m\times n}$, and for any thin singular value decomposition $S = U\,\mathrm{diag}(\sigma_1(S),\dots,\sigma_r(S))V^\top$,*

$$\nabla F(S) = U\,\mathrm{diag}\left(\phi'(\sigma_1(S)),\dots,\phi'(\sigma_r(S))\right)V^\top.$$

Proof. Define $f\colon \mathbb{R}^r \to \mathbb{R}$ by $f(\gamma) := \sum_{i=1}^{r} \phi(|\gamma_i|)$. Since ϕ is convex and $\phi'(0) = 0$, it is nondecreasing, and hence f is convex and absolutely symmetric. The condition $\phi'(0) = 0$ also makes $t \mapsto \phi(|t|)$ differentiable at 0, so f is differentiable on $\mathbb{R}^r$ and $\nabla f(\sigma(S)) = (\phi'(\sigma_1(S)),\dots,\phi'(\sigma_r(S)))$. Applying Lewis (1995, Theorem 3.1) to f shows that $F = f \circ \sigma$ is differentiable and has the above gradient, which completes the proof. □

The following lemma gives the Fenchel conjugate used in the FTRL interpretation.

Lemma B.3 (Fenchel conjugate formula for singular-value functions)**.** *Let $\phi\colon [0,\infty) \to (-\infty,+\infty]$ be proper and convex, and write $\phi^*(a) := \sup_{x\ge 0}\{ax - \phi(x)\}$. Then the Fenchel conjugate of $S \mapsto \sum_{i=1}^{r} \phi(\sigma_i(S))$ is $W \mapsto \sum_{i=1}^{r} \phi^*(\sigma_i(W))$.*

Proof. Define $f\colon \mathbb{R}^r \to (-\infty,+\infty]$ and $F\colon \mathbb{R}^{m\times n} \to (-\infty,+\infty]$ by $f(\gamma) := \sum_{i=1}^{r} \phi(|\gamma_i|)$ and $F(S) := \sum_{i=1}^{r} \phi(\sigma_i(S))$. The function f is invariant under sign changes and coordinate permutations, and $F = f \circ \sigma$. Hence, Lewis (1995, Theorem 2.4) gives $F^* = f^* \circ \sigma$. For every $\mu \in \mathbb{R}^r$,

$$\begin{aligned} f^*(\mu) &= \sup_{\gamma\in\mathbb{R}^r} \sum_{i=1}^{r} \{\mu_i\gamma_i - \phi(|\gamma_i|)\} = \sum_{i=1}^{r} \sup_{t\in\mathbb{R}} \{\mu_i t - \phi(|t|)\} \\ &= \sum_{i=1}^{r} \sup_{x\ge 0} \{|\mu_i| x - \phi(x)\} = \sum_{i=1}^{r} \phi^*(|\mu_i|). \end{aligned}$$

Since $\sigma_i(W) \ge 0$, we obtain $F^*(W) = f^*(\sigma(W)) = \sum_{i=1}^{r} \phi^*(\sigma_i(W))$, which completes the proof. □

The following lemma for singular-value maps is used to control the Lipschitz constant of the gradient of the spectral potential.

Lemma B.4 (Lipschitz bound for singular-value maps, Andersson et al., 2016, Proposition 4.1 and Theorem 4.2)**.** *Let $h\colon [0,\infty) \to \mathbb{R}$ satisfy $h(0) = 0$ and be $\mathrm{Lip}(h)$-Lipschitz. Then the singular-value map $\mathcal{H}_h$ of (1) is well defined and, for all $X, Y \in \mathbb{R}^{m\times n}$,*

$$\|\mathcal{H}_h(X) - \mathcal{H}_h(Y)\|_{\mathrm{F}} \le \mathrm{Lip}(h)\|X - Y\|_{\mathrm{F}}.$$

C DISCOUNTED O2NC PROTOCOL AND GUARANTEE

This appendix gives the full protocol for the conversion in Section 3 and proves Lemma 3.1. Algorithm 2 formalizes the overview in Section 5.1 of Jiang et al. (2026a), using the protocol in their Appendix G and the output rule in their Proposition 25.

Algorithm 2: Discounted O2NC with a generic online learner

Require : horizon T, initial point W_0, discount $\beta \in (0,1)$, radius $D > 0$, online learner $\mathcal{A}$

1 Initialize $\mathcal{A}$;
2 **for** $t = 1, 2, \dots T$ **do**
3 Receive X_t satisfying $\|X_t\|_{\mathrm{op}} \le D$ from the online learner $\mathcal{A}$;
4 Set $W_t \leftarrow W_{t-1} + X_t$;
5 Set $\widetilde{W}_t \leftarrow W_{t-1} + u_t X_t$ for $u_t \sim \mathrm{Unif}([0,1])$;
6 Query the stochastic gradient oracle at $\widetilde{W}_t$ to obtain G_t satisfying Assumption 2.1 with $W = \widetilde{W}_t$;
7 Construct and feed the linear loss $\ell_t^{[\beta]}(X) \coloneqq \beta^{-t}\langle G_t, X\rangle$ to $\mathcal{A}$;
8 Form $\bar{W}_t$ and sample τ as in (6), then return $\bar{W}_\tau$;

The identity $\mathcal{L}(W_t) - \mathcal{L}(W_{t-1}) = \mathbb{E}[\langle G_t, X_t\rangle \mid W_{t-1}, X_t]$ established in Section 3 is what makes the linear loss constructed in Line 7 track the decrease of the objective.

Proof. Let $\mathcal{F}_0$ be the σ-algebra generated by the learner's initialization and its internal randomness, and, for $t \ge 1$, let $\mathcal{F}_t$ be the σ-algebra generated by $\mathcal{F}_0$ and all randomness through round t. By construction, W_t and X_t are $\mathcal{F}_{t-1}$-measurable. Let $E_t \coloneqq G_t - \nabla\mathcal{L}(\widetilde{W}_t)$ denote the oracle noise. Specializing the proof of Proposition 25 of Jiang et al. (2026a) to the operator and nuclear norms and using $\mathcal{L}(W_T) \ge \inf_W \mathcal{L}(W)$ gives

$$\mathbb{E}\Big[\|\nabla\mathcal{L}(\bar{W}_\tau)\|_*^{[\rho]}\Big] \le \frac{4\Delta_{\mathcal{L}}}{(1-\beta)\rho T} + \frac{1}{T}\mathbb{E}\left[\mathrm{Reg}_T^{[\beta]}(1) + (1-\beta)\sum_{t=1}^{T-1}\mathrm{Reg}_t^{[\beta]}(1)\right] + \frac{1}{T}\mathbb{E}\left[\left\|\sum_{t=1}^{T}\beta^{T-t}E_t\right\|_*\right] + \frac{1-\beta}{T}\sum_{t=1}^{T-1}\mathbb{E}\left[\left\|\sum_{s=1}^{t}\beta^{t-s}E_s\right\|_*\right].$$

Each E_t is $\mathcal{F}_t$-measurable. Conditional on $\mathcal{F}_{t-1}$ and u_t, Assumption 2.1 gives $\mathbb{E}[E_t \mid \mathcal{F}_{t-1}, u_t] = 0$ and $\mathbb{E}[\|E_t\|_{\mathrm{F}}^2 \mid \mathcal{F}_{t-1}, u_t] \le \sigma^2$. The tower property therefore gives $\mathbb{E}[E_t \mid \mathcal{F}_{t-1}] = 0$ and $\mathbb{E}[\|E_t\|_{\mathrm{F}}^2] \le \sigma^2$, so $(E_t)_{t\ge1}$ is a matrix-valued martingale difference sequence. Consequently, the conditional orthogonality of the martingale differences gives, for every $t \ge 1$,

$$\mathbb{E}\left[\left\|\sum_{s=1}^{t}\beta^{t-s}E_s\right\|_{\mathrm{F}}^2\right] = \sum_{s=1}^{t}\beta^{2(t-s)}\mathbb{E}\big[\|E_s\|_{\mathrm{F}}^2\big] \le \sigma^2\frac{1-\beta^{2t}}{1-\beta^2} \le \frac{\sigma^2}{1-\beta^2}.$$

The inequality $\|\cdot\|_* \le \sqrt{r}\,\|\cdot\|_{\mathrm{F}}$ and Jensen's inequality give, for every $t \ge 1$,

$$\mathbb{E}\left[\left\|\sum_{s=1}^{t}\beta^{t-s}E_s\right\|_*\right] \le \sqrt{r}\left(\mathbb{E}\left[\left\|\sum_{s=1}^{t}\beta^{t-s}E_s\right\|_{\mathrm{F}}^2\right]\right)^{1/2} \le \frac{\sqrt{r}\,\sigma}{\sqrt{1-\beta^2}}.$$

Therefore,

$$\frac{1}{T}\mathbb{E}\left[\left\|\sum_{t=1}^{T}\beta^{T-t}E_t\right\|_*\right] + \frac{1-\beta}{T}\sum_{t=1}^{T-1}\mathbb{E}\left[\left\|\sum_{s=1}^{t}\beta^{t-s}E_s\right\|_*\right] \le \left(\frac{1}{T} + \frac{(1-\beta)(T-1)}{T}\right)\frac{\sqrt{r}\,\sigma}{\sqrt{1-\beta^2}} = \left(1-\beta+\frac{\beta}{T}\right)\frac{\sqrt{r}\,\sigma}{\sqrt{1-\beta^2}}.$$

Combining the preceding bounds completes the proof. □

D DEFERRED PROOFS FROM SECTION 4

This appendix provides the proofs of Theorem 4.2 and Lemmas 4.3–4.7.

Proof of Lemma 4.3. Write $\widetilde{\Phi}_{h,C}(M) = \sum_{i=1}^{r} \phi(\sigma_i(M))$ with $\phi(x) := C\phi_h(x/C)$. Since h is continuous, nonnegative, and nondecreasing with $h(0) = 0$, the function ϕ is convex, nondecreasing, and differentiable, and its derivative satisfies $\phi'(x) = h(x/C)$ and $\phi'(0) = 0$. Hence, by the convexity of separable singular-value functions (Lemma B.1), the function $\widetilde{\Phi}_{h,C}$ is convex on $\mathbb{R}^{m\times n}$, and by the standard singular-value gradient formula (Lemma B.2), its gradient is $\nabla\widetilde{\Phi}_{h,C}(M) = U\,\mathrm{diag}\left(h(\sigma_i(M)/C)\right)V^\top = \mathcal{H}_h(M/C)$, which completes the proof. □

Proof of Lemma 4.4. By the definition of the discounted regret in (7) and $M_t = (1-\beta)\sum_{s=1}^{t}\beta^{t-s}G_s$, we have

$$\mathrm{Reg}_t^{[\beta]}(D) = \sum_{s=1}^{t}\beta^{t-s}\langle G_s, X_s\rangle + \max_{\|X\|_{\mathrm{op}}\le D}\left\langle -\frac{M_t}{1-\beta}, X\right\rangle = \sum_{s=1}^{t}\beta^{t-s}\langle G_s, X_s\rangle + \frac{D}{1-\beta}\|M_t\|_*,$$

where we used the fact that the maximum in the operator–nuclear duality is attained at $X = -D\,\mathrm{polar}(M_t)$. Substituting the updates $X_s = -D\nabla\widetilde{\Phi}_{h,C_s}(M_{s-1})$ from Lemma 4.3 into the last equality and multiplying by $1-\beta$ gives

$$(1-\beta)\,\mathrm{Reg}_t^{[\beta]}(D) = -D\sum_{s=1}^{t}(1-\beta)\beta^{t-s}\langle G_s, \nabla\widetilde{\Phi}_{h,C_s}(M_{s-1})\rangle + D\|M_t\|_*. \tag{16}$$

We evaluate the first term in (16). By using $(1-\beta)G_s = M_s - M_{s-1} + (1-\beta)M_{s-1}$ and the definition of the Bregman divergence, we have

$$\begin{aligned}
&(1-\beta)\langle G_s, \nabla\widetilde{\Phi}_{h,C_s}(M_{s-1})\rangle \\
&= \langle M_s - M_{s-1}, \nabla\widetilde{\Phi}_{h,C_s}(M_{s-1})\rangle + (1-\beta)\langle M_{s-1}, \nabla\widetilde{\Phi}_{h,C_s}(M_{s-1})\rangle \\
&= \widetilde{\Phi}_{h,C_s}(M_s) - \widetilde{\Phi}_{h,C_s}(M_{s-1}) - B_{\widetilde{\Phi}_{h,C_s}}(M_s\|M_{s-1}) + (1-\beta)\langle M_{s-1}, \nabla\widetilde{\Phi}_{h,C_s}(M_{s-1})\rangle.
\end{aligned} \tag{17}$$

Now, by $M_0 = 0$ and $\widetilde{\Phi}_{h,C_1}(0) = 0$, we have

$$\sum_{s=1}^{t}\beta^{t-s}\left(\widetilde{\Phi}_{h,C_s}(M_s) - \widetilde{\Phi}_{h,C_s}(M_{s-1})\right) = \widetilde{\Phi}_{h,C_t}(M_t) + \sum_{s=1}^{t-1}\beta^{t-s-1}\left(\beta\widetilde{\Phi}_{h,C_s}(M_s) - \widetilde{\Phi}_{h,C_{s+1}}(M_s)\right). \tag{18}$$

We also have

$$\sum_{s=1}^{t}\beta^{t-s}\langle M_{s-1}, \nabla\widetilde{\Phi}_{h,C_s}(M_{s-1})\rangle = \sum_{s=1}^{t-1}\beta^{t-s-1}\langle M_s, \nabla\widetilde{\Phi}_{h,C_{s+1}}(M_s)\rangle, \tag{19}$$

where we used $M_0 = 0$. Substituting (17)–(19) into the right-hand side of (16) and dividing by $1-\beta$ completes the proof. □

Proof of Lemma 4.5. By (9) and the definition of $\Delta(h)$ in Assumption 4.1,

$$\Phi(M) - \widetilde{\Phi}_{h,C}(M) = C\sum_{i=1}^{r}\left(\frac{\sigma_i(M)}{C} - \phi_h\left(\frac{\sigma_i(M)}{C}\right)\right) = C\sum_{i=1}^{r}\int_0^{\sigma_i(M)/C}(1-h(u))\,\mathrm{d}u \le rC\Delta(h).$$

This completes the proof. □

Proof of Lemma 4.6. By the Lipschitzness of singular-value maps (Lemma B.4) applied to h satisfying Assumption 4.1, for any $M, M' \in \mathbb{R}^{m\times n}$,

$$\|\mathcal{H}_h(M/C) - \mathcal{H}_h(M'/C)\|_{\mathrm{F}} \le \frac{\mathrm{Lip}(h)}{C}\|M - M'\|_{\mathrm{F}}, \tag{20}$$

and thus $\nabla\widetilde{\Phi}_{h,C} = \mathcal{H}_h(\cdot/C)$ is $(\mathrm{Lip}(h)/C)$-Lipschitz. By applying the fundamental theorem of calculus to $u \mapsto \widetilde{\Phi}_{h,C}(M + u(M' - M))$ on $[0,1]$ and then using the Cauchy–Schwarz inequality and (20), for any $M, M' \in \mathbb{R}^{m\times n}$ we have

$$\begin{aligned} B_{\widetilde{\Phi}_{h,C}}(M'\|M) &= \int_0^1 \left\langle \nabla\widetilde{\Phi}_{h,C}(M + u(M' - M)) - \nabla\widetilde{\Phi}_{h,C}(M),\, M' - M \right\rangle \mathrm{d}u \\ &\le \int_0^1 \|\nabla\widetilde{\Phi}_{h,C}(M + u(M' - M)) - \nabla\widetilde{\Phi}_{h,C}(M)\|_{\mathrm{F}}\|M' - M\|_{\mathrm{F}}\,\mathrm{d}u \\ &\le \frac{\mathrm{Lip}(h)}{2C}\|M' - M\|_{\mathrm{F}}^2, \end{aligned}$$

which completes the proof. □

Proof of Lemma 4.7. The left-hand side of the desired inequality decomposes over the singular values of M, so it suffices to show, for each singular value λ of M, that

$$C'\phi_h(\lambda/C') - \beta C\phi_h(\lambda/C) - (1-\beta)\lambda\, h(\lambda/C') \le 0.$$

Write $x := \lambda/C'$ and $\theta := \beta C/C' \in (0,1]$, so that $\lambda/C = \beta x/\theta$. Then, dividing the above inequality by C', we can see that it suffices to show that

$$F(\theta) := \phi_h(x) - \theta\phi_h(\beta x/\theta) - (1-\beta)xh(x) \le 0.$$

This directly follows from the convexity of ϕ_h and $\phi_h(0) = 0$. Indeed, these two properties imply $\phi_h(\beta x) = \phi_h(\theta\cdot(\beta x/\theta) + (1-\theta)0) \le \theta\phi_h(\beta x/\theta)$. Again by the convexity of ϕ_h and $\phi_h' = h$, we have $\phi_h(\beta x) \ge \phi_h(x) + (\beta x - x)\phi_h'(x) = \phi_h(x) + (\beta x - x)h(x)$. Combining these two inequalities, we obtain

$$F(\theta) = \phi_h(x) - \theta\phi_h(\beta x/\theta) - (1-\beta)xh(x) \le \phi_h(x) - \phi_h(\beta x) - (1-\beta)xh(x) \le 0,$$

which completes the proof. □

Proof of Theorem 4.2. We upper bound the right-hand side of (12) in Lemma 4.4. Since the first two terms of (12) is at most $DrC_t\Delta(h)/(1-\beta)$ by Lemma 4.5 and $V_s \le 0$ by Lemma 4.7, we obtain (10). For the second claim, let $C_s = C$ for all $s \in [t]$ and take expectations under Assumption 2.1. Bounding each Bregman divergence in (10) by Lemma 4.6, we have

$$\mathbb{E}\left[\sum_{s=1}^t \beta^{t-s} B_{\widetilde{\Phi}_{h,C_s}}(M_s\|M_{s-1})\right] \le \frac{\mathrm{Lip}(h)}{2C}\,\mathbb{E}\left[\sum_{s=1}^t \beta^{t-s}\|M_s - M_{s-1}\|_{\mathrm{F}}^2\right] \le \frac{2(1-\beta)\Gamma^2\,\mathrm{Lip}(h)}{C}.$$

Here, the last inequality follows from $M_s - M_{s-1} = (1-\beta)(G_s - M_{s-1})$, the bound $\mathbb{E}[\|G_s - M_{s-1}\|_{\mathrm{F}}^2] \le 4\Gamma^2$, which follows from Assumption 2.1 and Jensen's inequality, and $(1-\beta)\sum_{s=1}^t \beta^{t-s} \le 1$. Substituting this bound into (10) gives (11), which completes the proof. □

E Deferred proofs from Section 5

This appendix provides the proofs deferred from Section 5.

E.1 Newton–Schulz iterations at a general Taylor order

The analysis in Section 5 specializes to the degree-five Newton–Schulz polynomial used in Algorithm 1. We begin by defining the Newton–Schulz iteration obtained by truncating the Taylor expansion at an arbitrary order $\kappa \ge 1$ and then show that the same penalty–stability tradeoff holds for every such order. For an integer $\kappa \ge 1$, recall from (3) in Section 2 the polynomial p_κ and coefficients c_s, and define

$$f_\kappa(x) := xp_\kappa(x^2) = x\sum_{s=0}^{\kappa} c_s(1-x^2)^s, \qquad x \in [0,1].$$

The corresponding Newton–Schulz iteration is $x_{j+1} = f_\kappa(x_j)$, where f_κ has degree $2\kappa+1$. The slope of f_κ at the origin is

$$A_\kappa := f_\kappa'(0) = (2\kappa+1)c_\kappa = \frac{(2\kappa+1)(2\kappa)!}{4^\kappa(\kappa!)^2}.$$

For an integer $q \geq 0$, define

$$h_{\kappa,q} := f_\kappa^{\circ q} \quad \text{on } [0,1], \qquad h_{\kappa,q}(x) := 1 \quad \text{for } x \geq 1.$$

The map considered in Section 5 is recovered by setting $\kappa = 2$, in which case

$$f_2(x) = \frac{15}{8}x - \frac{5}{4}x^3 + \frac{3}{8}x^5, \qquad A_2 = \frac{15}{8}, \qquad h_{2,q} = h_q.$$

The following lemma generalizes Lemma 5.2 to every $\kappa \geq 1$. Thus, for every Taylor order κ, the penalty term in (11) decreases at rate A_κ^{-q}, while the stability term grows at rate A_κ^q. Consequently, the argument in Section 5 applies with A_κ in place of A, and our analysis extends to the Newton–Schulz polynomials obtained by truncating the Taylor expansion at any finite order.

The following lemma establishes the bounds for every $\kappa \geq 1$ and yields Lemma 5.2 when $\kappa = 2$.

Lemma E.1 (Extension of Lemma 5.2)**.** *For every integer $\kappa \geq 1$ and $q \geq 0$,*

$$\mathrm{Lip}(h_{\kappa,q}) = A_\kappa^q, \qquad \frac{1}{2A_\kappa^q} \leq \Delta(h_{\kappa,q}) \leq \frac{1}{A_\kappa^q + 1} \leq \frac{1}{A_\kappa^q}.$$

Proof. Fix an integer $\kappa \geq 1$ and write $P_\kappa(u) := \sum_{s=0}^{\kappa} c_s u^s$. Then $f_\kappa(x) = xP_\kappa(1-x^2)$. With $u := 1-x^2$, differentiation and the identity $2(s+1)c_{s+1} = (2s+1)c_s$ give

$$\begin{aligned} f_\kappa'(x) &= P_\kappa(u) - 2(1-u)P_\kappa'(u) \\ &= \sum_{s=0}^{\kappa} c_s u^s - 2(1-u)\sum_{s=1}^{\kappa} s c_s u^{s-1} \\ &= (2\kappa+1)c_\kappa u^\kappa + \sum_{s=0}^{\kappa-1}\big((2s+1)c_s - 2(s+1)c_{s+1}\big)u^s \\ &= (2\kappa+1)c_\kappa u^\kappa = A_\kappa(1-x^2)^\kappa. \end{aligned}$$

Therefore $0 \leq f_\kappa'(x) \leq A_\kappa$ on $[0,1]$, with maximum A_κ at $x = 0$. Since $f_\kappa' \geq 0$, $f_\kappa(0) = 0$, and $f_\kappa(1) = 1$, the map f_κ sends $[0,1]$ into itself, and hence so does every $h_{\kappa,q}$. The case $q = 0$ is immediate, and for $q \geq 1$ the chain rule $h_{\kappa,q}'(x) = \prod_{j=0}^{q-1} f_\kappa'(h_{\kappa,j}(x))$ gives $0 \leq h_{\kappa,q}'(x) \leq A_\kappa^q$ on $[0,1]$. At $x = 0$ every factor equals $f_\kappa'(0) = A_\kappa$ because $h_{\kappa,j}(0) = 0$, and hence $h_{\kappa,q}'(0) = A_\kappa^q$. Extending $h_{\kappa,q}$ by the constant 1 on $[1,\infty)$ leaves this bound unchanged, so $\mathrm{Lip}(h_{\kappa,q}) = A_\kappa^q$.

For the lower bound on $\Delta(h_{\kappa,q})$, the properties $h_{\kappa,q}(0) = 0$, $\mathrm{Lip}(h_{\kappa,q}) = A_\kappa^q$, and $h_{\kappa,q} \leq 1$ give $h_{\kappa,q}(x) \leq \min\{A_\kappa^q x, 1\}$. Therefore

$$\Delta(h_{\kappa,q}) = \int_0^\infty (1 - h_{\kappa,q}(x))\,\mathrm{d}x \geq \int_0^{A_\kappa^{-q}} (1 - A_\kappa^q x)\,\mathrm{d}x = \frac{1}{A_\kappa^q} - \frac{1}{2A_\kappa^q} = \frac{1}{2A_\kappa^q}.$$

For the upper bound on $\Delta(h_{\kappa,q})$, define for $L \geq 1$ and $x \in [0,1]$

$$g_L(x) := 1 - (1-x)^L.$$

This family maps $[0,1]$ into itself and satisfies the composition identity $g_L \circ g_M = g_{LM}$ for $L, M \geq 1$. We first show that $f_\kappa(x) \geq g_{A_\kappa}(x)$ on $[0,1]$. Both $f_\kappa'(x) = A_\kappa(1-x^2)^\kappa$ and $g_{A_\kappa}'(x) = A_\kappa(1-x)^{A_\kappa - 1}$ are positive for $x \in [0,1)$, so $(f_\kappa - g_{A_\kappa})'(x)$ has the same sign as

$$\psi_\kappa(x) := \log \frac{f_\kappa'(x)}{g_{A_\kappa}'(x)} = \kappa \log(1+x) + (\kappa + 1 - A_\kappa)\log(1-x),$$

whose derivative is

$$\psi_\kappa'(x) = \frac{A_\kappa - 1 - (2\kappa + 1 - A_\kappa)x}{1-x^2}.$$

To determine the sign of the numerator, we bound A_κ. Since $f_\kappa(1) - f_\kappa(0) = 1$, integrating $f'_\kappa(x) = A_\kappa(1-x^2)^\kappa$ over $[0,1]$ gives $A_\kappa^{-1} = \int_0^1 (1-x^2)^\kappa \, \mathrm{d}x$. Combining this with $(1-x)^\kappa < (1-x^2)^\kappa < 1$ on $(0,1)$ gives

$$\frac{1}{\kappa+1} = \int_0^1 (1-x)^\kappa \, \mathrm{d}x < \int_0^1 (1-x^2)^\kappa \, \mathrm{d}x = \frac{1}{A_\kappa} < 1,$$

so $1 < A_\kappa < \kappa + 1$. Hence the numerator is positive at $x = 0$ and negative at $x = 1$, and it has a unique zero in $(0,1)$. The denominator is positive on $[0,1)$, so ψ_κ first increases and then decreases. Since $\psi_\kappa(0) = 0$ and $\psi_\kappa(x) \to -\infty$ as $x \to 1$, there is a unique $\bar{x} \in (0,1)$ such that

$$\psi_\kappa(x) \geq 0 \quad \text{for } x \in [0, \bar{x}], \qquad \psi_\kappa(x) \leq 0 \quad \text{for } x \in [\bar{x}, 1).$$

Hence $f_\kappa - g_{A_\kappa}$ is nondecreasing on $[0, \bar{x}]$ and nonincreasing on $[\bar{x}, 1]$, and since $f_\kappa(0) - g_{A_\kappa}(0) = f_\kappa(1) - g_{A_\kappa}(1) = 0$, we conclude $f_\kappa(x) \geq g_{A_\kappa}(x)$ on $[0,1]$.

We next prove $h_{\kappa,q} \geq g_{A_\kappa^q}$ for every $q \geq 0$ by induction on q. The base case $h_{\kappa,0} = g_1$ is immediate. Assuming $h_{\kappa,q} \geq g_{A_\kappa^q}$, the monotonicity of f_κ, the bound $f_\kappa \geq g_{A_\kappa}$, and the composition identity give

$$h_{\kappa,q+1} = f_\kappa \circ h_{\kappa,q} \geq f_\kappa \circ g_{A_\kappa^q} \geq g_{A_\kappa} \circ g_{A_\kappa^q} = g_{A_\kappa^{q+1}},$$

which completes the induction. Finally, $h_{\kappa,q} \geq g_{A_\kappa^q}$ and $h_{\kappa,q}(x) = 1$ for $x \geq 1$ give

$$\Delta(h_{\kappa,q}) = \int_0^1 (1 - h_{\kappa,q}(x)) \, \mathrm{d}x \leq \int_0^1 (1-x)^{A_\kappa^q} \, \mathrm{d}x = \frac{1}{A_\kappa^q + 1} \leq \frac{1}{A_\kappa^q},$$

which completes the proof. □

E.2 Proof of Theorem 5.1

Proof of Theorem 5.1. By the definition of the depth in Theorem 5.1,

$$\frac{G_{\mathrm{op}}\sqrt{r}}{\Gamma\sqrt{1-\beta}} \leq A^q < \frac{AG_{\mathrm{op}}\sqrt{r}}{\Gamma\sqrt{1-\beta}}.$$

Then, by $C = G_{\mathrm{op}}$ together with $\Delta(h_q) \leq A^{-q}$ and $\mathrm{Lip}(h_q) = A^q$ from Lemma 5.2, the first term in (11) of Theorem 4.2 is bounded by

$$\frac{rG_{\mathrm{op}}}{(1-\beta)A^q} \leq \frac{rG_{\mathrm{op}}}{(1-\beta)\left(G_{\mathrm{op}}\sqrt{r}/(\Gamma\sqrt{1-\beta})\right)} = \Gamma\sqrt{\frac{r}{1-\beta}}.$$

The second term satisfies

$$\frac{2A^q\Gamma^2}{G_{\mathrm{op}}} < \frac{2A\left(G_{\mathrm{op}}\sqrt{r}/(\Gamma\sqrt{1-\beta})\right)\Gamma^2}{G_{\mathrm{op}}} = 2A\Gamma\sqrt{\frac{r}{1-\beta}}.$$

Adding these two bounds yields (13), which completes the proof. □

E.3 Proof of Theorem 5.4

Proof of Theorem 5.4. If $\varepsilon > B_{\mathrm{NS}}$, the definition of B_{NS} and $A = 15/8$ give $B_{\mathrm{NS}} \geq (1 + 2A)\sqrt{r}\,\Gamma > 3\sqrt{r}\,\Gamma$. Hence, every point W satisfies

$$\|\nabla\mathcal{L}(W)\|_*^{[\rho]} \leq \|\nabla\mathcal{L}(W)\|_* \leq \sqrt{r}\,\|\nabla\mathcal{L}(W)\|_{\mathrm{F}} \leq \sqrt{r}\,\Gamma < \frac{\varepsilon}{3}.$$

Consequently, $\mathbb{E}[\|\nabla\mathcal{L}(\bar{W}_\tau)\|_*^{[\rho]}] < \varepsilon/3$, so (21) holds in this case.

It remains to consider $\varepsilon \leq B_{\mathrm{NS}}$. Then $1 - \beta = (\varepsilon/3B_{\mathrm{NS}})^2 \leq 1/9$. Since $\Gamma \leq \sqrt{r}G_{\mathrm{op}}$ by the normalization following Assumption 2.3 and $\varepsilon \leq B_{\mathrm{NS}}$, the maximum in the depth specified in Theorem 5.4 is its second argument. Using $\sqrt{1-\beta} = \varepsilon/(3B_{\mathrm{NS}})$, this depth equals the depth required in Theorem 5.1. Hence, $\mathrm{Reg}_t^{[\beta]}(D) = D\,\mathrm{Reg}_t^{[\beta]}(1)$ gives

$$\mathbb{E}[\mathrm{Reg}_t^{[\beta]}(1)] \leq (1 + 2A)\Gamma\sqrt{\frac{r}{1-\beta}}.$$

Substituting this into the regret term of Lemma 3.1 and using $(1+(1-\beta)(T-1))/T = 1-\beta+\beta/T$ gives

$$\frac{1}{T}\mathbb{E}\left[\mathrm{Reg}_T^{[\beta]}(1)+(1-\beta)\sum_{t=1}^{T-1}\mathrm{Reg}_t^{[\beta]}(1)\right] \le \left(1-\beta+\frac{\beta}{T}\right)(1+2A)\Gamma\sqrt{\frac{r}{1-\beta}}.$$

Since $\sqrt{1-\beta^2} \ge \sqrt{1-\beta}$, it holds that

$$\frac{1-\beta+\beta/T}{\sqrt{1-\beta^2}} \le \frac{1-\beta+\beta/T}{\sqrt{1-\beta}} = \sqrt{1-\beta}+\frac{\beta}{T\sqrt{1-\beta}},$$

and $B_{\mathrm{NS}} = \sqrt{r}\,((1+2A)\Gamma+\sigma)$ gives

$$\frac{1}{T}\mathbb{E}\left[\mathrm{Reg}_T^{[\beta]}(1)+(1-\beta)\sum_{t=1}^{T-1}\mathrm{Reg}_t^{[\beta]}(1)\right]+\left(1-\beta+\frac{\beta}{T}\right)\frac{\sqrt{r}\,\sigma}{\sqrt{1-\beta^2}} \le B_{\mathrm{NS}}\left(\sqrt{1-\beta}+\frac{\beta}{T\sqrt{1-\beta}}\right).$$

Thus Lemma 3.1 gives

$$\mathbb{E}\left[\|\nabla\mathcal{L}(\bar{W}_\tau)\|_*^{[\rho]}\right] \le \frac{4\Delta_{\mathcal{L}}}{(1-\beta)\rho T}+B_{\mathrm{NS}}\left(\sqrt{1-\beta}+\frac{\beta}{T\sqrt{1-\beta}}\right).$$

Here $B_{\mathrm{NS}}\sqrt{1-\beta} = \varepsilon/3$, the first term is $36\Delta_{\mathcal{L}}B_{\mathrm{NS}}^2/(\rho\varepsilon^2 T)$, and the third term is $3\beta B_{\mathrm{NS}}^2/(\varepsilon T)$. Since $B_{\mathrm{NS}}^2 = O(r(\Gamma+\sigma)^2)$, we obtain

$$\mathbb{E}\left[\|\nabla\mathcal{L}(\bar{W}_\tau)\|_*^{[\rho]}\right] \le \frac{\varepsilon}{3}+O\left(\frac{\Delta_{\mathcal{L}}\,r(\Gamma+\sigma)^2}{\rho\varepsilon^2 T}+\frac{r(\Gamma+\sigma)^2}{\varepsilon T}\right). \tag{21}$$

Requiring each of the two T-dependent terms to be at most $\varepsilon/3$ gives the complexity bound in Theorem 5.4, which completes the proof. □

E.4 Deferred proofs from Section 5.3

Here, we prove the results of Section 5.3 under Assumptions 2.1 and 5.5. Write $g_s := \nabla\mathcal{L}(\widetilde{W}_s)$ for the gradient at the query point of round s, and let $\mathcal{Q}_s$ be the σ-algebra of the history up to and including the choice of $\widetilde{W}_s$, so that g_s and M_{s-1} are $\mathcal{Q}_s$-measurable. Thus, Assumption 2.1 with $W = \widetilde{W}_s$ gives $\mathbb{E}[G_s \mid \mathcal{Q}_s] = g_s$ and $\mathbb{E}[\|G_s - g_s\|_{\mathrm{F}}^2 \mid \mathcal{Q}_s] \le \sigma^2$.

Proof of Lemma 5.6. For any $p \in \mathcal{P}(W;\rho)$, the triangle inequality and Assumption 5.5 give

$$\|\nabla\mathcal{L}(W)\|_* \le \|\mathbb{E}_{Y\sim p}[\nabla\mathcal{L}(Y)]\|_* + \mathbb{E}_{Y\sim p}[\|\nabla\mathcal{L}(W)-\nabla\mathcal{L}(Y)\|_*] \le \|\mathbb{E}_{Y\sim p}[\nabla\mathcal{L}(Y)]\|_* + L_{\mathrm{op}}\rho,$$

using $\mathbb{E}_{Y\sim p}[\|Y-W\|_{\mathrm{op}}] \le \rho$. Taking the infimum over p completes the proof. □

Lemma E.2. *Under Assumption 2.1, for every $s \ge 1$,*

$$\mathbb{E}\left[B_{\widetilde{\Phi}_{h_q,G_{\mathrm{op}}}}(M_s\|M_{s-1}) \,\middle|\, \mathcal{Q}_s\right] \le 2(1-\beta)\|g_s - M_{s-1}\|_* + \frac{A^q(1-\beta)^2}{2G_{\mathrm{op}}}\sigma^2. \tag{22}$$

Proof. Let $\Psi := \widetilde{\Phi}_{h_q,G_{\mathrm{op}}}$ and $\widehat{M}_s := \beta M_{s-1}+(1-\beta)g_s$ for simplicity. First, the momentum M_s is related to $\widehat{M}_s$ by

$$M_s = \widehat{M}_s + (1-\beta)(G_s - g_s), \qquad \widehat{M}_s = \beta M_{s-1}+(1-\beta)g_s.$$

By the definition of the Bregman divergence, the points $M_{s-1}, \widehat{M}_s, M_s$ satisfy

$$B_\Psi(M_s\|M_{s-1}) = B_\Psi(\widehat{M}_s\|M_{s-1}) + B_\Psi(M_s\|\widehat{M}_s) + \langle\nabla\Psi(\widehat{M}_s)-\nabla\Psi(M_{s-1}), M_s-\widehat{M}_s\rangle. \tag{23}$$

The last inner product has zero conditional expectation, since $\mathbb{E}[M_s-\widehat{M}_s \mid \mathcal{Q}_s] = (1-\beta)\mathbb{E}[G_s-g_s \mid \mathcal{Q}_s] = 0$.

For $B_\Psi(\widehat{M}_s\|M_{s-1})$, using $\widehat{M}_s - M_{s-1} = (1-\beta)(g_s - M_{s-1})$, we have

$$\begin{aligned} B_\Psi(\widehat{M}_s\|M_{s-1}) &= \Psi(\widehat{M}_s) - \Psi(M_{s-1}) - \langle \nabla\Psi(M_{s-1}), \widehat{M}_s - M_{s-1}\rangle \\ &\le \left|\Psi(\widehat{M}_s) - \Psi(M_{s-1})\right| + \left|\langle \nabla\Psi(M_{s-1}), \widehat{M}_s - M_{s-1}\rangle\right| \\ &\le 2\|\widehat{M}_s - M_{s-1}\|_* = 2(1-\beta)\|g_s - M_{s-1}\|_*, \end{aligned}$$

where the last inequality follows from operator–nuclear duality and $\|\nabla\Psi(M)\|_{\mathrm{op}} = \|\mathcal{H}_{h_q}(M/G_{\mathrm{op}})\|_{\mathrm{op}} \le 1$, which holds since $h_q \in [0,1]$.

For $B_\Psi(M_s\|\widehat{M}_s)$, Lemma 4.6 with $\mathrm{Lip}(h_q) = A^q$ gives

$$B_\Psi(M_s\|\widehat{M}_s) \le \frac{A^q}{2G_{\mathrm{op}}}\|M_s - \widehat{M}_s\|_{\mathrm{F}}^2 = \frac{A^q(1-\beta)^2}{2G_{\mathrm{op}}}\|G_s - g_s\|_{\mathrm{F}}^2.$$

Taking the conditional expectation of (23) and using $\mathbb{E}[\|G_s - g_s\|_{\mathrm{F}}^2 \mid \mathcal{Q}_s] \le \sigma^2$ gives (22), which completes the proof. □

Lemma E.3. *Under Assumptions 2.1 and 5.5, for every $s \ge 1$,*

$$\mathbb{E}[\|g_s - M_{s-1}\|_*] \le \beta^{s-1}\|\nabla\mathcal{L}(W_0)\|_* + \frac{2L_{\mathrm{op}}D}{1-\beta} + \sigma\sqrt{\frac{r(1-\beta)}{1+\beta}}.$$

Proof. For $s \ge 2$, since $\widetilde{W}_s = W_{s-1} + u_s X_s$, $\widetilde{W}_{s-1} = W_{s-2} + u_{s-1}X_{s-1}$, and $W_{s-1} = W_{s-2} + X_{s-1}$ with $u_s \in [0,1]$ in Algorithm 2, $\widetilde{W}_s - \widetilde{W}_{s-1} = (1-u_{s-1})X_{s-1} + u_s X_s$. Since $\|X_t\|_{\mathrm{op}} \le D$, this and Assumption 5.5 give

$$\|\widetilde{W}_s - \widetilde{W}_{s-1}\|_{\mathrm{op}} \le 2D, \qquad \|g_s - g_{s-1}\|_* \le 2L_{\mathrm{op}}D. \tag{24}$$

The definition of M_{s-1} and $G_j = g_j + (G_j - g_j)$ give

$$g_s - M_{s-1} = \left(g_s - (1-\beta)\sum_{j=1}^{s-1}\beta^{s-1-j}g_j\right) - (1-\beta)\sum_{j=1}^{s-1}\beta^{s-1-j}(G_j - g_j). \tag{25}$$

The first term of (25) can be rewritten as

$$g_s - (1-\beta)\sum_{j=1}^{s-1}\beta^{s-1-j}g_j = \beta^{s-1}g_1 + \sum_{j=2}^{s}\beta^{s-j}(g_j - g_{j-1}).$$

Since $X_1 = 0$ we have $\widetilde{W}_1 = W_0$ and $g_1 = \nabla\mathcal{L}(W_0)$, so (24) yields

$$\begin{aligned} \left\|g_s - (1-\beta)\sum_{j=1}^{s-1}\beta^{s-1-j}g_j\right\|_* &\le \beta^{s-1}\|\nabla\mathcal{L}(W_0)\|_* + 2L_{\mathrm{op}}D\sum_{j=2}^{s}\beta^{s-j} \\ &\le \beta^{s-1}\|\nabla\mathcal{L}(W_0)\|_* + \frac{2L_{\mathrm{op}}D}{1-\beta}. \end{aligned} \tag{26}$$

For the second term of (25), $\mathbb{E}[\|G_s - g_s\|_{\mathrm{F}}^2 \mid \mathcal{Q}_s] \le \sigma^2$ and $\mathbb{E}[G_s - g_s \mid \mathcal{Q}_s] = 0$ give

$$\begin{aligned} \mathbb{E}\left[\left\|(1-\beta)\sum_{j=1}^{s-1}\beta^{s-1-j}(G_j - g_j)\right\|_{\mathrm{F}}^2\right] &= (1-\beta)^2\sum_{j=1}^{s-1}\beta^{2(s-1-j)}\mathbb{E}\left[\|G_j - g_j\|_{\mathrm{F}}^2\right] \\ &\le \frac{1-\beta}{1+\beta}\sigma^2. \end{aligned}$$

Thus, Jensen's inequality and $\|Z\|_* \le \sqrt{r}\|Z\|_{\mathrm{F}}$ give

$$\mathbb{E}\left[\left\|(1-\beta)\sum_{j=1}^{s-1}\beta^{s-1-j}(G_j - g_j)\right\|_*\right] \le \sigma\sqrt{\frac{r(1-\beta)}{1+\beta}}. \tag{27}$$

Combining (25), (26), and (27) completes the proof. □

Proof of Theorem 5.7. If $\varepsilon > \frac{3}{2}\sqrt{r}\,\Gamma$, then Jensen's inequality and Assumption 2.1 imply that every point W satisfies

$$\|\nabla\mathcal{L}(W)\|_* \le \sqrt{r}\,\|\nabla\mathcal{L}(W)\|_{\mathrm{F}} \le \sqrt{r}\,\Gamma < \frac{2}{3}\varepsilon.$$

Consequently, (30) holds in this case.

It remains to consider $\varepsilon \le \frac{3}{2}\sqrt{r}\,\Gamma$. Let $\Psi := \widetilde{\Phi}_{h_q, G_{\mathrm{op}}}$ for simplicity. The choice $C_s = G_{\mathrm{op}}$ satisfies $C_{s+1} \ge \beta C_s$, so (10) of Theorem 4.2 applies with $h = h_q$. Using $\Delta(h_q) \le A^{-q}$ from Lemma 5.2 and $\mathrm{Reg}_t^{[\beta]}(D) = D\,\mathrm{Reg}_t^{[\beta]}(1)$, dividing by D and taking expectations gives

$$\begin{aligned}
\mathbb{E}[\mathrm{Reg}_t^{[\beta]}(1)] &\le \frac{rG_{\mathrm{op}}}{(1-\beta)A^q} + \frac{1}{1-\beta}\sum_{s=1}^{t}\beta^{t-s}\mathbb{E}[B_\Psi(M_s\|M_{s-1})] \\
&\le \frac{rG_{\mathrm{op}}}{(1-\beta)A^q} + \frac{A^q(1-\beta)\sigma^2}{2G_{\mathrm{op}}}\sum_{s=1}^{t}\beta^{t-s} + 2\sum_{s=1}^{t}\beta^{t-s}\mathbb{E}[\|g_s - M_{s-1}\|_*] \\
&\le \frac{rG_{\mathrm{op}}}{(1-\beta)A^q} + \frac{A^q(1-\beta)\sigma^2}{2G_{\mathrm{op}}}\sum_{s=1}^{t}\beta^{t-s} \\
&\quad + 2\sum_{s=1}^{t}\beta^{t-s}\left(\beta^{s-1}\|\nabla\mathcal{L}(W_0)\|_* + \frac{2L_{\mathrm{op}}D}{1-\beta} + \sigma\sqrt{\frac{r(1-\beta)}{1+\beta}}\right),
\end{aligned}$$

where in the second inequality we used Lemma E.2 and the tower property, and in the last inequality we used Lemma E.3. Using $\sum_{s=1}^{t}\beta^{t-s} \le 1/(1-\beta)$ and $\sum_{s=1}^{t}\beta^{t-s}\beta^{s-1} = t\beta^{t-1}$, we further bound this as

$$\mathbb{E}[\mathrm{Reg}_t^{[\beta]}(1)] \le \frac{rG_{\mathrm{op}}}{(1-\beta)A^q} + \frac{A^q\sigma^2}{2G_{\mathrm{op}}} + 2t\beta^{t-1}\|\nabla\mathcal{L}(W_0)\|_* + \frac{4L_{\mathrm{op}}D}{(1-\beta)^2} + \frac{2\sqrt{r}\,\sigma}{\sqrt{1-\beta^2}}. \tag{28}$$

The term $2t\beta^{t-1}\|\nabla\mathcal{L}(W_0)\|_*$ contributes to the bound of Lemma 3.1 as

$$\begin{aligned}
&\frac{1}{T}\left(2T\beta^{T-1} + 2(1-\beta)\sum_{t=1}^{T-1}t\beta^{t-1}\right)\|\nabla\mathcal{L}(W_0)\|_* \\
&\le \frac{1}{T}\left(\frac{2}{1-\beta} + \frac{2}{1-\beta}\right)\|\nabla\mathcal{L}(W_0)\|_* = \frac{4\|\nabla\mathcal{L}(W_0)\|_*}{T(1-\beta)},
\end{aligned}$$

where the inequality uses $T(1-\beta)\beta^{T-1} \le (1-\beta)\sum_{t=1}^{T}\beta^{t-1} = 1-\beta^T \le 1$ and $\sum_{t=1}^{T-1}t\beta^{t-1} \le (1-\beta)^{-2}$.

Set $\rho := \varepsilon/(12L_{\mathrm{op}})$, so that $D = (1-\beta)\rho/(4\beta)$. Substituting (28) into Lemma 3.1 and using $(1+(1-\beta)(T-1))/T = 1-\beta+\beta/T$ gives

$$\begin{aligned}
\mathbb{E}\left[\|\nabla\mathcal{L}(\bar{W}_\tau)\|_*^{[\rho]}\right] &\le \frac{4\Delta_{\mathcal{L}}}{(1-\beta)\rho T} + \frac{4\|\nabla\mathcal{L}(W_0)\|_*}{T(1-\beta)} \\
&\quad + \left(1-\beta+\frac{\beta}{T}\right)\left(\frac{rG_{\mathrm{op}}}{(1-\beta)A^q} + \frac{4L_{\mathrm{op}}D}{(1-\beta)^2} + \frac{3\sqrt{r}\,\sigma}{\sqrt{1-\beta^2}} + \frac{A^q\sigma^2}{2G_{\mathrm{op}}}\right).
\end{aligned}$$

Then, using Lemma 5.6 and substituting $D = (1-\beta)\rho/(4\beta)$ gives

$$\begin{aligned}
\mathbb{E}\left[\|\nabla\mathcal{L}(\bar{W}_\tau)\|_*\right] &\le \frac{4\Delta_{\mathcal{L}}}{(1-\beta)\rho T} + \frac{4\|\nabla\mathcal{L}(W_0)\|_*}{T(1-\beta)} + \left(1+\frac{\beta}{T(1-\beta)}\right)\frac{rG_{\mathrm{op}}}{A^q} \\
&\quad + L_{\mathrm{op}}\rho\left(1+\frac{1}{\beta}+\frac{1}{T(1-\beta)}\right) + 3\sqrt{r}\sigma\frac{1-\beta+\beta/T}{\sqrt{1-\beta^2}} + \left(1-\beta+\frac{\beta}{T}\right)\frac{A^q\sigma^2}{2G_{\mathrm{op}}}.
\end{aligned}$$

We next bound the six terms one by one. The choices of ρ and β give $L_{\mathrm{op}}\rho = \varepsilon/12$ and $1/(1-\beta) = \max\{2, 144r\sigma^2/\varepsilon^2\} \le 2 + 144r\sigma^2/\varepsilon^2$. Thus, the first term satisfies

$$\frac{4\Delta_{\mathcal{L}}}{(1-\beta)\rho T} = O\left(\frac{L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon T} + \frac{r\sigma^2 L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon^3 T}\right).$$

For the second term, we first bound $\|\nabla\mathcal{L}(W_0)\|_*$. Applying the fundamental theorem of calculus along the segment from W to W', followed by operator–nuclear duality and Assumption 5.5, gives

$$\mathcal{L}(W') \le \mathcal{L}(W) + \langle \nabla\mathcal{L}(W), W' - W\rangle + \frac{L_{\mathrm{op}}}{2}\|W' - W\|_{\mathrm{op}}^2. \tag{29}$$

Take $W = W_0$ and $W' = W_0 - \|\nabla\mathcal{L}(W_0)\|_* \operatorname{polar}(\nabla\mathcal{L}(W_0))/L_{\mathrm{op}}$. Then $\langle \nabla\mathcal{L}(W_0), W' - W_0\rangle = -\|\nabla\mathcal{L}(W_0)\|_*^2/L_{\mathrm{op}}$ and $\|W' - W_0\|_{\mathrm{op}} = \|\nabla\mathcal{L}(W_0)\|_*/L_{\mathrm{op}}$, so (29) gives $\mathcal{L}(W') \le \mathcal{L}(W_0) - \|\nabla\mathcal{L}(W_0)\|_*^2/(2L_{\mathrm{op}})$. Combining this with $\mathcal{L}(W') \ge \mathcal{L}(W_0) - \Delta_{\mathcal{L}}$ yields $\|\nabla\mathcal{L}(W_0)\|_* \le \sqrt{2L_{\mathrm{op}}\Delta_{\mathcal{L}}} \le \varepsilon + L_{\mathrm{op}}\Delta_{\mathcal{L}}/\varepsilon$ and then

$$\frac{4\|\nabla\mathcal{L}(W_0)\|_*}{(1-\beta)T} \le \frac{4\sqrt{2L_{\mathrm{op}}\Delta_{\mathcal{L}}}}{T}\left(2 + \frac{144r\sigma^2}{\varepsilon^2}\right) = O\left(\frac{\varepsilon}{T} + \frac{L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon T} + \frac{r\sigma^2}{\varepsilon T} + \frac{r\sigma^2 L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon^3 T}\right).$$

For the third term, $\Gamma \le \sqrt{r}\,G_{\mathrm{op}}$ and $\varepsilon \le \frac{3}{2}\sqrt{r}\,\Gamma$ imply $12rG_{\mathrm{op}}/\varepsilon \ge 8 > 1$. Thus, $q = \lceil \log_A(12rG_{\mathrm{op}}/\varepsilon)\rceil$ and $12rG_{\mathrm{op}}/\varepsilon \le A^q < 12ArG_{\mathrm{op}}/\varepsilon$. It follows that

$$\left(1 + \frac{\beta}{T(1-\beta)}\right)\frac{rG_{\mathrm{op}}}{A^q} \le \frac{\varepsilon}{12} + \frac{\varepsilon}{12T}\left(2 + \frac{144r\sigma^2}{\varepsilon^2}\right) = \frac{\varepsilon}{12} + O\left(\frac{\varepsilon}{T} + \frac{r\sigma^2}{\varepsilon T}\right).$$

For the fourth term, $1-\beta \le 1/2$ implies $1/\beta \le 2$. Together with $L_{\mathrm{op}}\rho = \varepsilon/12$, this gives

$$L_{\mathrm{op}}\rho\left(1 + \frac{1}{\beta} + \frac{1}{T(1-\beta)}\right) \le \frac{\varepsilon}{4} + \frac{\varepsilon}{12T}\left(2 + \frac{144r\sigma^2}{\varepsilon^2}\right) \le \frac{\varepsilon}{4} + O\left(\frac{\varepsilon}{T} + \frac{r\sigma^2}{\varepsilon T}\right).$$

For the fifth term, the definition of β ensures $3\sqrt{r}\sigma\sqrt{1-\beta} \le \varepsilon/4$. Moreover, $\sqrt{1-\beta^2} \ge \sqrt{1-\beta}$, $1/\sqrt{1-\beta} \le \sqrt{2} + 12\sqrt{r}\sigma/\varepsilon$, and $\sqrt{r\sigma^2} \le (\varepsilon + r\sigma^2/\varepsilon)/2$ give

$$3\sqrt{r}\sigma\frac{1-\beta+\beta/T}{\sqrt{1-\beta^2}} \le 3\sqrt{r}\sigma\sqrt{1-\beta} + \frac{3\sqrt{r}\sigma}{T\sqrt{1-\beta}} \le \frac{\varepsilon}{4} + O\left(\frac{\varepsilon}{T} + \frac{r\sigma^2}{\varepsilon T}\right).$$

For the sixth term, $(1-\beta)\sigma^2 \le \varepsilon^2/(144r)$ and the upper bound on A^q above give

$$\left(1 - \beta + \frac{\beta}{T}\right)\frac{A^q\sigma^2}{2G_{\mathrm{op}}} \le \frac{\varepsilon}{12} + O\left(\frac{r\sigma^2}{\varepsilon T}\right).$$

Therefore,

$$\mathbb{E}\left[\|\nabla\mathcal{L}(\bar{W}_\tau)\|_*\right] \le \frac{2}{3}\varepsilon + O\left(\frac{\varepsilon}{T} + \frac{L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon T} + \frac{r\sigma^2}{\varepsilon T} + \frac{r\sigma^2 L_{\mathrm{op}}\Delta_{\mathcal{L}}}{\varepsilon^3 T}\right). \tag{30}$$

Requiring the $1/T$ terms to be at most $\varepsilon/3$ gives the complexity bound in Theorem 5.7, which completes the proof. □

F Deferred details on the FTRL interpretation

This appendix describes exact-polar Muon as FTL, proves Theorem 5.3, and specializes the regularizer to finite Newton–Schulz.

Exact-polar Muon as follow-the-leader. For the gradients $(G_s)_{s=1}^t$ returned at the O2NC query points, we have $M_t = (1-\beta)\beta^t \sum_{s=1}^t \beta^{-s} G_s$. Hence, Muon with the exact-polar computation, in which the Newton–Schulz iteration is replaced by the exact polar factor, can be written as

$$X_{t+1} = -D\operatorname{polar}(M_t) = -D\operatorname{polar}\left(\sum_{s=1}^t \beta^{-s} G_s\right) \in \operatorname*{arg\,min}_{\|X\|_{\mathrm{op}} \le D} \sum_{s=1}^t \langle \beta^{-s} G_s, X\rangle, \tag{31}$$

where we used $\operatorname{polar}(cS) = \operatorname{polar}(S)$ for $c > 0$ and operator–nuclear duality. This implies that the Muon update with the exact-polar computation is equivalent to FTL over the operator-norm ball for the loss sequence $(\ell_s^{[\beta]})_{s=1}^t$, as noted by Jiang et al. (2026a, Section 5.2).

Since FTL need not achieve sublinear regret, the discounted O2NC framework alone does not provide a general stationarity guarantee for Muon with the exact-polar update on nonsmooth nonconvex objectives. This is consistent with the counterexamples of Parshakova et al. (2026), who construct nonsmooth convex Lipschitz objectives on which Muon with the exact-polar update and subgradients evaluated at the current iterate fails to converge. To obtain convergence guarantees for nonsmooth nonconvex optimization, Jiang et al. (2026a) instead replace FTL with online algorithms that enforce stability explicitly. In contrast, applying Theorem 5.3 with $h = h_q$ shows that the finite Newton–Schulz update in Algorithm 1 is itself FTRL, with a spectral regularizer induced by the smoothed potential.

F.1 Proof of Theorem 5.3

We first compute the two Fenchel conjugates that define the regularizer $R_{h,C}$ in Theorem 5.3.

Lemma F.1. *Extend ϕ_h to a proper convex function on $\mathbb{R}$ by setting $\phi_h(x) = +\infty$ for $x < 0$, and let $\phi_h^*(a) \coloneqq \sup_{x\geq 0}\{ax - \phi_h(x)\}$ be its Fenchel conjugate. Under Assumption 4.1, $\phi_h^*(a) < \infty$ if and only if $a \leq 1$.*

Proof. For $a \leq 1$ and $x \geq 0$,

$$ax - \phi_h(x) \leq x - \phi_h(x) = \int_0^x (1 - h(u))\,\mathrm{d}u \leq \int_0^\infty (1 - h(u))\,\mathrm{d}u = \Delta(h),$$

so $\phi_h^*(a) \leq \Delta(h) < \infty$. For $a > 1$, the bound $\phi_h(x) \leq x$ gives $ax - \phi_h(x) \geq (a-1)x \to \infty$ as $x \to \infty$, so $\phi_h^*(a) = +\infty$, which completes the proof. □

Lifting the scalar conjugate to matrices via Lemma B.3 gives the Fenchel conjugate of the smoothed potential.

Lemma F.2. *Under Assumption 4.1, for every $C > 0$, the Fenchel conjugate of $\widetilde{\Phi}_{h,C}$ is*

$$\widetilde{\Phi}_{h,C}^*(W) = C\sum_{i=1}^r \phi_h^*(\sigma_i(W)).$$

Proof. Apply the Fenchel conjugate formula for singular-value functions (Lemma B.3) to $\phi(x) \coloneqq C\phi_h(x/C)$. This expresses the Fenchel conjugate of $\widetilde{\Phi}_{h,C}$ as $W \mapsto \sum_{i=1}^r \phi^*(\sigma_i(W))$. The change of variables $x = Cu$ gives $\phi^*(a) = \sup_{x\geq 0}\{ax - C\phi_h(x/C)\} = C\phi_h^*(a)$, which completes the proof. □

We are now ready to prove Theorem 5.3.

Proof of Theorem 5.3. By Lemma 4.3, the update (8) satisfies $X_t^h = -D\nabla\widetilde{\Phi}_{h,C_t}(M_{t-1})$. It remains to show that X_t^h minimizes the FTRL objective. The definition of M_{t-1} gives $M_{t-1} = (1-\beta)\beta^{t-1}\sum_{s=1}^{t-1}\beta^{-s}G_s$. Hence the objective in (15) is $\big(\langle M_{t-1}, X\rangle + R_{h,C_t}(X)\big)/\big((1-\beta)\beta^{t-1}\big)$. Substituting $X = -DW$ and using $R_{h,C_t}(X) = D\widetilde{\Phi}_{h,C_t}^*(X/D)$ together with the evenness of $\widetilde{\Phi}_{h,C_t}^*$ (which holds since $\sigma_i(-W) = \sigma_i(W)$) gives

$$\langle M_{t-1}, X\rangle + R_{h,C_t}(X) = -D\left(\langle M_{t-1}, W\rangle - \widetilde{\Phi}_{h,C_t}^*(W)\right).$$

Thus minimizing the left-hand side over $\|X\|_{\mathrm{op}} \leq D$ is the same as maximizing $\langle M_{t-1}, W\rangle - \widetilde{\Phi}_{h,C_t}^*(W)$ over $\|W\|_{\mathrm{op}} \leq 1$. By the Fenchel–Young inequality,

$$\langle M_{t-1}, W\rangle - \widetilde{\Phi}_{h,C_t}^*(W) \leq \widetilde{\Phi}_{h,C_t}(M_{t-1}),$$

with equality if and only if $W = \nabla\widetilde{\Phi}_{h,C_t}(M_{t-1})$. The minimizer is therefore $X = -D\nabla\widetilde{\Phi}_{h,C_t}(M_{t-1}) = X_t^h$, which completes the proof. □

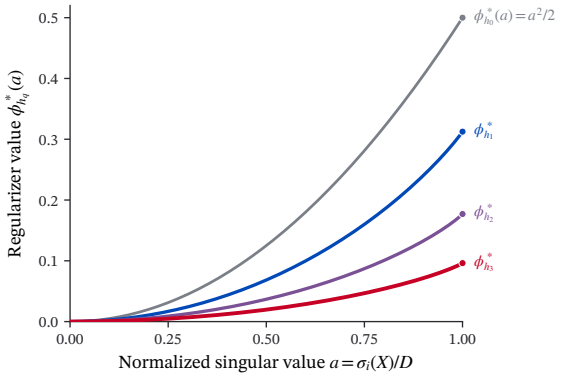


Figure 2: The scalar contribution $\phi^*_{h_q}(a)$ to the FTRL regularizer for depths $q = 0, \dots, 3$, where $a = \sigma_i(X)/D \in [0, 1]$. At depth zero, $\phi^*_{h_0}(a) = a^2/2$. The curves are pointwise nonincreasing in q, each marked endpoint equals $\phi^*_{h_q}(1) = \Delta(h_q)$, and these endpoint values converge to zero as $q \to \infty$.

Closed form of the regularizer. The regularizer $R_{h,C}$ of Theorem 5.3 is a sum of the scalar conjugate ϕ^*_h over the singular values. In terms of the generalized inverse $h^{-1}(v) := \inf\{x \geq 0 : h(x) \geq v\}$ with $\inf \emptyset := +\infty$, the scalar conjugate is

$$\phi^*_h(a) = \begin{cases} 0, & a \leq 0, \\ \displaystyle\int_0^a h^{-1}(v)\,\mathrm{d}v, & 0 \leq a \leq 1, \\ +\infty, & a > 1. \end{cases} \tag{32}$$

Indeed, for $a \leq 0$ the map $x \mapsto ax - \phi_h(x)$ is nonincreasing, so $\phi^*_h(a) = 0$, and Lemma F.1 gives $\phi^*_h(a) = +\infty$ for $a > 1$. For $a \in [0, 1]$, the integrand of $ax - \phi_h(x) = \int_0^x (a - h(u))\,\mathrm{d}u$ is positive exactly when $u < h^{-1}(a)$, so

$$\phi^*_h(a) = \int_0^{h^{-1}(a)} (a - h(u))\,\mathrm{d}u = \int_0^a h^{-1}(v)\,\mathrm{d}v,$$

where the second equality exchanges the order of integration over $\{(u, v) : h(u) < v \leq a\}$. Taking $a = 1$ gives

$$\phi^*_h(1) = \int_0^{h^{-1}(1)} (1 - h(u))\,\mathrm{d}u = \int_0^{\infty} (1 - h(u))\,\mathrm{d}u = \Delta(h),$$

where the second equality uses $1 - h(u) = 0$ for $u > h^{-1}(1)$. Thus, since ϕ^*_h is nondecreasing, $0 \leq \phi^*_h(a) \leq \Delta(h)$ for every $a \in [0, 1]$.

F.2 Specialization to Newton–Schulz h_q

Write $Y = M_t/G_{\mathrm{op}} = U\,\mathrm{diag}(x_i)V^\top$ for a thin singular value decomposition, with $x_i \in [0, 1]$ because $\|Y\|_{\mathrm{op}} \leq 1$. Then $(YY^\top)^k Y = U\,\mathrm{diag}(x_i^{2k+1})V^\top$. Using this for $k = 0, 1, 2$, we see that the Newton–Schulz step of Line 7 acts as

$$\frac{15}{8}Y - \frac{5}{4}(YY^\top)Y + \frac{3}{8}(YY^\top)^2 Y = U\,\mathrm{diag}\left(f(x_i)\right)V^\top.$$

Since f maps $[0, 1]$ into itself, the singular values stay in $[0, 1]$, so q steps replace each x_i by $f^{\circ q}(x_i) = h_q(x_i)$ from (4), and therefore the q-step iteration equals the singular-value map $\mathcal{H}_{h_q}(Y)$ of (5), as stated in Section 2.2.

For the constant normalization $C_t = G_{\mathrm{op}}$, Theorem 5.3 gives

$$X_t = -D\nabla\widetilde{\Phi}_{h_q, G_{\mathrm{op}}}(M_{t-1}) \in \operatorname*{arg\,min}_{\|X\|_{\mathrm{op}} \leq D} \left\{ \left\langle \sum_{s=1}^{t-1} \beta^{-s} G_s, X \right\rangle + \frac{R_{h_q, G_{\mathrm{op}}}(X)}{(1-\beta)\beta^{t-1}} \right\}.$$

Thus the update that applies q Newton–Schulz steps to the momentum is the FTRL update for the discounted linear losses.

Vanishing regularizer as $q \to \infty$. By the closed form (32), $\phi^*_{h_q}$ is nondecreasing with $0 \le \phi^*_{h_q}(a) \le \phi^*_{h_q}(1) = \Delta(h_q)$ on $[0,1]$. Since $\Delta(h_q) \to 0$ as $q \to \infty$ by (14) of Lemma 5.2, the FTRL update approaches the follow-the-leader update $-D\,\mathrm{polar}(M_{t-1})$ of exact-polar Muon. Figure 2 shows $\phi^*_{h_q}$ for depths $q = 0, \dots, 3$.

G Guarantees for other spectral maps

Section 5 focuses on the finite Newton–Schulz map, while the analysis of Section 4 applies to any spectral map satisfying Assumption 4.1. We illustrate this generality using a recently proposed smooth relaxation of the polar map (Mustafi et al., 2026; Feoktistov et al., 2026) as an example. The relaxation inherits the stationarity guarantees for nonsmooth nonconvex objectives from Sections 3 and 4, where the bound balances the approximation error $\Delta(h)$ against the Lipschitz constant $\mathrm{Lip}(h)$ as for finite Newton–Schulz. A smaller $\Delta(h)$ forces a larger $\mathrm{Lip}(h)$, and we show that this tradeoff is unavoidable for every map satisfying Assumption 4.1.

G.1 A smooth relaxation of the polar map

For a smoothing parameter $a > 0$, define

$$\widetilde{h}_a(x) := \frac{ax}{\sqrt{1 + a^2x^2}}, \qquad x \in [0, \infty).$$

This smooth approximation of the polar map is used in recent work (Feoktistov et al., 2026; Mustafi et al., 2026). We first check that $\widetilde{h}_a$ satisfies Assumption 4.1. It is continuous with $\widetilde{h}_a(0) = 0$, and its derivative

$$\widetilde{h}'_a(x) = \frac{a}{(1 + a^2x^2)^{3/2}}$$

is positive, so $\widetilde{h}_a$ is strictly increasing, with $0 \le \widetilde{h}_a(x) < 1$ on $[0, \infty)$. Since $\widetilde{h}'_a$ is decreasing on $[0, \infty)$, the Lipschitz constant is

$$\mathrm{Lip}(\widetilde{h}_a) = \widetilde{h}'_a(0) = a.$$

The approximation error is

$$\Delta(\widetilde{h}_a) = \int_0^\infty \left(1 - \frac{ax}{\sqrt{1 + a^2x^2}}\right) \mathrm{d}x = \lim_{z \to \infty} \left[x - \frac{1}{a}\sqrt{1 + a^2x^2}\right]_0^z = \frac{1}{a},$$

where we used $x - \frac{1}{a}\sqrt{1 + a^2x^2} = -\frac{1}{a^2x + a\sqrt{1+a^2x^2}} \to 0$ as $x \to \infty$. Combining these gives

$$\mathrm{Lip}(\widetilde{h}_a) = a, \qquad \Delta(\widetilde{h}_a) = \frac{1}{a} = \frac{1}{\mathrm{Lip}(\widetilde{h}_a)},$$

which is the analogue of Lemma 5.2 for $\widetilde{h}_a$. Balancing the approximation error against the Lipschitz constant as in Section 5 then yields guarantees such as Theorems 5.1 and 5.4 for $\widetilde{h}_a$, now controlled by the smoothing parameter a rather than the iteration depth q.

G.2 The tradeoff is intrinsic to general spectral maps

This inverse relationship between $\Delta(h)$ and $\mathrm{Lip}(h)$ is not special to the two maps above. For any h satisfying Assumption 4.1, the properties $h(0) = 0$, $h \le 1$, and $\mathrm{Lip}(h)$-Lipschitz continuity give $h(x) \le \min\{\mathrm{Lip}(h)\, x,\, 1\}$, and therefore

$$\Delta(h) = \int_0^\infty (1 - h(x))\, \mathrm{d}x \ge \int_0^{1/\mathrm{Lip}(h)} (1 - \mathrm{Lip}(h)\, x)\, \mathrm{d}x = \frac{1}{2\,\mathrm{Lip}(h)}.$$

The approximation error therefore cannot fall below $1/(2\,\mathrm{Lip}(h))$, so $\Delta(h)$ and $\mathrm{Lip}(h)$ cannot both be made small. The tradeoff between approximation and stability exploited by finite Newton–Schulz is thus intrinsic to every map satisfying Assumption 4.1, not an artifact of the Newton–Schulz iteration.

H NUMERICAL EXPERIMENTS

We test the finite Newton–Schulz update analyzed in Section 5 on a synthetic nonsmooth nonconvex objective. By Lemma 5.2, increasing the Newton–Schulz depth q reduces the approximation error $\Delta(h_q)$ relative to the exact polar map but increases the Lipschitz constant $\mathrm{Lip}(h_q)$, and the regret bound of Theorem 5.1 balances these two effects. The purpose of the experiment is not to benchmark the optimization performance of the individual methods, but to examine qualitatively whether this depth-dependent penalty–stability tradeoff appears in the optimization behavior. For this purpose, all methods evaluate the gradient at the current iterate rather than at the randomized query points of the O2NC conversion (Section 3).

For $k = 1, \dots, N$, let $Z_k \in \mathbb{R}^{d\times d}$ be a standard Gaussian random matrix and set $J_k = SZ_kS$ with $S = \mathrm{diag}(s_1, \dots, s_d)$ and $s_i = 10^{-(i-1)/(d-1)}$. For $W \in \mathbb{R}^{d\times d}$, we minimize an anisotropic variant of the objective employed in Jiang et al. (2026a),

$$\mathcal{L}(W) = \frac{1}{N}\sum_{k=1}^{N} \psi(\langle J_k, W\rangle), \qquad \psi(z) = |z|\big(1 - a\cos(\omega z)\big),$$

with $d = 20$, $N = 100$, $a = 0.9$, and $\omega = 3$, and we select the subgradient 0 at $z = 0$.

All methods run on the same problem instance with learning rates 0.01 and 0.05 and momentum parameter $\beta = 0.9$. We compare the following methods:

- Muon (SVD): the exact polar factor of the momentum matrix, computed by a singular value decomposition.
- Muon (NS): the update $\mathcal{H}_{h_q}(M_t/G_{\mathrm{op}})$ with depth $q \in \{0, 2, 5, 10\}$, where the learning rate plays the role of the radius D in Section 5, and $q = 0$ uses the normalized momentum itself. We fix $G_{\mathrm{op}} = 3.2$, which exceeds every gradient operator norm observed in the runs.
- Pion and Leon: Algorithms 2 and 3 of Jiang et al. (2026a) with step size $\eta = 1$ in their notation, second-moment parameter 0.9, and a numerical regularization of 10^{-8} in the preconditioner. Pion averages 10 perturbation samples per step.

Figure 3 shows how the loss evolves for each depth q and learning rate. At the larger learning rate, the exact-polar update continues to fluctuate around a high loss value, and the largest depth $q = 10$ also exhibits this instability. On the other hand, the shallowest depth $q = 0$ is stable but slow, needing many more gradient evaluations to make comparable progress. The intermediate depths $q = 2$ and $q = 5$ avoid both of these failures: they descend quickly and remain stable. This matches qualitatively the penalty–stability tradeoff discussed in Section 5, where increasing the depth gives a better approximation to the polar map but increases the Lipschitz constant that governs the stability of the update. In this sense, the experiment demonstrates a benefit of the finite Newton–Schulz iteration beyond approximating the exact polar map. Pion and Leon are also stable, and Pion is slower in wall-clock time since each of its steps computes several perturbed polar factors.

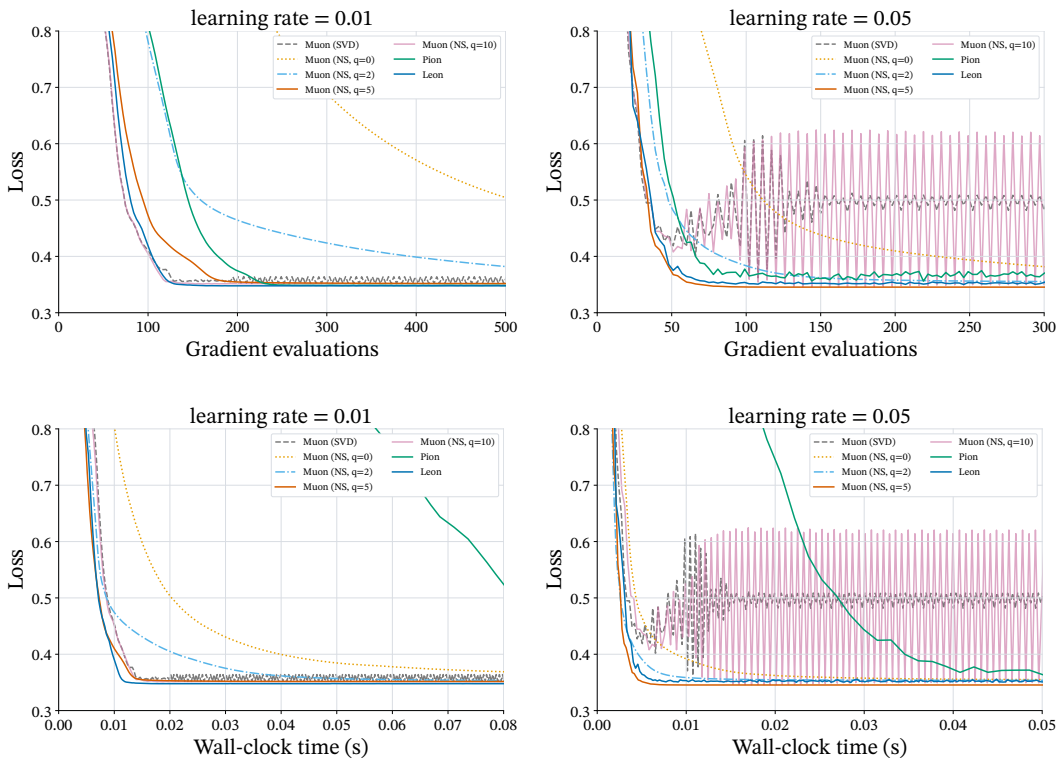


Figure 3: Loss versus gradient evaluations (top) and wall-clock time (bottom) with learning rates 0.01 (left) and 0.05 (right).