DISTANCE-AWARE MUON: ADAPTIVE STEP SCALING FOR NORMALIZED OPTIMIZATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Muon and related normalized optimizers determine a geometry-aware update direction but leave its magnitude to be tuned. Existing stochastic guarantees for star-convex Muon set this magnitude using the unknown distance to a minimizer. We formulate the missing scale as a one-dimensional online-learning problem and introduce Distance-Free Muon, which uses weighted follow-the-regularized-leader to learn the corresponding radius from past stochastic estimates of directional derivatives. For smooth star-convex objectives in general norm geometries, it achieves $\widetilde{O}(T^{-1/3})$ expected last-iterate suboptimality, matching the known-distance method in its dependence on T. The method requires neither the distance to a minimizer nor the noise level and assumes neither a bounded domain nor a bounded trajectory. A complementary predictable stepsize rule gives an $O(T^{-1/4})$ bound on the expected gradient norm for general smooth non-convex objectives. Controlled matrix experiments quantify both the finite-horizon cost of learning the radius and its benefit when a fixed radius is transferred across distance scales. For practical additive variants beyond the analyzed updates, separate GPT-124M/WikiText-103 studies show that FTRL rules match tuned Muon, while a current-gradient stepsize-search proxy with a Nesterov-style momentum direction improves an otherwise identical fixed-scale baseline tuned on a disjoint seed.

1 INTRODUCTION

Muon applies approximate matrix orthogonalization to momentum-based updates, producing a norm-dependent direction whose magnitude is set by a stepsize (Jordan et al., 2024; Large et al., 2024; Liu et al., 2025). The non-Euclidean trust-region interpretation of Kovalev (2025) makes this remaining decision explicit. In the star-convex case, the prescribed step is proportional to a distance or localization quantity involving a global minimizer. Such information is unavailable before optimization. We ask:

> *Can the stepsize required by star-convex Muon theory be learned from previously observed stochastic feedback without knowing the distance to a minimizer?*

The motivation parallels distance adaptation in gradient methods. Even for convex objectives, a rate-optimal stepsize can require the unknown initial distance to a minimizer. D-Adaptation replaces that input with observable information and demonstrates the value of this principle in neural-network training (Defazio & Mishchenko, 2023). Related approaches include parameter-free SGD, DoG, DoWG, and DADA (Carmon & Hinder, 2022; Ivgi et al., 2023; Khaled et al., 2023; Moshtaghifar et al., 2025). We seek an analogous adaptive radius for the norm-dependent linear minimization oracle (LMO) direction and momentum used by Muon. We use *Distance-Aware Muon* for the overall adaptive-scaling framework and *Distance-Free Muon* (DF-Muon) for the star-convex algorithm analyzed below.

Why star-convexity? We study the first-order condition

$$f(x) - f(x^\star) \le \langle \nabla f(x), x - x^\star \rangle.$$

This is differentiable star-convexity, equivalently 1-quasar-convexity with respect to $x^\star$ (Hinder et al., 2020). It connects a gradient to progress toward a global minimizer while imposing no convexity

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.

condition between arbitrary pairs of points. This structure exposes how the distance to a minimizer sets the scale in Muon theory. Our general non-convex extension instead targets stationarity.

Studies of one-point geometry along SGD trajectories and in smoothed neural-network objectives provide empirical motivation for this structure (Zhou et al., 2019; Kleinberg et al., 2018). Our Transformer experiments study training behavior independently of the star-convex theorem.

Learning the scale set by the distance to a minimizer. Existing stochastic star-convex Muon guarantees set the scale from the distance to a minimizer. Our result learns this radius from past stochastic estimates of directional derivatives while preserving last-iterate convergence. For a fixed reference point c, it places the comparison point $c + R_k s_k$ on the ray $\{c + Rs_k : R \geq 0\}$ in the LMO direction. With exact gradients, $D_c = \|x^\star - c\|$ is the natural fixed radius because $c + D_c s_k$ is at least as good as $x^\star$ in the linear model. Weighted follow-the-regularized-leader (FTRL) (Shalev-Shwartz & Singer, 2007; McMahan, 2011) selects R_k from past feedback and competes in cumulative surrogate loss with the fixed choice $R = D_c$. This extends online stepsize learning through convex surrogate losses (Zhuang et al., 2019) to radius selection along LMO directions.

The learned radius changes both the trajectory and the momentum error entering its own feedback, so the FTRL regret bound alone is insufficient. We fix the direction and radius before the fresh gradient, then couple the FTRL bound with weighted control of the momentum error.

Contributions.

1. **Distance-free last-iterate Muon.** For smooth star-convex objectives and an unbiased bounded-variance oracle, we give a weighted FTRL radius rule with a one-gradient-delayed Nesterov-style momentum direction and any deterministic reference point c. Writing $B_c = \max\{\|x_0 - c\|, \|x^\star - c\|\}$ and $s_\sigma = \rho_{\text{norm}}\sigma$, the guarantee is
$$\mathbb{E}[f(x_{T+1}) - f(x^\star)] = \widetilde{O}\left(\frac{LB_c^2 + B_c s_\sigma}{T^{1/3}} + \frac{s_\sigma^2/L}{T^{2/3}}\right).$$
The inputs are L, T, and c, while B_c and s_σ enter only the bound. For $c = x_0$, B_c is the initial distance to a minimizer. The resulting $\widetilde{O}(T^{-1/3})$ convergence rate matches that of the corresponding method of Kovalev (2025), which is supplied with this distance.
2. **Complementary non-convex guarantee.** For general smooth lower-bounded objectives, the method returns a uniformly random iterate $\widehat{x}$ satisfying $\mathbb{E}\|\nabla f(\widehat{x})\|_* = O(T^{-1/4})$. Its assumptions are smoothness, lower boundedness, and the stochastic oracle model. We retain the full stability estimate and noise-tuned refinements in the appendix. An additional exact-gradient distance-free result gives a $\widetilde{O}(LD^2/T)$ suboptimality bound.
3. **Controlled and empirical evidence.** Exact matrix experiments use a momentum parameter satisfying the conditions of Theorem B.7 and show a small adaptation cost relative to a known-distance baseline on the calibration task, together with better transfer than a fixed radius calibrated at smaller distances. On GPT-124M, additive FTRL variants match tuned Muon closely. A current-gradient stepsize-search proxy improves further with Nesterov-style momentum, rectangular scaling, and weight decay.

Table 1: Closest stochastic star-convex Muon comparison in the common unregularized norm-LMO setting. Inputs refer to the published rate-calibrated parameter choices.

Method	Normalized norm-LMO	D input	σ input	Bounded domain/ trajectory	Output	Leading rate
Known-distance Muon (Kovalev, 2025, Cor. 4)	Yes	Yes	Yes	No	Last iterate	$\widetilde{O}(T^{-1/3})$
Distance-Free Muon	Yes	**No**	**No**	No	Last iterate	$\widetilde{O}(T^{-1/3})$

Kovalev's rate-calibrated corollary sets the step from a distance bound and chooses its parameters using the noise level; the underlying theorem itself does not require a noise input. Both methods use L, and neither assumes a bounded domain or trajectory.

Our contribution is not a faster exponent, but removing the unavailable solution-distance and noise inputs while preserving the known-distance last-iterate exponent for the same normalized geometry.

2 Problem Setup and Oracle Information Pattern

We minimize a continuously differentiable function $f : \mathcal{E} \to \mathbb{R}$ on a finite-dimensional real vector space $\mathcal{E}$. The space is equipped with a norm $\|\cdot\|$ and dual norm

$$\|g\|_* := \sup_{\|u\|\le 1} \langle g, u\rangle.$$

Fix a Borel-measurable LMO selection

$$s(m) \in \arg\min_{\|s\|\le 1} \langle m, s\rangle, \qquad \langle m, s(m)\rangle = -\|m\|_*.$$

For different choices of the norm, this formulation recovers Euclidean normalization, sign directions, and the spectral–nuclear matrix geometry underlying Muon (Kovalev, 2025).

For Muon, $\mathcal{E} = \prod_{j=1}^{J} \mathbb{R}^{d_j \times p_j}$ is the space of parameter-matrix tuples. The product spectral norm is $\|X\| = \max_j \|X^{(j)}\|_{\mathrm{op}}$, with dual $\|M\|_* = \sum_j \|M^{(j)}\|_*$, where each summand is a nuclear norm. For a compact SVD $M^{(j)} = U_j \Sigma_j V_j^\top$, the LMO gives $s(M)^{(j)} = -U_j V_j^\top$ when $M^{(j)} \neq 0$. At zero we use a fixed choice in the unit ball.

Assumption 2.1 (Smoothness with respect to $\|\cdot\|$)**.** There is a constant $L > 0$ such that, for all $x, y \in \mathcal{E}$,

$$\|\nabla f(y) - \nabla f(x)\|_* \le L\|y - x\|.$$

Consequently,

$$f(y) \le f(x) + \langle \nabla f(x), y - x\rangle + \frac{L}{2}\|y - x\|^2.$$

For the stochastic results, x_0 is deterministic and the query point is fixed before a fresh sample is drawn.

Assumption 2.2 (Stochastic first-order oracle)**.** Let $\mathcal{H}$ be a pre-query sigma-field, let x be $\mathcal{H}$-measurable, and draw ξ independently of $\mathcal{H}$. The oracle output $g(x;\xi)$ satisfies almost surely

$$\mathbb{E}[g(x;\xi) \mid \mathcal{H}] = \nabla f(x), \qquad \mathbb{E}[\|g(x;\xi) - \nabla f(x)\|_2^2 \mid \mathcal{H}] \le \sigma^2.$$

For a finite norm-equivalence constant ρ_{norm}, assume $\|v\|_* \le \rho_{\mathrm{norm}}\|v\|_2$ and write $s_\sigma := \rho_{\mathrm{norm}}\sigma$.

This oracle model assumes only the stated second-moment bound and permits unbounded support, gradients, iterates, and domains.

For suboptimality results we additionally assume the following one-point structure.

Assumption 2.3 (Star-convexity)**.** There is $x^\star \in \arg\min_x f(x)$ such that

$$f(x) - f(x^\star) \le \langle \nabla f(x), x - x^\star\rangle \qquad \text{for all } x \in \mathcal{E}.$$

Every differentiable convex function with a minimizer satisfies this condition. More generally, it compares each point only with $x^\star$. The general non-convex result relies on the smoothness and oracle assumptions alone.

Information available to the stepsize rule. A normalized trust-region step has the form

$$s_k = s(m_k), \qquad x_{k+1} = c_k + R_k s_k, \tag{1}$$

where momentum and the LMO over the unit norm ball choose the direction. The update geometry specifies the center, and the one-dimensional rule learns the radius. Our stochastic methods use the information order

$$\mathcal{H}_{k-1} \longrightarrow (s_k, R_k) \longrightarrow G_k = g(x_k;\xi_k) \longrightarrow \mathcal{H}_k. \tag{2}$$

Thus s_k and R_k are pre-query measurable. In particular, $\mathbb{E}[\langle G_k - \nabla f(x_k), s_k\rangle R_k \mid \mathcal{H}_{k-1}] = 0$ whenever the product is integrable. This elementary cancellation is the reason for the predictable timing.

The full-gradient and stochastic methods use the same one-dimensional comparison with a fixed radius but different update rules. With exact gradients, the current linear term can enter a convex smoothness majorant. With stochastic feedback, the radius is learned from past one-dimensional surrogate losses.

3 DISTANCE-FREE MUON UNDER STAR-CONVEXITY

We first address the scale set by the distance to a minimizer under Assumptions 2.1, 2.2, and 2.3. Fix a deterministic reference point c. Define, for analysis only,

$$D_c = \|x^\star - c\|, \qquad B_c = \max\{\|x_0 - c\|, D_c\}.$$

Here D_c is the distance from the reference point c to a minimizer. Choosing $c = x_0$ gives $D_c = B_c = \|x_0 - x^\star\|$, the initial distance to a minimizer. The algorithm receives the fixed point c, while D_c and B_c are analysis quantities.

3.1 FROM AN UNKNOWN DISTANCE TO A LEARNED RADIUS

For an exact-gradient LMO direction $s_k^\star = s(\nabla f(x_k))$, duality gives

$$\langle \nabla f(x_k), c + D_c s_k^\star - x^\star \rangle = -D_c \|\nabla f(x_k)\|_* - \langle \nabla f(x_k), x^\star - c \rangle \le 0. \tag{3}$$

Thus the comparison point $c + D_c s_k^\star$ on the ray in the LMO direction is no worse than the minimizer in the linear model. Star-convexity converts this comparison into objective progress. This explains both why the distance from c to a minimizer is useful and why its absence is the central scale-selection problem.

We use FTRL to select a nonnegative radius R_k and set

$$x_{k+1} = (1-\beta)x_k + \beta(c + R_k s_k), \qquad x_{k+1} - x_k = \beta(c - x_k) + \eta_k s_k, \quad \eta_k = \beta R_k. \tag{4}$$

The first term regularizes toward c. With $c = x_0$ it pulls toward initialization, and with $c = 0$ it has a weight-decay form. The full displacement combines this term with the directional step $\eta_k s_k$, where $\eta_k = \beta R_k$.

For a horizon $T \ge 1$, set

$$\begin{aligned} \beta &= \min\left\{\frac{1}{512}, \frac{2\log(T+1)}{3T}\right\}, \qquad \alpha = 8\beta^{2/3}, \qquad r = 1 - \frac{\beta}{2}, \\ w_k &= r^{T-k}, \qquad H = \frac{w_1}{1-r}, \qquad q_R = \frac{9}{4} L \frac{\beta^2}{\alpha}. \end{aligned} \tag{5}$$

Weighted FTRL minimizes the sum of past one-dimensional surrogate losses $\widehat{\ell}_i(R) = w_i(\beta R \langle G_i, s_i \rangle + q_R R^2)$ plus the initial regularizer $q_R H R^2$. Its closed-form prediction is

$$R_k = \operatorname*{arg\,min}_{R \ge 0} \left\{ q_R H R^2 + \sum_{i<k} \widehat{\ell}_i(R) \right\} = \left[-\frac{\beta A_{k-1}}{2 q_R (H + W_{k-1})} \right]_+, \tag{6}$$

where $A_{k-1} = \sum_{i<k} w_i \langle G_i, s_i \rangle$ and $W_{k-1} = \sum_{i<k} w_i$. Negative accumulated directional derivatives yield a positive radius, while the quadratic regularizer controls its magnitude. The weights $w_1, \dots, w_T$ emphasize later iterations, as required by the last-iterate analysis. The guarantee compares cumulative loss with any fixed radius, including $R = D_c$, while allowing the learned sequence R_k to vary freely.

Algorithm 1 Distance-free Muon with a learned radius

Require: Initial point x_0, fixed reference c, smoothness L, horizon T
1: Set the parameters in (5)
2: Draw $G_0 = g(x_0; \xi_0)$ and set $x_1 = x_0$, $m_1 = G_0$, $A_0 = W_0 = 0$
3: **for** $k = 1, \dots, T$ **do**
4: Set $n_k = (1-\alpha)m_k + \alpha G_{k-1}$ and $s_k = s(n_k)$
5: Compute R_k by (6)
6: Set $x_{k+1} = (1-\beta)x_k + \beta(c + R_k s_k)$
7: Draw $G_k = g(x_k; \xi_k)$ at x_k
8: Set $m_{k+1} = (1-\alpha)m_k + \alpha G_k$
9: Set $A_k = A_{k-1} + w_k \langle G_k, s_k \rangle$ and $W_k = W_{k-1} + w_k$
10: **end for**
11: Return x_{T+1}

The algorithm uses a one-gradient-delayed Nesterov-style momentum direction, which is available before G_k is drawn. The method uses $T+1$ gradient calls in total. Its initial radius is zero. For general c, the first update can still move toward c.

3.2 Last-iterate guarantee

Theorem 3.1 (Distance-free last-iterate convergence)**.** *Under Assumptions 2.1, 2.2, and 2.3, Algorithm 1 satisfies*

$$\mathbb{E}[f(x_{T+1}) - f(x^\star)] = \widetilde{O}\left(\frac{LB_c^2 + B_c s_\sigma}{T^{1/3}} + \frac{s_\sigma^2/L}{T^{2/3}}\right). \tag{7}$$

The algorithm takes only L, T, and c as input. The quantities B_c and s_σ appear only in the bound. No bounded-domain or bounded-trajectory assumption is required. Appendix B.4 gives the finite-time inequality and explicit constants.

From FTRL regret to last-iterate convergence. The update and star-convexity give a one-step inequality for $\mathcal{L}_k = f(x_k) - f(x^\star) + 6\beta L\|x_k - c\|^2$. It consists of $r\mathcal{L}_k$, the surrogate-loss difference between R_k and D_c, and a momentum-error term. Multiplication by w_k and summation cancel the intermediate potentials because $rw_k = w_{k-1}$. FTRL bounds the accumulated surrogate-loss difference, and the weighted momentum estimate controls the delayed-direction error. The choice of q_R absorbs the remaining radius-dependent terms, leaving a bound on the final function-value gap.

The $\widetilde{O}(T^{-1/3})$ dependence on the number of iterations matches the corresponding result of Kovalev (2025), which is supplied with the distance; Table 1 states the comparison explicitly. With exact gradients, the one-dimensional minimization of a smoothness upper model in Appendix E gives $\widetilde{O}(LD^2/T)$ with the distance retained only in the bound.

4 Extension to General Non-Convex Objectives

The same online-learning mechanism extends to general non-convex objectives. Let $f^\star := \inf_x f(x) > -\infty$ and $\Delta_0 := f(x_0) - f^\star$. A minimizer need not exist. We use the current iterate as center and learn a stepsize from one-dimensional losses, with stationarity as the target. The fixed-stepsize method of Kovalev (2025) already avoids any distance parameter. Here the stepsize itself is adapted online.

Fix a horizon $T \geq 1$, momentum parameter $\alpha \in (0, 1]$, and $H_0 := 32$. Set

$$h_T := \log\left(1 + \frac{T}{H_0}\right), \qquad q := \frac{L}{2} + \max\left\{\frac{\sqrt{2}L}{\alpha}, 4Lh_T\right\}. \tag{8}$$

At round k, define $A_{k-1} := \sum_{i=1}^{k-1}\langle G_i, s_i\rangle$ and predict

$$\eta_k = \left[-\frac{A_{k-1}}{2q(H_0 + k - 1)}\right]_+. \tag{9}$$

This is the closed-form FTRL minimizer of the past one-dimensional losses $\ell_i(\eta) = \langle G_i, s_i\rangle\eta + q\eta^2$ over $\eta \geq 0$, with initial regularizer $qH_0\eta^2$. The quantity $\langle G_i, s_i\rangle$ estimates the directional derivative. At a fixed denominator, more negative accumulated values give a larger step, while positive values give zero. Across iterations, both the accumulated evidence and the quadratic penalty change.

Algorithm 2 Non-convex FTRL stepsize extension

Require: x_0, horizon T, momentum α, smoothness L
1: Draw $G_0 = g(x_0; \xi_0)$ and set $x_1 = x_0$, $m_1 = G_0$, $A_0 = 0$
2: **for** $k = 1, \ldots, T$ **do**
3: Set $s_k = s(m_k)$ and compute η_k by (9)
4: Set $x_{k+1} = x_k + \eta_k s_k$
5: Draw the only fresh gradient $G_k = g(x_k; \xi_k)$ at x_k
6: Set $m_{k+1} = (1-\alpha)m_k + \alpha G_k$ and $A_k = A_{k-1} + \langle G_k, s_k\rangle$
7: **end for**
8: Return $\widehat{x} = x_\tau$ for independent uniform $\tau \in \{1, \ldots, T\}$

The initialization costs one gradient and each displayed round costs one. Since $\eta_1 = 0$, the first update is zero.

Theorem 4.1 (Oracle inequality relative to any fixed stepsize)**.** *Under Assumptions 2.1 and 2.2, Algorithm 2 satisfies, for every deterministic fixed stepsize $\eta > 0$,*

$$\frac{1}{T}\sum_{k=1}^{T}\mathbb{E}\|\nabla f(x_k)\|_* \le \frac{\Delta_0 + \mathbb{E}\mathcal{S}_T}{\eta T} + C_{\alpha,T}\eta + \frac{2s_\sigma}{\alpha T} + 2s_\sigma\sqrt{\frac{\alpha}{2-\alpha}}, \tag{10}$$

where $\mathcal{S}_T$ is a quadratic feedback term and

$$\mathbb{E}\mathcal{S}_T \le (e-1)\left(\Delta_0 + \frac{s_\sigma^2}{2L}\right).$$

The explicit $C_{\alpha,T}$ and a sharper momentum-dependent bound on the quadratic feedback term are given in Appendix C. The fixed stepsize η is an analysis choice used for comparison.

Choosing parameters that depend on the horizon T, but not on unknown problem quantities, yields the following fully explicit rate.

Corollary 4.2 (Stationarity with horizon-dependent parameters)**.** *If $T \ge 16$ and $\alpha = T^{-1/2}$, then*

$$\mathbb{E}\|\nabla f(\widehat{x})\|_* \le 2\sqrt{7L\left(e\Delta_0 + \frac{e-1}{2L}s_\sigma^2\right)}\,T^{-1/4} + 2s_\sigma T^{-1/4} + 2s_\sigma T^{-1/2}.$$

The schedule depends on L, T, and the fixed constant $H_0 = 32$. It is independent of a minimizer, noise level, stepsize cap, trajectory bound, and gradient bound.

Appendix C gives the full oracle inequality, stability proof, an anytime discussion, and schedules tuned to the target accuracy and noise level. Up to logarithmic factors, the tuned bound matches the oracle complexity of fixed-stepsize Muon when its parameters are chosen using the problem constants (Kovalev, 2025). The oracle inequality adapts to every fixed stepsize while retaining the same worst-case stochastic order.

5 Experiments

We first isolate the cost and potential benefit of learning the radius set by the unknown distance to a minimizer with the exact star-convex update. A general non-convex matrix test evaluates the stepsize extension. Transformer studies evaluate additive FTRL and current-gradient stepsize search.

5.1 Star-convex radius adaptation with an exact LMO

On 8×8 matrices we use the convex quadratic $f(X) = \frac{1}{4}\|X - X^\star\|_F^2$, which is 4-smooth with respect to the spectral norm, whose dual is the nuclear norm. This objective is a special case of star-convexity. We set $c = X_0 = 0$, vary $\|X^\star\|_{\mathrm{op}} \in \{0.5, 2, 8\}$ and Gaussian noise $\sigma \in \{0, 0.25, 1\}$, and run Algorithm 1 with an exact SVD-based LMO. The inputs include L but omit the distance and noise level. A matched known-distance baseline sets $R_k = D = \|X^\star\|_{\mathrm{op}}$ while keeping the direction, contraction, and momentum schedule fixed. This comparison isolates the cost of learning the radius.

An exploratory calibration over theorem-admissible momentum parameters selected $\alpha = 0.15$ on seeds 0–23 for the same quadratic. This value lies near the upper boundary allowed by the finite-time inequalities and was then frozen across evaluation seeds 100–111, distances, noise levels, and objective families in this sensitivity and transfer study. We use $\kappa = 1/4$ and $q_R = (9/4)L\beta^2/\alpha$. At $D = 2$, $\sigma = 0.25$, and $T = 8192$, the learned-radius method has last-iterate gap 0.01395 ± 0.00113, compared with 0.01076 ± 0.00128 for the known-distance baseline with $R = D$. Thus learning the radius incurs a factor-1.30 finite-horizon cost in this setting without using D or σ.

With the same momentum parameter, a single fixed radius calibrated on independent quadratics at $D \in \{1, 2\}$ selects $R = 4$ and is then frozen across test distances. At $D = 8$, $\sigma = 0.25$, and $T = 8192$, the learned rule has gap 0.23739 ± 0.03115, versus 6.42523 ± 1.28230 for transferred $R = 4$ and 0.02527 ± 0.00232 for the known-distance baseline with $R = D$. Thus learning helps

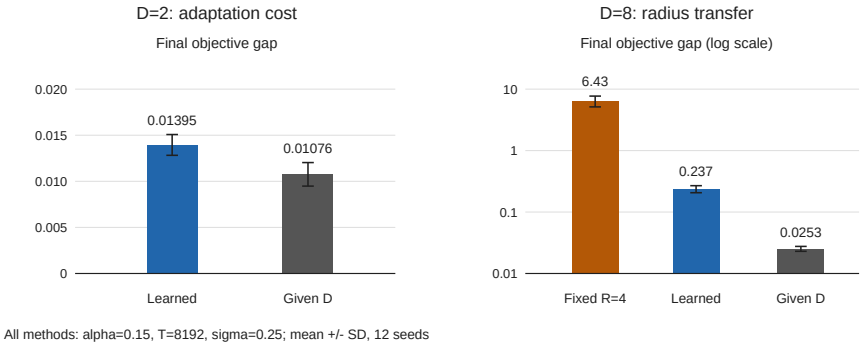


Figure 1: Exact contracted quadratic with theorem-admissible $\alpha = 0.15$, $T = 8192$, and $\sigma = 0.25$ over 12 evaluation seeds (mean $\pm$ SD). Left: the learned-radius gap is 1.30 times the given-D control at $D = 2$. Right: at $D = 8$, learning improves substantially over the fixed radius $R = 4$ selected on separate problems, while the given-D control remains strongest.

large-distance transfer but remains behind the known-distance baseline. The transferred radius wins at some smaller distances and throughout the tested anisotropic radial family. Appendix H.4 gives the calibration protocol and additional cells.

The calibrated value $\alpha = 0.15$ differs from the horizon-dependent schedule used in the asymptotic corollary. The experiment therefore probes finite-horizon constants rather than changing the algorithmic information assumptions or the convergence exponent. Appendix H.1 gives the complete protocol and admissibility check.

5.2 Predictable stepsize adaptation with an exact LMO

On 16×16 matrices, we use the smooth non-convex quadratic-plus-cosine objective in Appendix I, with known smoothness constant $L = 17.6$ with respect to the spectral norm and an exact LMO. We vary the initial distance to the minimizer, $D \in \{1, 2\}$, noise level $\sigma \in \{0.5, 1\}$, and horizon $T \in \{2^9, \ldots, 2^{15}\}$. The algorithm treats D and σ as unknown. Its main baseline is fresh-momentum Muon with a constant radius tuned separately for each problem instance and horizon on disjoint seeds. Appendix I gives the objective, oracle, average gradient norm, matched-timing baseline, and full curves.

Table 2: First tested horizon after which the upper end of the 95% ratio interval remains below one, and the ratio at $T = 32768$. Lower is better.

D	σ	Crossover horizon	Ratio at $T = 32768$
1	0.5	4096	$0.5167\ [0.5120, 0.5214]$
1	1	2048	$0.4755\ [0.4716, 0.4794]$
2	0.5	16384	$0.7989\ [0.7809, 0.8170]$
2	1	8192	$0.7099\ [0.6940, 0.7259]$

At short horizons, the predictable FTRL method has a larger average gradient norm than the offline-tuned baseline. The comparison reverses as T grows, with the crossover occurring later when $D = 2$. At $T = 32768$, the predictable FTRL method reduces the average gradient norm by about 20%–52% across the four cells and has a lower average gradient norm on each of the five problem instances. The crossover shows that the adaptive stepsize outperforms the tuned constant-stepsize baseline at longer horizons.

5.3 GPT-124M on WikiText-103

We train GPT-2 Small for 10,000 stochastic-gradient evaluations, tune on seed 777, and evaluate on three matched seeds. Muon-based methods update matrix parameters, while auxiliary AdamW

Table 3: Matched GPT-124M results at 10,000 stochastic-gradient evaluations, mean $\pm$ sample standard deviation over three seeds. FTRL rows use additive updates.

Method	Validation loss	Δ vs. Muon	Time (min)
Muon	3.47492 ± 0.00710	0	14.676 ± 0.344
Practical weighted-FTRL Muon	3.47632 ± 0.00842	$+0.00140$	15.490 ± 0.049
Non-convex DF-Muon	3.47602 ± 0.01042	$+0.00110$	15.239 ± 0.133
AdamW	3.63699 ± 0.01471	$+0.16207$	5.521 ± 0.199

handles the remaining parameters. The study tests the stepsize-learning principle in an additive Muon recipe beyond the analyzed contracted update.

Practical weighted-FTRL Muon retains the weighted FTRL update for the radius and one-gradient-delayed Nesterov-style momentum direction of Algorithm 1, but applies $\eta_k = \beta R_k$ through an additive update. Its quadratic regularization coefficient uses $\kappa = 64$ and an empirical $\widehat{L}$ tuned once and frozen for evaluation. The experiment with the theoretical contracted update uses the theoretical value $\kappa = 1/4$. The non-convex FTRL variant uses its own empirical coefficient. Appendix H.5 gives the full parameters and gradient handling.

The update, direction, quadratic coefficient, and training-recipe differences from the theorem are itemized in Appendix 9.

Both additive FTRL variants achieve mean validation loss within 0.002 of tuned Muon at 10,000 stochastic-gradient evaluations (Table 3). AdamW has higher final loss but substantially shorter runtime. Replaying the weighted-FTRL radius schedule from the tuning seed gives similar losses on the three evaluation seeds (Appendix H.6), suggesting that online adaptation provides little advantage in this setting.

Current-gradient stepsize-search study. We also evaluate DF-Muon-Proxy under the same GPT-124M data, a budget of 10,000 stochastic-gradient evaluations, and three evaluation seeds, with a tuned Muon baseline rerun within that study. Unlike FTRL, the proxy uses the current stochastic gradient to score candidate stepsizes in a bounded interval. The original proxy configuration matches tuned Muon in mean validation loss (3.47504 versus 3.47495). Adding only Nesterov-style momentum lowers the proxy loss to 3.46104. Adding rectangular scaling and matrix weight decay 0.01 in successive ablations lowers it to 3.44288 ± 0.00693, versus 3.47495 ± 0.00690 for the tuned Muon baseline in the same study. The best-performing variant therefore combines stepsize search with three additional design choices.

In this matched comparison, the fixed scale was selected from eight predeclared values on the disjoint tuning seed 777. The adaptive proxy is ahead on every evaluation seed (Table 4). This supports adaptation relative to that tuned grid baseline, not relative to every possible fixed scale. Appendix H.7 gives the stepwise ablations, and Appendix H.8 gives the matched protocol and per-seed results.

6 Related Work

Muon and non-Euclidean LMO methods. Muon and related methods apply approximate matrix orthogonalization to momentum or gradient estimates in neural-network training (Jordan et al., 2024; Large et al., 2024; Liu et al., 2025; Bernstein & Newhouse, 2024). The trust-region interpretation of Kovalev (2025) gives fixed-radius non-convex and star-convex guarantees. Pethick et al. (2025) study stochastic optimization with LMOs over norm balls, including unconstrained deep-learning problems. In the star-convex setting, the prescribed stepsize depends on the distance to a minimizer or on a localization quantity. Our main result learns the corresponding scale from stochastic estimates of directional derivatives while retaining the $\widetilde{O}(T^{-1/3})$ dependence on the number of iterations for last-iterate convergence. Gluon broadens the geometric framework, and Muon-MVR improves stochastic rates using variance reduction and mean-squared smoothness (Riabinin et al., 2025; Qian et al., 2025). These developments address aspects of Muon different from the distance-dependent scale studied here.

Table 4: Matched GPT-124M control in which only the stepsize rule changes (three paired evaluation seeds; mean $\pm$ SD).

Method	Validation loss
Adaptive Proxy	3.44382 ± 0.00736
Fixed $\eta = 0.005$	3.47102 ± 0.00559
Paired difference	-0.02721 ± 0.00343

Distance adaptation and online learning. D-Adaptation learns an unknown initial distance to a minimizer from optimization feedback, with extensions to SGD and Adam (Defazio & Mishchenko, 2023). Parameter-free SGD, DoG, DoWG, and DADA adapt scale similarly (Carmon & Hinder, 2022; Ivgi et al., 2023; Khaled et al., 2023; Moshtaghifar et al., 2025). Online surrogate stepsize learning casts stepsize selection as one-dimensional online convex optimization and provides regret guarantees relative to a fixed stepsize (Zhuang et al., 2019). Coin-betting methods provide related parameter-free guarantees (McMahan & Orabona, 2014; Orabona & Pál, 2016; Cutkosky & Orabona, 2018). Our weighted quadratic FTRL rule instead compares its cumulative loss with the fixed radius D_c while allowing R_k to vary.

Structured non-convex objectives. Star- and quasar-convexity allow suboptimality analysis under one-point geometry (Kleinberg et al., 2018; Zhou et al., 2019; Hinder et al., 2020). Trajectory-based and local observations in neural networks motivate this structure, while our global assumption remains theoretical for Transformers. Other algorithms for stochastic quasar-convex optimization achieve better rates under their respective assumptions and output rules (Jin, 2020; Gower et al., 2021; Fu et al., 2023; Saad et al., 2025). Our result addresses distance-free scale selection for a delayed Muon direction with one fresh gradient per round. Further comparisons appear in Appendix T.

7 Conclusion and Limitations

Our results show that the unknown distance to a minimizer is not necessary to retain the stochastic last-iterate exponent of known-distance Muon. Weighted FTRL learns the corresponding radius from past stochastic estimates of directional derivatives, while one-gradient-delayed Nesterov-style momentum preserves predictability under a bounded-variance oracle. The method requires no bounded-domain or bounded-trajectory assumption. The same online-learning mechanism yields a stationarity guarantee for general smooth non-convex objectives.

Exact experiments expose the finite-horizon tradeoff of learning this scale. With theorem-admissible $\alpha = 0.15$, the learned rule comes within a factor of 1.30 of the known-distance control on the calibration task and substantially improves over a fixed radius transferred to a larger distance, while the given-distance control remains stronger. Because α was selected exploratorily on the same task, these results support finite-horizon adaptation rather than a universal tuning claim.

On GPT-124M, additive FTRL implementations with $\kappa = 64$ train comparably to tuned Muon. An FTRL schedule transferred from the tuning seed performs similarly to online radius selection in this setting. A current-gradient proxy with Nesterov-style momentum attains lower validation loss in a matched study, with further gains from rectangular scaling and weight decay. These findings motivate an analysis of additive updates and matched experiments that isolate online stepsize adaptation from the surrounding training recipe.

Reproducibility Statement

Section 2 states the assumptions, geometry, and stochastic-oracle information pattern, while Sections 3 and 4 give the algorithms and theorem statements. Complete proofs are provided in Appendices B and D. The objectives, baselines, parameter choices, tuning/evaluation splits, random seeds, and aggregation procedures used in the controlled and GPT experiments are documented in Appendices H, I, and S; the latter also records the relevant software and dataset assets and their licenses.

AI Use Statement

We used generative AI tools to help formulate and critique mathematical claims, suggest proof steps and draft proof text, design experiments, write and review research code, interpret numerical results, locate related work, and draft and edit the manuscript. The authors take responsibility for all mathematical claims, implementations, experimental results, and text in the submission, including AI-assisted material.

References

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

Anuj Apte, Pranav Deshpande, Niraj Kumar, Shouvanik Chakrabarti, and Junhyung Lyle Kim. Anytime training with schedule-free spectral optimization. *arXiv preprint arXiv:2605.23061*, 2026.

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. *Mathematical Programming*, 199:165–214, 2023. doi: 10.1007/s10107-022-01814-1. URL https://arxiv.org/abs/1912.02365.

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

Yair Carmon and Oliver Hinder. Making sgd parameter-free. In *Conference on learning theory*, pp. 2360–2389. PMLR, 2022.

Abhishek Chakraborty and Angelia Nedić. Randomized feasibility methods for constrained optimization with adaptive step sizes. *arXiv preprint arXiv:2601.20076*, 2026.

Andrew R Conn, Nicholas IM Gould, and Philippe L Toint. *Trust region methods*. SIAM, 2000.

Ashok Cutkosky and Francesco Orabona. Black-box reductions for parameter-free online learning in banach spaces. In *Conference On Learning Theory*, pp. 1493–1529. PMLR, 2018.

Aaron Defazio and Konstantin Mishchenko. Learning-rate-free learning by d-adaptation. In *International Conference on Machine Learning*, pp. 7449–7479. PMLR, 2023.

Yury Demidovich, Grigory Malinovsky, Igor Sokolov, and Peter Richtárik. A guide through the zoo of biased sgd. *Advances in Neural Information Processing Systems*, 36:23158–23171, 2023.

Yury Demidovich, Petr Ostroukhov, Grigory Malinovsky, Samuel Horváth, Martin Takáč, Peter Richtárik, and Eduard Gorbunov. Methods with local steps and random reshuffling for generally smooth non-convex federated optimization. In *International Conference on Learning Representations*, volume 2025, pp. 80916–80983, 2025.

Nikita Doikov, Konstantin Mishchenko, and Yurii Nesterov. Super-universal regularized newton method. *SIAM Journal on Optimization*, 34(1):27–56, 2024.

Qiang Fu, Dongchu Xu, and Ashia Camage Wilson. Accelerated stochastic optimization methods under quasar-convexity. In *Proceedings of the Fortieth International Conference on Machine Learning*, pp. 10431–10460. PMLR, 2023.

Eduard Gorbunov, Nazarii Tupitsa, Sayantan Choudhury, Alen Aliev, Peter Richtárik, Samuel Horváth, and Martin Takáč. Methods for convex (L_0, L_1)-smooth optimization: Clipping, acceleration, and adaptivity. In *International Conference on Learning Representations*, volume 2025, pp. 97261–97305, 2025.

Robert M. Gower, Othmane Sebbouh, and Nicolas Loizou. Sgd for structured nonconvex functions: Learning rates, minibatching and interpolation. In *Proceedings of the Twenty Fourth International Conference on Artificial Intelligence and Statistics*, pp. 1315–1323. PMLR, 2021.

Kaja Gruntkowska, Alexander Gaponov, Zhirayr Tovmasyan, and Peter Richtárik. Error feedback for muon and friends. *arXiv preprint arXiv:2510.00643*, 2025.

Oliver Hinder, Aaron Sidford, and Nimit Sohoni. Near-optimal methods for minimizing star-convex functions and beyond. In *Proceedings of the Thirty Third Conference on Learning Theory*, pp. 1894–1938. PMLR, 2020.

Maor Ivgi, Oliver Hinder, and Yair Carmon. Dog is sgd's best friend: A parameter-free dynamic step size schedule. In *International Conference on Machine Learning*, pp. 14465–14499. PMLR, 2023.

Jikai Jin. On the convergence of first order methods for quasar-convex optimization. *arXiv preprint arXiv:2010.04937*, 2020.

Keller Jordan, Yuchen Jin, Vlado Boza, You Jiacheng, Franz Cesista, Laker Newhouse, and Jeremy Bernstein. Muon: An optimizer for hidden layers in neural networks, 2024. *URL https://kellerjordan.github.io/posts/muon/*, 6(3):4, 2024.

Ahmed Khaled and Peter Richtárik. Better theory for SGD in the nonconvex world. *arXiv preprint arXiv:2002.03329*, 2020.

Ahmed Khaled, Konstantin Mishchenko, and Chi Jin. Dowg unleashed: An efficient universal parameter-free gradient descent method. *Advances in Neural Information Processing Systems*, 36: 6748–6769, 2023.

Bobby Kleinberg, Yuanzhi Li, and Yang Yuan. An alternative view: When does sgd escape local minima? In *International conference on machine learning*, pp. 2698–2707. PMLR, 2018.

Dmitry Kovalev. Understanding gradient orthogonalization for deep learning via non-euclidean trust-region optimization. *arXiv preprint arXiv:2503.12645*, 2025.

Dmitry Kovalev and Ekaterina Borodich. Non-euclidean sgd for structured optimization: Unified analysis and improved rates. *arXiv preprint arXiv:2511.11466*, 2025.

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

Tim Large, Yang Liu, Minyoung Huh, Hyojin Bahng, Phillip Isola, and Jeremy Bernstein. Scalable optimization in the modular norm. *Advances in Neural Information Processing Systems*, 37: 73501–73548, 2024.

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.

Zijian Liu and Zhengyuan Zhou. Stochastic nonsmooth convex optimization with heavy-tailed noises: High-probability bound, in-expectation rate and initial distance adaptation. *arXiv preprint arXiv:2303.12277*, 2023.

Brendan McMahan. Follow-the-regularized-leader and mirror descent: Equivalence theorems and L1 regularization. In *Proceedings of the Fourteenth International Conference on Artificial Intelligence and Statistics*, pp. 525–533. JMLR Workshop and Conference Proceedings, 2011.

H Brendan McMahan and Francesco Orabona. Unconstrained online linear learning in hilbert spaces: Minimax algorithms and normal approximations. In *Conference on Learning Theory*, pp. 1020–1039. PMLR, 2014.

Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer sentinel mixture models. *arXiv preprint arXiv:1609.07843*, 2016.

Mohammad Moshtaghifar, Anton Rodomanov, Daniil Vankov, and Sebastian Stich. Dada: Dual averaging with distance adaptation. *arXiv preprint arXiv:2501.10258*, 2025.

Yu Nesterov. Accelerating the cubic regularization of newton's method on convex problems. *Mathematical Programming*, 112(1):159–181, 2008.

Yurii E. Nesterov. A method for solving the convex programming problem with convergence rate $\mathcal{O}(1/k^2)$. *Doklady Akademii Nauk SSSR*, 269(3):543–547, 1983.

Francesco Orabona and Dávid Pál. Coin betting and parameter-free online learning. *Advances in Neural Information Processing Systems*, 29, 2016.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, pp. 1310–1318. PMLR, 2013.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. Pytorch: An imperative style, high-performance deep learning library. *Advances in neural information processing systems*, 32, 2019.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained lmos. *arXiv preprint arXiv:2502.07529*, 2025.

Boris T Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964.

Xun Qian, Hussein Rammal, Dmitry Kovalev, and Peter Richtárik. Muon is provably faster with momentum variance reduction. *arXiv preprint arXiv:2512.16598*, 2025.

Artem Riabinin, Egor Shulgin, Kaja Gruntkowska, and Peter Richtárik. Gluon: Making muon & scion great again!(bridging theory and practice of lmo-based optimizers for llms). *arXiv preprint arXiv:2505.13416*, 2025.

El Mehdi Saad, Wei-Cheng Lee, and Francesco Orabona. New lower bounds for non-convex stochastic optimization through divergence decomposition. In *Proceedings of the Thirty Eighth Conference on Learning Theory*, pp. 5054–5107. PMLR, 2025.

Abdurakhmon Sadiev, Yury Demidovich, Igor Sokolov, Grigory Malinovsky, Sarit Khirirat, and Peter Richtárik. Improved convergence in parameter-agnostic error feedback through momentum. *arXiv preprint arXiv:2511.14501*, 2025.

Shai Shalev-Shwartz and Yoram Singer. A primal-dual perspective of online learning algorithms. *Machine Learning*, 69(2):115–142, 2007.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. *arXiv preprint arXiv:1905.11881*, 2019.

Yi Zhou, Junjie Yang, Huishuai Zhang, Yingbin Liang, and Vahid Tarokh. Sgd converges to global minimum in deep learning via star-convex path. *arXiv preprint arXiv:1901.00451*, 2019.

Zhenxun Zhuang, Ashok Cutkosky, and Francesco Orabona. Surrogate losses for online learning of stepsizes in stochastic non-convex optimization. In *Proceedings of the 36th International Conference on Machine Learning*, volume 97, pp. 7664–7672. PMLR, 2019. URL https://proceedings.mlr.press/v97/zhuang19a.html.

A Guide to the Supplement

Appendix B gives the complete fixed-reference algorithm with a one-gradient-delayed Nesterov-style momentum direction and proves Theorem 3.1. Appendices C and D retain the full non-convex oracle inequality, stability estimates, and schedule refinements. Appendix E states the additional exact-gradient result, followed by its full algorithm and proof. Appendices H and I give the matrix and Transformer protocols. Appendix H.4 tests a fixed radius calibrated on separate problems, while Appendices H.6 and H.7 report distinct practical GPT studies. The current-gradient proxy is not the predictable FTRL algorithm.

Additional algorithms and results. The remaining supplement retains the complementary Distance-Adaptive and Scale-Calibrated Muon results, with their complete algorithms, assumptions, and proofs. Appendix J also gives the initialization-centered EMA version of the predictable distance-free method. It is distinct from the fixed-reference algorithm with delayed Nesterov-style momentum in the main text. The subsequent trajectory-radius and local-certificate methods are analyzed under their respective localization or stability conditions. These are additional results, not prerequisites for either main theorem. The exact-gradient distance-free result also uses its own one-dimensional rule and is not a faster-rate corollary of the main stochastic algorithm.

Additional experiments. Appendix S retains the DF-Muon-Proxy, Distance-Adaptive, and Scale-Calibrated experiments, including the earlier GPT-124M and ViT-Tiny studies and their diagnostics. These experiments use different algorithms and training protocols from the current FTRL study in Appendix H.5. They are reported separately, with their original numerical results.

B Star-Convex DF-Muon: Full Algorithm and Proof

This appendix proves Theorem 3.1 for the fixed-reference method with a one-gradient-delayed Nesterov-style momentum direction. We use Assumptions 2.1, 2.2, and 2.3, and write $\mathcal{F}_{k-1}$ for the information before G_k is drawn. The effective dual-norm noise level is $s_\sigma = \rho_{\mathrm{norm}}\sigma$. No bounded domain, bounded gradients, bounded iterates, bounded path, or bounded noise support is assumed.

Fix the measurable LMO selection from Section 2, so

$$s(m) \in \arg\min_{\|s\|\le 1}\langle m, s\rangle, \qquad \langle m, s(m)\rangle = -\|m\|_*. \tag{11}$$

For a deterministic reference point c, define the analysis-only quantity

$$B_c = \max\{\|x_0 - c\|, \|x^\star - c\|\}. \tag{12}$$

B.1 Algorithm

Fix $T \ge 1$, $0 < \alpha \le 1$, $0 < \beta \le 1/64$, and $\kappa > 0$. Put

$$r = 1 - \frac{\beta}{2}, \qquad w_k = r^{T-k}, \qquad H = \frac{w_1}{1-r}. \tag{13}$$

Take $\rho = 1$, $M = 6$, $c_0 = 1 + 2\rho + M = 9$, and

$$q = \frac{\kappa c_0 L \beta^2}{\alpha}. \tag{14}$$

Draw G_0 at x_0. Set $x_1 = x_0$, $m_1 = G_0$, and $A_0 = W_0 = 0$. At round $k = 1, \dots, T$, perform the following operations.

1. Before drawing G_k, form the one-gradient-delayed Nesterov-style momentum direction
$$n_k = (1-\alpha)m_k + \alpha G_{k-1}, \qquad s_k = s(n_k). \tag{15}$$
2. Compute
$$R_k = \left(-\frac{\beta A_{k-1}}{2q(H + W_{k-1})}\right)_+. \tag{16}$$
Set the next iterate
$$x_{k+1} = (1-\beta)x_k + \beta(c + R_k s_k). \tag{17}$$

3. Draw the only fresh gradient G_k at x_k.
4. Update

$$m_{k+1} = (1-\alpha)m_k + \alpha G_k, \quad A_k = A_{k-1} + w_k\langle G_k, s_k\rangle, \quad W_k = W_{k-1} + w_k. \tag{18}$$

The method returns x_{T+1} and uses $T+1$ stochastic gradients. Every quantity defining x_{k+1} is $\mathcal{F}_{k-1}$-measurable.

For $c = 0$, define the matrix stepsize $\eta_k = \beta R_k$. Then

$$x_{k+1} = (1-\beta)x_k + \eta_k s_k. \tag{19}$$

Whenever $R_k > 0$, this is exactly decoupled matrix weight decay with coefficient $1/R_k$, because $\beta = \eta_k/R_k$. Thus the shrinkage-to-step ratio is selected through the pre-query radius rather than fixed in advance. If $R_k = 0$, the update still applies the explicit shrinkage $1-\beta$.

B.2 Basic inequalities

Write

$$y_k = x_k - c, \quad u^\star = x^\star - c, \quad D_c = \|u^\star\|, \quad \Delta_k = f(x_k) - f(x^\star), \quad \epsilon_k^n = \|n_k - \nabla f(x_k)\|_*. \tag{20}$$

Lemma B.1 (Smooth self-bounding)**.** *For every x,*

$$\|\nabla f(x)\|_*^2 \le 2L(f(x) - f(x^\star)). \tag{21}$$

Also $\Delta_1 \le 2LB_c^2$.

Proof. Let $v = s(\nabla f(x))$. Assumption 2.1 gives

$$f\left(x + \frac{\|\nabla f(x)\|_*}{L} v\right) \le f(x) - \frac{\|\nabla f(x)\|_*^2}{2L}.$$

Global minimality proves (21). Since $\nabla f(x^\star) = 0$, smoothness also gives $\Delta_1 \le (L/2)\|x_0 - x^\star\|^2$. The triangle inequality and (12) give $\|x_0 - x^\star\| \le 2B_c$. **Q.E.D.**

Lemma B.2 (Finite moments)**.** *For every finite T, all variables in* (15)–(18) *have finite second moments. In particular, every scalar product and conditional expectation used below is integrable.*

Proof. All norms on a finite-dimensional space are equivalent. Suppose x_k has a finite second moment. Smoothness and $\nabla f(x^\star) = 0$ then give a finite second moment for $\nabla f(x_k)$ with respect to $\|\cdot\|_2$. Assumption 2.2 gives the same conclusion for G_k. The LMO directions have norm at most one. Hence A_k, m_{k+1}, and n_{k+1} have finite second moments whenever the variables at the start of round k do. The denominator in (16) is deterministic and at least $2qH > 0$, so R_k has a finite second moment. The update (17) then gives the claim for x_{k+1}. The initialization at x_0 starts the induction. **Q.E.D.**

Lemma B.3 (One-step comparison)**.** *Let*

$$\mathcal{L}_k = \Delta_k + M\beta L\|y_k\|^2, \qquad \phi_k(R) = \beta R\langle \nabla f(x_k), s_k\rangle + qR^2. \tag{22}$$

Then

$$\begin{aligned}\mathcal{L}_{k+1} + L\|x_{k+1} - x_k\|^2 + \frac{\beta}{2}\Delta_k + (q - c_0 L\beta^2)R_k^2 \\ \le r\mathcal{L}_k + \phi_k(R_k) - \phi_k(D_c) + qD_c^2 + 2\beta D_c\epsilon_k^n.\end{aligned} \tag{23}$$

Proof. The displacement is $x_{k+1} - x_k = \beta(R_k s_k - y_k)$. Decompose

$$R_k s_k - y_k = (u^\star - y_k) + (D_c s_k - u^\star) + (R_k - D_c)s_k.$$

Star-convexity bounds the first inner product with $\nabla f(x_k)$ by $-\Delta_k$. The LMO property gives $\langle n_k, D_c s_k - u^\star\rangle \le 0$, and duality gives

$$\langle \nabla f(x_k), D_c s_k - u^\star\rangle \le 2D_c\epsilon_k^n. \tag{24}$$

Consequently,

$$\beta\langle\nabla f(x_k), R_k s_k - y_k\rangle \le -\beta\Delta_k + 2\beta D_c\epsilon_k^n + \beta(R_k - D_c)\langle\nabla f(x_k), s_k\rangle. \tag{25}$$

Convexity of the squared norm and the triangle inequality give

$$\|y_{k+1}\|^2 \le (1-\beta)\|y_k\|^2 + \beta R_k^2, \qquad \|R_k s_k - y_k\|^2 \le 2R_k^2 + 2\|y_k\|^2. \tag{26}$$

Apply the descent inequality, add $M\beta L\|y_{k+1}\|^2 + L\|x_{k+1} - x_k\|^2$, and use (25) and (26). The coefficient of $\|y_k\|^2$ is valid because

$$M\beta L(1-\beta) + 3L\beta^2 \le M\beta L\left(1 - \frac{\beta}{2}\right),$$

where $M = 6$. Finally,

$$\begin{aligned}&\beta(R_k - D_c)\langle\nabla f(x_k), s_k\rangle + c_0 L\beta^2 R_k^2 \\ &\quad = \phi_k(R_k) - \phi_k(D_c) + qD_c^2 - (q - c_0 L\beta^2)R_k^2.\end{aligned}$$

This proves (23). **Q.E.D.**

Lemma B.4 (Pre-query one-dimensional comparison)**.** *Let $b_k = \langle G_k, s_k\rangle$. Then*

$$\mathbb{E}\sum_{k=1}^T w_k\big(\phi_k(R_k) - \phi_k(D_c)\big) \le qHD_c^2 + \frac{4\alpha\beta}{\kappa c_0}\sum_{k=1}^T w_k\mathbb{E}\Delta_k + 4\beta q\sum_{k=1}^T w_k\mathbb{E}R_k^2 + \frac{4\alpha s_\sigma^2}{\kappa c_0 L}. \tag{27}$$

Proof. Set $\widehat{\ell}_k(R) = w_k(\beta b_k R + qR^2)$ and

$$F_k(R) = qHR^2 + \sum_{i=1}^k \widehat{\ell}_i(R), \qquad F_0(R) = qHR^2.$$

The closed form (16) is precisely $R_{k+1} = \arg\min_{R\ge 0} F_k(R)$. The be-the-leader inequality gives

$$\sum_{k=1}^T \widehat{\ell}_k(R_{k+1}) \le qHR^2 + \sum_{k=1}^T \widehat{\ell}_k(R) \qquad (R \ge 0).$$

It remains to add the differences $\widehat{\ell}_k(R_k) - \widehat{\ell}_k(R_{k+1})$. Pathwise, this gives

$$\sum_{k=1}^T\big(\widehat{\ell}_k(R_k) - \widehat{\ell}_k(R)\big) \le qHR^2 + \frac{\beta^3}{q}\sum_{k=1}^T w_k b_k^2 + 4\beta q\sum_{k=1}^T w_k R_k^2. \tag{28}$$

Indeed, F_k is $2q(H + W_k)$-strongly convex. The variational inequalities for the constrained minimizers R_k and R_{k+1}, followed by completing the one-dimensional quadratic, bound the sum of those differences by

$$\sum_{k=1}^T \frac{w_k^2(\beta b_k + 2qR_k)^2}{q(H + W_k)}.$$

The same calculation applies at the boundary $R = 0$ because the one-sided derivative has the required sign. Using $H + W_k \ge 2w_k/\beta$ and expanding the square gives (28). The pre-query timing and Assumption 2.2 imply that the expected FTRL regret against the deterministic value D_c equals the expected regret formed with $\nabla f(x_k)$. By Lemma B.1,

$$\mathbb{E}[b_k^2 \mid \mathcal{F}_{k-1}] \le 4L\Delta_k + 2s_\sigma^2.$$

Now use $q = \kappa c_0 L\beta^2/\alpha$ and $\sum_k w_k \le 2/\beta$. **Q.E.D.**

B.3 Delayed Nesterov-Style Momentum Tracking

Put $\zeta_k = G_k - \nabla f(x_k)$, $q_m = 1 - \alpha$, and $a_i = \|x_i - x_{i-1}\|$. Define the EMA error $\epsilon_k^m = \|m_k - \nabla f(x_k)\|_*$.

Lemma B.5 (Weighted EMA error)**.** *Suppose $\beta \le \alpha/4$ and let $\delta = \alpha - \beta/2$. Then*

$$2\beta D_c \sum_{k=1}^{T} w_k \mathbb{E}\epsilon_k^m \le \frac{q}{4} \sum_{k=1}^{T} w_k \mathbb{E}R_k^2 + \frac{4608}{49\kappa c_0} \frac{L\beta D_c^2}{\alpha} + 4D_c s_\sigma \sqrt{\frac{\alpha}{2-\alpha}} + \frac{2\beta D_c s_\sigma}{\delta} r^T + \frac{4\beta L D_c \|x_0 - c\|}{\delta} r^T. \tag{29}$$

Proof. The EMA recursion gives, for $k \ge 2$,

$$\nabla f(x_k) - m_k = q_m(\nabla f(x_{k-1}) - m_{k-1}) + \nabla f(x_k) - \nabla f(x_{k-1}) - \alpha\zeta_{k-1}.$$

For $i < j$, the vector ζ_i is $\mathcal{F}_{j-1}$-measurable and $\mathbb{E}[\zeta_j \mid \mathcal{F}_{j-1}] = 0$. Hence $\mathbb{E}\langle \zeta_i, \zeta_j \rangle_2 = 0$. Unrolling the recursion, using this orthogonality in the inner product associated with $\|\cdot\|_2$, and then applying the norm comparison gives

$$\mathbb{E}\epsilon_k^m \le L \sum_{i=2}^{k} q_m^{k-i} \mathbb{E}a_i + s_\sigma q_m^{k-1} + s_\sigma \sqrt{\frac{\alpha}{2-\alpha}}. \tag{30}$$

The reference recursion implies

$$\sum_{i=2}^{T} w_i \mathbb{E}a_i \le 2\|x_0 - c\| r^{T-1} + \frac{3\beta}{r} \sum_{k=1}^{T} w_k \mathbb{E}R_k. \tag{31}$$

Since $q_m < r$, exchanging the convolution sums gives

$$\sum_{k=1}^{T} w_k \sum_{i=2}^{k} q_m^{k-i} \mathbb{E}a_i \le \frac{2\|x_0 - c\| r^T}{\delta} + \frac{3\beta}{\delta} \sum_{k=1}^{T} w_k \mathbb{E}R_k. \tag{32}$$

Also

$$\sum_{k=1}^{T} w_k q_m^{k-1} \le \frac{r^T}{\delta}, \qquad \sum_{k=1}^{T} w_k \le \frac{2}{\beta}.$$

Multiply (30) by $2\beta D_c w_k$ and sum. Weighted Cauchy–Schwarz and Young's inequality bound the radius-dependent term by

$$\frac{q}{4} \sum_{k=1}^{T} w_k \mathbb{E}R_k^2 + \frac{72\alpha L\beta D_c^2}{\kappa c_0 \delta^2}.$$

The condition $\beta \le \alpha/4$ gives $\delta \ge 7\alpha/8$, which proves (29). **Q.E.D.**

Lemma B.6 (Nesterov correction)**.** *Suppose $\beta \le \alpha/4$. Then*

$$2\beta D_c \sum_{k=1}^{T} w_k \mathbb{E}\epsilon_k^n \le 2\beta D_c \sum_{k=1}^{T} w_k \mathbb{E}\epsilon_k^m + \frac{q}{8} \sum_{k=1}^{T} w_k \mathbb{E}R_k^2 + \frac{144\alpha^3 \beta L D_c^2}{\kappa c_0 r^2} + 4\alpha D_c s_\sigma + 4\alpha\beta L D_c \|x_0 - c\| r^{T-1}. \tag{33}$$

Proof. At $k = 1$, $n_1 = m_1 = G_0$. For $k \ge 2$,

$$\epsilon_k^n = \|(1-\alpha)m_k + \alpha G_{k-1} - \nabla f(x_k)\|_* \tag{34}$$

$$\le \epsilon_k^m + \alpha \|\zeta_{k-1}\|_* + \alpha L a_k. \tag{35}$$

Jensen's inequality gives $\mathbb{E}\|\zeta_{k-1}\|_* \le s_\sigma$. Use (31) for the movement term. Its radius-dependent part is

$$\frac{6\alpha\beta^2 D_c L}{r} \sum_{k=1}^{T} w_k \mathbb{E}R_k.$$

Since $\sum_k w_k \le 2/\beta$, weighted Cauchy–Schwarz bounds this by

$$\frac{6\sqrt{2}\alpha\beta^{3/2}D_cL}{r}\left(\sum_{k=1}^{T} w_k \mathbb{E}R_k^2\right)^{1/2}.$$

Young's inequality with quadratic coefficient $q/8$ gives

$$\frac{q}{8}\sum_{k=1}^{T} w_k \mathbb{E}R_k^2 + \frac{144\alpha^3\beta LD_c^2}{\kappa c_0 r^2}.$$

The remaining noise and initialization terms are bounded, respectively, by

$$4\alpha D_c s_\sigma \qquad \text{and} \qquad 4\alpha\beta L D_c \left\|x_0 - c\right\| r^{T-1}.$$

Q.E.D.

B.4 Finite-Time Guarantee and Horizon-Dependent Schedule

Theorem B.7 (Last-iterate bound with delayed Nesterov-style momentum)**.** *Suppose Assumptions 2.1, 2.2, and 2.3 hold. Run* (15)–(18) *with*

$$0 < \alpha \le 1, \qquad 0 < \beta \le \min\left\{\frac{\alpha}{4}, \frac{1}{64}\right\}, \qquad \kappa > \max\left\{\frac{8\alpha}{c_0}, \frac{\alpha}{5/8 - 4\beta}\right\}. \tag{36}$$

Let

$$C_{\mathrm{ctrl}} = 3\kappa c_0 + \frac{4608}{49\kappa c_0}. \tag{37}$$

Then

$$\begin{aligned}\mathbb{E}\Delta_{T+1} \le{}& r^T(\Delta_1 + M\beta L B_c^2) + C_{\mathrm{ctrl}}\frac{L\beta B_c^2}{\alpha} + 4B_c s_\sigma\sqrt{\frac{\alpha}{2-\alpha}} \\ &+ \left(\frac{2\beta B_c s_\sigma}{\delta} + \frac{4\beta L B_c^2}{\delta}\right) r^T + \frac{4\alpha s_\sigma^2}{\kappa c_0 L} \\ &+ \frac{144\alpha^3\beta L B_c^2}{\kappa c_0 r^2} + 4\alpha B_c s_\sigma + 4\alpha\beta L B_c^2 r^{T-1}.\end{aligned} \tag{38}$$

No bounded-domain, bounded-gradient, bounded-iterate, bounded-path, or noise-tail assumption is used.

Proof. Lemma B.2 justifies all conditional expectations below.

Multiply (23) by w_k and sum. The Lyapunov terms telescope because $rw_k = w_{k-1}$ in the shifted sum. Apply Lemma B.4, then Lemmas B.5 and B.6. The radius coefficient on the left satisfies

$$q - c_0 L\beta^2 = q\left(1 - \frac{\alpha}{\kappa}\right).$$

The one-dimensional comparison uses $4\beta q$, the EMA bound uses $q/4$, and the Nesterov correction uses $q/8$. Hence the remaining radius coefficient is

$$q\left(\frac{5}{8} - 4\beta - \frac{\alpha}{\kappa}\right) > 0.$$

The weighted gap coefficient is

$$\beta\left(\frac{1}{2} - \frac{4\alpha}{\kappa c_0}\right) > 0.$$

Both inequalities follow from (36). Drop these and the other nonnegative terms. Finally,

$$H + \sum_{k=1}^{T} w_k \le \frac{3}{\beta},$$

The penalty at the fixed radius D_c therefore satisfies

$$3\kappa c_0 L\beta D_c^2/\alpha \le 3\kappa c_0 L\beta B_c^2/\alpha.$$

We also use $\|x_0 - c\| \le B_c$ and $D_c \|x_0 - c\| \le B_c^2$. Collecting the remaining terms proves (38). **Q.E.D.**

Corollary B.8 (Reduced horizon-dependent shrinkage)**.** *Set*

$$\boxed{\beta_T = \min\left\{\frac{1}{512}, \frac{2\log(T+1)}{3T}\right\}, \qquad \alpha_T = 8\beta_T^{2/3}, \qquad \kappa = \frac{1}{4}.} \tag{39}$$

Then (36) *holds for every* $T \geq 1$*. Whenever the logarithmic branch of* (39) *is active,*

$$\begin{aligned}\mathbb{E}\Delta_{T+1} \leq{}& (2 + M\beta_T)LB_c^2(T+1)^{-1/3} + \frac{C_{\mathrm{ctrl}}}{8}LB_c^2\beta_T^{1/3} + 8\sqrt{2}B_c s_\sigma \beta_T^{1/3} \\ &+ \left(\frac{32}{127}B_c s_\sigma + \frac{64}{127}LB_c^2\right)\beta_T^{1/3}(T+1)^{-1/3} + \frac{32}{\kappa c_0}\frac{s_\sigma^2}{L}\beta_T^{2/3} \\ &+ \frac{73728}{\kappa c_0 r^2}LB_c^2\beta_T^3 + 32B_c s_\sigma \beta_T^{2/3} + 32LB_c^2\beta_T^{5/3}r^{T-1}.\end{aligned} \tag{40}$$

Consequently,

$$\mathbb{E}[f(x_{T+1}) - f(x^\star)] = \widetilde{O}\left(\frac{LB_c^2 + B_c s_\sigma}{T^{1/3}} + \frac{s_\sigma^2/L}{T^{2/3}}\right). \tag{41}$$

The algorithm receives neither B_c *nor* s_σ*.*

Proof. Write $z = \beta_T^{1/3}$. Since $\beta_T \leq 1/512$, we have $z \leq 1/8$, $\alpha_T = 8z^2 \leq 1/8$, and

$$\delta = \alpha_T - \frac{\beta_T}{2} = \alpha_T\left(1 - \frac{z}{16}\right) \geq \frac{127}{128}\alpha_T.$$

Moreover,

$$\frac{8\alpha_T}{c_0} \leq \frac{1}{9} < \frac{1}{4}, \qquad \frac{\alpha_T}{5/8 - 4\beta_T} \leq \frac{16}{79} < \frac{1}{4},$$

so the condition on κ in (36) holds. Also

$$\beta_T \leq \frac{\alpha_T}{4}, \qquad \frac{\beta_T}{\delta} \leq \frac{16}{127}\beta_T^{1/3}, \qquad \frac{\beta_T}{\alpha_T} = \frac{1}{8}\beta_T^{1/3}, \qquad \sqrt{\frac{\alpha_T}{2 - \alpha_T}} \leq \sqrt{8}\,\beta_T^{1/3}.$$

Lemma B.1 gives $\Delta_1 \leq 2LB_c^2$. If the logarithmic branch is active, then

$$r^T \leq \exp\left(-\frac{\beta_T T}{2}\right) = (T+1)^{-1/3}.$$

Substitute these inequalities and $\alpha_T = 8\beta_T^{2/3}$ into Theorem B.7. This gives (40). The last three terms are of smaller order than the displayed leading terms, which proves (41). **Q.E.D.**

B.5 Connection to the reported implementations

The exact matrix diagnostic in Appendix H.1 uses the proved update with $c = x_0 = 0$, the stated horizon-dependent schedule, and an exact LMO. The Transformer implementation retains weighted FTRL and the delayed direction, but uses an additive update and the training components listed in Appendix H.5. Its fixed matrix weight decay is not the adaptive coefficient $1/R_k$ in (19). The last-iterate theorem is not asserted for that implementation.

C Full Results for the Predictable Non-Convex Method

This appendix gives the full algorithm, constants, stability estimates, and schedules behind Theorem 4.1. For a general smooth non-convex objective, we keep the trust-region center at the current iterate and adapt the stepsize along the normalized direction. We retain the DF label for this FTRL-based stepsize adaptation because it uses the same predictable one-dimensional mechanism as the star-convex method. The fixed-stepsize non-convex baseline of Kovalev (2025) already has no distance parameter.

At round k, the direction $s_k = s(m_k)$ and stepsize η_k are measurable with respect to the history before the k-th oracle call. We therefore call the update predictable. Algorithm 1 of Kovalev (2025)

instead steps with the freshly updated m_{k+1}. Our timing permits one fresh gradient per round and keeps the one-dimensional comparison unbiased.

Throughout this section, assume $f^\star := \inf_x f(x) > -\infty$. The infimum need not be attained. Set $s_\sigma := \rho_{\mathrm{norm}}\sigma$. We assume neither star-convexity nor the existence of a minimizer. In particular, no distance to a minimizer is used. We also make no bounded-gradient, bounded-domain, or bounded-trajectory assumption.

C.1 Predictable Adaptive Stepsize

Fix a horizon $T \geq 1$, momentum parameter $\alpha \in (0, 1]$, and the regularization offset

$$H_0 := 32, \qquad h_T := \log\left(1 + \frac{T}{H_0}\right).$$

Choose the quadratic regularization coefficient

$$q := \frac{L}{2} + \max\left\{\frac{\sqrt{2}L}{\alpha}, 4Lh_T\right\}, \qquad c := q - \frac{L}{2} > 0. \tag{42}$$

The first term in the maximum balances momentum drift. The logarithmic term ensures the stability bound for the quadratic feedback term without a bounded-gradient assumption. The value $H_0 = 32$ is a sufficient universal proof constant rather than a problem parameter. Because q depends on T, the stated method is horizon-dependent rather than anytime.

For realized one-dimensional feedback

$$b_k := \langle G_k, s_k \rangle,$$

define the quadratic surrogate loss

$$\ell_k(\eta) := b_k\eta + q\eta^2, \qquad \eta \geq 0. \tag{43}$$

The adaptive stepsize rule is FTRL with initial regularizer $qH_0\eta^2$. If $A_{k-1} := \sum_{i=1}^{k-1} b_i$, its closed-form prediction is

$$\eta_k = \left(-\frac{A_{k-1}}{2q(H_0 + k - 1)}\right)_+. \tag{44}$$

Thus η_k is fixed before G_k is drawn.

Algorithm 3 Predictable non-convex DF-Muon

Require: $x_0 \in \mathcal{E}$, horizon T, momentum $\alpha \in (0, 1]$, smoothness L

Draw $G_0 = g(x_0; \xi_0)$. Set $x_1 \leftarrow x_0$, $m_1 \leftarrow G_0$, and $A_0 \leftarrow 0$
Set $H_0 = 32$ and compute q by (42)
for $k = 1, 2, \ldots, T$ **do**
  Choose the direction $s_k \leftarrow s(m_k)$
  Compute the predictable stepsize η_k by (44)
  Store the proposal $x_{k+1} = x_k + \eta_k s_k$
  Draw the only fresh gradient of the round, $G_k = g(x_k; \xi_k)$
  Apply the stored update $x_k \mapsto x_{k+1}$
  Set $m_{k+1} \leftarrow (1 - \alpha)m_k + \alpha G_k$
  Set $A_k \leftarrow A_{k-1} + \langle G_k, s_k \rangle$
end for
Draw τ uniformly from $\{1, \ldots, T\}$, independently of the oracle, and return $\widehat{x} = x_\tau$

The algorithm uses one initialization gradient and one new stochastic gradient in each displayed round. The first stepsize is $\eta_1 = 0$, so the first update is zero. The total number of stochastic-gradient evaluations is $T + 1$.

Remark C.1 (No β and no recentering). The non-convex update is centered at x_k. It is not obtained by naively setting $\beta = 0$ in the star-convex DF-Muon recursion, which would freeze that recursion. Rather, the scalar variable η is the stepsize along the normalized direction. It coincides with the trust-region radius in Algorithm 1 of Kovalev (2025). Consequently β, the recentered ray, and any distance-based fixed choice D disappear from both the algorithm and the proof.

C.1.1 STATIONARITY GUARANTEES

Let

$$\Delta_0 := f(x_0) - f^\star$$

and define the quadratic feedback term

$$\mathcal{S}_T := \frac{2}{q} \sum_{k=1}^{T} \frac{b_k^2}{H_0 + k}. \tag{45}$$

This quantity is introduced only for the analysis. It is not an input to the algorithm.

Theorem C.2 (Oracle inequality for every fixed stepsize)**.** *Let Assumptions 2.1 and 2.2 hold, assume $f^\star > -\infty$, and run Algorithm 3. Define*

$$C_{\alpha,T} := q\left(1 + \frac{H_0}{T}\right) + \frac{2L^2}{c\alpha^2}. \tag{46}$$

Then, for every deterministic fixed stepsize $\eta > 0$,

$$\frac{1}{T}\sum_{k=1}^{T} \mathbb{E}\left\|\nabla f(x_k)\right\|_* \le \frac{\Delta_0 + \mathbb{E}\mathcal{S}_T}{\eta T} + C_{\alpha,T}\eta + \frac{2s_\sigma}{\alpha T} + 2s_\sigma\sqrt{\frac{\alpha}{2-\alpha}}. \tag{47}$$

Consequently, minimizing the right-hand side over fixed stepsizes gives

$$\mathbb{E}\left\|\nabla f(\widehat{x})\right\|_* \le 2\sqrt{\frac{C_{\alpha,T}(\Delta_0 + \mathbb{E}\mathcal{S}_T)}{T}} + \frac{2s_\sigma}{\alpha T} + 2s_\sigma\sqrt{\frac{\alpha}{2-\alpha}}. \tag{48}$$

The algorithm does not use the value of η that minimizes this bound. If $\Delta_0 + \mathbb{E}\mathcal{S}_T = 0$, the optimized expression is understood as the infimum obtained when $\eta \downarrow 0$.

The next theorem bounds the only data-dependent term in Theorem C.2 under the bounded-variance oracle itself. No bounded gradient assumption is added.

Theorem C.3 (Stability bound for the quadratic feedback term)**.** *Under the conditions of Theorem C.2, let*

$$\Gamma_0 := \Delta_0 + \frac{s_\sigma^2}{2L}, \qquad \theta := \frac{4L}{q}.$$

Then

$$\begin{aligned}\mathbb{E}\mathcal{S}_T &\le \Gamma_0\left[\prod_{k=1}^{T}\left(1 + \frac{\theta}{H_0 + k}\right) - 1\right] \\ &\le \Gamma_0\left[\left(1 + \frac{T}{H_0}\right)^{\theta} - 1\right] \le (e-1)\Gamma_0.\end{aligned} \tag{49}$$

The last inequality follows from $q \ge 4Lh_T$, which is built into (42)*. Moreover, with*

$$\kappa := 2\sqrt{2}(e-1),$$

the same argument gives the sharper momentum-dependent estimate

$$\mathbb{E}\mathcal{S}_T \le \Gamma_0 \min\left\{e - 1, \kappa\alpha h_T\right\}. \tag{50}$$

Combining the two theorems gives a completely explicit bound. Define

$$\Xi_0 := e\Delta_0 + \frac{e-1}{2L}s_\sigma^2. \tag{51}$$

Corollary C.4 (Explicit predictable non-convex bound)**.** *Under Assumptions 2.1 and 2.2, Algorithm 3 satisfies*

$$\mathbb{E}\left\|\nabla f(\widehat{x})\right\|_* \le 2\sqrt{\frac{C_{\alpha,T}\Xi_0}{T}} + \frac{2s_\sigma}{\alpha T} + 2s_\sigma\sqrt{\frac{\alpha}{2-\alpha}}. \tag{52}$$

The same upper bound also holds for expected best-iterate stationarity (an analysis quantity), since

$$\mathbb{E}\min_{1\le k\le T}\left\|\nabla f(x_k)\right\|_* \le \min_{1\le k\le T}\mathbb{E}\left\|\nabla f(x_k)\right\|_* \le \mathbb{E}\left\|\nabla f(\widehat{x})\right\|_*.$$

Corollary C.5 (Horizon-dependent stochastic schedule)**.** *Suppose* $T \geq 16$ *and set*

$$\alpha = T^{-1/2}.$$

Then the curvature in (42) *is*

$$q = \frac{L}{2} + \sqrt{2}L\sqrt{T},$$

and $C_{\alpha,T} \leq 7L\sqrt{T}$*. Consequently,*

$$\mathbb{E}\,\|\nabla f(\widehat{x})\|_* \leq 2\sqrt{7L\Xi_0}\,T^{-1/4} + 2s_\sigma T^{-1/4} + 2s_\sigma T^{-1/2}. \tag{53}$$

In particular,

$$\mathbb{E}\,\|\nabla f(\widehat{x})\|_* = O\Big(\big(\sqrt{L\Delta_0} + \rho_{\mathrm{norm}}\sigma\big)T^{-1/4}\Big),$$

which gives an $O(\varepsilon^{-4})$ *iteration complexity for stochastic non-convex optimization. This dependence is known to be unimprovable in general under an unbiased bounded-variance oracle (Arjevani et al., 2023). More explicitly, with*

$$B_0 := 2\sqrt{7L\Xi_0} + 4s_\sigma,$$

any integer

$$T \geq \max\left\{16, \left\lceil (B_0/\varepsilon)^4 \right\rceil\right\}$$

ensures $\mathbb{E}\,\|\nabla f(\widehat{x})\|_* \leq \varepsilon$ *and uses* $T+1$ *stochastic-gradient evaluations. For a prescribed horizon, the update uses* T *and* L*, but not* D*,* σ*,* Δ_0*, or* ε *directly. Selecting a horizon from an* ε*-guarantee of course uses the displayed complexity relation.*

Corollary C.6 (Accuracy- and noise-tuned stochastic schedule)**.** *Let* $T \geq 32$*, let* $\varepsilon > 0$*, and write* $s := s_\sigma$*. Choose*

$$\alpha_{\varepsilon,T} := \min\left\{1, \frac{\varepsilon^2}{16s^2}, \frac{1}{2\sqrt{2}h_T}\right\}, \tag{54}$$

where the middle entry is interpreted as $+\infty$ *when* $s = 0$*. There is a universal numerical constant* K_0 *such that*

$$T \geq K_0 \max\left\{1, \frac{L\Delta_0}{\alpha_{\varepsilon,T}\varepsilon^2}, \frac{s}{\alpha_{\varepsilon,T}\varepsilon}, \frac{s^2 h_T}{\varepsilon^2}\right\} \tag{55}$$

implies

$$\mathbb{E}\,\|\nabla f(\widehat{x})\|_* \leq \varepsilon.$$

Consequently, up to logarithmic factors, it is sufficient to use

$$T = \widetilde{O}\left(\max\left\{1, \frac{s}{\varepsilon}, \frac{s^3}{\varepsilon^3}, \frac{L\Delta_0}{\varepsilon^2}, \frac{L\Delta_0 s^2}{\varepsilon^4}\right\}\right). \tag{56}$$

Thus, in the unregularized case, the one-gradient predictable method matches all four parameter-dependent terms in the oracle complexity for expected stationarity of the fixed-stepsize method in Kovalev (2025, Corollary 2), up to logarithmic factors. This is a rate comparison, not an equivalence of the realized algorithms: the direction timing differs, and the present result does not cover a nonsmooth regularizer. Its oracle complexity is $T+1$*. Unlike Corollary C.5, this schedule uses* s_σ *and the target accuracy.*

Corollary C.7 (Deterministic special case)**.** *If* $\sigma = 0$ *and* $\alpha \in (0, 1]$ *is fixed, then*

$$\min_{1\leq k\leq T} \|\nabla f(x_k)\|_* = O_\alpha\left(\sqrt{\frac{L\Delta_0(1+\log(T+1))}{T}}\right). \tag{57}$$

Thus the same adaptive stepsize rule has the usual $\widetilde{O}(T^{-1/2})$ *deterministic stationarity order without a trajectory-bound assumption. Relative to the optimized fixed-stepsize bound, the present rule carries a* $\sqrt{\log(T+1)}$ *overhead.*

C.1.2 Comparison and scope

For the unregularized case, Algorithm 1 of Kovalev (2025) uses a prescribed fixed stepsize and already requires neither a minimizer nor a distance to one. The contribution here is instead an online oracle inequality indexed by every deterministic fixed stepsize. It is an inequality for the derived stationarity upper bound, not a pathwise comparison with the trajectory of the best fixed-stepsize run in hindsight.

The horizon-dependent schedule in Corollary C.5 attains the worst-case $O(\varepsilon^{-4})$ iteration complexity without using D, σ, Δ_0, or ε in the update. If the noise level and target accuracy are known, Corollary C.6 matches all parameter-dependent terms in Kovalev (2025, Corollary 2) up to logarithmic factors. This comparison concerns oracle complexity for achieving a prescribed expected gradient norm. The momentum timing differs, and our result does not cover a nonsmooth regularizer. Since the method uses L and the prescribed horizon T, it is not fully parameter-free or anytime.

DA-Muon adapts from the explored trajectory radius and uses a bounded-trajectory condition in its analysis. The current-centered distance-free method instead adapts from one-dimensional feedback and needs no trajectory bound, but its quadratic regularization coefficient depends on the horizon. Both methods use one sample per round, without sample splitting.

D Proofs for Predictable Non-Convex DF-Muon

Let

$$\mathcal{F}_k := \sigma(G_0, G_1, \dots, G_k), \qquad \zeta_k := G_k - \nabla f(x_k).$$

Then x_k, m_k, s_k, η_k are $\mathcal{F}_{k-1}$-measurable and

$$\mathbb{E}[\zeta_k \mid \mathcal{F}_{k-1}] = 0, \qquad \mathbb{E}[\|\zeta_k\|_*^2 \mid \mathcal{F}_{k-1}] \le s_\sigma^2. \tag{58}$$

D.1 Basic geometric estimates

Lemma D.1 (Self-bounding gradient)**.** *Under Assumption 2.1, if $f^\star := \inf_x f(x) > -\infty$, then for every $x \in \mathcal{E}$,*

$$\|\nabla f(x)\|_*^2 \le 2L\big(f(x) - f^\star\big). \tag{59}$$

The infimum need not be attained.

Proof of Lemma D.1. If $\nabla f(x) = 0$, the claim is immediate. Otherwise choose a norming direction u with $\|u\| = 1$ and $\langle \nabla f(x), u\rangle = \|\nabla f(x)\|_*$. Smoothness at

$$y = x - \frac{\|\nabla f(x)\|_*}{L} u$$

gives

$$f(y) \le f(x) - \frac{\|\nabla f(x)\|_*^2}{2L}.$$

Since $f^\star \le f(y)$, rearrangement proves (59). Attainment of $f^\star$ is not used. **Q.E.D.**

Lemma D.2 (LMO alignment)**.** *Let*

$$e_k := \|m_k - \nabla f(x_k)\|_* .$$

Then

$$\langle \nabla f(x_k), s_k\rangle \le -\|\nabla f(x_k)\|_* + 2e_k. \tag{60}$$

Proof. By the LMO identity and duality,

$$\begin{aligned}\langle \nabla f(x_k), s_k\rangle &= -\|m_k\|_* + \langle \nabla f(x_k) - m_k, s_k\rangle \\ &\le -\|m_k\|_* + e_k \le -\|\nabla f(x_k)\|_* + 2e_k.\end{aligned}$$

Q.E.D.

Lemma D.3 (Finite moments)**.** *For every finite k, the random variables $\|x_k - x_0\|^2$, $\|G_k\|_*^2$, $\|m_k\|_*^2$, b_k^2, and η_k^2 have finite expectation. Hence all conditional expectations and tower-property steps below are legitimate.*

Proof. The claim follows by induction. It holds at the deterministic point x_0 by Assumption 2.2, since

$$\mathbb{E}\,\|G_0\|_*^2 \le 2\,\|\nabla f(x_0)\|_*^2 + 2s_\sigma^2 < \infty.$$

If it holds through round $k-1$, then A_{k-1} is a finite sum of square-integrable one-dimensional feedback values, so the explicit formula (44) gives $\mathbb{E}\eta_k^2 < \infty$. Since $\|s_k\| \le 1$, the update gives $\|x_{k+1} - x_0\| \le \|x_k - x_0\| + \eta_k$. Smoothness implies $\|\nabla f(x_{k+1})\|_* \le \|\nabla f(x_0)\|_* + L\,\|x_{k+1} - x_0\|$, and the oracle variance bound then gives a finite second moment for G_{k+1}. The momentum recursion, duality $|b_k| \le \|G_k\|_*$, and another application of $(a+b)^2 \le 2a^2 + 2b^2$ complete the induction. **Q.E.D.**

D.2 Pathwise FTRL regret

Lemma D.4 (Quadratic FTRL regret)**.** *For every realized sequence $b_1, \dots, b_T \in \mathbb{R}$, every $t \le T$, and every fixed $\eta \ge 0$,*

$$\sum_{k=1}^{t} \big(\ell_k(\eta_k) - \ell_k(\eta)\big) \le qH_0\eta^2 + \sum_{k=1}^{t} \frac{(b_k + 2q\eta_k)^2}{q(H_0 + k)} \tag{61}$$

$$\le qH_0\eta^2 + \mathcal{S}_t + \frac{8q}{H_0 + 1} \sum_{k=1}^{t} \eta_k^2, \tag{62}$$

where

$$\mathcal{S}_t := \frac{2}{q} \sum_{k=1}^{t} \frac{b_k^2}{H_0 + k}.$$

For the choices $H_0 = 32$ and (42),

$$\frac{8q}{H_0 + 1} \le \frac{c}{2}. \tag{63}$$

Proof. Define

$$F_k(\eta) := qH_0\eta^2 + \sum_{i=1}^{k} \ell_i(\eta), \qquad \eta \ge 0.$$

The explicit update (44) is precisely $\eta_k \in \arg\min_{\eta \ge 0} F_{k-1}(\eta)$, and η_{k+1} minimizes F_k. The standard be-the-leader inequality gives

$$\sum_{k=1}^{t} \big(\ell_k(\eta_k) - \ell_k(\eta)\big) \le qH_0\eta^2 + \sum_{k=1}^{t} \big(\ell_k(\eta_k) - \ell_k(\eta_{k+1})\big). \tag{64}$$

The function F_k is $2q(H_0 + k)$-strongly convex. Its variational inequality at the constrained minimizer η_{k+1}, evaluated at η_k, and the optimality of η_k for F_{k-1} imply

$$q(H_0 + k)(\eta_k - \eta_{k+1})^2 \le \ell_k(\eta_k) - \ell_k(\eta_{k+1}).$$

On the other hand, quadratic expansion gives

$$\ell_k(\eta_k) - \ell_k(\eta_{k+1}) \le |b_k + 2q\eta_k|\,|\eta_k - \eta_{k+1}|.$$

Therefore

$$|\eta_k - \eta_{k+1}| \le \frac{|b_k + 2q\eta_k|}{q(H_0 + k)}.$$

Substitution into (64) proves (61). The inequality $(a+b)^2 \le 2a^2 + 2b^2$ and $H_0 + k \ge H_0 + 1$ give (62).

Finally, $c \geq \sqrt{2}L/\alpha$ and

$$\frac{q}{c} = 1 + \frac{L}{2c} \leq 1 + \frac{\alpha}{2\sqrt{2}} \leq 1 + \frac{1}{2\sqrt{2}}.$$

Thus

$$\frac{16q}{c} \leq 16\left(1 + \frac{1}{2\sqrt{2}}\right) < 33 = H_0 + 1,$$

which is equivalent to (63). **Q.E.D.**

D.3 Objective and quadratic-feedback stability

Lemma D.5 (Prefix objective control)**.** *Set $\mathcal{S}_0 := 0$. For every $t \in \{0, 1, \ldots, T\}$,*

$$\mathbb{E}\big[f(x_{t+1}) - f^\star\big] + \frac{c}{2}\sum_{k=1}^{t} \mathbb{E}\eta_k^2 \leq \Delta_0 + \mathbb{E}\mathcal{S}_t. \tag{65}$$

Proof. For $t = 0$, the claim is the identity $f(x_1) - f^\star = \Delta_0$. Fix now $t \geq 1$. Because x_k, s_k, η_k are $\mathcal{F}_{k-1}$-measurable, conditional unbiasedness gives

$$\mathbb{E}[\eta_k b_k \mid \mathcal{F}_{k-1}] = \eta_k \langle \nabla f(x_k), s_k \rangle.$$

Smoothness and $\|s_k\| \leq 1$ therefore imply

$$\begin{aligned}\mathbb{E}[f(x_{k+1}) - f(x_k)] &\leq \mathbb{E}\left[\eta_k b_k + \frac{L}{2}\eta_k^2\right] \\ &= \mathbb{E}\left[\ell_k(\eta_k) - c\eta_k^2\right].\end{aligned}$$

Sum through t and apply Lemma D.4 with $\eta = 0$. The final stability term in (62) is absorbed using (63), which proves (65). **Q.E.D.**

Proof of Theorem C.3. Conditional on $\mathcal{F}_{k-1}$,

$$\begin{aligned}\mathbb{E}[b_k^2 \mid \mathcal{F}_{k-1}] &= \langle \nabla f(x_k), s_k \rangle^2 + \mathbb{E}[\langle \zeta_k, s_k \rangle^2 \mid \mathcal{F}_{k-1}] \\ &\leq \|\nabla f(x_k)\|_*^2 + s_\sigma^2 \\ &\leq 2L(f(x_k) - f^\star) + s_\sigma^2,\end{aligned}$$

where the conditional cross term vanishes by (58) and the predictability of s_k, and Lemma D.1 was used last. Lemma D.5 with $t = k - 1$ gives

$$\mathbb{E}b_k^2 \leq 2L\big(\Delta_0 + \mathbb{E}\mathcal{S}_{k-1}\big) + s_\sigma^2.$$

Consequently,

$$\mathbb{E}\mathcal{S}_k - \mathbb{E}\mathcal{S}_{k-1} \leq \frac{4L/q}{H_0 + k}\left(\Delta_0 + \frac{s_\sigma^2}{2L} + \mathbb{E}\mathcal{S}_{k-1}\right).$$

With $U_k := \Gamma_0 + \mathbb{E}\mathcal{S}_k$ and $\theta = 4L/q$, this becomes

$$U_k \leq U_{k-1}\left(1 + \frac{\theta}{H_0 + k}\right), \qquad U_0 = \Gamma_0.$$

Iteration proves the product bound. Moreover,

$$\sum_{k=1}^{T} \frac{1}{H_0 + k} \leq \log\left(1 + \frac{T}{H_0}\right) = h_T,$$

so the product is at most $e^{\theta h_T}$. Since $q \geq 4Lh_T$, $\theta h_T \leq 1$, proving the remaining inequalities in (49).

For $u \in [0, 1]$, convexity of the exponential gives $e^u - 1 \leq (e - 1)u$. In addition, $q \geq \sqrt{2}L/\alpha$ implies

$$\theta h_T = \frac{4Lh_T}{q} \leq 2\sqrt{2}\alpha h_T.$$

Combining this estimate with both the exponential product bound and the coarse bound $e^{\theta h_T} - 1 \leq e - 1$ proves (50). **Q.E.D.**

D.4 MOMENTUM TRACKING AND STATIONARITY

Lemma D.6 (Predictable momentum tracking)**.** *For Algorithm 3,*

$$\sum_{k=1}^{T} \mathbb{E}e_k \le \frac{s_\sigma}{\alpha} + T s_\sigma \sqrt{\frac{\alpha}{2-\alpha}} + \frac{L}{\alpha} \sum_{k=1}^{T} \mathbb{E}\eta_k. \tag{66}$$

Proof. Set $r_m := 1 - \alpha$. The initialization gives $m_1 - \nabla f(x_1) = \zeta_0$. For $k \ge 1$,

$$m_{k+1} - \nabla f(x_{k+1}) = r_m\big(m_k - \nabla f(x_k)\big) + \nabla f(x_k) - \nabla f(x_{k+1}) + \alpha\zeta_k.$$

Unrolling this recursion, applying smoothness to its drift terms, and using orthogonality of Euclidean martingale differences gives

$$\mathbb{E}e_k \le L \sum_{i=1}^{k-1} r_m^{k-1-i} \mathbb{E}\eta_i + s_\sigma r_m^{k-1} + s_\sigma \sqrt{\frac{\alpha}{2-\alpha}}. \tag{67}$$

Indeed, the variance of the moving-average noise is bounded by

$$\sigma^2 \left(r_m^{2(k-1)} + \alpha^2 \sum_{j=0}^{k-2} r_m^{2j} \right) \le \sigma^2 \left(r_m^{2(k-1)} + \frac{\alpha}{2-\alpha} \right),$$

after which the norm comparison and Jensen's inequality apply. Sum (67) over k, use $\sum_{j\ge 0} r_m^j = 1/\alpha$, and enlarge the final drift sum to include η_T. **Q.E.D.**

Proof of Theorem C.2. The proof of Lemma D.5, now followed by the FTRL regret bound against an arbitrary fixed $\eta > 0$, gives

$$\mathbb{E}[f(x_{T+1}) - f(x_1)] + \frac{c}{2} \sum_{k=1}^{T} \mathbb{E}\eta_k^2 \le \eta \sum_{k=1}^{T} \mathbb{E}b_k + q(T + H_0)\eta^2 + \mathbb{E}\mathcal{S}_T.$$

By conditional unbiasedness and Lemma D.2,

$$\mathbb{E}b_k \le -\mathbb{E}\left\|\nabla f(x_k)\right\|_* + 2\mathbb{E}e_k.$$

Using $f(x_{T+1}) \ge f^\star$ and Lemma D.6, we obtain

$$\eta \sum_{k=1}^{T} \mathbb{E}\left\|\nabla f(x_k)\right\|_* + \frac{c}{2} \sum_{k=1}^{T} \mathbb{E}\eta_k^2 \le \Delta_0 + q(T + H_0)\eta^2 + \mathbb{E}\mathcal{S}_T + \frac{2\eta s_\sigma}{\alpha} + 2\eta T s_\sigma \sqrt{\frac{\alpha}{2-\alpha}} + \frac{2\eta L}{\alpha} \sum_{k=1}^{T} \mathbb{E}\eta_k.$$

For every k, Young's inequality gives

$$\frac{2\eta L}{\alpha}\eta_k \le \frac{c}{2}\eta_k^2 + \frac{2\eta^2 L^2}{c\alpha^2}.$$

Absorb the quadratic energy, divide by ηT, and use the definition (46) to prove (47). Minimizing $a/\eta + C_{\alpha,T}\eta$ over $\eta > 0$, with $a = (\Delta_0 + \mathbb{E}\mathcal{S}_T)/T$, proves (48). **Q.E.D.**

Proof of Corollary C.4. Theorem C.3 implies

$$\Delta_0 + \mathbb{E}\mathcal{S}_T \le e\Delta_0 + \frac{e-1}{2L} s_\sigma^2 = \Xi_0.$$

Substitute this into (48). Uniform sampling gives

$$\mathbb{E}\left\|\nabla f(\widehat{x})\right\|_* = \frac{1}{T} \sum_{k=1}^{T} \mathbb{E}\left\|\nabla f(x_k)\right\|_*.$$

The two minimum statements follow from $\mathbb{E}\min_k Z_k \le \min_k \mathbb{E}Z_k \le T^{-1}\sum_k \mathbb{E}Z_k$ for nonnegative Z_k. **Q.E.D.**

Proof of Corollary C.5. For $T \geq 16$ and $\alpha = T^{-1/2}$, the elementary inequality

$$4 \log\left(1 + \frac{T}{32}\right) \leq \sqrt{2T}$$

shows that the first branch in the maximum in (42) is active. Thus $c = \sqrt{2}L\sqrt{T}$ and

$$C_{\alpha,T} = \left(\frac{L}{2} + \sqrt{2}L\sqrt{T}\right)\left(1 + \frac{32}{T}\right) + \sqrt{2}L\sqrt{T} \leq 7L\sqrt{T}.$$

Moreover,

$$2s_\sigma\sqrt{\frac{\alpha}{2-\alpha}} \leq 2s_\sigma T^{-1/4}, \qquad \frac{2s_\sigma}{\alpha T} = 2s_\sigma T^{-1/2}.$$

Substitute these estimates into Corollary C.4. **Q.E.D.**

Proof of Corollary C.6. Write $\alpha := \alpha_{\varepsilon,T}$ and $\kappa = 2\sqrt{2}(e-1)$. By construction, $\alpha h_T \leq 1/(2\sqrt{2})$, and hence the first branch in the maximum in (42) is active. Since $T \geq 32$,

$$C_{\alpha,T} \leq 2\left(\frac{L}{2} + \frac{\sqrt{2}L}{\alpha}\right) + \frac{\sqrt{2}L}{\alpha} \leq \frac{6L}{\alpha}.$$

The refined stability estimate (50) and $\kappa\alpha h_T \leq e - 1$ give

$$\Delta_0 + \mathbb{E}\mathcal{S}_T \leq e\Delta_0 + \frac{\kappa\alpha h_T}{2L}s^2.$$

Substitution into (48), followed by $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$, yields

$$\mathbb{E}\left\|\nabla f(\widehat{x})\right\|_* \leq A\sqrt{\frac{L\Delta_0}{\alpha T}} + Bs\sqrt{\frac{h_T}{T}} + \frac{2s}{\alpha T} + 2s\sqrt{\alpha}, \tag{68}$$

where

$$A := 2\sqrt{6e}, \qquad B := 2\sqrt{3\kappa}.$$

The choice (54) makes the last term at most $\varepsilon/2$. A universal constant

$$K_0 \geq \max\{36A^2, 36B^2, 12\}$$

in (55) makes each of the first three terms in (68) at most $\varepsilon/6$, proving the claimed accuracy.

Finally,

$$\frac{1}{\alpha} = \max\left\{1, \frac{16s^2}{\varepsilon^2}, 2\sqrt{2}h_T\right\}.$$

Expanding (55) therefore produces the four non-logarithmic terms in (56), together with h_T times terms already present there. The latter are absorbed by the $\widetilde{O}$ notation. For example, $(s/\varepsilon)^2 \leq \max\{s/\varepsilon, s^3/\varepsilon^3\}$. More explicitly, if M_ε denotes the maximum in (56), then a sufficiently large universal multiple of $M_\varepsilon[1 + \log(2 + M_\varepsilon)]^2$ satisfies the implicit condition (55). This resolves the occurrence of $h_T = \log(1 + T/H_0)$ and proves the stated $\widetilde{O}$ claim. **Q.E.D.**

Proof of Corollary C.7. When $s_\sigma = 0$, $\Xi_0 = e\Delta_0$ and the two noise terms in (52) vanish. For fixed α, (42) and (46) give

$$C_{\alpha,T} = O_\alpha\big(L(1 + \log(T+1))\big).$$

Apply Corollary C.4. Since the oracle and iterates are then deterministic, the expectation signs may be removed. **Q.E.D.**

E Full-Gradient Counterpart

With exact gradients, the radius decision may use the current linear term. The deterministic counterpart keeps the recentered ray

$$z_k(R) = x_0 + (1-\beta)(x_k - x_0) + \beta R s_k$$

and selects $R_k \geq 0$ by minimizing a regularized smoothness upper model along this ray. The model is a linear term plus squared norms of affine functions of R and a positive quadratic regularizer. It is therefore a convex one-dimensional problem. Unlike an objective-value line search, it requires no trial evaluations of $f(z_k(R))$.

The method also maintains a D-adaptation-style lower certificate $d_k \leq D$, but the proof does not require $d_k \to D$. The theorem permits any nonnegative certificate weights. The canonical fully specified choice is $\omega_k \equiv 1$. The radius search is over all $R \geq 0$, so its one-step inequality may be evaluated at the fixed choice $R = D$. Appendix F gives the exact majorant and update.

Theorem E.1 (Deterministic distance-free last iterate, concrete schedule)**.** *Under Assumptions 2.1 and 2.3, run deterministic DF-Muon with*

$$\alpha = \frac{1}{2}, \quad \rho = \lambda = 1, \quad M = 6, \quad \omega_k = 1, \quad \beta = \min\left\{\frac{1}{2}, \frac{2\log(T+1)}{T}\right\}.$$

For $T \geq 10$,

$$f(x_T) - f(x^\star) \leq \frac{LD^2}{2(T+1)} + 70LD^2\frac{\log(T+1)}{T} = \widetilde{O}\left(\frac{LD^2}{T}\right),$$

where $D = \|x_0 - x^\star\|$. The algorithm does not use $D = \|x_0 - x^\star\|$ and assumes neither bounded iterates, bounded gradients, a bounded domain, nor a bounded initial sublevel set.

The deterministic and stochastic methods share the LMO over the unit norm ball, recentered geometry, and comparison with $R = D$. Their one-dimensional updates differ because only the full-gradient method may use the current gradient before choosing the radius. Consequently, the stochastic theorem cannot be obtained by inserting minibatch gradients into the deterministic majorant.

F Full Results for Deterministic Distance-Free Muon

This appendix gives the exact majorant, algorithm, and finite-time theorem behind Theorem E.1. DF-Muon does not take the distance to a minimizer as input and does not assume bounded iterates, bounded gradients, a bounded domain, or a bounded initial sublevel set. It chooses the trust-region radius by a regularized one-dimensional search on a smoothness majorant, rather than by evaluating the true objective along the ray.

Muon-type trust-region analyses typically require a radius proportional to the unknown distance from the initialization to a global minimizer (Kovalev, 2025). Let

$$D := \|x_0 - x^\star\|.$$

Distance-Free Muon does not take D as an input. Instead, it uses three components: a one-dimensional lower certificate for D, the Muon linear-minimization direction, and a one-dimensional radius search whose analysis uses the fixed feasible choice $R = D$.

Distance certificate. The method maintains a scalar proxy d_k, inspired by the D-Adaptation principle of constructing lower certificates for the unknown distance to a minimizer (Defazio & Mishchenko, 2023). For arbitrary weights $\omega_k \geq 0$, let

$$\begin{aligned} S_{k+1} &= S_k + \omega_k g_k, \qquad B_{k+1} = B_k - \omega_k\langle g_k, x_k - x_0\rangle, \\ \widehat{d}_{k+1} &= \begin{cases} [B_{k+1}]_+/\|S_{k+1}\|_*, & \text{if } \|S_{k+1}\|_* > 0, \\ 0, & \text{if } \|S_{k+1}\|_* = 0, \end{cases} \qquad d_{k+1} = \max\{d_k, \widehat{d}_{k+1}\}. \end{aligned} \tag{69}$$

As shown in Appendix G, if $0 \leq d_0 \leq D$, then $d_k \leq D$ for all k. Unlike standard D-adaptation, our proof does not require d_k to recover D. The proxy only regularizes the radius search below.

Muon direction and recentered ray. Given exponential momentum

$$m_{k+1} = (1-\alpha)m_k + \alpha g_k,$$

Distance-Free Muon keeps the Muon direction

$$s_k \in \arg\min_{\|s\|\le 1} \langle m_{k+1}, s\rangle. \tag{70}$$

In the spectral–nuclear matrix geometry, this is exactly the orthogonalized momentum direction used by Muon-type methods. The trust-region path is recentered toward the initialization:

$$c_k := x_0 + (1-\beta)(x_k - x_0), \qquad z_k(R) := c_k + \beta R s_k. \tag{71}$$

Equivalently, with $y_k := x_k - x_0$,

$$z_k(R) - x_k = \beta(Rs_k - y_k), \qquad z_k(R) - x_0 = (1-\beta)y_k + \beta R s_k.$$

One-dimensional radius search. Instead of setting $R = D$, which is unknown, the method chooses R_k by minimizing a regularized smoothness majorant along the recentered ray. Define

$$\begin{aligned}\mathcal{Q}_k(R) := f(x_k) &+ \beta\langle g_k, Rs_k - y_k\rangle + \frac{L\beta^2}{2}\|Rs_k - y_k\|^2 \\ &+ M\beta L\|(1-\beta)y_k + \beta R s_k\|^2 + \rho L\beta^2\|Rs_k - y_k\|^2 \\ &+ \frac{\lambda L\beta^2}{2}(R - d_{k+1})^2.\end{aligned} \tag{72}$$

Then

$$R_k \in \arg\min_{R\ge 0} \mathcal{Q}_k(R), \qquad x_{k+1} = z_k(R_k). \tag{73}$$

The objective $\mathcal{Q}_k$ is convex in the one-dimensional variable R. It is a linear term plus squared norms of affine functions of R, together with a quadratic regularizer. Although D is unknown to the algorithm, the search is over all $R \ge 0$. The proof can therefore compare the chosen radius with the fixed feasible choice $R = D$, recovering the scale used in fixed-radius trust-region analyses without passing D to the method.

Algorithm 4 Distance-Free Muon

Require: Initial point $x_0 \in \mathcal{E}$, horizon T, initial proxy $d_0 = 0$, $\alpha, \beta \in (0,1)$, $\rho, \lambda, M > 0$

```
m_0 ← ∇f(x_0), S_0 ← 0, B_0 ← 0
for k = 0, 1, ..., T − 1 do
    g_k ← ∇f(x_k)
    Choose ω_k ≥ 0 and update d_{k+1} by (69)
    Update m_{k+1} and choose s_k by (70)
    Define c_k, z_k(R) by (71)
    Choose R_k by (73)
    x_{k+1} ← z_k(R_k)
end for
return x_T
```

The guarantee below holds for every nonnegative weight sequence. For a concrete implementation, we use the default choice $\omega_k \equiv 1$. Other nonnegative choices change only the lower certificate and do not change the theorem.

Theorem F.1 (Last-iterate convergence of Distance-Free Muon)**.** *Let Assumptions 2.1 and 2.3 hold. Let Algorithm 4 be run with $d_0 = 0$, or more generally with any $0 \le d_0 \le D$, where $D = \|x_0 - x^\star\|$. Suppose $M \ge 2(1+2\rho)$, $\alpha, \beta \in (0,1)$, and $\alpha > \beta/2$. Then, for every $T \ge 1$,*

$$f(x_T) - f(x^\star) \le \left(1 - \frac{\beta}{2}\right)^T \big(f(x_0) - f(x^\star)\big) + C_{\alpha,\beta,\rho,\lambda,M} L\beta D^2,$$

where

$$C_{\alpha,\beta,\rho,\lambda,M} := 2C_0 + \frac{4(1-\alpha)^2}{\rho(\alpha - \beta/2)^2}, \qquad C_0 := 1 + 2\rho + \frac{\lambda}{2} + M.$$

Consequently, choosing

$$\beta = \min\left\{\alpha, \frac{2\log(T+1)}{T}\right\}$$

gives, whenever $2\log(T+1)/T \le \alpha$,

$$f(x_T) - f(x^\star) = \widetilde{O}\left(\frac{L\|x_0 - x^\star\|^2}{T}\right),$$

where the hidden constants depend only on α, ρ, λ, M, *but not on* D.

Appendix G gives the proof. The unknown distance $D = \|x_0 - x^\star\|$ enters only through the fixed analysis choice $R = D$. No bounded domain, bounded-gradient, bounded-iterate, or bounded-sublevel assumption is needed. The constants ρ, λ, M regularize the radius search. They do not estimate a problem scale. One valid choice is $\rho = 1$, $\lambda = 1$, and $M = 6$. Compared with Kovalev (2025), the algorithm no longer needs a distance or boundedness parameter. Using a smoothness majorant also keeps the one-dimensional subproblem convex, unlike a direct search over objective values along the ray.

G Proofs for Distance-Free Muon

This appendix proves Theorem F.1. Throughout the section, let

$$f^\star := f(x^\star), \qquad D := \|x_0 - x^\star\|, \qquad y_k := x_k - x_0, \qquad u^\star := x^\star - x_0.$$

Thus $D = \|u^\star\|$. We also write

$$\Delta_k := f(x_k) - f^\star.$$

G.1 Validity of the scalar D-certificate

The scalar proxy d_k is inspired by D-adaptation, but the proof below only uses the one-sided lower-certificate property $d_k \le D$.

Lemma G.1 (Validity of the D-certificate)**.** *Suppose Assumption 2.3 holds and* $0 \le d_0 \le D$. *Then the sequence generated by Algorithm 4 satisfies*

$$d_k \le D \qquad \text{for all } k \ge 0.$$

Proof. By Assumption 2.3, for every i,

$$f(x_i) - f^\star \le \langle \nabla f(x_i), x_i - x^\star\rangle.$$

Since $x_i - x^\star = (x_i - x_0) - (x^\star - x_0) = y_i - u^\star$, we obtain

$$f(x_i) - f^\star \le \langle \nabla f(x_i), y_i\rangle - \langle \nabla f(x_i), u^\star\rangle.$$

Multiplying by $\omega_i \ge 0$ and summing from $i = 0$ to k gives

$$\sum_{i=0}^{k} \omega_i (f(x_i) - f^\star) \le \sum_{i=0}^{k} \omega_i \langle \nabla f(x_i), y_i\rangle - \left\langle \sum_{i=0}^{k} \omega_i \nabla f(x_i), u^\star \right\rangle.$$

The left-hand side is nonnegative because $x^\star$ is a global minimizer. Using the definitions

$$S_{k+1} := \sum_{i=0}^{k} \omega_i \nabla f(x_i), \qquad B_{k+1} := -\sum_{i=0}^{k} \omega_i \langle \nabla f(x_i), x_i - x_0\rangle,$$

we therefore have

$$0 \le -B_{k+1} - \langle S_{k+1}, u^\star\rangle.$$

Equivalently,

$$B_{k+1} \le -\langle S_{k+1}, u^\star\rangle \le \|S_{k+1}\|_* \, \|u^\star\| = D\|S_{k+1}\|_*.$$

If $\|S_{k+1}\|_* > 0$, then

$$\widehat{d}_{k+1} = \frac{[B_{k+1}]_+}{\|S_{k+1}\|_*} \le D.$$

If $\|S_{k+1}\|_* = 0$, then by definition $\widehat{d}_{k+1} = 0 \le D$. Since $d_{k+1} = \max\{d_k, \widehat{d}_{k+1}\}$ and $d_0 \le D$, induction gives the claim. **Q.E.D.**

G.2 Existence and convexity of the radius search

Lemma G.2 (Existence and convexity of the one-dimensional radius search)**.** *For every k, the one-dimensional minimization problem defining R_k in Algorithm 4 admits a solution. Moreover, $R \mapsto \mathcal{Q}_k(R)$ is convex on $[0, \infty)$.*

Proof. If $L = 0$, then by smoothness ∇f is constant on $\mathcal{E}$. Since f has an unconstrained global minimizer, this constant is zero, so $f \equiv f^\star$, and the claims are trivial. Hence assume $L > 0$.

For fixed k, the map

$$R \mapsto z_k(R) = x_0 + (1-\beta)(x_k - x_0) + \beta R s_k$$

is continuous. Therefore all terms in $\mathcal{Q}_k(R)$ are continuous on $[0, \infty)$. Since $\lambda > 0$, $L > 0$, and $\beta > 0$, the term

$$\frac{\lambda L \beta^2}{2}(R - d_{k+1})^2$$

is coercive as $R \to +\infty$. Hence $\mathcal{Q}_k(R) \to +\infty$ as $R \to +\infty$, and $\mathcal{Q}_k$ attains its minimum on the closed set $[0, \infty)$.

Convexity follows because

$$R \mapsto \|R s_k - y_k\|^2, \qquad R \mapsto \|(1-\beta) y_k + \beta R s_k\|^2$$

are squared norms of affine functions of R, hence convex. The remaining nonconstant terms are linear or quadratic in R. Thus $R \mapsto \mathcal{Q}_k(R)$ is convex. **Q.E.D.**

G.3 One-step majorized inequality

Define the Lyapunov function

$$\mathcal{L}_k := \Delta_k + M\beta L \|y_k\|^2.$$

The quadratic term is created by the regularized radius search and is used to control the momentum error. It is not a boundedness assumption.

Also define

$$e_k := \nabla f(x_k) - m_{k+1}.$$

Lemma G.3 (One-step majorized inequality)**.** *Let Assumptions 2.1 and 2.3 hold, and let Algorithm 4 be run with $0 \le d_0 \le D$. If $D > 0$ and*

$$M \ge 2(1 + 2\rho),$$

then, for every $k \ge 0$,

$$\mathcal{L}_{k+1} + \rho L \|x_{k+1} - x_k\|^2 \le \left(1 - \frac{\beta}{2}\right) \mathcal{L}_k + C_0 L \beta^2 D^2 + 2\beta D \|e_k\|_*,$$

where

$$C_0 := 1 + 2\rho + \frac{\lambda}{2} + M.$$

Proof. By L-smoothness and the definition of $\mathcal{Q}_k$, for every $R \ge 0$,

$$f(z_k(R)) \le f(x_k) + \beta \langle g_k, R s_k - y_k \rangle + \frac{L\beta^2}{2} \|R s_k - y_k\|^2.$$

At the chosen radius R_k, this implies

$$\begin{aligned} f(x_{k+1}) + M\beta L \|y_{k+1}\|^2 + \rho L \|x_{k+1} - x_k\|^2 \\ \le \mathcal{Q}_k(R_k) - \frac{\lambda L \beta^2}{2}(R_k - d_{k+1})^2. \end{aligned}$$

Since the last term on the right-hand side is nonpositive,

$$f(x_{k+1}) + M\beta L \|y_{k+1}\|^2 + \rho L \|x_{k+1} - x_k\|^2 \le \mathcal{Q}_k(R_k).$$

By optimality of R_k in the radius search and since $R = D$ is feasible,

$$\mathcal{Q}_k(R_k) \le \mathcal{Q}_k(D).$$

Therefore

$$\begin{aligned} f(x_{k+1}) + M\beta L\|y_{k+1}\|^2 + \rho L\|x_{k+1} - x_k\|^2 \\ \le f(x_k) + \beta\langle g_k, Ds_k - y_k\rangle + \frac{L\beta^2}{2}\|Ds_k - y_k\|^2 \\ + M\beta L\|(1-\beta)y_k + \beta Ds_k\|^2 + \rho L\beta^2\|Ds_k - y_k\|^2 \\ + \frac{\lambda L\beta^2}{2}(D - d_{k+1})^2. \end{aligned}$$

By Lemma G.1, $0 \le d_{k+1} \le D$, hence

$$(D - d_{k+1})^2 \le D^2.$$

We now control the linear term. Decompose

$$Ds_k - y_k = (u^\star - y_k) + (Ds_k - u^\star).$$

By Assumption 2.3,

$$\langle g_k, u^\star - y_k\rangle = \langle \nabla f(x_k), x^\star - x_k\rangle \le -\Delta_k.$$

Write $g_k = m_{k+1} + e_k$. Since

$$s_k \in \arg\min_{\|s\|\le 1}\langle m_{k+1}, s\rangle$$

and $\|u^\star/D\| \le 1$, we have

$$\langle m_{k+1}, Ds_k\rangle \le \langle m_{k+1}, u^\star\rangle.$$

Therefore

$$\langle m_{k+1}, Ds_k - u^\star\rangle \le 0.$$

Moreover,

$$\|Ds_k - u^\star\| \le 2D.$$

Consequently,

$$\begin{aligned} \beta\langle g_k, Ds_k - y_k\rangle &\le -\beta\Delta_k + \beta\langle e_k, Ds_k - u^\star\rangle \\ &\le -\beta\Delta_k + 2\beta D\|e_k\|_*. \end{aligned}$$

Since $\|s_k\| \le 1$,

$$\|Ds_k - y_k\|^2 \le 2D^2 + 2\|y_k\|^2.$$

Also, by convexity of $v \mapsto \|v\|^2$,

$$\|(1-\beta)y_k + \beta Ds_k\|^2 \le (1-\beta)\|y_k\|^2 + \beta D^2.$$

Substituting the preceding inequalities gives

$$\begin{aligned} \mathcal{L}_{k+1} + \rho L\|x_{k+1} - x_k\|^2 \\ \le (1-\beta)\Delta_k + M\beta L(1-\beta)\|y_k\|^2 \\ + L\beta^2(1+2\rho)\|y_k\|^2 \\ + L\beta^2\left(1 + 2\rho + \frac{\lambda}{2}\right)D^2 + ML\beta^2 D^2 + 2\beta D\|e_k\|_*. \end{aligned}$$

Because $M \ge 2(1+2\rho)$,

$$L\beta^2(1+2\rho)\|y_k\|^2 \le \frac{\beta}{2}M\beta L\|y_k\|^2.$$

Also,

$$(1-\beta)\Delta_k \le \left(1 - \frac{\beta}{2}\right)\Delta_k.$$

Therefore

$$\begin{aligned} \mathcal{L}_{k+1} + \rho L\|x_{k+1} - x_k\|^2 \\ \le \left(1 - \frac{\beta}{2}\right)\left(\Delta_k + M\beta L\|y_k\|^2\right) + C_0 L\beta^2 D^2 + 2\beta D\|e_k\|_*, \end{aligned}$$

which is the desired inequality. **Q.E.D.**

G.4 MOMENTUM-ERROR CONTROL

Let

$$a_i := \|x_i - x_{i-1}\|, \qquad r := 1 - \frac{\beta}{2}, \qquad q := 1 - \alpha.$$

Lemma G.4 (Weighted momentum-error bound)**.** *Suppose Assumption 2.1 holds and*

$$\alpha > \frac{\beta}{2}.$$

Then, for every $T \geq 1$,

$$\sum_{k=0}^{T-1} r^{T-1-k}\|e_k\|_* \leq \frac{(1-\alpha)L}{\alpha - \beta/2}\sqrt{\frac{2}{\beta}}\left(\sum_{i=1}^{T} r^{T-i}a_i^2\right)^{1/2}.$$

Proof. Since $m_0 = \nabla f(x_0)$, we have $e_0 = 0$. For $k \geq 1$,

$$\begin{aligned} e_k &= \nabla f(x_k) - m_{k+1} \\ &= (1-\alpha)\big(\nabla f(x_k) - m_k\big) \\ &= q\big(\nabla f(x_k) - \nabla f(x_{k-1}) + e_{k-1}\big). \end{aligned}$$

By L-smoothness,

$$\|e_k\|_* \leq qLa_k + q\|e_{k-1}\|_*.$$

Unrolling this recursion gives

$$\|e_k\|_* \leq qL\sum_{i=1}^{k} q^{k-i}a_i.$$

Therefore

$$\begin{aligned} \sum_{k=0}^{T-1} r^{T-1-k}\|e_k\|_* &\leq qL\sum_{k=1}^{T-1} r^{T-1-k}\sum_{i=1}^{k} q^{k-i}a_i \\ &= qL\sum_{i=1}^{T-1} a_i \sum_{k=i}^{T-1} r^{T-1-k}q^{k-i}. \end{aligned}$$

Since $q < r$, or equivalently $\alpha > \beta/2$,

$$\sum_{k=i}^{T-1} r^{T-1-k}q^{k-i} \leq \frac{r^{T-i}}{r-q}.$$

Hence

$$\sum_{k=0}^{T-1} r^{T-1-k}\|e_k\|_* \leq \frac{qL}{r-q}\sum_{i=1}^{T-1} r^{T-i}a_i.$$

By Cauchy–Schwarz,

$$\sum_{i=1}^{T-1} r^{T-i}a_i \leq \left(\sum_{i=1}^{T-1} r^{T-i}\right)^{1/2}\left(\sum_{i=1}^{T-1} r^{T-i}a_i^2\right)^{1/2}.$$

Since

$$\sum_{i=1}^{T-1} r^{T-i} \leq \frac{1}{1-r} = \frac{2}{\beta},$$

and extending the second sum to $i = T$ only increases it, we obtain

$$\sum_{k=0}^{T-1} r^{T-1-k}\|e_k\|_* \leq \frac{qL}{r-q}\sqrt{\frac{2}{\beta}}\left(\sum_{i=1}^{T} r^{T-i}a_i^2\right)^{1/2}.$$

Substituting $q = 1 - \alpha$ and $r - q = \alpha - \beta/2$ proves the claim. **Q.E.D.**

G.5 Proof of Theorem F.1

Proof of Theorem F.1. If $D = 0$, then $x_0 = x^\star$. Since $x^\star$ is an unconstrained global minimizer, $\nabla f(x^\star) = 0$. The one-dimensional radius search is minimized at $R_0 = 0$. Hence $x_1 = x^\star$. Inductively, $x_k = x^\star$ for all k, so the claim is immediate. Assume $D > 0$.

Let

$$A_T := \sum_{i=1}^{T} r^{T-i} a_i^2, \qquad r := 1 - \frac{\beta}{2}.$$

Multiplying Lemma G.3 by r^{T-1-k} and summing from $k = 0$ to $T - 1$, we get

$$\mathcal{L}_T + \rho L A_T \le r^T \mathcal{L}_0 + C_0 L \beta^2 D^2 \sum_{k=0}^{T-1} r^{T-1-k} + 2\beta D \sum_{k=0}^{T-1} r^{T-1-k} \|e_k\|_*.$$

Since $1 - r = \beta/2$,

$$\sum_{k=0}^{T-1} r^{T-1-k} \le \frac{2}{\beta}.$$

Therefore

$$\mathcal{L}_T + \rho L A_T \le r^T \mathcal{L}_0 + 2C_0 L \beta D^2 + 2\beta D \sum_{k=0}^{T-1} r^{T-1-k} \|e_k\|_*.$$

By Lemma G.4,

$$\sum_{k=0}^{T-1} r^{T-1-k} \|e_k\|_* \le \frac{(1-\alpha)L}{\alpha - \beta/2} \sqrt{\frac{2}{\beta}} A_T^{1/2}.$$

Thus

$$2\beta D \sum_{k=0}^{T-1} r^{T-1-k} \|e_k\|_* \le \frac{2\sqrt{2}(1-\alpha) L D \sqrt{\beta}}{\alpha - \beta/2} A_T^{1/2}.$$

Using Young's inequality,

$$ab \le \frac{\rho L}{2} a^2 + \frac{b^2}{2\rho L},$$

with

$$a = A_T^{1/2}, \qquad b = \frac{2\sqrt{2}(1-\alpha) L D \sqrt{\beta}}{\alpha - \beta/2},$$

we obtain

$$2\beta D \sum_{k=0}^{T-1} r^{T-1-k} \|e_k\|_* \le \frac{\rho L}{2} A_T + \frac{4(1-\alpha)^2 L \beta D^2}{\rho(\alpha - \beta/2)^2}.$$

Consequently,

$$\mathcal{L}_T + \frac{\rho L}{2} A_T \le r^T \mathcal{L}_0 + 2C_0 L \beta D^2 + \frac{4(1-\alpha)^2 L \beta D^2}{\rho(\alpha - \beta/2)^2}.$$

Dropping the nonnegative term $(\rho L/2) A_T$, and using

$$\mathcal{L}_T \ge f(x_T) - f^\star, \qquad \mathcal{L}_0 = f(x_0) - f^\star,$$

gives

$$f(x_T) - f^\star \le r^T (f(x_0) - f^\star) + \left[2C_0 + \frac{4(1-\alpha)^2}{\rho(\alpha - \beta/2)^2} \right] L \beta D^2.$$

Substituting $r = 1 - \beta/2$ proves the first claim.

For the explicit rate, choose

$$\beta = \min \left\{ \alpha, \frac{2 \log(T+1)}{T} \right\}.$$

Whenever $2\log(T+1)/T \le \alpha$, we have $\beta = 2\log(T+1)/T$, and hence

$$\left(1-\frac{\beta}{2}\right)^T \le \exp\left(-\frac{\beta T}{2}\right) = \frac{1}{T+1}.$$

Moreover, since $\beta \le \alpha$,

$$\alpha - \frac{\beta}{2} \ge \frac{\alpha}{2}.$$

Therefore

$$f(x_T) - f^\star \le \frac{f(x_0)-f^\star}{T+1} + 2\left[2C_0 + \frac{16(1-\alpha)^2}{\rho\alpha^2}\right]\frac{LD^2\log(T+1)}{T}.$$

It remains only to bound the initial gap by the same LD^2 scale. Since $x^\star$ is an unconstrained global minimizer and f is differentiable, $\nabla f(x^\star) = 0$. By star-convexity, duality, and smoothness,

$$\begin{aligned} f(x_0) - f(x^\star) &\le \langle \nabla f(x_0), x_0 - x^\star\rangle \le \|\nabla f(x_0)\|_* D \\ &= \|\nabla f(x_0) - \nabla f(x^\star)\|_* D \le LD^2. \end{aligned}$$

Thus

$$f(x_T) - f^\star = \widetilde{O}\left(\frac{LD^2}{T}\right), \qquad D = \|x_0 - x^\star\|,$$

with constants depending only on α, ρ, λ, and M, but not on D. **Q.E.D.**

Remark G.5 (What is distance-free?). The algorithm never receives $D = \|x_0 - x^\star\|$ as an input. The proof uses D only through the fixed feasible choice $R = D$ in the radius search. The scalar proxy d_k is a D-adaptation-style lower certificate and is used only through the valid inequality $d_k \le D$. Thus the result avoids both the fixed-radius choice $\eta = \beta D$ and the need to prove that d_k discovers D.

Remark G.6 (Majorized versus objective-value radius search). The radius search minimizes a regularized smoothness majorant rather than the true one-dimensional loss $R \mapsto f(z_k(R))$. This avoids a global search over the objective value along the ray and yields a convex one-dimensional subproblem. The proof still compares the chosen radius with the fixed choice $R = D$, which is why the distance parameter appears only in the analysis.

H Star-Convex and Transformer Experimental Details

H.1 Exact Star-Convex Matrix Experiment

We use 8×8 matrices and

$$f(X) = \frac{\mu}{2}\|X - X^\star\|_F^2, \qquad \mu = \frac{1}{2}.$$

For any perturbation H, $\|\nabla f(X+H) - \nabla f(X)\|_* = \mu\|H\|_*$. Since $\|H\|_* \le 8\|H\|_{\mathrm{op}}$, the function is 4-smooth with respect to the spectral norm. We set $c = X_0 = 0$ and construct $X^\star$ with operator norm $D \in \{0.5, 2, 8\}$. The oracle adds an independent Gaussian matrix with Frobenius standard deviation $\sigma \in \{0, 0.25, 1\}$. The corresponding nuclear-norm noise bound is $s_\sigma = \sqrt{8}\,\sigma$.

Algorithm 1 uses $\kappa = 1/4$, momentum parameter $\alpha = 0.15$, the prescribed coefficient $q_R = (9/4)L\beta^2/\alpha$, and an exact SVD-based LMO. The value $\alpha = 0.15$ satisfies the conditions of finite-time Theorem B.7. The algorithm receives neither D nor σ. There is no clipping, warmup, cosine decay, additional weight decay, radius floor, or radius cap. The contraction in Algorithm 1 is retained. The reported evaluation uses $T = 8192$ and seeds 100–111, disjoint from momentum-parameter and fixed-radius calibration.

The known-distance baseline uses the same one-gradient-delayed Nesterov-style momentum direction and the same contracted update. It replaces the learned radius R_k by D at every round. Thus its stepsize is $\eta_k = \beta D$. The baseline is designed to isolate the cost of learning the unknown radius. It is

not intended to reproduce Algorithm 2 of Kovalev (2025), which uses a direction formed from the current stochastic gradient.

Table 5: Last-iterate objective gap at $D = 2$, $\sigma = 0.25$, and $T = 8192$ with $\alpha = 0.15$. Entries are mean $\pm$ sample standard deviation over 12 evaluation seeds disjoint from momentum-parameter selection.

Method	Final objective gap
DF-Muon, learned radius	0.01395 ± 0.00113
Same update with $R = D$	0.01076 ± 0.00128

The learned-radius gap is 1.30 times that of the known-distance baseline. Both methods use the same momentum parameter, direction rule, contracted update, and matched noise streams. Only the radius rule changes. Their different trajectories can nevertheless produce different realized directions.

H.2 Momentum-Parameter Calibration and Finite-Time Admissibility

We varied only α on the quadratic with $D = 2$, $\sigma = 0.25$, and $T = 8192$, recomputing $q_R = (9/4)L\beta^2/\alpha$ as prescribed. Calibration on seeds 0–23 considered $\{0.08, 0.10, 0.12\}$ and then extended the grid to $\{0.13, 0.14, 0.15\}$ after observing improvement. The selected value 0.15 is therefore exploratory rather than a preregistered universal choice. It is covered by the finite-time Theorem B.7: $\beta \le 1/512$, $\beta < \alpha/4$, and, with $c_0 = 9$,

$$\max\left\{\frac{8\alpha}{c_0}, \frac{\alpha}{5/8 - 4\beta}\right\} \le \max\left\{\frac{0.15 \cdot 8}{9}, \frac{0.15}{5/8 - 4/512}\right\} < \frac{1}{4} = \kappa.$$

The horizon-dependent corollary uses $\alpha_T = 8\beta_T^{2/3}$. The calibrated value $\alpha = 0.15$ satisfies the conditions of Theorem B.7 but does not replace this rate schedule. Disjoint evaluation seeds prevent reuse of the same random draws but do not constitute cross-task validation.

H.3 Non-Convex Star-Convex Diagnostic

To probe a non-convex member of the star-convex class, we use

$$f(X) = g(\|X - X^\star\|_F), \qquad g(u) = u(1 - e^{-u}).$$

For $u \ge 0$, $ug'(u) - g(u) = u^2 e^{-u} \ge 0$, which establishes the first-order star-convex condition about $X^\star$. Since $g''(u) = (2 - u)e^{-u}$ is negative for $u > 2$, the objective is not convex. The radial and tangential Hessian eigenvalues have absolute value at most 2, including their limits at zero. Thus $L = 16$ is a valid smoothness constant with respect to the spectral norm on 8×8 matrices.

We retain $c = X_0 = 0$, the Gaussian noise model, exact SVD direction, $\alpha = 0.15$, and $\kappa = 1/4$ from the quadratic experiment. We report $D \in \{2, 8\}$, $\sigma = 0.25$, $T = 8192$, and the same 12 evaluation seeds. Fixed-radius baselines use the same contracted update with either $R = D$ or the transferred fixed radius $R = 4$ selected below.

Table 6: Radial non-convex star-convex diagnostic with $\alpha = 0.15$, $\sigma = 0.25$, and $T = 8192$. Final gaps are mean $\pm$ sample standard deviation over 12 evaluation seeds.

D	Learned	Known $R = D$	Transferred $R = 4$
2	0.07447 ± 0.01312	0.00660 ± 0.00063	0.00158 ± 0.00030
8	4.00505 ± 0.38396	0.01705 ± 0.00106	5.03444 ± 0.52255

The learned rule remains behind the known-distance baseline. It beats the transferred fixed radius at $D = 8$ but not at $D = 2$, so the quadratic transfer result does not imply uniform dominance across objectives or scales.

H.4 Transfer of a Calibrated Fixed Radius

We test whether one fixed radius chosen on separate problems can replace online radius learning. The candidate set is $\{0.25, 0.5, 1, 2, 4, 8, 16, 32\}$. Calibration uses isotropic quadratics with $D \in \{1, 2\}$, $\sigma = 0.25$, $T = 4096$, and seeds 1000 through 1007. The criterion is the mean final gap divided by the initial gap, with smaller radius breaking ties. With $\alpha = 0.15$, it selects $R = 4$ before any test run. Evaluation uses seeds 100–111, $T = 8192$, the same contracted update, $\kappa = 1/4$, exact SVD directions, and matched noise streams. The baselines use either transferred $R = 4$ or the known-distance choice $R = D$. We include isotropic and anisotropic quadratics together with radial and anisotropic radial star-convex objectives.

Table 7: Transfer of the fixed radius $R = 4$ with $\alpha = 0.15$ at $D = 8$, $\sigma = 0.25$, and $T = 8192$. Final gaps are mean $\pm$ sample standard deviation over 12 evaluation seeds. The $R = D$ baseline receives the test distance.

Objective	Learned	Transferred $R = 4$	Known $R = D$
Quadratic	0.23739 ± 0.03115	6.42523 ± 1.28230	0.02527 ± 0.00232
Anisotropic quadratic	0.34567 ± 0.06196	3.17830 ± 0.88614	0.03643 ± 0.00409
Radial non-convex	4.00505 ± 0.38396	5.03444 ± 0.52255	0.01705 ± 0.00106
Anisotropic radial	3.72378 ± 0.47077	3.44617 ± 0.55152	0.02441 ± 0.00236

At $D = 8$, learning beats the transferred radius on three of the four displayed objectives, but not on the anisotropic radial objective. At $D = 0.5$ on the isotropic quadratic, transferred $R = 4$ also beats learning (0.00142 versus 0.00259). Thus the result demonstrates useful large-distance transfer rather than uniform superiority. The known-distance baseline with $R = D$ remains strongest in all four displayed cells. The complete per-seed results, calibration decision, and source hashes accompany the experiment.

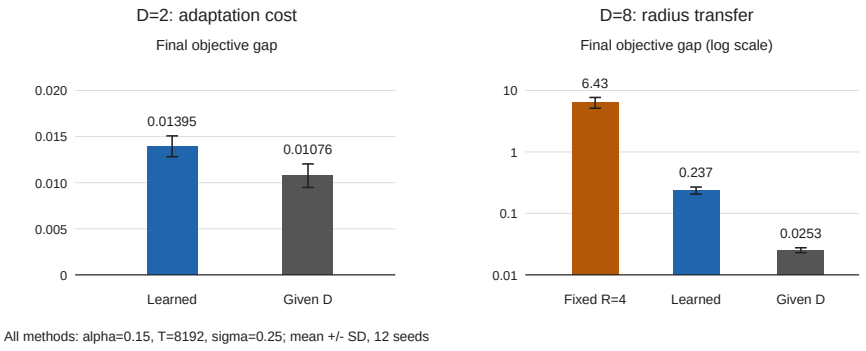


Figure 2: Exact contracted quadratic with the admissible momentum parameter $\alpha = 0.15$ at $T = 8192$ and $\sigma = 0.25$ over 12 evaluation seeds, reported as mean $\pm$ SD. Left: adaptation cost at $D = 2$. Right: transfer to $D = 8$. The fixed radius $R = 4$ was selected on separate problems.

H.5 Transformer Configuration

We use the raw WikiText-103 train and validation splits with the GPT-2 tokenizer. The model has 12 layers, 12 heads, width 768, sequence length 512, batch size 16, zero dropout, and BF16 autocast. Each final run uses 10,000 stochastic-gradient evaluations. Tuning uses seed 777 and final evaluation uses seeds 42, 1337, and 2024.

Muon-based methods update parameters with at least two dimensions except token and position embeddings, normalization parameters, and the output head. Auxiliary AdamW updates all remaining parameters. All three Muon-based methods use five Newton–Schulz iterations, rectangular adjustment, global gradient clipping at one, and 500-step warmup followed by cosine decay.

Table 8: The four methods in the practical FTRL Transformer study. The separate current-gradient proxy study is reported below.

Method	Stepsize rule	Direction and training components
Practical weighted-FTRL Muon	$\eta_k = \beta R_k$, with R_k chosen by weighted FTRL	One-gradient-delayed Nesterov-style momentum direction, additive update, matrix weight decay 0.01, Newton–Schulz, rectangular adjustment, clipping, warmup–cosine
Non-convex DF-Muon	η_k chosen by non-convex FTRL	Delayed Nesterov-style momentum direction, matrix weight decay 0.01, Newton–Schulz, rectangular adjustment, clipping, warmup–cosine
Muon	Tuned matrix learning rate 0.005	Muon direction with the same matrix weight decay, Newton–Schulz, rectangular adjustment, clipping, warmup–cosine
AdamW	Tuned learning rate 0.0007071	AdamW update for all parameters

Table 9: Theory versus Transformer implementation for the weighted FTRL update for the radius. The practical implementation is not covered by our analysis.

Component	Theorem 3.1	Transformer implementation
Update	Contracted toward c	Additive Muon step with fixed matrix weight decay
Direction	Exact LMO over the unit norm ball	Five-step Newton–Schulz polar approximation
Quadratic coefficient	$\kappa = 1/4$ with certified L	$\kappa = 64$ with empirical $\widehat{L}$ tuned once
Setting	Isolated Muon recursion	Full recipe with clipping, warmup, cosine decay, and auxiliary AdamW

Practical weighted-FTRL Muon uses

$$q_R = \kappa c_0 \widehat{L} \frac{\beta^2}{\alpha} \quad \text{with} \quad \kappa = 64, \qquad c_0 = 9,$$

and selects $\widehat{L} = 0.25$ on seed 777. Thus the practical method keeps $\kappa = 64$. The value $\widehat{L} = 0.25$ is an empirical calibration and is distinct from the theorem choice $\kappa = 1/4$. The non-convex implementation uses $\widehat{L} = 64$. These are not certified global smoothness constants. The weighted FTRL update for the radius and one-gradient-delayed Nesterov-style momentum direction come from the theory, but the additive parameter update and empirical coefficient are not covered by Theorem 3.1. The training components of non-convex DF-Muon are likewise not covered by Corollary 4.2. We make no star-convexity claim for the Transformer loss.

Actual FTRL parameters and feedback. The 10,000 stochastic-gradient evaluations comprise one initialization evaluation and $T_{\mathrm{FTRL}} = 9999$ subsequent rounds. Both practical FTRL variants use $\alpha = 0.05$. The weighted rule uses

$$\beta = \frac{2\log(T_{\mathrm{FTRL}} + 1)}{T_{\mathrm{FTRL}}} \simeq 0.00184225, \qquad q_R = 576\widehat{L}\beta^2/\alpha \simeq 0.00977441.$$

This equality records the implementation used for the reported runs, not the theoretical coefficient in (5). Its initial quadratic weight is $H = w_1/(1-r) \simeq 0.108203$. Thus the radius formula is retained, but the horizon-dependent schedule (5) is not used in these GPT runs. The non-convex rule uses $H_0 = 32$ and $q \simeq 1842.19336$ from (8), with the practical α and $\widehat{L}$. Both implementations form their delayed Nesterov-style momentum directions from clipped gradient history. The weighted rule evaluates its directional-derivative estimate before clipping, whereas the non-convex rule uses clipped feedback. The latter also replaces the EMA-only direction in Algorithm 2 by a delayed Nesterov-style momentum direction. The scalar feedback uses the implemented, rectangular-adjusted directions.

For both rules the applied matrix update has the form

$$X_{k+1} = (1 - 0.01\,\widetilde{\eta}_k)X_k + \widetilde{\eta}_k \widetilde{s}_k, \qquad \widetilde{\eta}_k = \tau_k \eta_k,$$

where τ_k is the common warmup–cosine multiplier and $\widetilde{s}_k$ is the implemented matrix direction. There is no contraction toward a reference point beyond this fixed matrix weight decay. Auxiliary AdamW uses base learning rate 0.001, momentum parameters $(0.9, 0.95)$, and weight decay 0.1.

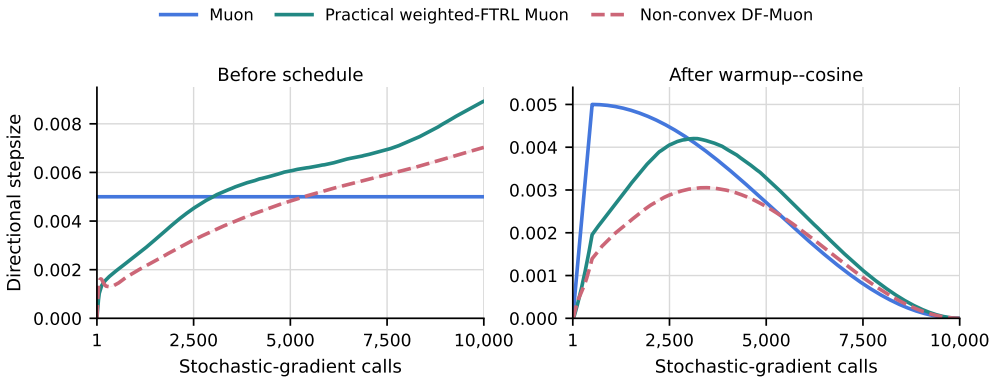


Figure 3: Stepsizes for the three Muon-based methods. Left: the stepsize before the shared schedule. Right: the effective stepsize after warmup and cosine decay. Curves are means over three seeds and shaded regions show one sample standard deviation. The raw radius R_k of the weighted method is not directly comparable with the stepsize variable of the non-convex method.

Table 10: Matched GPT-124M results after 10,000 stochastic-gradient evaluations. Top: paired validation losses using the same initialization and data streams. Bottom: validation loss and H100 runtime, shown as mean ± sample standard deviation over three seeds.

Seed	Muon	Weighted FTRL	Non-convex DF
42	3.46752	3.46666	3.46531
1337	3.48168	3.48209	3.48611
2024	3.47555	3.48023	3.47664

Method	Validation loss	Δ vs. Muon	Time (min)
Muon	3.47492 ± 0.00710	0	14.676 ± 0.344
Practical weighted-FTRL Muon	3.47632 ± 0.00842	$+0.00140$	15.490 ± 0.049
Non-convex DF-Muon	3.47602 ± 0.01042	$+0.00110$	15.239 ± 0.133
AdamW	3.63699 ± 0.01471	$+0.16207$	5.521 ± 0.199

H.6 Frozen-Schedule Replay for Practical FTRL

This diagnostic asks whether the online weighted-FTRL radius decisions improve GPT training beyond a radius schedule obtained on another seed. Using the practical additive recipe of Appendix H.5, we record all 9999 post-initialization radii from seed 777 and freeze that sequence. On seeds 42, 1337, and 2024, we compare online FTRL with replay of the frozen sequence. Both methods retain independently evolving delayed directions, identical data streams within each seed, $\kappa = 64$, $\widehat{L} = 0.25$, and all other training components. These six runs are separate from Table 10 and are not pooled with it.

Table 11: Practical GPT weighted-FTRL online versus seed-777 radius replay. Lower final validation loss is better. The three evaluation seeds are paired.

Seed	Online FTRL	Frozen replay	Online minus replay
42	3.46539	3.46254	+0.00285
1337	3.48217	3.48070	+0.00147
2024	3.47603	3.47685	−0.00081

The mean paired online-minus-replay difference is $+0.00117 \pm 0.00185$, where the second number is the sample standard deviation of the three paired differences. This experiment does not show an online-adaptation advantage on this GPT recipe. It also does not prove equivalence or test a reduction in tuning burden. The replay is a schedule transferred from the tuning seed, not an independent learned optimizer and not the contracted update analyzed in Theorem 3.1.

H.7 Current-Gradient Proxy and Nesterov-Style Momentum Ablations at 10,000 Calls

The practical DF-Muon-Proxy is distinct from weighted FTRL. It forms a direction using the current stochastic gradient and chooses a nominal matrix scale by minimizing a one-step Frobenius-proxy score over a bounded interval. The score uses current gradient–direction alignment, squared direction norm, and a quadratic preference for the preceding scale. The resulting candidate is smoothed with coefficient 0.7 and multiplied by the shared warmup–cosine schedule. The candidate interval is $[0.006, 0.03]$, the initial scale is 0.015, and the step and prior score coefficients are both 0.1. There are 21 grid candidates and six local refinements. No additional gradient calls are used by the one-dimensional search. These hyperparameters are frozen across the ablations. The direction momentum coefficient is 0.9 for every proxy row. This differs from the tuned Muon baseline's momentum coefficient 0.95, so the best-performing proxy–Muon comparison is not a one-factor test of scale adaptation.

All rows below use GPT-2 Small, WikiText-103, 10,000 stochastic-gradient evaluations, batch size 16, BF16, 500-step warmup, cosine decay, global gradient clipping at one, five Newton–Schulz iterations, and auxiliary AdamW on the same parameter split. Muon is tuned on seed 777. Evaluation uses matched initialization and data streams on seeds 42, 1337, and 2024. The four proxy rows add one component at a time: current-gradient Nesterov-style momentum, rectangular adjustment, and matrix weight decay 0.01. The old proxy row reproduces the historical direction path.

Table 12: Separate current-gradient proxy study with 10,000 stochastic-gradient evaluations. Final GPT-124M validation losses are mean ± sample standard deviation over three matched evaluation seeds. The best-performing proxy variant and tuned Muon differ in both scalar rule and momentum coefficient.

Method	Validation loss	Time (min)
Tuned Muon	3.47495 ± 0.00690	13.79 ± 0.01
Original DF-Muon-Proxy	3.47504 ± 0.00965	14.97 ± 0.04
Proxy + Nesterov	3.46104 ± 0.01165	14.98 ± 0.04
Proxy + Nesterov + rectangular scaling	3.44968 ± 0.00836	14.96 ± 0.05
Proxy + Nesterov + scaling + weight decay	3.44288 ± 0.00693	14.97 ± 0.03

The paired mean differences for the three successive additions are -0.01400, -0.01136, and -0.00680 validation loss. The best-performing proxy variant minus the tuned Muon baseline in the same study is -0.03207 on average, with a sample standard deviation of 0.00194 for the three paired differences. For the best-performing proxy variant, the candidate selected before smoothing is at the floor, cap, or interior in 39.53%, 22.42%, or 38.05% of calls, respectively. These fractions do not describe the smoothed applied scale. The original proxy nearly ties tuned Muon at this budget, despite the stronger result of the augmented variant. This ablation alone does not isolate the value of adapting the scale. The otherwise identical tuned fixed-scale baseline is reported in Appendix H.8. Neither experiment validates the predictable FTRL analysis.

Table 13: Per-seed validation losses for the separate proxy study with 10,000 stochastic-gradient evaluations. Columns follow the ablation order of Table 12.

Seed	Tuned Muon	Original proxy	+ Nesterov	+ Scaling	+ Weight decay
42	3.46813	3.46404	3.44795	3.44179	3.43745
1337	3.48193	3.48205	3.47027	3.45843	3.45069
2024	3.47479	3.47903	3.46490	3.44881	3.44050

H.8 Matched Fixed-Scale Baseline for the Practical Proxy

This separate study holds the best-performing proxy configuration above fixed and changes only the matrix stepsize rule. The fixed-scale branch uses the same current-gradient Nesterov-style direction, five Newton–Schulz iterations, rectangular adjustment, matrix weight decay 0.01, auxiliary AdamW, clipping, data stream, and 500-call warmup followed by cosine decay. It sets the candidate floor, initial value, and cap to the same constant. The adaptive branch retains the interval $[0.006, 0.03]$, initial value 0.015, smoothing coefficient 0.7, and the same score parameters as the best-performing variant above. Both branches use GPT-2 Small on WikiText-103, batch size 16, BF16, and $10{,}000$ stochastic-gradient evaluations.

Before evaluation, we select the constant on seed 777 by final validation loss over the predeclared grid $\{0.00125, 0.0025, 0.005, 0.0075, 0.01, 0.015, 0.02, 0.03\}$. The selected value is 0.005. The evaluation seeds 42, 1337, and 2024 are not used for scale selection. This study is separate from the older ablation table and from the practical FTRL study.

Table 14: Matched GPT-124M stepsize-rule comparison at $10{,}000$ stochastic-gradient evaluations. Lower final validation loss is better. The fixed scale was selected on seed 777 only.

Seed	Adaptive Proxy	Fixed $\eta = 0.005$	Adaptive minus fixed
42	3.43669	3.46461	-0.02792
1337	3.45139	3.47487	-0.02348
2024	3.44337	3.47359	-0.03022
Mean $\pm$ SD	3.44382 ± 0.00736	3.47102 ± 0.00559	-0.02721 ± 0.00343

The adaptive candidate hits the floor, cap, or interior on approximately 39%, 22%, or 38% of calls. These fractions concern candidate selection before smoothing, not the applied schedule-multiplied step. The matched comparison favors the adaptive method on all three evaluation seeds relative to this eight-point fixed-scale baseline, which was tuned on a disjoint seed. It does not establish an advantage over an unrestricted optimal fixed scale and does not cover the predictable FTRL method or its star-convex theorem.

I Exact Synthetic Evaluation of Predictable DF-Muon

This appendix gives the complete protocol behind Section 5.2. The experiment uses the exact predictable update and exact LMO over the unit norm ball analyzed in Theorem 4.1. No clipping, radius cap, moving-average smoothing, warmup schedule, or cosine schedule is used.

I.1 Problem construction

We use the objective in (74) on $d \times d$ matrices with $d = 16$. For every problem seed, Q and R are drawn as independent orthogonal matrices. The initial point is $X_0 = DQ$, and the objective and gradient are

$$f(X) = \frac{\lambda}{2}\|X\|_F^2 + \frac{\mu}{\omega^2}\sum_{i,j}\left[1 - \cos\big(\omega(R^\top X)_{ij}\big)\right], \tag{74}$$

$$\nabla f(X) = \lambda X + \frac{\mu}{\omega} R \sin\big(\omega R^\top X\big), \tag{75}$$

where the sine is entrywise. Both terms in f are nonnegative, so $X^\star = 0$ is a global minimizer. Since $\mu > \lambda$, the periodic term also creates regions of negative curvature.

The primal norm is the operator norm and the dual norm is the nuclear norm. Using $\|A\|_* \le \sqrt{d}\|A\|_F$ and $\|A\|_F \le \sqrt{d}\|A\|_{\mathrm{op}}$ gives

$$\|\nabla f(Y) - \nabla f(X)\|_* \le d(\lambda + \mu)\|Y - X\|_{\mathrm{op}}.$$

Thus $L = d(\lambda + \mu) = 17.6$ is a known valid smoothness constant. If $M = U\Sigma V^\top \neq 0$, the exact LMO is $s(M) = -UV^\top$. Hence $\|s(M)\|_{\mathrm{op}} = 1$ and $\langle M, s(M)\rangle = -\|M\|_*$. We set $s(0) = 0$.

The oracle returns $G(X) = \nabla f(X) + Z$, where the entries of Z are independent $\mathcal{N}(0, (\sigma/d)^2)$ variables. It is unbiased and $\mathbb{E}\|Z\|_F^2 = \sigma^2$. Because Q is orthogonal, $\|X_0 - X^\star\|_{\mathrm{op}} = D$ exactly. The values $D = 1$ and $D = 2$ therefore change only the initial distance scale for a fixed problem instance. They are used to test sensitivity to an unknown scale. Neither value is passed to an update rule. The fixed baselines exploit the cell identity only through their offline per-cell radius tuning.

I.2 Methods and fixed-radius tuning

All methods use $\alpha = T^{-1/2}$. Predictable DF-Muon and the post-query ablation use $H_0 = 32$. For the tested horizons, (42) reduces to

$$q = \frac{L}{2} + \sqrt{2}L\sqrt{T}.$$

We compare four methods.

1. **Predictable DF-Muon.** This is Algorithm 3. It forms s_k and $\eta_k = (-A_{k-1}/(2q(H_0 + k - 1)))_+$ before drawing G_k.
2. **Matched-timing fixed.** This method uses the same past-measurable $s_k = s(m_k)$ as Predictable DF-Muon, but replaces η_k by a tuned constant stepsize. It isolates stepsize adaptation while keeping the information timing fixed.
3. **Fresh-momentum Muon.** This is the stronger practical baseline. It draws G_k, updates $m_{k+1} = (1 - \alpha)m_k + \alpha G_k$, forms $s(m_{k+1})$, and then takes the tuned constant-radius step. Thus the fresh sample affects the current direction, as in standard Muon timing.
4. **Post-query FTRL.** This ablation draws G_k and uses the current one-dimensional feedback $b_k = \langle G_k, s_k\rangle$ in the current radius. It is deliberately non-predictable and is not covered by our analysis.

The two fixed-radius baselines are tuned independently. We first search $\eta = 2^p/L$ for integer $p \in \{-10, -9, \dots, 4\}$ and then refine around the best exponent in quarter steps. The chosen radius minimizes the average true gradient norm on ten tuning noise seeds. Tuning is repeated for every method, problem instance, T, D, and σ. None of those tuning seeds is used for evaluation. The search optimum is interior in every cell, so no result is caused by a truncated tuning range.

We evaluate seven horizons $T \in \{512, 1024, \dots, 32768\}$, four (D, σ) cells, five independently oriented problem instances, and ten fresh noise seeds per instance. Methods share noise streams within an instance and evaluation seed. For each instance, we first average over its ten noise seeds. Ratios are then formed within the instance, and the reported mean and two-sided 95% Student-t interval are computed across the five problem instances. The fifty trajectories are therefore not treated as fifty independent problem instances.

I.3 Average gradient norm and crossover

For every run we record the average gradient norm

$$\overline{g}_T := \frac{1}{T}\sum_{k=1}^{T} \|\nabla f(X_k)\|_*,$$

If τ is independent and uniformly distributed over $\{1, \dots, T\}$, then $\mathbb{E}\overline{g}_T = \mathbb{E}\|\nabla f(X_\tau)\|_*$. Figure 4 shows this quantity, and Figure 5 reports the paired ratio against fresh-momentum Muon.

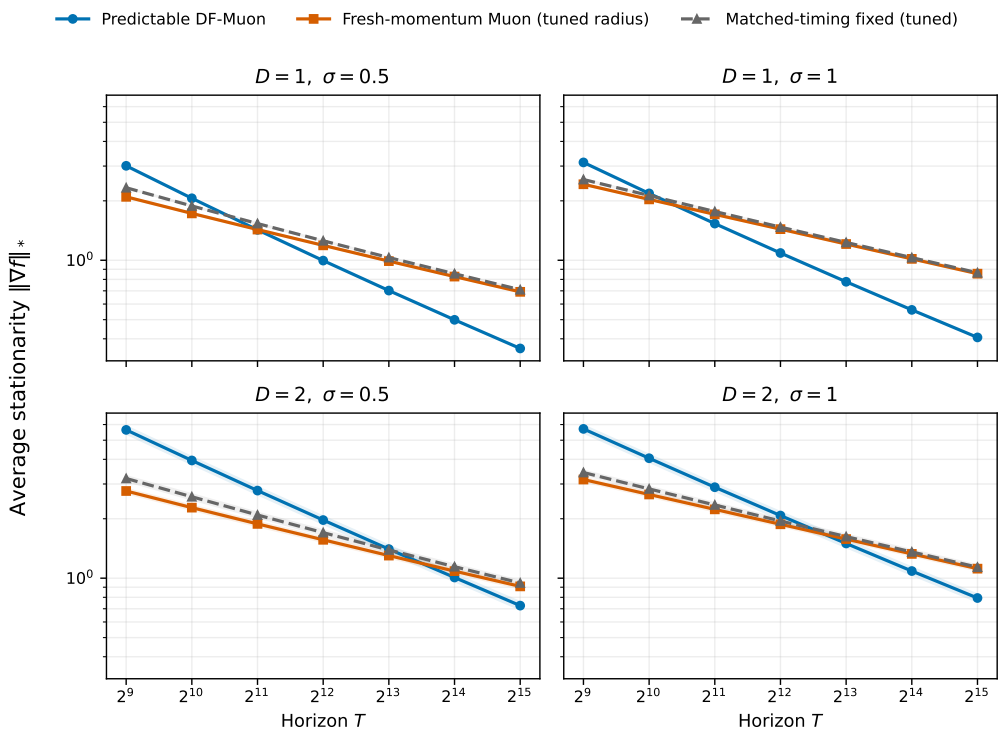


Figure 4: Average gradient norm for the exact predictable method and the two separately tuned fixed-radius baselines. Shaded regions are 95% Student-t intervals across five problem instances after averaging ten noise seeds within each instance.

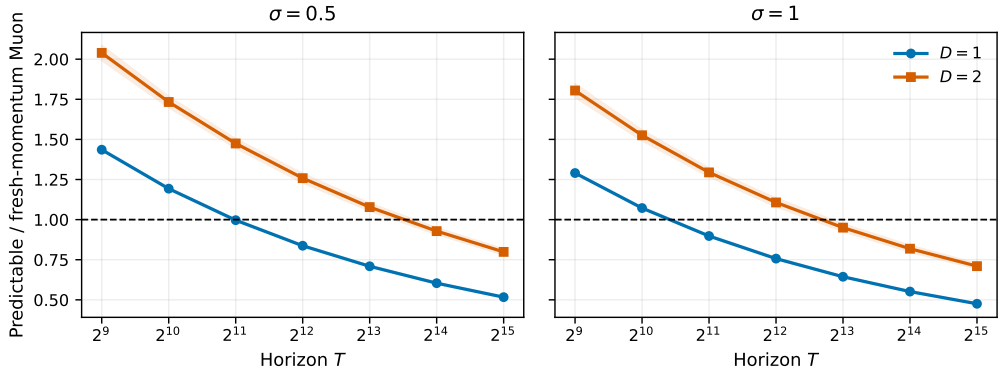


Figure 5: Predictable DF-Muon relative to separately tuned fresh-momentum Muon. The metric is $T^{-1}\sum_{k=1}^{T}\|\nabla f(X_k)\|_*$. Values below one favor the predictable FTRL method. Ten noise seeds are averaged within each problem instance, followed by a mean and a 95% Student-t interval across five independently oriented instances. All short-horizon losses are shown.

Predictable DF-Muon is not uniformly best. At $T = 512$, its mean ratio to the fresh-momentum baseline ranges from 1.29 to 2.04. The ratio decreases steadily with the horizon. The corresponding crossover horizons are listed in Table 2. By $T = 32768$ the ratios lie between 0.4755 and 0.7989, and the predictable method has a lower average gradient norm on all five instances in every cell. Increasing D delays the crossover. Changing σ over the tested range has a smaller effect and slightly advances it.

The explicit horizon-dependent theorem bound is conservative on this problem. Its ratio to the observed average gradient norm ranges from 5.93 to 21.49 over the full grid and grows with T. We include this comparison to calibrate the finite-time constants, not as evidence that the empirical curve follows a tight upper bound.

I.4 Timing and predictability ablations

Figure 6 removes the fresh-direction advantage by comparing against the matched-timing fixed baseline. The predictable FTRL method crosses this baseline at $T = 2048$ for both $D = 1$ cells, at $T = 8192$ for $(D, \sigma) = (2, 1)$, and at $T = 16384$ for $(D, \sigma) = (2, 0.5)$. The fresh-momentum fixed method is itself better than matched timing in every tested cell. At $T = 32768$ the remaining difference is about 1.1%–4.1%.

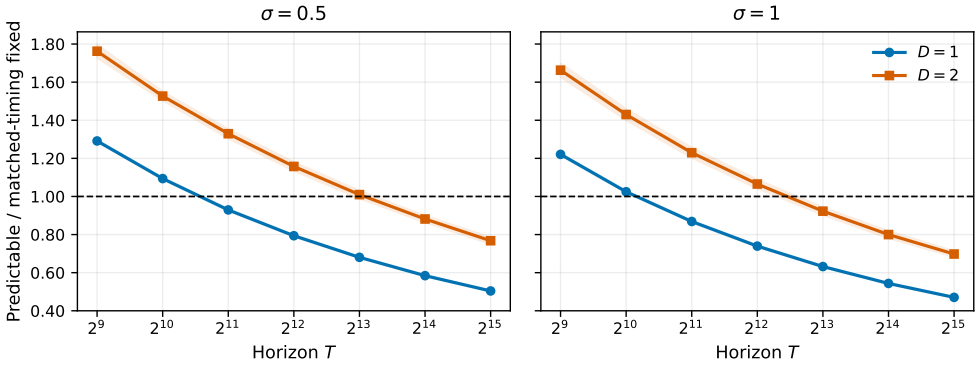


Figure 6: Predictable DF-Muon relative to the separately tuned fixed-radius baseline with matched lagged-direction timing. Values below one favor the predictable FTRL method.

The post-query ablation asks how much is lost by enforcing the information order needed by the proof. As shown in Figure 7, post-query FTRL is only 0.02%–0.64% better on this grid. This small difference does not validate post-query reuse: its radius is correlated with the current noise, so it lacks the conditional-unbiasedness step used in the theorem.

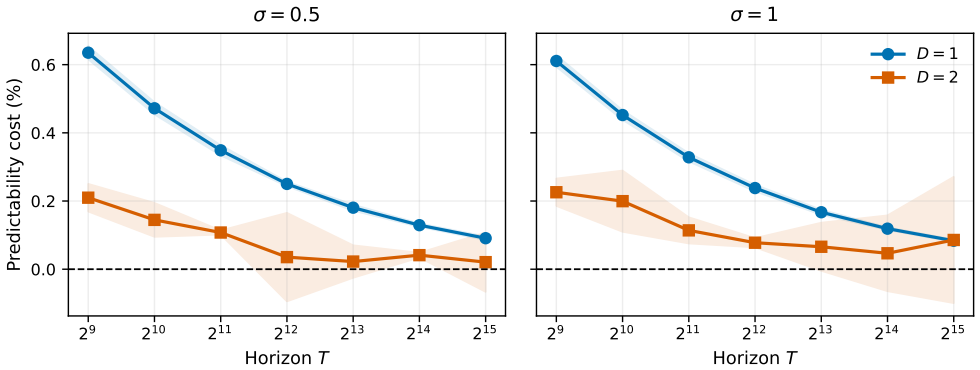


Figure 7: Empirical cost of predictability, measured relative to the non-predictable post-query FTRL ablation. The ablation uses the current sample to choose the current radius and is not covered by the theory.

I.5 Learned radius

Figure 8 shows the dimensionless radius $L\eta_k$ at the longest horizon. The predictable FTRL method generates a rise-and-decay schedule rather than estimating one constant radius. It reaches its peak

between rounds 287 and 444, then cools gradually. The only zero-radius round is the required initialization round $\eta_1 = 0$. The observed zero fraction is $1/32768$ in every cell.

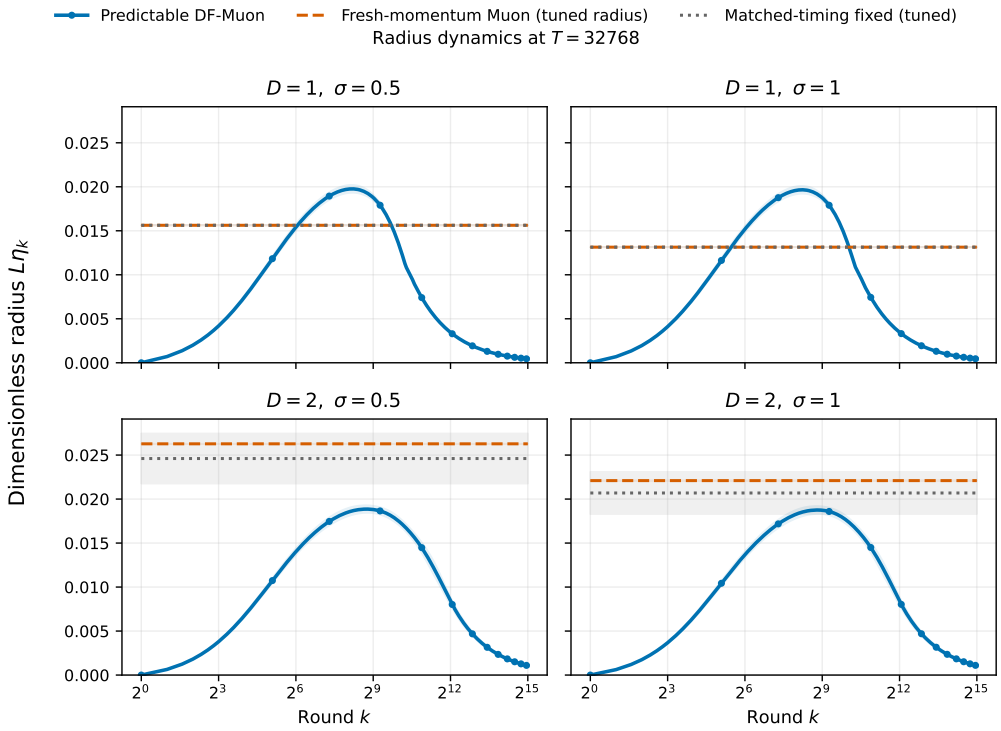


Figure 8: Learned radius at $T = 32768$. Horizontal lines are the separately tuned constant-radius references. Proximity to either reference is not a goal of the method.

Table 15: Predictable radius diagnostics at $T = 32768$. Ratios use the tuned fresh-momentum radius in the corresponding cell.

D	σ	Peak round	Peak $L\eta$	Peak/fresh	Mean/fresh
1	0.5	287	0.01976	1.265	0.130
1	1	303	0.01966	1.496	0.155
2	0.5	420	0.01886	0.718	0.149
2	1	444	0.01876	0.849	0.178

The trajectories are nearly unchanged when only σ changes. Doubling D shifts the peak later and lengthens the decay. For $D = 1$, the peak exceeds the tuned fresh-momentum radius. For $D = 2$, it remains below it. This confirms that the method is not simply reconstructing the offline-tuned constant. The final radius becomes a small fraction of that reference as the horizon grows, as summarized in Figure 9.

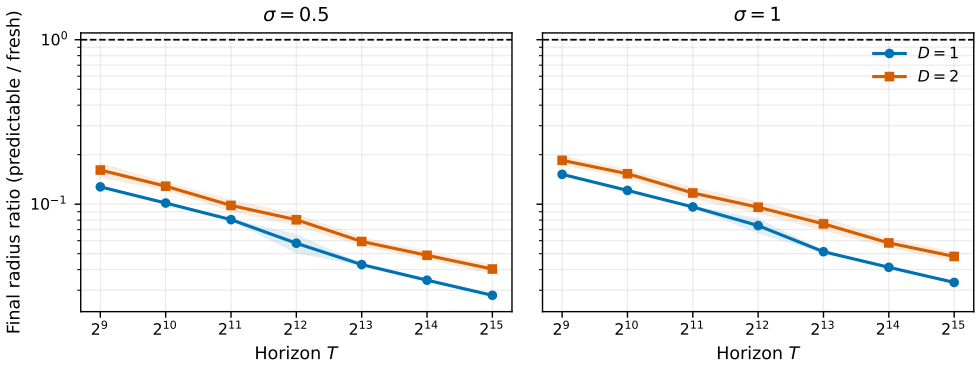


Figure 9: Final predictable radius divided by the tuned fresh-momentum radius. The vertical scale is logarithmic. This is a scale diagnostic, not a performance metric: a smaller final radius is not by itself better.

I.6 Checks and scope

The final run contains 5600 evaluation trajectories and no non-finite run. All 140 evaluation cells completed, the candidate-grid optima were interior, and recomputation from seed-level outputs reproduced every reported aggregate exactly. The LMO test includes a nonzero matrix of magnitude 10^{-30} to ensure that small inputs are not incorrectly rounded to the zero direction.

This experiment directly checks the general smooth non-convex FTRL method and its one-gradient predictable information order. It does not test the star-convex last-iterate theorem, the deterministic distance-free method, or large-model behavior. The star-convex diagnostic and practical Transformer implementations are described separately in Appendix H.

J Recentered Predictable Method with EMA Momentum

This appendix gives the initialization-centered EMA version of the predictable distance-free method, including its finite-time theorem, constants, and deterministic comparison. Its direction is $s(m_k)$ and its reference point is x_0. The fixed-reference method with a one-gradient-delayed Nesterov-style momentum direction from Theorem 3.1 has a separate proof in Appendix B. The results below apply to the EMA algorithm and parameters specified here.

The deterministic distance-free method chooses its radius from a majorant that contains the current gradient. Reusing the same minibatch to choose the radius and to evaluate its linear term breaks the conditional unbiasedness used in stochastic analysis. We therefore choose the direction and radius before drawing the fresh sample. The sample is then used once to update the momentum and the one-dimensional state for later rounds.

The center and ray are unchanged:

$$x_{k+1} = x_0 + (1-\beta)(x_k - x_0) + \beta R_k s_k, \qquad R_k \geq 0.$$

The method still competes in the proof with the unknown radius

$$D := \|x_0 - x^\star\|,$$

but neither D nor an estimate of D is passed to the algorithm.

J.1 Predictable Radius Update

Fix a horizon $T \geq 1$ and parameters $\alpha, \beta \in (0,1)$. Let

$$r := 1 - \frac{\beta}{2}, \qquad w_k := r^{T-k}, \quad k = 1, \dots, T.$$

Choose $\rho_{\text{search}} > 0$ and

$$M \geq 2(1 + 2\rho_{\text{search}}), \qquad c_0 := 1 + 2\rho_{\text{search}} + M, \qquad \kappa := 64.$$

The quadratic regularization coefficient is

$$q_R := \frac{\kappa c_0 L \beta^2}{\alpha}. \tag{76}$$

Initialize

$$H_0 := \frac{w_1}{1-r}, \qquad A_0 = W_0 := 0.$$

At round k, before observing G_k, set

$$R_k := \left[-\frac{\beta A_{k-1}}{2q_R(H_0 + W_{k-1})} \right]_+ . \tag{77}$$

Equivalently, R_k minimizes over $R \geq 0$ the one-dimensional objective

$$q_R H_0 R^2 + \sum_{i=1}^{k-1} w_i \left(\beta \langle G_i, s_i \rangle R + q_R R^2 \right). \tag{78}$$

FTRL therefore serves only as the radius update inside the distance-free method.

The initialization uses one gradient, followed by one fresh gradient in each of the T rounds. The total is $T+1$ stochastic-gradient evaluations. No round uses two gradients, a split batch, or a validation batch. The first radius is $R_1 = 0$, so the first update is zero.

J.1.1 Convergence

Write $s_\sigma := \rho_{\text{norm}} \sigma$. The proof in Appendix K first establishes finite second moments, then applies conditional unbiasedness to the one-dimensional comparison. It also proves the FTRL regret and the weighted momentum bounds used below.

Algorithm 5 DF-Muon with a predictable stochastic radius rule

Require: $x_0 \in \mathcal{E}$, horizon T, $\alpha, \beta \in (0,1)$, $\rho_{\text{search}} > 0$, $M \geq 2(1 + 2\rho_{\text{search}})$
1: Draw $G_0 = g(x_0; \xi_0)$. Set $x_1 \leftarrow x_0$ and $m_1 \leftarrow G_0$
2: Set $A_0 = W_0 = 0$, $r = 1 - \beta/2$, $w_k = r^{T-k}$, $H_0 = w_1/(1-r)$, and q_R by (76)
3: **for** $k = 1, 2, \ldots, T$ **do**
4: Choose $s_k = s(m_k)$ using the fixed LMO tie-breaking rule
5: Compute R_k by (77)
6: Store $x_{k+1} = x_0 + (1-\beta)(x_k - x_0) + \beta R_k s_k$
7: Draw the only fresh gradient of the round, $G_k = g(x_k; \xi_k)$
8: Apply the stored update $x_k \mapsto x_{k+1}$
9: $m_{k+1} \leftarrow (1-\alpha)m_k + \alpha G_k$
10: $A_k \leftarrow A_{k-1} + w_k \langle G_k, s_k \rangle$ and $W_k \leftarrow W_{k-1} + w_k$
11: **end for**
12: **return** x_{T+1}

Theorem J.1 (Predictable stochastic distance-free method)**.** *Let Assumptions 2.1, 2.3, and 2.2 hold. Run Algorithm 5 with*

$$0 < \alpha \leq 1, \qquad 0 < \beta \leq \min\left\{\frac{\alpha}{4}, \frac{1}{64}\right\}.$$

Let $r = 1 - \beta/2$, $\delta = \alpha - \beta/2$, and

$$C_{\text{ctrl}} := 3\kappa c_0 + \frac{4608}{49\kappa c_0}.$$

Then

$$\mathbb{E}[f(x_{T+1}) - f(x^\star)] \leq r^T\big(f(x_0) - f(x^\star)\big) + C_{\text{ctrl}}\frac{L\beta D^2}{\alpha} + 4Ds_\sigma\sqrt{\frac{\alpha}{2-\alpha}} + \frac{2\beta D s_\sigma}{\delta}r^T + \frac{4\alpha s_\sigma^2}{\kappa c_0 L}. \tag{79}$$

The parameters of the algorithm do not use D.

For deterministic gradients, any fixed $\alpha \in (0,1]$ and $\beta \asymp \log(T)/T$ give the last-iterate order $\widetilde{O}(LD^2/T)$. With stochastic gradients, α must decrease with the horizon to remove the momentum noise floor, but it need not depend on D.

Corollary J.2 (Horizon-dependent stochastic schedule)**.** *Set*

$$\beta = \min\left\{\frac{1}{512}, \frac{2\log(T+1)}{T}\right\}, \qquad \alpha = 8\beta^{2/3}.$$

Then the conditions of Theorem J.1 hold for every $T \geq 1$, and

$$\mathbb{E}[f(x_{T+1}) - f(x^\star)] \leq LD^2\exp\left(-\frac{\beta T}{2}\right) + \frac{C_{\text{ctrl}}}{8}LD^2\beta^{1/3} + \left(8\sqrt{2} + \frac{1}{2}\right)Ds_\sigma\beta^{1/3} + \frac{32}{\kappa c_0}\frac{s_\sigma^2}{L}\beta^{2/3}. \tag{80}$$

The second branch in the definition of β is active exactly from $T = 9365$ onward, so for those horizons $\exp(-\beta T/2) = 1/(T+1)$. Therefore, asymptotically,

$$\mathbb{E}[f(x_{T+1}) - f(x^\star)] = \widetilde{O}\left(\frac{LD^2 + D\rho_{\text{norm}}\sigma}{T^{1/3}} + \frac{\rho_{\text{norm}}^2\sigma^2/L}{T^{2/3}}\right).$$

This schedule depends only on T, not on D or σ.

J.1.2 Relation to deterministic DF-Muon

The two methods share the Muon geometry, the recentered ray, the domain $R \geq 0$, and the fixed analysis choice $R = D$. Their radius updates differ because they use different information.

	Deterministic DF	**Predictable stochastic DF**
Radius update	fresh majorized search	cumulative one-dimensional FTRL
Direction	uses m_{k+1}	uses past-measurable m_k
Explicit D-certificate	yes	no
Distance input	no	no
Gradient cost	one per round	one per round, plus initialization
Deterministic order	$\widetilde{O}(1/T)$	$\widetilde{O}(1/T)$

The iterates generally differ even when $\sigma = 0$. The common object is the distance-free recentered method, not the exact one-dimensional recursion.

K Proofs for Predictable Stochastic Distance-Free Muon

Throughout this section,

$$D := \|x_0 - x^\star\|, \qquad s_\sigma := \rho_{\mathrm{norm}}\sigma, \qquad r := 1 - \frac{\beta}{2}, \qquad \delta := \alpha - \frac{\beta}{2}.$$

All other scalar quantities are as defined in Section J.

K.1 A self-bounding estimate

Lemma K.1 (Self-bounding in the dual norm)**.** *Under Assumption 2.1, for every $x \in \mathcal{E}$,*

$$\|\nabla f(x)\|_*^2 \le 2L\big(f(x) - f(x^\star)\big).$$

Under Assumptions 2.1 and 2.3,

$$f(x_0) - f(x^\star) \le LD^2.$$

Proof. Because $x^\star$ is an unconstrained differentiable minimizer, $\nabla f(x^\star) = 0$. If $\nabla f(x) = 0$, the first claim is immediate. Otherwise, choose u with $\|u\| = 1$ and $\langle \nabla f(x), u\rangle = \|\nabla f(x)\|_*$. Smoothness at

$$y = x - \frac{\|\nabla f(x)\|_*}{L} u$$

gives

$$f(y) \le f(x) - \frac{\|\nabla f(x)\|_*^2}{2L}.$$

Since $f(x^\star) \le f(y)$, the first claim follows. For the second claim, star-convexity and smoothness give

$$f(x_0) - f(x^\star) \le \langle \nabla f(x_0), x_0 - x^\star\rangle \le \|\nabla f(x_0)\|_* D \le LD^2.$$

Q.E.D.

K.2 Filtration and predictability

Let $\mathcal{F}_k := \sigma(G_0, \dots, G_k)$. Then $x_k, m_k, s_k, R_k, A_{k-1}, W_{k-1}$ are $\mathcal{F}_{k-1}$-measurable, whereas G_k is sampled after these quantities have been fixed. Let

$$\zeta_k := G_k - \nabla f(x_k).$$

Because x_k is $\mathcal{F}_{k-1}$-measurable and ξ_k is independent of $\mathcal{F}_{k-1}$, Assumption 2.2 gives

$$\mathbb{E}[\zeta_k \mid \mathcal{F}_{k-1}] = 0, \qquad \mathbb{E}[\|\zeta_k\|_*^2 \mid \mathcal{F}_{k-1}] \le s_\sigma^2. \tag{81}$$

Lemma K.2 (Finite moments)**.** *Under Assumptions 2.1 and 2.2, for every $k = 1, \dots, T$,*

$$\mathbb{E}\|x_k - x^\star\|^2 < \infty, \qquad \mathbb{E}\|\nabla f(x_k)\|_*^2 < \infty, \qquad \mathbb{E}\|G_k\|_*^2 < \infty, \qquad \mathbb{E}R_k^2 < \infty.$$

Consequently, for every deterministic $R \ge 0$,

$$\langle \zeta_k, s_k\rangle (R_k - R)$$

is integrable.

Proof. We use induction. The base point $x_1 = x_0$ is deterministic and $R_1 = 0$. Since $\nabla f(x^\star) = 0$, smoothness gives

$$\|\nabla f(x_k)\|_* \le L\|x_k - x^\star\|.$$

Moreover, by (81) and $(a+b)^2 \le 2a^2 + 2b^2$,

$$\mathbb{E}\|G_k\|_*^2 \le 2\mathbb{E}\|\nabla f(x_k)\|_*^2 + 2s_\sigma^2.$$

Thus a finite second moment of $x_k - x^\star$ implies finite second moments of $\nabla f(x_k)$ and G_k. Since $\|s_k\| \le 1$, the scalar $b_k = \langle G_k, s_k\rangle$ also has a finite second moment.

Suppose the asserted moments are finite through round k. The update gives

$$\|x_{k+1} - x^\star\| \le D + (1-\beta)\|x_k - x_0\| + \beta R_k,$$

so $x_{k+1} - x^\star$ has a finite second moment. Also, $A_k = \sum_{i=1}^k w_i b_i$ has a finite second moment. For $k < T$,

$$R_{k+1} = \left[-\frac{\beta A_k}{2q_R(H_0 + W_k)}\right]_+.$$

The denominator is deterministic and strictly positive, so $\mathbb{E}R_{k+1}^2 < \infty$. This closes the induction.

Finally,

$$|\langle \zeta_k, s_k\rangle(R_k - R)| \le \|\zeta_k\|_*(R_k + R),$$

which is integrable by Cauchy–Schwarz, (81), and $\mathbb{E}R_k^2 < \infty$. **Q.E.D.**

Lemma K.3 (Predictable one-dimensional comparison)**.** *For every deterministic $R \ge 0$,*

$$\mathbb{E}\left[\sum_{k=1}^T w_k \beta\langle G_k - \nabla f(x_k), s_k\rangle(R_k - R)\right] = 0.$$

Proof. For every k, s_k, R_k, and R are $\mathcal{F}_{k-1}$-measurable. The conditional expectation of the kth summand is therefore zero by (81). Sum and apply the tower property. **Q.E.D.**

K.3 One-step inequality

Define

$$y_k := x_k - x_0, \qquad u^\star := x^\star - x_0, \qquad \Delta_k := f(x_k) - f(x^\star),$$
$$\mathcal{L}_k := \Delta_k + M\beta L\|y_k\|^2, \qquad e_k := \nabla f(x_k) - m_k.$$

The true one-dimensional model is

$$\phi_k(R) := \beta\langle \nabla f(x_k), s_k\rangle R + q_R R^2. \tag{82}$$

Lemma K.4 (One-step competitive inequality)**.** *Under Assumptions 2.1 and 2.3, for every $k = 1, \dots, T$,*

$$\begin{aligned}\mathcal{L}_{k+1} + \rho_{\text{search}} L\|x_{k+1} - x_k\|^2 + \frac{\beta}{2}\Delta_k + (q_R - c_0 L\beta^2)R_k^2 \\ \le r\mathcal{L}_k + \phi_k(R_k) - \phi_k(D) + q_R D^2 + 2\beta D\|e_k\|_*.\end{aligned} \tag{83}$$

Proof. Smoothness and $x_{k+1} - x_k = \beta(R_k s_k - y_k)$ imply

$$f(x_{k+1}) \le f(x_k) + \beta\langle \nabla f(x_k), R_k s_k - y_k\rangle + \frac{L\beta^2}{2}\|R_k s_k - y_k\|^2.$$

Decompose

$$R_k s_k - y_k = (u^\star - y_k) + (Ds_k - u^\star) + (R_k - D)s_k.$$

Star-convexity gives

$$\langle \nabla f(x_k), u^\star - y_k\rangle \le -\Delta_k.$$

Since s_k minimizes $s \mapsto \langle m_k, s\rangle$ over the unit ball and $\|u^\star/D\| \le 1$ when $D > 0$,

$$\langle m_k, Ds_k - u^\star\rangle \le 0.$$

For $D = 0$ the same expression is zero. Hence

$$\langle \nabla f(x_k), Ds_k - u^\star \rangle \le \|e_k\|_* \|Ds_k - u^\star\| \le 2D\|e_k\|_*.$$

It follows that

$$\beta\langle \nabla f(x_k), R_k s_k - y_k \rangle \le -\beta\Delta_k + 2\beta D\|e_k\|_* + \beta(R_k - D)\langle \nabla f(x_k), s_k \rangle. \tag{84}$$

The recentered recursion gives

$$y_{k+1} = (1-\beta)y_k + \beta R_k s_k.$$

By convexity of the squared norm and $\|s_k\| \le 1$,

$$\|y_{k+1}\|^2 \le (1-\beta)\|y_k\|^2 + \beta R_k^2. \tag{85}$$

Moreover,

$$\|R_k s_k - y_k\|^2 \le 2R_k^2 + 2\|y_k\|^2. \tag{86}$$

Add $M\beta L\|y_{k+1}\|^2$ and $\rho_{\text{search}} L\|x_{k+1} - x_k\|^2$ to the left of the smoothness inequality, then use (84)–(86). This yields

$$\begin{aligned}
&\mathcal{L}_{k+1} + \rho_{\text{search}} L\|x_{k+1} - x_k\|^2 + \frac{\beta}{2}\Delta_k \\
&\qquad \le r\mathcal{L}_k + \beta(R_k - D)\langle \nabla f(x_k), s_k \rangle + c_0 L\beta^2 R_k^2 + 2\beta D\|e_k\|_*.
\end{aligned}$$

The coefficient of $\|y_k\|^2$ is valid because

$$M\beta L(1-\beta) + L\beta^2(1 + 2\rho_{\text{search}}) \le M\beta L\left(1 - \frac{\beta}{2}\right)$$

when $M \ge 2(1 + 2\rho_{\text{search}})$. Finally, by (82),

$$\begin{aligned}
&\beta(R_k - D)\langle \nabla f(x_k), s_k \rangle + c_0 L\beta^2 R_k^2 \\
&\qquad = \phi_k(R_k) - \phi_k(D) + q_R D^2 - (q_R - c_0 L\beta^2)R_k^2.
\end{aligned}$$

Move the last term to the left. **Q.E.D.**

K.4 One-dimensional FTRL regret

For $b_k := \langle G_k, s_k \rangle$, define

$$\widehat{\ell}_k(R) := w_k(\beta b_k R + q_R R^2), \qquad R \ge 0.$$

Algorithm 5 is FTRL with initial regularizer $q_R H_0 R^2$ for these one-dimensional surrogate losses.

Lemma K.5 (Pathwise FTRL regret)**.** *For every realized sequence $b_1, \dots, b_T \in \mathbb{R}$ and every $R \ge 0$,*

$$\sum_{k=1}^{T}\big(\widehat{\ell}_k(R_k) - \widehat{\ell}_k(R)\big) \le q_R H_0 R^2 + \sum_{k=1}^{T} \frac{w_k^2(\beta b_k + 2q_R R_k)^2}{q_R(H_0 + W_k)}. \tag{87}$$

Furthermore,

$$H_0 + W_k \ge \frac{w_k}{1-r} = \frac{2w_k}{\beta}, \tag{88}$$

and hence

$$\sum_{k=1}^{T}\big(\widehat{\ell}_k(R_k) - \widehat{\ell}_k(R)\big) \le q_R H_0 R^2 + \frac{\beta^3}{q_R}\sum_{k=1}^{T} w_k b_k^2 + 4\beta q_R \sum_{k=1}^{T} w_k R_k^2. \tag{89}$$

Proof. Let

$$F_k(R) := q_R H_0 R^2 + \sum_{i=1}^{k} \widehat{\ell}_i(R).$$

Then R_{k+1} minimizes F_k over $[0,\infty)$ and R_k minimizes F_{k-1}. Set $f_0(R) := q_R H_0 R^2$ and $f_k(R) := \widehat{\ell}_k(R)$ for $k \geq 1$. Since R_{k+1} minimizes $\sum_{i=0}^{k} f_i$, the be-the-leader inequality

$$\sum_{k=0}^{T} f_k(R_{k+1}) \leq \sum_{k=0}^{T} f_k(R)$$

holds for every $R \geq 0$. To see this, apply the claim for horizon $T-1$ at $R = R_{T+1}$, add $f_T(R_{T+1})$, and use the optimality of R_{T+1} for $\sum_{k=0}^{T} f_k$. The base case is the optimality of R_1 for f_0.

Since $R_1 = 0$ and $f_0(R_1) = 0$,

$$\sum_{k=1}^{T} \widehat{\ell}_k(R_{k+1}) \leq q_R H_0 R^2 + \sum_{k=1}^{T} \widehat{\ell}_k(R).$$

Therefore,

$$\sum_{k=1}^{T} \big(\widehat{\ell}_k(R_k) - \widehat{\ell}_k(R)\big) \leq q_R H_0 R^2 + \sum_{k=1}^{T} \big(\widehat{\ell}_k(R_k) - \widehat{\ell}_k(R_{k+1})\big). \tag{90}$$

The function F_k is $2q_R(H_0 + W_k)$-strongly convex on $[0,\infty)$. At its constrained minimizer,

$$F_k'(R_{k+1})(R - R_{k+1}) \geq 0 \qquad \text{for all } R \geq 0.$$

Strong convexity with $R = R_k$ gives

$$F_k(R_k) - F_k(R_{k+1}) \geq q_R(H_0 + W_k)(R_k - R_{k+1})^2.$$

Since R_k minimizes F_{k-1},

$$F_{k-1}(R_k) \leq F_{k-1}(R_{k+1}),$$

and consequently

$$q_R(H_0 + W_k)(R_k - R_{k+1})^2 \leq \widehat{\ell}_k(R_k) - \widehat{\ell}_k(R_{k+1}).$$

Because $\widehat{\ell}_k$ is quadratic,

$$\begin{aligned}\widehat{\ell}_k(R_k) - \widehat{\ell}_k(R_{k+1}) &= w_k(\beta b_k + 2q_R R_k)(R_k - R_{k+1}) - w_k q_R (R_k - R_{k+1})^2 \\ &\leq w_k |\beta b_k + 2q_R R_k|\, |R_k - R_{k+1}|.\end{aligned}$$

Thus

$$|R_k - R_{k+1}| \leq \frac{w_k |\beta b_k + 2q_R R_k|}{q_R(H_0 + W_k)}.$$

Substitution into (90) proves (87).

Finally,

$$W_k = w_k \frac{1 - r^k}{1 - r}, \qquad H_0 = \frac{w_1}{1 - r} = \frac{w_k r^{k-1}}{1 - r},$$

so

$$H_0 + W_k = \frac{w_k(1 - r^k + r^{k-1})}{1 - r} \geq \frac{w_k}{1 - r}.$$

Using this bound in (87) and $(a+b)^2 \leq 2a^2 + 2b^2$ gives (89). **Q.E.D.**

Lemma K.6 (Expected one-dimensional comparison)**.** *Assume $0 < \alpha \leq 1$ and set $R = D$ in Lemma K.5. Then*

$$\mathbb{E} \sum_{k=1}^{T} w_k \big(\phi_k(R_k) - \phi_k(D)\big) \leq q_R H_0 D^2 + \frac{\beta}{16} \sum_{k=1}^{T} w_k \mathbb{E}\Delta_k + 4\beta q_R \sum_{k=1}^{T} w_k \mathbb{E} R_k^2 + \frac{4\alpha s_\sigma^2}{\kappa c_0 L}. \tag{91}$$

Proof. By Lemma K.3, the expected realized FTRL regret against D equals the expected true-model comparison. Conditional on $\mathcal{F}_{k-1}$,

$$\begin{aligned}\mathbb{E}[\langle G_k, s_k\rangle^2 \mid \mathcal{F}_{k-1}] &\le 2\langle \nabla f(x_k), s_k\rangle^2 + 2\mathbb{E}[\langle \zeta_k, s_k\rangle^2 \mid \mathcal{F}_{k-1}] \\ &\le 2\|\nabla f(x_k)\|_*^2 + 2s_\sigma^2 \\ &\le 4L\Delta_k + 2s_\sigma^2,\end{aligned}$$

where the last inequality is Lemma K.1. Using (89), (76), and $\sum_{k=1}^T w_k \le 2/\beta$,

$$\begin{aligned}\frac{\beta^3}{q_R}\sum_{k=1}^T w_k \mathbb{E}\langle G_k, s_k\rangle^2 &\le \frac{4\alpha\beta}{\kappa c_0}\sum_{k=1}^T w_k \mathbb{E}\Delta_k + \frac{4\alpha s_\sigma^2}{\kappa c_0 L} \\ &\le \frac{\beta}{16}\sum_{k=1}^T w_k \mathbb{E}\Delta_k + \frac{4\alpha s_\sigma^2}{\kappa c_0 L},\end{aligned}$$

where $\kappa = 64$, $c_0 \ge 1$, and $\alpha \le 1$ were used in the last line. **Q.E.D.**

K.5 Momentum tracking and the recentered filter

Let

$$q_m := 1 - \alpha, \qquad \delta := r - q_m = \alpha - \frac{\beta}{2}, \qquad a_i := \|x_i - x_{i-1}\|.$$

Lemma K.7 (Pointwise stochastic momentum tracking)**.** *If $\delta > 0$, then for every $k \ge 1$,*

$$\mathbb{E}\|e_k\|_* \le L\sum_{i=2}^k q_m^{k-i}\mathbb{E}a_i + s_\sigma q_m^{k-1} + s_\sigma\sqrt{\frac{\alpha}{2-\alpha}}. \tag{92}$$

Proof. The warm start gives $e_1 = \nabla f(x_0) - G_0 = -\zeta_0$. For $k \ge 2$,

$$e_k = q_m e_{k-1} + \nabla f(x_k) - \nabla f(x_{k-1}) - \alpha\zeta_{k-1}.$$

Unrolling,

$$e_k = q_m^{k-1}e_1 + \sum_{i=2}^k q_m^{k-i}\big(\nabla f(x_i) - \nabla f(x_{i-1})\big) - \alpha\sum_{i=2}^k q_m^{k-i}\zeta_{i-1}.$$

The drift has dual norm at most $L\sum_{i=2}^k q_m^{k-i}a_i$. The remaining terms form a weighted sum of Euclidean martingale differences. For $i < j$, the earlier noise is $\mathcal{F}_{j-1}$-measurable and

$$\mathbb{E}\langle \zeta_i, \zeta_j\rangle = \mathbb{E}\langle \zeta_i, \mathbb{E}[\zeta_j \mid \mathcal{F}_{j-1}]\rangle = 0.$$

Thus all Euclidean cross terms vanish, and

$$\begin{aligned}\mathbb{E}\left\|-q_m^{k-1}\zeta_0 - \alpha\sum_{i=2}^k q_m^{k-i}\zeta_{i-1}\right\|_2^2 &\le \sigma^2\left(q_m^{2(k-1)} + \alpha^2\sum_{j=0}^{k-2} q_m^{2j}\right) \\ &\le \sigma^2\left(q_m^{2(k-1)} + \frac{\alpha}{2-\alpha}\right).\end{aligned}$$

Apply $\|v\|_* \le \rho_{\mathrm{norm}}\|v\|_2$, Jensen's inequality, and $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$. **Q.E.D.**

Lemma K.8 (Weighted stable filter)**.** *For the recentered recursion of Algorithm 5,*

$$\sum_{i=2}^T w_i a_i \le \frac{3\beta}{r}\sum_{k=1}^T w_k R_k. \tag{93}$$

Proof. Since $y_1 = 0$ and $y_i = (1-\beta)y_{i-1} + \beta R_{i-1}s_{i-1}$,

$$y_{i-1} = \beta \sum_{j=1}^{i-2} (1-\beta)^{i-2-j} R_j s_j.$$

Moreover,

$$a_i = \beta \|R_{i-1}s_{i-1} - y_{i-1}\| \le \beta R_{i-1} + \beta \|y_{i-1}\|.$$

The first contribution satisfies

$$\beta \sum_{i=2}^{T} w_i R_{i-1} = \frac{\beta}{r} \sum_{j=1}^{T-1} w_j R_j.$$

For the second contribution, expand y_{i-1} and exchange sums:

$$\begin{aligned}\sum_{i=2}^{T} w_i \|y_{i-1}\| &\le \beta \sum_{j=1}^{T-2} R_j \sum_{i=j+2}^{T} w_i (1-\beta)^{i-2-j} \\ &\le \frac{2}{r} \sum_{j=1}^{T-2} w_j R_j.\end{aligned}$$

Indeed, for $j \le T-2$,

$$\begin{aligned}\sum_{i=j+2}^{T} w_i (1-\beta)^{i-2-j} &= \frac{w_j}{r^2} \sum_{\ell=0}^{T-j-2} \left(\frac{1-\beta}{r}\right)^{\ell} \\ &\le \frac{w_j}{r^2} \frac{1}{1-(1-\beta)/r} = \frac{2w_j}{\beta r},\end{aligned}$$

where $r - (1-\beta) = \beta/2$. Combining the two contributions proves (93). **Q.E.D.**

Lemma K.9 (Weighted momentum contribution)**.** *Suppose* $0 < \beta \le \alpha/4$*. Then*

$$\begin{aligned}2\beta D \sum_{k=1}^{T} w_k \mathbb{E}\|e_k\|_* \le \frac{q_R}{4} \sum_{k=1}^{T} w_k \mathbb{E} R_k^2 + \frac{4608}{49\kappa c_0} \frac{L\beta D^2}{\alpha} \\ + 4Ds_\sigma \sqrt{\frac{\alpha}{2-\alpha}} + \frac{2\beta D s_\sigma}{\delta} r^T. \qquad (94)\end{aligned}$$

Proof. Sum (92) with weights w_k. The initial-noise convolution satisfies

$$\sum_{k=1}^{T} w_k q_m^{k-1} = \frac{r^T - q_m^T}{r - q_m} \le \frac{r^T}{\delta},$$

and $\sum_{k=1}^{T} w_k \le 2/\beta$. After multiplication by $2\beta D$, the two stochastic contributions give the final two terms of (94).

For the drift, exchange sums and use $q_m < r$:

$$\begin{aligned}\sum_{k=1}^{T} w_k \sum_{i=2}^{k} q_m^{k-i} \mathbb{E} a_i &\le \frac{r}{\delta} \sum_{i=2}^{T} w_i \mathbb{E} a_i \\ &\le \frac{3\beta}{\delta} \sum_{k=1}^{T} w_k \mathbb{E} R_k,\end{aligned}$$

where Lemma K.8 was used. Its contribution is at most

$$\frac{6\beta^2 DL}{\delta} \sum_{k=1}^{T} w_k \mathbb{E} R_k.$$

By Jensen's inequality, weighted Cauchy–Schwarz, and $\sum_k w_k \le 2/\beta$,

$$\frac{6\beta^2 DL}{\delta}\sum_{k=1}^{T} w_k \mathbb{E}R_k \le \frac{6\sqrt{2}\,\beta^{3/2}DL}{\delta}\left(\sum_{k=1}^{T} w_k \mathbb{E}R_k^2\right)^{1/2}.$$

Young's inequality with coefficient $q_R/4$ yields

$$\frac{6\sqrt{2}\,\beta^{3/2}DL}{\delta}\left(\sum_{k=1}^{T} w_k \mathbb{E}R_k^2\right)^{1/2} \le \frac{q_R}{4}\sum_{k=1}^{T} w_k \mathbb{E}R_k^2 + \frac{72\alpha L\beta D^2}{\kappa c_0 \delta^2}.$$

Since $\beta \le \alpha/4$ implies $\delta \ge 7\alpha/8$, the last term is at most

$$\frac{4608}{49\kappa c_0}\frac{L\beta D^2}{\alpha}.$$

Q.E.D.

K.6 Proof of Theorem J.1

Proof. Multiply (83) by $w_k = r^{T-k}$ and sum from $k = 1$ to T. The Lyapunov terms telescope:

$$\sum_{k=1}^{T}\big(w_k\mathcal{L}_{k+1} - r w_k \mathcal{L}_k\big) = \mathcal{L}_{T+1} - r^T\mathcal{L}_1.$$

Since $x_1 = x_0$, $\mathcal{L}_1 = f(x_0) - f(x^\star)$. Hence

$$\begin{aligned}
&\mathbb{E}\mathcal{L}_{T+1} + \rho_{\text{search}} L\sum_{k=1}^{T} w_k \mathbb{E}\|x_{k+1} - x_k\|^2 + \frac{\beta}{2}\sum_{k=1}^{T} w_k \mathbb{E}\Delta_k \\
&\quad + (q_R - c_0 L\beta^2)\sum_{k=1}^{T} w_k \mathbb{E}R_k^2 \\
&\quad \le r^T\Delta_0 + \mathbb{E}\sum_{k=1}^{T} w_k\big(\phi_k(R_k) - \phi_k(D)\big) + q_R D^2\sum_{k=1}^{T} w_k \\
&\quad\quad + 2\beta D\sum_{k=1}^{T} w_k \mathbb{E}\|e_k\|_*.
\end{aligned}$$

Apply Lemmas K.6 and K.9. The comparison lemma uses at most $\beta/16$ of the gap term and $4\beta q_R$ of the radius term. The momentum lemma uses $q_R/4$ of the radius term. These terms can be absorbed because

$$q_R - c_0 L\beta^2 = q_R\left(1 - \frac{\alpha}{\kappa}\right) \ge \frac{63}{64}q_R,$$

while $4\beta q_R \le q_R/16$, and therefore

$$(q_R - c_0 L\beta^2) - 4\beta q_R - \frac{q_R}{4} \ge \frac{43}{64}q_R > 0.$$

The remaining gap coefficient is $\beta/2 - \beta/16 = 7\beta/16 > 0$. Drop the remaining nonnegative terms. Finally,

$$\sum_{k=1}^{T} w_k + H_0 = \frac{2}{\beta}(1 - r^T + r^{T-1}) \le \frac{3}{\beta},$$

so

$$q_R D^2\left(\sum_{k=1}^{T} w_k + H_0\right) \le 3\kappa c_0 \frac{L\beta D^2}{\alpha}.$$

Use $\mathcal{L}_{T+1} \ge \Delta_{T+1}$ and collect the terms. **Q.E.D.**

Corollary K.10 (Deterministic order of the predictable FTRL method)**.** *If $\sigma = 0$ and $\alpha \in (0, 1]$ is fixed, choose*

$$\beta = \min\left\{\frac{\alpha}{4}, \frac{1}{64}, \frac{2\log(T+1)}{T}\right\}.$$

Whenever the last term is active,

$$f(x_{T+1}) - f(x^\star) \le \frac{LD^2}{T+1} + \frac{2C_{\mathrm{ctrl}}}{\alpha}\frac{LD^2\log(T+1)}{T}.$$

Hence $f(x_{T+1}) - f(x^\star) = \widetilde{O}(LD^2/T)$.

Proof. Use $(1-\beta/2)^T \le e^{-\beta T/2} = 1/(T+1)$, Lemma K.1, and Theorem J.1. **Q.E.D.**

K.7 Proof of Corollary J.2

Proof. For $\beta \le 1/512$, the choice $\alpha = 8\beta^{2/3}$ gives $\alpha \le 1/8$, $\beta \le \alpha/4$, and $\delta \ge (127/128)\alpha$. Apply Theorem J.1, use

$$\sqrt{\frac{\alpha}{2-\alpha}} \le \sqrt{\alpha}, \qquad r^T \le \exp\left(-\frac{\beta T}{2}\right), \qquad \frac{2\beta}{\delta}r^T \le \frac{256}{127}\frac{\beta}{\alpha} = \frac{32}{127}\beta^{1/3} \le \frac{1}{2}\beta^{1/3},$$

and bound the initial gap by Lemma K.1. Substituting $\alpha = 8\beta^{2/3}$ gives (80) for every $T \ge 1$. When $2\log(T+1)/T \le 1/512$, we have $\beta = 2\log(T+1)/T$ and $\exp(-\beta T/2) = 1/(T+1)$, which gives the stated asymptotic rate. **Q.E.D.**

The horizon-dependent schedule does not use D or σ. The rate implies that, up to logarithmic factors, it is sufficient that

$$T = \widetilde{O}\left(\max\left\{\left(\frac{LD^2}{\varepsilon}\right)^3, \left(\frac{D\rho_{\mathrm{norm}}\sigma}{\varepsilon}\right)^3, \left(\frac{\rho_{\mathrm{norm}}^2\sigma^2}{L\varepsilon}\right)^{3/2}\right\}\right)$$

to obtain expected error at most ε. This bounds the required number of rounds. It is not a tuning rule.

L Additional Trajectory-Based Radius Rules

This appendix records trajectory-based alternatives to the predictable methods in the main text. They are useful comparisons, but their rate statements require trajectory or running-radius boundedness that is not assumed by the main results. Distance-Adaptive Muon uses the standard Muon center $z_k = x_k$, but chooses the trust-region radius from the running maximum distance of the iterates from initialization.

L.1 Deterministic trajectory-based radius rule

Starting from an initial scale proxy $r > 0$, the method maintains

$$\bar{r}_{k+1} = \max\{\bar{r}_k, \|x_k - x_0\|\}, \qquad \eta_k = \frac{\bar{r}_{k+1}}{\sqrt{k+1}}.$$

Thus the step scale increases only when the iterates move farther from the initialization, while still retaining the usual $1/\sqrt{k}$-type decay.

Algorithm 6 Distance-Adaptive Muon

Require: Initial point $x_0 \in \mathcal{E}$, $m_0 = \nabla f(x_0)$, $\bar{r}_0 = r > 0$, and $\alpha \in (0, 1)$

1: **for** $k = 0, 1, \ldots, T-1$ **do**
2: $\quad \bar{r}_{k+1} \leftarrow \max\{\bar{r}_k, \|x_k - x_0\|\}$
3: $\quad \eta_k \leftarrow \bar{r}_{k+1}/\sqrt{k+1}$
4: $\quad m_{k+1} \leftarrow (1-\alpha)m_k + \alpha \nabla f(x_k)$
5: $\quad$ Choose
$$u_{k+1} \in \arg\max_{\|u\| \le 1} \langle m_{k+1}, u \rangle$$
6: $\quad x_{k+1} \leftarrow x_k - \eta_k u_{k+1}$
7: **end for**

The algorithm does not require a diameter or the distance to a minimizer. The following assumption is used only to control the trajectory radius in the analysis.

Assumption L.1 (Bounded DA-Muon trajectory)**.** For each deterministic or stochastic non-convex DA variant considered in this section and each time horizon T, there exists a deterministic $D_{\mathrm{tr}} < \infty$ such that its iterates satisfy

$$\|x_k - x_0\| \le D_{\mathrm{tr}}, \qquad k = 0, \ldots, T.$$

For a stochastic variant, the displayed bound holds almost surely over the complete oracle trajectory.

Theorem L.2 (Stationarity of Distance-Adaptive Muon)**.** *Let Assumptions 2.1 and L.1 hold, assume that $f^\star := \inf_{x \in \mathcal{E}} f(x) > -\infty$, and choose the analysis-only trajectory bound D_{tr} so that $D_{\mathrm{tr}} \ge r$. Then, for all $T \ge 1$, the iterates of Algorithm 6 satisfy*

$$\min_{0 \le k \le T-1} \|\nabla f(x_{k+1})\|_* \le \frac{\sqrt{2}\big(f(x_0) - f^\star\big)}{r\sqrt{T}} + \frac{2LC_\alpha(T)D_{\mathrm{tr}}}{\sqrt{T}} \left(\frac{D_{\mathrm{tr}}}{r}\right)^{2/T} \log\left(\frac{eD_{\mathrm{tr}}}{r}\right),$$

where $C_\alpha(T) = O_\alpha(1 + \log T)$. The explicit expression for $C_\alpha(T)$ and the proof are given in Appendix M.

Theorem L.2 gives an $O(\log T/\sqrt{T})$ stationarity guarantee for smooth non-convex objectives. The trajectory bound D_{tr} appears only in the analysis: the update rule uses only the observed radius $\bar{r}_k$. The dependence on the initial radius $r = \bar{r}_0$ is standard in distance-adaptive methods and reflects the cost of starting from a user-specified lower scale.

The next subsection retains the same center and running-radius state, but implements the trajectory-based radius rule with stochastic gradients and a variance-calibrated schedule.

L.2 Variance-calibrated stochastic trajectory-based radius adaptation

We next replace the exact gradient in DA-Muon by one fresh stochastic gradient per round, while keeping the same center and running-radius state. Fix $\vartheta \in [0, 1]$ and define

$$\psi_k := 1 + \vartheta\sqrt{k+1}, \qquad \alpha_k^{\mathrm{DA}} := \frac{1}{1+\psi_k} = \frac{1}{2 + \vartheta\sqrt{k+1}}. \tag{95}$$

Starting from $m_0 = 0$ and $\bar{r}_0 = r > 0$, the stochastic variant uses

$$\bar{r}_{k+1} := \max\{\bar{r}_k, \|x_k - x_0\|\}, \qquad \eta_k := \frac{\bar{r}_{k+1}}{\sqrt{k+1}\sqrt{\psi_k}}, \tag{96}$$

$$G_k := g(x_k; \xi_k), \qquad m_1 := G_0, \tag{97}$$

$$m_{k+1} := (1 - \alpha_k^{\mathrm{DA}})m_k + \alpha_k^{\mathrm{DA}} G_k, \qquad k \ge 1, \tag{98}$$

$$u_{k+1} := u(m_{k+1}), \qquad x_{k+1} := x_k - \eta_k u_{k+1}. \tag{99}$$

The first momentum parameter is one. All subsequent momentum parameters are given by (95). The query point, running radius, and stepsize are fixed before G_k is drawn, while the fresh sample is used once in the momentum and direction. The population gradient is never evaluated.

The parameter ϑ controls a crossover rather than a terminal horizon. At $\vartheta = 0$,

$$\eta_k = \frac{\bar{r}_{k+1}}{\sqrt{k+1}}, \qquad \alpha_k^{\mathrm{DA}} = \frac{1}{2}.$$

For $\vartheta > 0$ and k larger than order ϑ^{-2},

$$\eta_k \asymp \frac{\bar{r}_{k+1}}{\sqrt{\vartheta}\,(k+1)^{3/4}}, \qquad \alpha_k^{\mathrm{DA}} \asymp \frac{1}{\vartheta\sqrt{k+1}}.$$

The two asymptotic regimes come from the schedule alone. Every update still uses G_k unless the oracle variance is zero.

Proposition L.3 (Exact zero-variance iterate reduction)**.** *Suppose Assumption 2.2 holds with $\sigma = 0$, set $\vartheta = 0$, and use the common measurable LMO selection from Section 2. Then the stochastic variant* (95)–(99) *produces the same trajectory almost surely as Algorithm 6 with momentum parameter $1/2$. The momenta coincide from m_1 onward, as do all running radii, directions, and iterates.*

Write

$$\Delta_0 := f(x_0) - f^\star, \qquad s_\sigma := \rho_{\mathrm{norm}}\sigma,$$

and define

$$H_T := \sum_{j=1}^{T} \frac{1}{j}, \qquad B_{\vartheta,T} := 1 + \vartheta\sqrt{T}, \qquad S_{\vartheta,T} := \sum_{j=1}^{T} \frac{1}{\sqrt{j}\,(1+\vartheta\sqrt{j})}, \tag{100}$$
$$\Lambda_T(q) := q^{2/T}\log(eq), \qquad q \geq 1.$$

Theorem L.4 (Stochastic variant of the DA radius rule)**.** *Let Assumptions 2.1, 2.2, and L.1 hold, assume $f^\star > -\infty$, and enlarge the analysis-only bound so that $D_{\mathrm{tr}} \geq r$. Then, for every $T \geq 2$ and every $\vartheta \in [0,1]$, the iterates* (95)–(99) *satisfy*

$$\mathbb{E}\left[\min_{0\leq k<T} \|\nabla f(x_{k+1})\|_*\right] \leq \frac{\sqrt{B_{\vartheta,T}}}{\sqrt{T}}\left[\frac{\sqrt{2}\Delta_0}{r} + 27LD_{\mathrm{tr}}H_T\Lambda_T(D_{\mathrm{tr}}/r) + \frac{2\sqrt{6}D_{\mathrm{tr}}s_\sigma}{r}S_{\vartheta,T}\right]. \tag{101}$$

The expectation is unconditional over the complete oracle trajectory. The updates use neither D_{tr}, Δ_0, nor the terminal horizon. The proof is given in Appendix M.2.

Theorem L.4 holds for any deterministic ϑ. To interpolate between the deterministic and stochastic regimes, choose the variance-calibrated value

$$\vartheta_\sigma := \frac{s_\sigma}{Lr + s_\sigma} = \frac{\rho_{\mathrm{norm}}\sigma}{Lr + \rho_{\mathrm{norm}}\sigma}. \tag{102}$$

Theorem L.4 allows any fixed ϑ. The choice ϑ_σ is the variance-calibrated specialization.

Corollary L.5 (Deterministic–stochastic interpolation)**.** *Run the stochastic DA variant with $\vartheta = \vartheta_\sigma$. Let*

$$\ell_T := 1 + 2\log(1+\sqrt{T}),$$
$$C_T^{\mathrm{DA}} := \frac{\sqrt{2}\Delta_0}{r} + 27LD_{\mathrm{tr}}H_T\Lambda_T(D_{\mathrm{tr}}/r),$$
$$A_T^{\mathrm{DA}}(\sigma) := C_T^{\mathrm{DA}} + \frac{2\sqrt{6}D_{\mathrm{tr}}}{r}\min\left\{2s_\sigma\sqrt{T}, (Lr+s_\sigma)\ell_T\right\}.$$

Then the finite-time bound

$$\boxed{\mathbb{E}\left[\min_{0\leq k<T} \|\nabla f(x_{k+1})\|_*\right] \leq A_T^{\mathrm{DA}}(\sigma)\left(T^{-1/2} + \sqrt{\vartheta_\sigma}\,T^{-1/4}\right).} \tag{103}$$

holds for every $T \geq 2$. At $\sigma = 0$, it reduces exactly to $C_T^{\mathrm{DA}} T^{-1/2}$, and the iterates reduce to deterministic DA-Muon by Proposition L.3. If one deterministic D_{tr} controls the trajectories over the horizons under consideration, then, for fixed problem parameters,

$$\sigma = 0: \quad \widetilde{O}(T^{-1/2}), \qquad \sigma > 0: \quad \widetilde{O}(T^{-1/4}).$$

For each fixed horizon and a common valid trajectory bound, the right-hand side of (103) *converges continuously to the displayed zero-variance bound as $\sigma \downarrow 0$. Such a common finite-horizon bound always exists uniformly over ϑ. Appendix M.2 records a deterministic, although asymptotically loose, construction.*

Scope. The calibration requires L and a valid upper bound on the noise level. It does not estimate an unknown or time-varying variance. An overestimate is safe, but shifts the crossover to an earlier iteration. The finite-time theorem may use a horizon-specific $D_{\mathrm{tr},T}$, whereas the asymptotic statement needs a common or otherwise controlled trajectory bound. The guarantee is in expectation for the best iterate and assumes an exact LMO over the unit norm ball. Its main point is the horizon-free interpolation and exact recovery of deterministic DA-Muon at zero variance.

The non-convex trajectory results give stationarity rather than an objective gap. We next turn to star-convex objectives, where the global geometry supports last-iterate objective-gap guarantees.

L.3 Stochastic star-convex trajectory-based radius adaptation

We now apply the current-centered DA radius rule to a star-convex objective. The state $\bar{r}_k$, EMA momentum, and norm-ball LMO are unchanged. Star-convexity allows constant momentum and stepsize parameters, and the analysis targets the last iterate. The update remains current-centered and uses no DF-style one-dimensional search.

Let Assumptions 2.1, 2.3, and 2.2 hold, use the fixed measurable LMO selection $s(m)$ from Section 2, and write

$$D_\star := \|x_0 - x^\star\|, \qquad s_\sigma := \rho_{\mathrm{norm}} \sigma, \qquad \chi := 1 + \frac{D_\star}{r}. \tag{104}$$

Starting from $\bar{r}_0 = r > 0$, define

$$\bar{r}_{k+1} := \max\{\bar{r}_k, \|x_k - x_0\|\}, \qquad \eta_k := \beta \bar{r}_{k+1}. \tag{105}$$

Algorithm 7 Stochastic star-convex variant of DA-Muon

Require: $x_0 \in \mathcal{E}$, initial scale $r > 0$, horizon $T \geq 1$, $\alpha \in (0, 1]$, and $\beta \in (0, 1)$
Set $\bar{r}_0 \leftarrow r$, draw $G_0 \leftarrow g(x_0; \xi_0)$, and $m_1 \leftarrow G_0$
for $k = 0, 1, \ldots, T-1$ **do**
  $\bar{r}_{k+1} \leftarrow \max\{\bar{r}_k, \|x_k - x_0\|\}$ and $\eta_k \leftarrow \beta \bar{r}_{k+1}$
  if $k \geq 1$ **then**
    Draw $G_k \leftarrow g(x_k; \xi_k)$
    $m_{k+1} \leftarrow (1 - \alpha) m_k + \alpha G_k$
  end if
  $s_{k+1} \leftarrow s(m_{k+1})$
  $x_{k+1} \leftarrow x_k + \eta_k s_{k+1}$
end for
return x_T

The method uses exactly T fresh stochastic gradients. Its center, momentum direction, and running-radius state are those of DA-Muon. Only the radius schedules and returned iterate differ from the non-convex variant in Section L.2.

For a fixed horizon, define the a posteriori root-mean-square running radius

$$\mathcal{R}_T := \max_{0 \leq k < T} \left(\mathbb{E}\bar{r}_{k+1}^2\right)^{1/2}. \tag{106}$$

This quantity is always finite under the three shared assumptions. It is not an additional hypothesis. Indeed, pathwise, $\bar{r}_{k+2} \le (1+\beta)\bar{r}_{k+1}$, and hence

$$r \le \mathcal{R}_T \le r(1+\beta)^T. \tag{107}$$

The automatic upper bound is deliberately not substituted into the theorem: the actual root-mean-square radius can be much smaller.

L.3.1 Trajectory-dependent and explicit bounds

Theorem L.6 (Stochastic star-convex DA-Muon)**.** *Let Assumptions 2.1, 2.3, and 2.2 hold, and let x_0 be deterministic. Run Algorithm 7 with*

$$0 < \alpha \le 1, \qquad 0 < \beta \le \frac{\alpha}{4}. \tag{108}$$

Set

$$\theta := \frac{\beta}{\chi}, \qquad p := 1-\theta, \qquad q := 1-\alpha, \qquad \delta := \alpha - \theta > 0, \tag{109}$$

and $\Delta_k := f(x_k) - f(x^\star)$. The following pathwise last-iterate bound holds:

$$\Delta_T \le \prod_{j=0}^{T-1}(1-c_j)\Delta_0 + \sum_{k=0}^{T-1}\left[2\beta\bar{r}_{k+1}\|e_{k+1}\|_* + \frac{L\beta^2}{2}\bar{r}_{k+1}^2\right]\prod_{j=k+1}^{T-1}(1-c_j), \tag{110}$$

$$c_k := \frac{\beta\bar{r}_{k+1}}{D_\star + \bar{r}_{k+1}}, \qquad e_{k+1} := m_{k+1} - \nabla f(x_k).$$

Moreover, the explicit expected last-iterate bound is

$$\begin{aligned}\mathbb{E}\Delta_T &\le p^T\Delta_0 + \left(\frac{2(1-\alpha)}{\alpha} + \frac{1}{2}\right)L\beta\mathcal{R}_T^2\chi \\ &\quad + 2\mathcal{R}_T\chi s_\sigma\sqrt{\frac{\alpha}{2-\alpha}} + 2\beta\mathcal{R}_T s_\sigma\frac{p^T - q^T}{\delta} \qquad (111)\\ &\le p^T\Delta_0 + \frac{2L\beta\mathcal{R}_T^2\chi}{\alpha} + 2\mathcal{R}_T\chi s_\sigma\sqrt{\frac{\alpha}{2-\alpha}} + \frac{2\beta\mathcal{R}_T s_\sigma}{\delta}p^T. \qquad (112)\end{aligned}$$

The output is the last iterate. The algorithm uses none of $D_\star, \chi, \mathcal{R}_T, s_\sigma$, or a target accuracy. No bounded-gradient, bounded-support, trajectory-bound, higher-moment, or sub-Gaussian assumption is imposed. The proof is in Appendix N.

The pathwise form retains the actual DA contraction: a larger running radius immediately increases c_k. The closed form uses only $c_k \ge \beta r/(D_\star + r)$, which introduces the factor $\chi = 1 + D_\star/r$ and a transient dependence on the initial scale. This dependence is the price of preserving the current-centered trajectory-based method. Recentering the update would change the method into the DF family.

L.3.2 Horizon-dependent rate under a uniform second-moment bound

As for the stochastic last-iterate Scale-Calibrated theorem, a rate for a horizon-indexed family requires a stability condition uniform in the horizon.

Assumption L.7 (Uniform second-moment bound on the running radius for horizon-tuned DA-Muon)**.** For each $T \ge 1$, run Algorithm 7 with the parameters in (114). There is an analysis-only constant $\overline{R}_{\mathrm{DA}} < \infty$, independent of T, such that

$$\sup_{T\ge 1}\mathcal{R}_T \le \overline{R}_{\mathrm{DA}}. \tag{113}$$

The constant is not passed to the algorithm.

Corollary L.8 (Horizon-dependent stochastic DA schedule)**.** *For a prescribed horizon $T \ge 1$, choose*

$$\beta_T := \min\left\{\frac{1}{32}, \frac{\log^2(T+1)}{T}\right\}, \qquad \alpha_T := 8\beta_T^{2/3}. \tag{114}$$

Then the gain conditions in Theorem L.6 hold and, without any stability assumption,

$$\mathbb{E}\Delta_T \le \Delta_0 \exp\left(-\frac{\beta_T T}{\chi}\right) + \left[\frac{L\mathcal{R}_T^2\chi}{4} + \left(4\sqrt{2}+\frac{1}{3}\right)\mathcal{R}_T\chi s_\sigma\right]\beta_T^{1/3}. \tag{115}$$

Under Assumption L.7, this becomes

$$\mathbb{E}\Delta_T \le \Delta_0 \exp\left(-\frac{1}{\chi}\min\left\{\frac{T}{32}, \log^2(T+1)\right\}\right) + \widetilde{O}\left(\frac{L\overline{R}_{\mathrm{DA}}^2\chi + \overline{R}_{\mathrm{DA}}\chi s_\sigma}{T^{1/3}}\right). \tag{116}$$

For all sufficiently large T, the minimum in the exponential equals $\log^2(T+1)$. Thus every fixed problem instance has a $\widetilde{O}(T^{-1/3})$ expected last-iterate rate after an explicit scale-dependent burn-in. The parameters use only T.

L.3.3 Oracle-complexity comparison under a supplied radius bound

The next result compares Theorem L.6 with prescribed-radius analyses when an external stability certificate is available. It is a benchmark under a supplied radius bound rather than a tuning rule for the horizon-dependent method.

Corollary L.9 (Stability-certified accuracy schedule)**.** *Assume $s_\sigma > 0$. Suppose that $D_\star$ and a deterministic number $\overline{R} \ge r$ are supplied before the run, and that $\overline{R}$ certifies that the run with the parameters below satisfies $\mathcal{R}_T \le \overline{R}$. Let*

$$P := \overline{R}^2\chi, \qquad Q := \overline{R}\chi, \tag{117}$$

and choose

$$\alpha_\varepsilon := \min\left\{1, \frac{\varepsilon^2}{64Q^2s_\sigma^2}\right\}, \qquad \beta_\varepsilon := \min\left\{\frac{\alpha_\varepsilon}{4}, \frac{1}{32}, \frac{\alpha_\varepsilon\varepsilon}{20LP}\right\}, \tag{118}$$

and

$$\Lambda_\varepsilon := \max\left\{1, \frac{4\Delta_0}{\varepsilon}, \frac{8\overline{R}s_\sigma}{3\varepsilon}\right\}. \tag{119}$$

If

$$T \ge \frac{\chi}{\beta_\varepsilon}\log\Lambda_\varepsilon, \tag{120}$$

then $\mathbb{E}\Delta_T \le \varepsilon$. Consequently,

$$T = \widetilde{O}\left(\max\left\{\chi, \frac{L\overline{R}^2\chi^2}{\varepsilon}, \frac{\overline{R}^2\chi^3s_\sigma^2}{\varepsilon^2}, \frac{L\overline{R}^4\chi^4s_\sigma^2}{\varepsilon^3}\right\}\right). \tag{121}$$

The premise is non-circular: $\overline{R}$ is an external certificate fixed before the parameters are selected. Without it, one should use the horizon-dependent schedule and its a posteriori bound (115)*.*

When r, $\overline{R}$, and $D_\star$ are comparable, (121) reduces, up to constants and logarithms, to

$$1 + \frac{LD_\star^2}{\varepsilon} + \frac{D_\star^2s_\sigma^2}{\varepsilon^2} + \frac{LD_\star^4s_\sigma^2}{\varepsilon^3}. \tag{122}$$

In particular, the dominant mixed term scales as ε^{-3}, as in ordinary-momentum stochastic star-convex optimization.

If $s_\sigma = 0$, set $\alpha = 1$ in Theorem L.6. Then

$$f(x_T) - f(x^\star) \le \left(1 - \frac{\beta}{\chi}\right)^T\Delta_0 + \frac{L\beta\mathcal{R}_T^2\chi}{2}. \tag{123}$$

Under a certified bound $\mathcal{R}_T \le \overline{R}$, this gives $\widetilde{O}(L\overline{R}^2\chi^2/\varepsilon)$ complexity, recovering $\widetilde{O}(LD_\star^2/\varepsilon)$ at balanced scales.

L.3.4 Position among stochastic star-convex Muon results

All comparisons below use $s_\sigma = \rho_{\mathrm{norm}}\sigma$ and the expected suboptimality of the last iterate.

Method	Horizon rate	Localization/stability
DA-Muon (here)	$\widetilde{O}((L\overline{R}_{\mathrm{DA}}^2\chi + \overline{R}_{\mathrm{DA}}\chi s_\sigma)T^{-1/3})$	Uniform second-moment bound on the running-radius state. No distance or stability input (the initial proxy r remains)
Scale-Calibrated Muon	$O((L\overline{D}_{\mathrm{ms}}^2 + \overline{D}_{\mathrm{ms}}s_\sigma)T^{-1/3} + (s_\sigma^2/L)T^{-2/3})$	uniform mean-square boundedness of the trajectory. No stability-radius input
Predictable DF-Muon	$\widetilde{O}((LD_\star^2 + D_\star s_\sigma)T^{-1/3} + (s_\sigma^2/L)T^{-2/3})$	no trajectory localization and no distance input

At balanced scales, DA-Muon has the same leading $\widetilde{O}(T^{-1/3})$ order and the natural $LD_\star^2 + D_\star s_\sigma$ scale. It uses the simplest of the three methods and one gradient per update, but pays the factor χ and needs the stronger running-maximum stability condition.

The two stability assumptions can be compared directly. If

$$D_{\mathrm{ms},T} := \max_{0\le k\le T} \left(\mathbb{E}\|x_k - x^\star\|^2\right)^{1/2},$$

then Lemma N.1 and Minkowski's inequality give

$$D_{\mathrm{ms},T} \le D_\star + (1+\beta)\mathcal{R}_T. \tag{124}$$

Hence a uniform second-moment bound on the DA running maximum implies uniform mean-square boundedness of the iterates. The converse does not follow from second moments alone: the elementary bound is only

$$\mathcal{R}_T^2 \le r^2 + 4TD_{\mathrm{ms},T}^2. \tag{125}$$

The DA rate therefore rests on a stronger localization premise than the Scale-Calibrated rate. For $k < T$, Minkowski and $\|x_k - x^\star\| \le D_\star + \bar{r}_{k+1}$ give the first bound, while the final iterate follows from $\|x_T - x_0\| \le (1+\beta)\bar{r}_T$. For the second, use $\bar{r}_{k+1}^2 \le r^2 + \sum_{i=0}^k \|x_i - x_0\|^2$ and $\|x_i - x_0\|^2 \le 2\|x_i - x^\star\|^2 + 2D_\star^2$, then take expectations and use $D_\star \le D_{\mathrm{ms},T}$.

For the fixed-radius/weight-decay result of Kovalev (2025), let $D_{\mathrm{K}} := \max\{\|x_0\|, \|x^\star\|\}$. Direct tuning of the bound in Theorem 4 gives, up to logarithms,

$$1 + \frac{LD_{\mathrm{K}}^2}{\varepsilon} + \frac{D_{\mathrm{K}}^3 s_\sigma^3}{\varepsilon^3} + \frac{LD_{\mathrm{K}}^4 s_\sigma^2}{\varepsilon^3}. \tag{126}$$

There is a power-of-D_{K} discrepancy inside that reference: Corollary 4 displays $LD_{\mathrm{K}}^3 s_\sigma^2/\varepsilon^3$, whereas direct substitution into Theorem 4 gives $LD_{\mathrm{K}}^4 s_\sigma^2/\varepsilon^3$, the same power displayed in its bounded-domain Corollary 8. We use the expression obtained directly from Theorem 4. When D_{K} and the balanced DA scales are comparable, (122) has the same mixed term and the same ε^{-3} dependence. Its proved upper bound has a separate pure-noise term $D_\star^2 s_\sigma^2/\varepsilon^2$, whereas direct tuning of the bound in Theorem 4 gives $D_{\mathrm{K}}^3 s_\sigma^3/\varepsilon^3$. These are upper-bound comparisons. They do not establish a separation. Our theorem is unregularized and its accuracy benchmark assumes an external stability certificate, whereas the weight-decay result also covers a composite term and obtains stability from its regularizer.

For one $m \times n$ matrix block in spectral–nuclear geometry, $\rho_{\mathrm{norm}} = \sqrt{\min\{m,n\}}$. Thus every occurrence of s_σ becomes $\sqrt{\min\{m,n\}}\,\sigma$, while distances and running radii are measured in spectral norm. Unlike the local Scale-Calibrated method, DA-Muon does not evaluate $\|m_k\|_*$: the dual norm is used only in the analysis. The exact theorem assumes a valid global smoothness constant and an exact norm-ball LMO. Approximate Newton–Schulz orthogonalization is a practical variant and is not covered by our analysis.

M Proofs for Distance-Adaptive Muon

This appendix proves the deterministic and stochastic stationarity guarantees for the distance-adaptive trajectory-based radius rule. We first treat Theorem L.2, then prove the stochastic extension from Section L.2.

M.1 Deterministic trajectory-based radius rule

M.1.1 Auxiliary summation estimates

Lemma M.1 (Discrete convolution bound)**.** *Let $\alpha \in (0,1)$, set $q := 1-\alpha$, and let $k \geq 1$. Then*

$$\sum_{i=0}^{k-1} \frac{q^{k-i}}{\sqrt{i+1}} \leq \frac{q^{k/2}}{\alpha} + \frac{\sqrt{2}\,q}{\alpha\sqrt{k+2}}.$$

Proof. Changing variables $j = k-i$, we obtain

$$\sum_{i=0}^{k-1} \frac{q^{k-i}}{\sqrt{i+1}} = \sum_{j=1}^{k} \frac{q^j}{\sqrt{k-j+1}}.$$

Split the sum into $j \leq k/2$ and $j > k/2$. If $j \leq k/2$, then $k-j+1 \geq (k+2)/2$, and therefore

$$\sum_{1\leq j\leq k/2} \frac{q^j}{\sqrt{k-j+1}} \leq \frac{\sqrt{2}}{\sqrt{k+2}} \sum_{j=1}^{\infty} q^j = \frac{\sqrt{2}\,q}{\alpha\sqrt{k+2}}.$$

If $j > k/2$, then $1/\sqrt{k-j+1} \leq 1$, and hence

$$\sum_{k/2<j\leq k} \frac{q^j}{\sqrt{k-j+1}} \leq \sum_{j>k/2}^{\infty} q^j \leq \frac{q^{k/2}}{\alpha}.$$

Combining the two bounds proves the claim. **Q.E.D.**

Lemma M.2. *Let $a \in (0,1)$ and $t \geq 0$ be an integer. Then*

$$\sum_{k=0}^{t} \frac{a^k}{\sqrt{k+1}} \leq 1 + \frac{1}{a}\sqrt{\frac{\pi}{\log\left(\frac{1}{a}\right)}}.$$

Proof. The function $x \mapsto a^x/\sqrt{x+1}$ is nonincreasing on $[0,\infty)$. Hence

$$\sum_{k=0}^{t} \frac{a^k}{\sqrt{k+1}} \leq 1 + \int_0^t \frac{a^k}{\sqrt{k+1}}\,dk = 1 + \int_0^t \frac{e^{k\log a}}{\sqrt{k+1}}\,dk.$$

Since $a < 1$, we have $\log a < 0$. Define

$$\lambda := -\log a > 0.$$

Then

$$\sum_{k=0}^{t} \frac{a^k}{\sqrt{k+1}} \leq 1 + \int_0^{\infty} \frac{e^{-\lambda k}}{\sqrt{k+1}}\,dk.$$

Applying the change of variables $u = k+1$, we obtain

$$\begin{aligned}\sum_{k=0}^{t} \frac{a^k}{\sqrt{k+1}} &\leq 1 + e^{\lambda} \int_1^{\infty} \frac{e^{-\lambda u}}{\sqrt{u}}\,du \\ &\leq 1 + e^{\lambda} \int_0^{\infty} \frac{e^{-\lambda u}}{\sqrt{u}}\,du = 1 + e^{\lambda}\sqrt{\frac{\pi}{\lambda}},\end{aligned}$$

where the last integral is the standard Gamma-function integral. Substituting $\lambda = \log(1/a)$ and $e^{\lambda} = 1/a$ gives the desired bound. **Q.E.D.**

Lemma M.3. (Liu & Zhou, 2023, Lemma 30), (Moshtaghifar et al., 2025, Lemma A.1), (Chakraborty & Nedić, 2026, Lemma B.3) *Let $\{a_i\}_{i=0}^{T+1}$ be a positive, nondecreasing sequence. Then, for any integer $T \geq 1$,*

$$\min_{1\leq k\leq T} \frac{a_{k+1}}{\sum_{i=1}^{k} a_i} \leq \frac{1}{T}\left(\frac{a_{T+1}}{a_0}\right)^{1/T} \log\left(\frac{e\, a_{T+1}}{a_0}\right).$$

Lemma M.4 (Tail radius-ratio bound)**.** *Let $\{\bar{r}_i\}_{i=0}^{T+1}$ be a positive, nondecreasing sequence. Then, for any integer $T \geq 2$,*

$$\min_{\lfloor T/2\rfloor+1\leq k\leq T} \frac{\bar{r}_{k+1}}{\sum_{i=1}^{k} \bar{r}_i} \leq \frac{2}{T}\left(\frac{\bar{r}_{T+1}}{\bar{r}_0}\right)^{2/T} \log\left(\frac{e\, \bar{r}_{T+1}}{\bar{r}_0}\right).$$

Proof. Set $s = \lfloor T/2\rfloor$ and $M = T - s$. Define the shifted sequence

$$a_j := \bar{r}_{s+j}, \qquad j = 0, \ldots, M+1.$$

Applying Lemma M.3 to $\{a_j\}_{j=0}^{M+1}$ with horizon M gives

$$\min_{1\leq j\leq M} \frac{a_{j+1}}{\sum_{\ell=1}^{j} a_\ell} \leq \frac{1}{M}\left(\frac{a_{M+1}}{a_0}\right)^{1/M} \log\left(\frac{e\, a_{M+1}}{a_0}\right).$$

For $k = s + j$, we have

$$\sum_{i=1}^{k} \bar{r}_i \geq \sum_{i=s+1}^{k} \bar{r}_i = \sum_{\ell=1}^{j} a_\ell, \qquad \bar{r}_{k+1} = a_{j+1}.$$

Therefore,

$$\min_{s+1\leq k\leq T} \frac{\bar{r}_{k+1}}{\sum_{i=1}^{k} \bar{r}_i} \leq \frac{1}{M}\left(\frac{\bar{r}_{T+1}}{\bar{r}_s}\right)^{1/M} \log\left(\frac{e\, \bar{r}_{T+1}}{\bar{r}_s}\right).$$

Since $M \geq T/2$, $\bar{r}_s \geq \bar{r}_0$, and $\bar{r}_{T+1} \geq \bar{r}_0$, the right-hand side is at most

$$\frac{2}{T}\left(\frac{\bar{r}_{T+1}}{\bar{r}_0}\right)^{2/T} \log\left(\frac{e\, \bar{r}_{T+1}}{\bar{r}_0}\right).$$

This proves the claim. **Q.E.D.**

M.1.2 Trust-region optimality identity

Lemma M.5 (Trust-region optimality identity)**.** (Kovalev, 2025, Modification of Lemma 3). *Let $\eta \geq 0$, let $\Psi : \mathcal{E} \to \mathbb{R}$ be differentiable, and let $z_+ \in \mathcal{E}$ be a solution of*

$$\min_{x\in\mathcal{E}} \Psi(x) \qquad \textit{s.t.} \qquad \|x - z\| \leq \eta.$$

Then

$$\langle \nabla\Psi(z_+), z - z_+\rangle = \eta\|\nabla\Psi(z_+)\|_*.$$

Proof. Let

$$B_\eta(z) := \{x \in \mathcal{E} : \|x - z\| \leq \eta\}.$$

If $\eta = 0$, then $z_+ = z$, and the claim is immediate. Hence, assume $\eta > 0$. Since z_+ minimizes Ψ over $B_\eta(z)$, for every $x \in B_\eta(z)$ and every $\tau \in [0, 1]$, the point $z_+ + \tau(x - z_+)$ remains feasible. Therefore, the one-sided directional derivative at $\tau = 0$ is nonnegative:

$$\langle \nabla\Psi(z_+), x - z_+\rangle \geq 0, \qquad \forall x \in B_\eta(z).$$

If $\nabla\Psi(z_+) = 0$, the claimed identity is immediate. Otherwise, z_+ cannot lie in the interior of $B_\eta(z)$, and hence

$$\|z_+ - z\| = \eta.$$

Set $g := \nabla\Psi(z_+)$ and $w := z_+ - z$. The variational inequality above implies that, for all d with $\|d\| \le \eta$,

$$\langle g, d - w\rangle \ge 0, \qquad \text{or equivalently} \qquad \langle g, w\rangle \le \langle g, d\rangle.$$

Taking the infimum over $\|d\| \le \eta$, we obtain

$$\langle g, w\rangle \le \inf_{\|d\|\le\eta} \langle g, d\rangle = -\eta\|g\|_*.$$

On the other hand, by Hölder's inequality and $\|w\| = \eta$,

$$\langle g, w\rangle \ge -\|g\|_*\|w\| = -\eta\|g\|_*.$$

Thus

$$\langle g, w\rangle = -\eta\|g\|_*.$$

Since $w = z_+ - z$, this is equivalent to

$$\langle \nabla\Psi(z_+), z - z_+\rangle = \eta\|\nabla\Psi(z_+)\|_*,$$

as claimed. **Q.E.D.**

M.1.3 Momentum tracking

Lemma M.6 (Momentum tracking)**.** *Let Assumption 2.1 hold, and consider the recursion*

$$m_{k+1} = (1-\alpha)m_k + \alpha\nabla f(x_k), \qquad \alpha \in (0,1).$$

Assume that the iterates satisfy

$$\|x_{i+1} - x_i\| \le \eta_i.$$

Then, for all $k \ge 0$, with the summation interpreted as zero when $k = 0$,

$$\|m_{k+1} - \nabla f(x_k)\|_* \le (1-\alpha)^{k+1}\|m_0 - \nabla f(x_0)\|_* + L\sum_{i=0}^{k-1}(1-\alpha)^{k-i}\eta_i.$$

Moreover, if $m_0 = \nabla f(x_0)$ and

$$\eta_i = \frac{\bar r_{i+1}}{\sqrt{i+1}},$$

then, for all $k \ge 0$,

$$\|m_{k+1} - \nabla f(x_k)\|_* \le L\bar r_k\left[\frac{(1-\alpha)^{k/2}}{\alpha} + \frac{\sqrt{2}(1-\alpha)}{\alpha\sqrt{k+2}}\right].$$

Proof. For $k = 0$, the first claim follows directly from

$$m_1 - \nabla f(x_0) = (1-\alpha)(m_0 - \nabla f(x_0)).$$

If, in addition, $m_0 = \nabla f(x_0)$, then the left-hand side of the second claim is zero for $k = 0$, so the second claim also holds. We now consider $k \ge 1$.

We start from the recursion

$$m_{k+1} - \nabla f(x_k) = (1-\alpha)m_k + \alpha\nabla f(x_k) - \nabla f(x_k),$$

which can be rewritten as

$$m_{k+1} - \nabla f(x_k) = (1-\alpha)\big(m_k - \nabla f(x_{k-1})\big) + (1-\alpha)\big(\nabla f(x_{k-1}) - \nabla f(x_k)\big).$$

Unrolling the recursion yields

$$m_{k+1} - \nabla f(x_k) = (1-\alpha)^{k+1}(m_0 - \nabla f(x_0)) + \sum_{i=0}^{k-1}(1-\alpha)^{k-i}\big(\nabla f(x_i) - \nabla f(x_{i+1})\big).$$

Taking dual norms and applying the triangle inequality yields

$$\|m_{k+1} - \nabla f(x_k)\|_* \le (1-\alpha)^{k+1}\|m_0 - \nabla f(x_0)\|_* + \sum_{i=0}^{k-1}(1-\alpha)^{k-i}\|\nabla f(x_i) - \nabla f(x_{i+1})\|_*.$$

Using L-smoothness of f, we obtain

$$\|m_{k+1} - \nabla f(x_k)\|_* \leq (1-\alpha)^{k+1}\|m_0 - \nabla f(x_0)\|_* + L\sum_{i=0}^{k-1}(1-\alpha)^{k-i}\|x_i - x_{i+1}\|.$$

From the trust-region bound of Algorithm 6, we have

$$\|x_{i+1} - x_i\| \leq \eta_i,$$

which proves the first claim.

For the case $m_0 = \nabla f(x_0)$ and

$$\eta_i = \frac{\bar{r}_{i+1}}{\sqrt{i+1}},$$

the first claim simplifies to

$$\|m_{k+1} - \nabla f(x_k)\|_* \leq L\sum_{i=0}^{k-1}(1-\alpha)^{k-i}\frac{\bar{r}_{i+1}}{\sqrt{i+1}} \leq L\bar{r}_k\sum_{i=0}^{k-1}\frac{(1-\alpha)^{k-i}}{\sqrt{i+1}}.$$

Applying Lemma M.1 with $q = 1-\alpha$, we obtain

$$\|m_{k+1} - \nabla f(x_k)\|_* \leq L\bar{r}_k\left[\frac{(1-\alpha)^{k/2}}{\alpha} + \frac{\sqrt{2}(1-\alpha)}{\alpha\sqrt{k+2}}\right].$$

This proves the second claim. **Q.E.D.**

M.1.4 Proof of Theorem L.2

Proof. Using L-smoothness of f, we obtain

$$f(x_{k+1}) \leq f(x_k) + \langle \nabla f(x_k), x_{k+1} - x_k\rangle + \frac{L}{2}\|x_{k+1} - x_k\|^2. \tag{127}$$

The second term on the right-hand side can be split as

$$\begin{aligned}\langle \nabla f(x_k), x_{k+1} - x_k\rangle = &\langle m_{k+1}, x_{k+1} - x_k\rangle \\ &+ \langle \nabla f(x_{k+1}) - m_{k+1}, x_{k+1} - x_k\rangle \\ &+ \langle \nabla f(x_k) - \nabla f(x_{k+1}), x_{k+1} - x_k\rangle.\end{aligned}$$

We apply Hölder's inequality to the last two terms on the right-hand side and then apply Assumption 2.1 to obtain

$$\begin{aligned}\langle \nabla f(x_k), x_{k+1} - x_k\rangle \leq &\langle m_{k+1}, x_{k+1} - x_k\rangle + \|\nabla f(x_{k+1}) - m_{k+1}\|_*\|x_{k+1} - x_k\| \\ &+ L\|x_{k+1} - x_k\|^2.\end{aligned}$$

Substituting the preceding relation in (127) and using $\|x_{k+1} - x_k\| \leq \eta_k$, we get

$$f(x_{k+1}) \leq f(x_k) + \langle m_{k+1}, x_{k+1} - x_k\rangle + \eta_k\|\nabla f(x_{k+1}) - m_{k+1}\|_* + \frac{3L}{2}\eta_k^2.$$

By construction, x_{k+1} solves the linear trust-region problem

$$\min_{\|x - x_k\| \leq \eta_k}\langle m_{k+1}, x\rangle.$$

Indeed, Algorithm 6 chooses $u_{k+1} \in \arg\max_{\|u\|\leq 1}\langle m_{k+1}, u\rangle$ and sets $x_{k+1} = x_k - \eta_k u_{k+1}$. Applying Lemma M.5 with $\Psi(x) = \langle m_{k+1}, x\rangle$, $z = x_k$, $z_+ = x_{k+1}$, and $\eta = \eta_k$, we obtain

$$\langle m_{k+1}, x_{k+1} - x_k\rangle = -\eta_k\|m_{k+1}\|_*.$$

Therefore,

$$f(x_{k+1}) \leq f(x_k) - \eta_k\|m_{k+1}\|_* + \eta_k\|\nabla f(x_{k+1}) - m_{k+1}\|_* + \frac{3L}{2}\eta_k^2.$$

Applying the triangle inequality

$$-\|m_{k+1}\|_* \le \|\nabla f(x_{k+1}) - m_{k+1}\|_* - \|\nabla f(x_{k+1})\|_*$$

gives

$$f(x_{k+1}) \le f(x_k) - \eta_k \|\nabla f(x_{k+1})\|_* + 2\eta_k \|\nabla f(x_{k+1}) - m_{k+1}\|_* + \frac{3L}{2}\eta_k^2.$$

We apply the triangle inequality to the third term on the right-hand side:

$$\begin{aligned} f(x_{k+1}) &\le f(x_k) - \eta_k \|\nabla f(x_{k+1})\|_* + 2\eta_k \|\nabla f(x_k) - m_{k+1}\|_* \\ &\quad + 2\eta_k \|\nabla f(x_{k+1}) - \nabla f(x_k)\|_* + \frac{3L}{2}\eta_k^2 \\ &\le f(x_k) - \eta_k \|\nabla f(x_{k+1})\|_* + 2\eta_k \|\nabla f(x_k) - m_{k+1}\|_* \\ &\quad + 2\eta_k L \|x_{k+1} - x_k\| + \frac{3L}{2}\eta_k^2. \end{aligned}$$

Since Algorithm 6 sets $m_0 = \nabla f(x_0)$ and

$$\eta_i = \frac{\bar{r}_{i+1}}{\sqrt{i+1}},$$

the third and fourth quantities on the right-hand side can be bounded using Lemma M.6 and the trust-region bound $\|x_{k+1} - x_k\| \le \eta_k$. This yields

$$\begin{aligned} f(x_{k+1}) \le f(x_k) &- \eta_k \|\nabla f(x_{k+1})\|_* \\ &+ \frac{2\eta_k L \bar{r}_k}{\alpha}(1-\alpha)^{k/2} \\ &+ \frac{2\sqrt{2}(1-\alpha)\eta_k L}{\alpha}\frac{\bar{r}_k}{\sqrt{k+2}} + \frac{7L}{2}\eta_k^2. \end{aligned}$$

Note that

$$\frac{\bar{r}_k}{\sqrt{k+2}} \le \frac{\bar{r}_{k+1}}{\sqrt{k+1}} = \eta_k.$$

Hence the preceding relation simplifies to

$$\begin{aligned} f(x_{k+1}) \le f(x_k) &- \eta_k \|\nabla f(x_{k+1})\|_* \\ &+ \frac{2\eta_k L \bar{r}_k}{\alpha}(1-\alpha)^{k/2} \\ &+ \left[\frac{7}{2} + \frac{2\sqrt{2}(1-\alpha)}{\alpha}\right] L\eta_k^2. \end{aligned}$$

Rearranging terms and summing both sides from $k = 0$ to t, for all $t \ge 0$, gives

$$\begin{aligned} \sum_{k=0}^{t} \eta_k \|\nabla f(x_{k+1})\|_* \le f(x_0) - f(x_{t+1}) &+ \frac{2L}{\alpha}\sum_{k=0}^{t} \bar{r}_k \eta_k (1-\alpha)^{k/2} \\ &+ \left[\frac{7}{2} + \frac{2\sqrt{2}(1-\alpha)}{\alpha}\right] L \sum_{k=0}^{t} \eta_k^2. \end{aligned} \tag{128}$$

The left-hand side of (128) can be lower-bounded as

$$\begin{aligned} \sum_{k=0}^{t} \eta_k \|\nabla f(x_{k+1})\|_* &\ge \min_{0\le k\le t} \|\nabla f(x_{k+1})\|_* \sum_{k=0}^{t} \frac{\bar{r}_{k+1}}{\sqrt{k+1}} \\ &\ge \frac{\min_{0\le k\le t} \|\nabla f(x_{k+1})\|_*}{\sqrt{t+1}} \sum_{k=0}^{t} \bar{r}_{k+1}. \end{aligned} \tag{129}$$

The momentum-tracking term on the right-hand side of (128) can be upper-bounded using Lemma M.2 with $a = \sqrt{1-\alpha}$:

$$\frac{2L}{\alpha}\sum_{k=0}^{t}\bar{r}_k\eta_k(1-\alpha)^{k/2} \le \frac{2L\bar{r}_{t+1}^2}{\alpha}\sum_{k=0}^{t}\frac{(1-\alpha)^{k/2}}{\sqrt{k+1}} \le 2L\bar{r}_{t+1}^2 A(\alpha), \tag{130}$$

where

$$A(\alpha) = \frac{1}{\alpha}\left[1 + \frac{1}{\sqrt{1-\alpha}}\sqrt{\frac{2\pi}{\log\left(\frac{1}{1-\alpha}\right)}}\right].$$

The quadratic term on the right-hand side of (128) can be upper-bounded as

$$\sum_{k=0}^{t}\eta_k^2 \le \bar{r}_{t+1}^2\sum_{k=0}^{t}\frac{1}{k+1} \le \bar{r}_{t+1}^2(1+\log(t+1)). \tag{131}$$

Using $-f(x_{t+1}) \le -f^\star$, and applying (129), (130), and (131) to (128), we get

$$\begin{aligned}\min_{0\le k\le t}\|\nabla f(x_{k+1})\|_* &\le \frac{\sqrt{t+1}(f(x_0)-f^\star)}{\sum_{k=0}^{t}\bar{r}_{k+1}} + LC_\alpha(t+1)\frac{\sqrt{t+1}\bar{r}_{t+1}^2}{\sum_{k=0}^{t}\bar{r}_{k+1}} \\ &\le \frac{f(x_0)-f^\star}{r\sqrt{t+1}} + LC_\alpha(t+1)\frac{\sqrt{t+1}\bar{r}_{t+1}^2}{\sum_{k=0}^{t}\bar{r}_{k+1}},\end{aligned} \tag{132}$$

where we used $r \le \bar{r}_{k+1}$ for all $k \ge 0$, and

$$C_\alpha(t+1) := 2A(\alpha) + \left[\frac{7}{2} + \frac{2\sqrt{2}(1-\alpha)}{\alpha}\right](1+\log(t+1)).$$

For notational convenience, extend the radius sequence by one final update

$$\bar{r}_{T+1} := \max\{\bar{r}_T, \|x_T - x_0\|\}.$$

Since the trajectory bound is used only in the analysis, we may enlarge it and assume without loss of generality that $D_{\mathrm{tr}} \ge r$. Under Assumption L.1, this gives

$$\bar{r}_{T+1} \le D_{\mathrm{tr}}.$$

If $T = 1$, then (132) with $t = 0$ gives

$$\min_{0\le k\le 0}\|\nabla f(x_{k+1})\|_* \le \frac{f(x_0)-f^\star}{r} + LC_\alpha(1)D_{\mathrm{tr}},$$

which is bounded by the claimed estimate because $D_{\mathrm{tr}}/r \ge 1$ and $\log(eD_{\mathrm{tr}}/r) \ge 1$. Hence, it remains to consider $T \ge 2$.

Set $s = \lfloor T/2 \rfloor$. For every $k \in \{s+1, \dots, T\}$, applying (132) with $t = k-1$ gives

$$\min_{0\le j\le k-1}\|\nabla f(x_{j+1})\|_* \le \frac{f(x_0)-f^\star}{r\sqrt{k}} + LC_\alpha(k)\frac{\sqrt{k}\bar{r}_k^2}{\sum_{i=1}^{k}\bar{r}_i}. \tag{133}$$

Since $k \ge s+1 > T/2$, $k \le T$, $C_\alpha(k) \le C_\alpha(T)$, and $\bar{r}_k^2 \le D_{\mathrm{tr}}\bar{r}_{k+1}$, we obtain

$$\min_{0\le j\le k-1}\|\nabla f(x_{j+1})\|_* \le \frac{\sqrt{2}\big(f(x_0)-f^\star\big)}{r\sqrt{T}} + LC_\alpha(T)\sqrt{T}\,D_{\mathrm{tr}}\frac{\bar{r}_{k+1}}{\sum_{i=1}^{k}\bar{r}_i}.$$

Moreover,

$$\min_{0\le j\le T-1}\|\nabla f(x_{j+1})\|_* \le \min_{0\le j\le k-1}\|\nabla f(x_{j+1})\|_*$$

for every $k \le T$. Taking the minimum over $k \in \{s+1, \dots, T\}$, we get

$$\min_{0\le j\le T-1} \|\nabla f(x_{j+1})\|_* \le \frac{\sqrt{2}\big(f(x_0)-f^\star\big)}{r\sqrt{T}} + LC_\alpha(T)\sqrt{T}\, D_{\mathrm{tr}} \min_{s+1\le k\le T} \frac{\bar r_{k+1}}{\sum_{i=1}^k \bar r_i}. \tag{134}$$

Applying Lemma M.4 to the second term on the right-hand side of (134) yields

$$\begin{aligned}\min_{0\le j\le T-1} \|\nabla f(x_{j+1})\|_* &\le \frac{\sqrt{2}\big(f(x_0)-f^\star\big)}{r\sqrt{T}} + \frac{2LC_\alpha(T)D_{\mathrm{tr}}}{\sqrt{T}} \left(\frac{\bar r_{T+1}}{\bar r_0}\right)^{2/T} \log\!\left(\frac{e\,\bar r_{T+1}}{\bar r_0}\right) \\ &\le \frac{\sqrt{2}\big(f(x_0)-f^\star\big)}{r\sqrt{T}} + \frac{2LC_\alpha(T)D_{\mathrm{tr}}}{\sqrt{T}} \left(\frac{D_{\mathrm{tr}}}{r}\right)^{2/T} \log\!\left(\frac{eD_{\mathrm{tr}}}{r}\right),\end{aligned}$$

where we used $\bar r_{T+1} \le D_{\mathrm{tr}}$ and $\bar r_0 = r$. Hence, the claim follows. **Q.E.D.**

M.2 Stochastic trajectory-based radius rule

Throughout this subsection, write

$$h_k := \nabla f(x_k), \quad \zeta_k := G_k - h_k, \quad a_k := \alpha_k^{\mathrm{DA}} = \frac{1}{2+\vartheta\sqrt{k+1}}, \quad q_k := 1 - a_k.$$

The quantity a_0 is used only in the analysis: the actual first momentum gain in (97) is one.

M.2.1 Scalar estimates and exact reduction

Lemma M.7 (Elementary schedule sums)**.** *For every $T \ge 1$, $H_T \le 1 + \log T$ and $S_{0,T} \le 2\sqrt{T}$. For every $\vartheta \in (0,1]$,*

$$S_{\vartheta,T} \le 1 + \frac{2}{\vartheta}\log(1+\vartheta\sqrt{T}). \tag{135}$$

Proof. The harmonic estimate is standard. Since $x \mapsto x^{-1/2}(1+\vartheta\sqrt{x})^{-1}$ is nonincreasing,

$$\begin{aligned}S_{\vartheta,T} &\le 1 + \int_1^T \frac{dx}{\sqrt{x}(1+\vartheta\sqrt{x})} \\ &= 1 + \frac{2}{\vartheta}\log\frac{1+\vartheta\sqrt{T}}{1+\vartheta} \le 1 + \frac{2}{\vartheta}\log(1+\vartheta\sqrt{T}).\end{aligned}$$

At $\vartheta = 0$, the same comparison gives $S_{0,T} \le 1 + \int_1^T x^{-1/2}dx \le 2\sqrt{T}$. **Q.E.D.**

Lemma M.8 (Gain-ratio estimates)**.** *For every $k \ge 1$,*

$$\frac{a_{k-1}}{a_k} \le 1 + a_k. \tag{136}$$

Consequently, $a_{k-1} \le 2a_k$. Also $a_{k+1}^{-1} \le 2a_k^{-1}$ for every $k \ge 0$.

Proof. Put $c_k = a_k^{-1} = 2 + \vartheta\sqrt{k+1}$. At $\vartheta = 0$ the first claim is immediate. Otherwise it is equivalent to $c_k(c_k - c_{k-1}) \le c_{k-1}$. For $k \ge 1$, with $s = \sqrt{k+1}$ and $t = \sqrt{k}$,

$$c_k(c_k - c_{k-1}) = \frac{\vartheta(2+\vartheta s)}{s+t} \le \frac{2+s}{s+t} \le 2 \le c_{k-1}.$$

The first consequence uses $a_k \le 1/2$. Finally, $c_{k+1} \le 2c_k$ follows from $c_k \ge 2$ and monotonicity of the square root. **Q.E.D.**

Proof of Proposition L.3. When $\sigma = 0$, Assumption 2.2 gives $G_k = \nabla f(x_k)$ almost surely at every queried index. At $\vartheta = 0$, all subsequent coefficients equal $1/2$. If m_k^{det} denotes the momentum of Algorithm 6 with gain $1/2$, then

$$m_1^{\mathrm{det}} = \tfrac{1}{2}m_0^{\mathrm{det}} + \tfrac{1}{2}\nabla f(x_0) = \nabla f(x_0) = G_0 = m_1.$$

The common measurable LMO selection gives the same direction and x_1. Induction over the momentum, radius, direction, and iterate recursions proves the claim. A countable intersection of probability-one oracle events makes the equality simultaneous over all generated indices. **Q.E.D.**

Lemma M.9 (Common crude finite-horizon radius bound)**.** *For every $T \geq 1$, every stochastic DA trajectory satisfies, simultaneously for all $\vartheta \in [0,1]$,*

$$\max_{0\leq k\leq T} \|x_k - x_0\| \leq r \prod_{j=1}^{T} (1 + j^{-1/2}) \leq r \exp(2\sqrt{T}). \tag{137}$$

Proof. Since $\psi_k \geq 1$ and $\|u(m)\| \leq 1$, $\eta_k \leq \bar{r}_{k+1}/\sqrt{k+1}$ and hence $\bar{r}_{k+2} \leq \bar{r}_{k+1}(1+(k+1)^{-1/2})$. Iterate this inequality and use $\sum_{j=1}^{T} j^{-1/2} \leq 2\sqrt{T}$. **Q.E.D.**

M.2.2 Descent and momentum tracking

Lemma M.10 (DA-Muon one-step descent)**.** *Under Assumption 2.1, every stochastic DA round satisfies*

$$f(x_{k+1}) \leq f(x_k) - \eta_k \|h_{k+1}\|_* + 2\eta_k \|h_k - m_{k+1}\|_* + \frac{7L}{2}\eta_k^2. \tag{138}$$

Proof. Let $\delta_k = x_{k+1} - x_k$. Smoothness, $\|\delta_k\| \leq \eta_k$, and the norming-direction identity give

$$\begin{aligned} f(x_{k+1}) &\leq f(x_k) + \langle m_{k+1}, \delta_k \rangle + \langle h_{k+1} - m_{k+1}, \delta_k \rangle + \langle h_k - h_{k+1}, \delta_k \rangle + \tfrac{L}{2}\eta_k^2 \\ &\leq f(x_k) - \eta_k \|m_{k+1}\|_* + \eta_k \|h_{k+1} - m_{k+1}\|_* + \tfrac{3L}{2}\eta_k^2. \end{aligned}$$

Now use $-\|m_{k+1}\|_* \leq -\|h_{k+1}\|_* + \|h_{k+1} - m_{k+1}\|_*$ and $\|h_{k+1} - m_{k+1}\|_* \leq \|h_k - m_{k+1}\|_* + L\eta_k$. **Q.E.D.**

Let $\mathcal{F}_k = \sigma(\xi_0, \ldots, \xi_k)$ and let $\mathcal{F}_{-1}$ be trivial. Induction shows that x_k, $\bar{r}_{k+1}$, and η_k are $\mathcal{F}_{k-1}$-measurable. Thus Assumption 2.2 gives

$$\mathbb{E}[\zeta_k \mid \mathcal{F}_{k-1}] = 0, \qquad \mathbb{E}[\|\zeta_k\|_2^2 \mid \mathcal{F}_{k-1}] \leq \sigma^2. \tag{139}$$

Set $d_1 = 0$, $n_1 = \zeta_0$, and, for $k \geq 1$, define

$$d_{k+1} = q_k d_k + q_k(h_{k-1} - h_k), \qquad n_{k+1} = q_k n_k + a_k \zeta_k. \tag{140}$$

Lemma M.11 (Tracking decomposition)**.** *For every $k \geq 0$,*

$$m_{k+1} - h_k = d_{k+1} + n_{k+1}. \tag{141}$$

If $W_1 = 0$ and

$$W_{k+1} = q_k(W_k + \eta_{k-1}), \qquad k \geq 1, \tag{142}$$

then $\|d_{k+1}\|_* \leq L W_{k+1}$.

Proof. The initialization gives $m_1 - h_0 = \zeta_0 = d_1 + n_1$. For $k \geq 1$,

$$m_{k+1} - h_k = q_k(m_k - h_{k-1}) + q_k(h_{k-1} - h_k) + a_k \zeta_k,$$

which proves the decomposition by induction. Smoothness gives $\|d_{k+1}\|_* \leq q_k \|d_k\|_* + q_k L \eta_{k-1}$, proving the domination. **Q.E.D.**

Lemma M.12 (Variable-gain drift energy)**.** *Let $v_0, \ldots, v_t \geq 0$, and define $W_1 = 0$ and $W_{k+1} = q_k(W_k + v_{k-1})$ for $k = 1, \ldots, t+1$. Then*

$$2\sum_{k=1}^{t} v_k W_{k+1} \leq 5 \sum_{k=0}^{t} \frac{v_k^2}{a_k}. \tag{143}$$

Proof. Since $a_k + q_k = 1$, weighted Cauchy–Schwarz gives

$$(W_k + v_{k-1})^2 \leq \frac{W_k^2}{q_k} + \frac{v_{k-1}^2}{a_k}.$$

Consequently,

$$W_{k+1}^2 \leq q_k W_k^2 + \frac{q_k^2}{a_k} v_{k-1}^2, \qquad a_k W_k^2 \leq W_k^2 - W_{k+1}^2 + \frac{v_{k-1}^2}{a_k}.$$

Summing the second inequality for $k = 1, \dots, t+1$ yields

$$\sum_{k=1}^{t+1} a_k W_k^2 \le \sum_{k=1}^{t+1} \frac{v_{k-1}^2}{a_k}. \tag{144}$$

Young's inequality and Lemma M.8 now give

$$\begin{aligned} 2\sum_{k=1}^{t} v_k W_{k+1} &\le \sum_{k=1}^{t} \frac{v_k^2}{a_k} + \sum_{k=1}^{t} a_k W_{k+1}^2 \\ &\le \sum_{k=0}^{t} \frac{v_k^2}{a_k} + 2\sum_{j=2}^{t+1} a_j W_j^2 \\ &\le \sum_{k=0}^{t} \frac{v_k^2}{a_k} + 2\sum_{i=0}^{t} \frac{v_i^2}{a_{i+1}} \le 5\sum_{k=0}^{t} \frac{v_k^2}{a_k}. \end{aligned}$$

Q.E.D.

Lemma M.13 (Stochastic momentum energy)**.** *Under Assumption 2.2, for every $k \ge 0$,*

$$\mathbb{E}\|n_{k+1}\|_2^2 \le 3\sigma^2 a_k, \qquad \mathbb{E}\|n_{k+1}\|_* \le \frac{\sqrt{3}s_\sigma}{\sqrt{\psi_k}}. \tag{145}$$

Proof. Let $V_{k+1} = \mathbb{E}\|n_{k+1}\|_2^2$. Since n_k is $\mathcal{F}_{k-1}$-measurable, the cross term vanishes by (139). Hence

$$V_1 \le \sigma^2, \qquad V_{k+1} \le q_k^2 V_k + a_k^2 \sigma^2 \quad (k \ge 1).$$

Because $a_0 \ge 1/3$, the claimed bound holds at $k = 0$. If $V_k \le 3\sigma^2 a_{k-1}$, Lemma M.8 gives

$$\begin{aligned} V_{k+1} &\le \sigma^2 a_k [3(1-a_k)^2(1+a_k) + a_k] \\ &= \sigma^2 a_k [3 - 2a_k - 3a_k^2 + 3a_k^3] \le 3\sigma^2 a_k, \end{aligned}$$

because $0 < a_k \le 1/2$. Norm comparison and Jensen's inequality yield

$$\mathbb{E}\|n_{k+1}\|_* \le \rho_{\mathrm{norm}} \sqrt{V_{k+1}} \le \sqrt{3} s_\sigma \sqrt{a_k} \le \frac{\sqrt{3}s_\sigma}{\sqrt{\psi_k}},$$

where $a_k = (1+\psi_k)^{-1} \le \psi_k^{-1}$. **Q.E.D.**

Lemma M.14 (Finite moments)**.** *Under Assumptions 2.1, 2.2, and L.1, all random variables used below are integrable.*

Proof. The trajectory assumption and smoothness imply $\|h_k\|_* \le \|h_0\|_* + LD_{\mathrm{tr}}$ almost surely. The oracle assumption and norm comparison imply $\mathbb{E}\|\zeta_k\|_*^2 \le s_\sigma^2$. Thus G_k has a finite dual second moment. The finite momentum recursions preserve square integrability, while $\eta_k \le D_{\mathrm{tr}}$ almost surely. The claim follows from Cauchy–Schwarz. **Q.E.D.**

M.2.3 Proof of Theorem L.4

Proof. Fix $n \in \{1, \dots, T\}$. Sum (138) for $k = 0, \dots, n-1$, use $f(x_n) \ge f^\star$, and apply Lemma M.11. With W_k from (142), this gives pathwise

$$\sum_{k=0}^{n-1} \eta_k \|h_{k+1}\|_* \le \Delta_0 + \frac{7L}{2}\sum_{k=0}^{n-1} \eta_k^2 + 2L\sum_{k=0}^{n-1} \eta_k W_{k+1} + 2\sum_{k=0}^{n-1} \eta_k \|n_{k+1}\|_*. \tag{146}$$

Here $W_1 = 0$. Lemma M.12, with $v_k = \eta_k$, and the schedule identities imply

$$\begin{aligned} 2L\sum_{k=0}^{n-1} \eta_k W_{k+1} &\le 5L \sum_{k=0}^{n-1} \frac{\eta_k^2}{a_k}, \\ \frac{\eta_k^2}{a_k} &= \frac{\bar{r}_{k+1}^2}{k+1} \frac{2+\vartheta\sqrt{k+1}}{1+\vartheta\sqrt{k+1}} \le \frac{2\bar{r}_{k+1}^2}{k+1}, \\ \eta_k^2 &\le \frac{\bar{r}_{k+1}^2}{k+1}. \end{aligned}$$

Since $\bar{r}_{k+1} \le \bar{r}_n$ for $k < n$,

$$\sum_{k=0}^{n-1} \eta_k \|h_{k+1}\|_* \le \Delta_0 + \frac{27L}{2}\bar{r}_n^2 H_n + 2\sum_{k=0}^{n-1} \eta_k \|n_{k+1}\|_*. \tag{147}$$

The weights $(\sqrt{k+1}\sqrt{\psi_k})^{-1}$ are nonincreasing. Hence

$$\min_{0\le k<n} \|h_{k+1}\|_* \le \frac{\sqrt{n}\sqrt{1+\vartheta\sqrt{n}}}{\sum_{i=1}^n \bar{r}_i} \left[\Delta_0 + \frac{27L}{2}\bar{r}_n^2 H_n + 2\sum_{k=0}^{n-1} \eta_k \|n_{k+1}\|_*\right]. \tag{148}$$

Extend the radius sequence only for the analysis by $\bar{r}_{T+1} = \max\{\bar{r}_T, \|x_T - x_0\|\}$. Then $\bar{r}_{T+1} \le D_{\mathrm{tr}}$ almost surely. Let $s = \lfloor T/2 \rfloor$ and restrict n to $\{s+1, \dots, T\}$. Since $n > T/2$, $\sum_{i=1}^n \bar{r}_i \ge nr$, and $1 + \vartheta\sqrt{n} \le B_{\vartheta,T}$, the objective term is at most

$$\frac{\sqrt{2B_{\vartheta,T}}}{r\sqrt{T}}\Delta_0. \tag{149}$$

For the deterministic tracking term, use $H_n \le H_T$ and $\bar{r}_n^2 \le D_{\mathrm{tr}}\bar{r}_{n+1}$. On each trajectory, let N be the smallest tail index minimizing the ratio in Lemma M.4. This finite tie-breaking rule makes N measurable. At $n = N$, that lemma gives

$$\begin{aligned}&\frac{27L}{2}H_n \frac{\sqrt{n}\sqrt{1+\vartheta\sqrt{n}}\,\bar{r}_n^2}{\sum_{i=1}^n \bar{r}_i} \\ &\quad\le \frac{27LD_{\mathrm{tr}}H_T\sqrt{B_{\vartheta,T}}}{\sqrt{T}}\Lambda_T(D_{\mathrm{tr}}/r).\end{aligned} \tag{150}$$

The stochastic term has one common pathwise upper bound for every tail prefix:

$$\begin{aligned}&\frac{2\sqrt{n}\sqrt{1+\vartheta\sqrt{n}}}{\sum_{i=1}^n \bar{r}_i}\sum_{k=0}^{n-1}\eta_k\|n_{k+1}\|_* \\ &\quad\le \frac{2\sqrt{2}D_{\mathrm{tr}}\sqrt{B_{\vartheta,T}}}{r\sqrt{T}}\sum_{k=0}^{T-1}\frac{\|n_{k+1}\|_*}{\sqrt{k+1}\sqrt{\psi_k}}.\end{aligned} \tag{151}$$

This domination is made before the radius-dependent prefix is selected, so the choice of N creates neither conditioning nor selection bias.

The full-trajectory minimum is no larger than the prefix minimum at N. Combining the preceding estimates, taking unconditional expectation, and applying Lemma M.13 gives

$$\mathbb{E}\sum_{k=0}^{T-1}\frac{\|n_{k+1}\|_*}{\sqrt{k+1}\sqrt{\psi_k}} \le \sqrt{3}s_\sigma\sum_{k=0}^{T-1}\frac{1}{\sqrt{k+1}\psi_k} = \sqrt{3}s_\sigma S_{\vartheta,T}.$$

All expectations are legitimate by Lemma M.14. Collecting the three terms proves (101). **Q.E.D.**

M.2.4 Proof of Corollary L.5

Proof. Let $s = s_\sigma$ and $g = Lr$. If $s = 0$, then $\vartheta_\sigma = 0$, the stochastic term in (101) vanishes, and the bound is exactly $C_T^{\mathrm{DA}}/\sqrt{T}$.

Suppose $s > 0$ and put $\vartheta = s/(g+s)$. Termwise domination by the $\vartheta = 0$ sum and Lemma M.7 give

$$\begin{aligned}sS_{\vartheta,T} &\le \min\left\{2s\sqrt{T}, s + \frac{2s}{\vartheta}\log(1+\vartheta\sqrt{T})\right\} \\ &\le \min\left\{2s\sqrt{T}, (g+s)\ell_T\right\}.\end{aligned} \tag{152}$$

Also,

$$\frac{\sqrt{B_{\vartheta,T}}}{\sqrt{T}} \le T^{-1/2} + \sqrt{\vartheta}\,T^{-1/4}. \tag{153}$$

Applying these inequalities to Theorem L.4 proves (103).

For fixed $s > 0$, the second branch above gives the stated $\widetilde{O}(T^{-1/4})$ order, provided the trajectory bound remains controlled. At $s = 0$ the exact $\widetilde{O}(T^{-1/2})$ form was proved above. For fixed T, the first branch tends to zero with s, while $\sqrt{\vartheta}$ also tends to zero. Lemma M.9 provides one deterministic D_{tr} valid for every ϑ, proving the finite-time continuity as $\sigma \downarrow 0$. **Q.E.D.**

N Proofs for Stochastic Star-Convex DA-Muon

Throughout the appendix, write

$$\Delta_k := f(x_k) - f(x^\star), \qquad d_k := \|x_k - x^\star\|, \qquad q := 1 - \alpha,$$

and use $D_\star, \chi, s_\sigma$ from (104). Recall

$$e_{k+1} := m_{k+1} - \nabla f(x_k). \tag{154}$$

N.1 Filtration, moments, and deterministic geometry

Let $\mathcal{F}_{-1}$ be the trivial sigma-field and $\mathcal{F}_k := \sigma(G_0, \dots, G_k)$. At iteration k, x_k and $\bar{r}_{k+1}$ are $\mathcal{F}_{k-1}$-measurable, whereas G_k is fresh. Put

$$\zeta_k := G_k - \nabla f(x_k). \tag{155}$$

Assumption 2.2 gives

$$\mathbb{E}[\zeta_k \mid \mathcal{F}_{k-1}] = 0, \qquad \mathbb{E}[\|\zeta_k\|_*^2 \mid \mathcal{F}_{k-1}] \le s_\sigma^2. \tag{156}$$

Lemma N.1 (Automatic running-radius bound)**.** *For every realization and every $k \ge 0$,*

$$r \le \bar{r}_{k+1} \le r(1+\beta)^k, \qquad \|x_{k+1} - x_k\| \le \beta \bar{r}_{k+1}, \qquad d_k \le D_\star + \bar{r}_{k+1}. \tag{157}$$

Consequently, $\mathcal{R}_T$ in (106) *is finite and satisfies* (107)*.*

Proof. The first lower bound follows from monotonicity and $\bar{r}_0 = r$. Since $\|s_{k+1}\| \le 1$,

$$\|x_{k+1} - x_0\| \le \|x_k - x_0\| + \beta \bar{r}_{k+1} \le (1+\beta)\bar{r}_{k+1}.$$

Therefore $\bar{r}_{k+2} \le (1+\beta)\bar{r}_{k+1}$, and induction proves the upper bound. The step bound is the update itself. Finally, $d_k \le D_\star + \|x_k - x_0\| \le D_\star + \bar{r}_{k+1}$. **Q.E.D.**

All random variables used below are integrable. Indeed, Lemma N.1 deterministically bounds every finite-horizon trajectory. Smoothness then bounds the population gradients, the oracle has a conditional second moment, and the momentum is a finite linear combination of oracle returns.

Because $x^\star$ is an unconstrained differentiable global minimizer, $\nabla f(x^\star) = 0$. Smoothness at $x^\star$ also gives

$$0 \le \Delta_0 \le \frac{L}{2} D_\star^2. \tag{158}$$

N.2 Pathwise star-convex progress

Lemma N.2 (One-step DA progress)**.** *For every $k = 0, \dots, T-1$, pathwise,*

$$\Delta_{k+1} \le (1 - c_k)\Delta_k + 2\beta \bar{r}_{k+1} \|e_{k+1}\|_* + \frac{L\beta^2}{2} \bar{r}_{k+1}^2, \qquad c_k := \frac{\beta \bar{r}_{k+1}}{D_\star + \bar{r}_{k+1}}. \tag{159}$$

Moreover,

$$c_k \ge \frac{\beta r}{D_\star + r} = \frac{\beta}{\chi} =: \theta. \tag{160}$$

Proof. Smoothness and $\eta_k = \beta \bar{r}_{k+1}$ give

$$\Delta_{k+1} \le \Delta_k + \eta_k \langle \nabla f(x_k), s_{k+1} \rangle + \frac{L}{2} \eta_k^2. \tag{161}$$

When $d_k > 0$, put $v_k := (x^\star - x_k)/d_k$. Otherwise put $v_k = 0$. Since $\|v_k\| \le 1$ and s_{k+1} minimizes the m_{k+1}-linear form,

$$\langle m_{k+1}, s_{k+1} \rangle \le \langle m_{k+1}, v_k \rangle.$$

Consequently,

$$\begin{aligned} \langle \nabla f(x_k), s_{k+1} \rangle &\le \langle \nabla f(x_k), v_k \rangle + \langle \nabla f(x_k) - m_{k+1}, s_{k+1} - v_k \rangle \\ &\le -\frac{\Delta_k}{d_k} + 2\|e_{k+1}\|_*. \end{aligned} \tag{162}$$

Here star-convexity was used. If $d_k = 0$, then $x_k = x^\star$ and $\Delta_k = 0$, so the same display holds with the quotient interpreted as zero. Lemma N.1 gives

$$\frac{\eta_k}{d_k} \ge \frac{\beta \bar{r}_{k+1}}{D_\star + \bar{r}_{k+1}}$$

whenever $d_k > 0$. Substitution proves (159). The map $a \mapsto a/(D_\star + a)$ is increasing and $\bar{r}_{k+1} \ge r$, proving (160). **Q.E.D.**

Iterating Lemma N.2 proves the pathwise last-iterate bound (110). This argument uses no probability and retains the realized trajectory-dependent contractions.

N.3 RMS STOCHASTIC MOMENTUM TRACKING

Lemma N.3 (Pointwise RMS momentum bound)**.** *For every $k = 0, \dots, T-1$,*

$$\mathbb{E}[\bar{r}_{k+1} \|e_{k+1}\|_*] \le \mathcal{R}_T \left[q^k s_\sigma + L\beta \mathcal{R}_T \frac{1-\alpha}{\alpha} + s_\sigma \sqrt{\frac{\alpha}{2-\alpha}} \right]. \tag{163}$$

When $\alpha = 1$, the drift term vanishes. The initialization term remains only at $k = 0$ and is represented by the convention $q^0 = 1$.

Proof. The initialization is $e_1 = \zeta_0$. For $k \ge 1$,

$$e_{k+1} = q e_k + q(\nabla f(x_{k-1}) - \nabla f(x_k)) + \alpha \zeta_k. \tag{164}$$

Unroll this recursion and split the result into drift and noise:

$$\begin{aligned} e_{k+1} &= q \sum_{i=1}^{k} q^{k-i} (\nabla f(x_{i-1}) - \nabla f(x_i)) + n_{k+1}, \\ n_{k+1} &:= q^k \zeta_0 + \alpha \sum_{i=1}^{k} q^{k-i} \zeta_i. \end{aligned}$$

Since the radius is nondecreasing, smoothness, the step bound, and Cauchy–Schwarz give

$$\begin{aligned} &\mathbb{E}\left[\bar{r}_{k+1} \left\| q \sum_{i=1}^{k} q^{k-i} (\nabla f(x_{i-1}) - \nabla f(x_i)) \right\|_* \right] \\ &\quad \le L\beta \sum_{i=1}^{k} q^{k-i+1} \mathbb{E}[\bar{r}_{k+1} \bar{r}_i] \le L\beta \mathcal{R}_T^2 \sum_{j=1}^{\infty} q^j = L\beta \mathcal{R}_T^2 \frac{1-\alpha}{\alpha}. \end{aligned}$$

The martingale-difference property makes all Euclidean cross terms in n_{k+1} vanish. Thus

$$\mathbb{E}\|n_{k+1}\|_*^2 \le s_\sigma^2 \left(q^{2k} + \frac{\alpha}{2-\alpha} \right). \tag{165}$$

Another application of Cauchy–Schwarz, followed by $\sqrt{a+b} \le \sqrt{a} + \sqrt{b}$, yields

$$\mathbb{E}[\bar{r}_{k+1} \|n_{k+1}\|_*] \le \mathcal{R}_T s_\sigma \left(q^k + \sqrt{\frac{\alpha}{2-\alpha}} \right).$$

Combining the drift and noise estimates proves the lemma. **Q.E.D.**

Lemma N.4 (Geometrically weighted momentum)**.** *Let $p := 1 - \theta$ and $\delta := p - q = \alpha - \theta > 0$. Then*

$$\sum_{k=0}^{T-1} p^{T-1-k}\mathbb{E}[\bar{r}_{k+1}\|e_{k+1}\|_*] \leq \mathcal{R}_T s_\sigma \frac{p^T - q^T}{\delta} + \frac{1-p^T}{\theta}\left[L\beta\mathcal{R}_T^2\frac{1-\alpha}{\alpha} + \mathcal{R}_T s_\sigma\sqrt{\frac{\alpha}{2-\alpha}}\right]. \tag{166}$$

Proof. Multiply (163) by p^{T-1-k} and sum. Use the exact identities

$$\sum_{k=0}^{T-1} p^{T-1-k}q^k = \frac{p^T - q^T}{p-q}, \qquad \sum_{k=0}^{T-1} p^{T-1-k} = \frac{1-p^T}{1-p}.$$

Q.E.D.

N.4 Explicit finite-horizon bound and schedules

Proof of Theorem L.6. By (160), (159) implies

$$\Delta_{k+1} \leq p\Delta_k + 2\beta\bar{r}_{k+1}\|e_{k+1}\|_* + \frac{L\beta^2}{2}\bar{r}_{k+1}^2.$$

Take expectations and unroll:

$$\mathbb{E}\Delta_T \leq p^T\Delta_0 + 2\beta\sum_{k=0}^{T-1} p^{T-1-k}\mathbb{E}[\bar{r}_{k+1}\|e_{k+1}\|_*] + \frac{L\beta^2}{2}\sum_{k=0}^{T-1} p^{T-1-k}\mathbb{E}\bar{r}_{k+1}^2. \tag{167}$$

Apply Lemma N.4, use $\mathbb{E}\bar{r}_{k+1}^2 \leq \mathcal{R}_T^2$, $\theta = \beta/\chi$, and $1 - p^T \leq 1$. The drift, stationary-noise, and smoothness terms become, respectively,

$$2L\beta\mathcal{R}_T^2\chi\frac{1-\alpha}{\alpha}, \qquad 2\mathcal{R}_T\chi s_\sigma\sqrt{\frac{\alpha}{2-\alpha}}, \qquad \frac{L\beta\mathcal{R}_T^2\chi}{2}.$$

This proves (111). Since $2(1-\alpha)/\alpha + 1/2 \leq 2/\alpha$ and $p^T - q^T \leq p^T$, (112) follows. Finally, $\theta \leq \beta \leq \alpha/4$ implies $\delta = \alpha - \theta \geq 3\alpha/4 > 0$. **Q.E.D.**

Proof of Corollary L.8. Since $\beta_T \leq 1/32$, $\alpha_T = 8\beta_T^{2/3} < 1$ and $\beta_T \leq \alpha_T/4$. In (112), use

$$p^T \leq \exp(-\beta_T T/\chi), \qquad \sqrt{\frac{\alpha_T}{2-\alpha_T}} \leq 2\sqrt{2}\,\beta_T^{1/3},$$
$$\delta \geq \frac{3}{4}\alpha_T, \qquad \frac{2\beta_T}{\delta} \leq \frac{1}{3}\beta_T^{1/3}, \qquad \frac{2L\beta_T}{\alpha_T} = \frac{L}{4}\beta_T^{1/3}.$$

This gives (115). Substitute $\mathcal{R}_T \leq \overline{R}_{\mathrm{DA}}$, use

$$\beta_T T = \min\left\{\frac{T}{32}, \log^2(T+1)\right\}, \qquad \beta_T^{1/3} \leq \frac{\log^{2/3}(T+1)}{T^{1/3}},$$

and absorb the logarithm into $\widetilde{O}(\cdot)$ to obtain (116). **Q.E.D.**

Proof of Corollary L.9. Under the certificate $\mathcal{R}_T \leq \overline{R}$, apply (112) with $P = \overline{R}^2\chi$ and $Q = \overline{R}\chi$. The gain definitions imply $\delta \geq 3\alpha_\varepsilon/4$ and

$$\frac{2L\beta_\varepsilon P}{\alpha_\varepsilon} \leq \frac{\varepsilon}{10}, \qquad 2Qs_\sigma\sqrt{\frac{\alpha_\varepsilon}{2-\alpha_\varepsilon}} \leq \frac{\varepsilon}{4}.$$

The horizon gives $p^T \le \exp(-\beta_\varepsilon T/\chi) \le \Lambda_\varepsilon^{-1}$. Hence the initial contraction is at most $\varepsilon/4$, while

$$\frac{2\beta_\varepsilon \overline{R} s_\sigma}{\delta} p^T \le \frac{2\overline{R} s_\sigma}{3} \Lambda_\varepsilon^{-1} \le \frac{\varepsilon}{4}.$$

The four terms sum to at most $17\varepsilon/20$.

Up to numerical constants,

$$\frac{1}{\beta_\varepsilon} \lesssim \max\left\{1, \frac{1}{\alpha_\varepsilon}, \frac{LP}{\alpha_\varepsilon \varepsilon}\right\}, \qquad \frac{1}{\alpha_\varepsilon} \lesssim 1 + \frac{Q^2 s_\sigma^2}{\varepsilon^2}.$$

Multiply by χ from (120) and substitute P, Q to obtain (121). **Q.E.D.**

Proof of the deterministic specialization (123). When $s_\sigma = 0$ and $\alpha = 1$, the initialization, drift, and stationary noise terms in (111) vanish. The remaining smoothness coefficient is $1/2$. **Q.E.D.**

O Additional Local-Certificate Methods

This appendix records a local-certificate alternative. Scale-Calibrated Muon sets the radius from current gradient–momentum alignment and achieves an $O(1/T)$ last-iterate rate under a bounded initial sublevel set. The assumption and theorem are independent of the main predictable results. Stochastic local-certificate variants are stated in Appendix P.

O.1 Local certificate: Scale-Calibrated Muon

Scale-Calibrated Muon is a star-convex scale rule that is distance-free at the algorithmic level: it does not use the distance to a minimizer, the diameter of a prescribed clipping domain, a trajectory radius, or the level-set radius D_{lev}. The radius D_{lev} appears only in the analysis through the bounded initial sublevel-set assumption below.

Assumption O.1 (Bounded initial sublevel set)**.** Let $x^\star$ be the global minimizer from Assumption 2.3. There exists $D_{\mathrm{lev}} < \infty$ such that

$$f(x) \le f(x_0) \quad \Longrightarrow \quad \|x - x^\star\| \le D_{\mathrm{lev}}.$$

Assumption O.1 is a standard initial-level-set localization condition in descent, trust-region, and regularized Newton-type analyses (Conn et al., 2000; Nesterov, 2008; Doikov et al., 2024). The descent inequality below implies that Algorithm 8 is monotone. The formal proof is given in Appendix Q. Hence its iterates remain in this set. The assumption is therefore local to the trajectory and weaker than assuming a bounded clipping domain.

Let $g_k := \nabla f(x_k)$. The method maintains the same exponential moving average as Muon,

$$m_{k+1} = (1-\alpha) m_k + \alpha g_k, \qquad \alpha \in (0,1),$$

initialized by $m_0 = g_0$. The Muon direction is chosen by the linear minimization oracle

$$u_k \in \arg\max_{\|u\| \le 1} \langle m_{k+1}, u \rangle, \qquad \langle m_{k+1}, u_k \rangle = \|m_{k+1}\|_*.$$

Because m_{k+1} may be stale, we certify its alignment with the current gradient:

$$\langle g_k, u_k \rangle = \langle m_{k+1}, u_k \rangle + \langle g_k - m_{k+1}, u_k \rangle \ge \|m_{k+1}\|_* - \|g_k - m_{k+1}\|_*.$$

We define

$$e_k := \|g_k - m_{k+1}\|_*, \qquad a_k := (\|m_{k+1}\|_* - e_k)_+ ,$$

and set

$$\eta_k := \frac{a_k}{L}, \qquad x_{k+1} = x_k - \eta_k u_k.$$

If $a_k > 0$, then the update direction $-u_k$ is certified as a descent direction for the current gradient. If the momentum direction is not certified, the positive part reduces the step to zero. By L-smoothness,

$$f(x_{k+1}) \le f(x_k) - \frac{a_k^2}{2L},$$

so the method is monotone and all iterates remain in the initial sublevel set.

Algorithm 8 Scale-Calibrated Muon

Require: Initial point $x_0 \in \mathcal{E}$, smoothness constant L, momentum parameter $\alpha \in (0,1)$

$g_0 \leftarrow \nabla f(x_0)$, $m_0 \leftarrow g_0$
for $k = 0, 1, \ldots, T-1$ **do**
  $g_k \leftarrow \nabla f(x_k)$
  $m_{k+1} \leftarrow (1-\alpha)m_k + \alpha g_k$
  $e_k \leftarrow \|g_k - m_{k+1}\|_*$
  Choose $u_k \in \arg\max_{\|u\|\le 1} \langle m_{k+1}, u\rangle$
  $a_k \leftarrow (\|m_{k+1}\|_* - e_k)_+$
  $\eta_k \leftarrow a_k / L$
  $x_{k+1} \leftarrow x_k - \eta_k u_k$
end for

Theorem O.2 (Last-iterate rate for Scale-Calibrated Muon)**.** *Let Assumptions 2.1, 2.3, and O.1 hold, and consider Algorithm 8. Then, for all $T \ge 1$,*

$$f(x_T) - f(x^\star) \le K_\alpha \frac{L D_{\mathrm{lev}}^2}{T},$$

where $K_\alpha > 0$ depends only on the momentum parameter α.

Appendix Q proves the result with a Lyapunov function that combines the objective gap and the squared momentum-tracking error. Section P.1 develops the one-gradient stochastic version. Its steps are no longer pathwise monotone, so the unconditional result controls non-convex stationarity rather than a localized star-convex objective gap.

DF-Muon, introduced next, uses a one-dimensional search over the trust-region radius and does not require the bounded initial sublevel-set assumption.

P Additional Stochastic Local-Certificate Methods

This appendix studies a different timing pattern from the predictable method in the main text. A local Scale-Calibrated method may use the fresh gradient in its momentum, direction, and radius. An augmented Lyapunov function handles the resulting correlation. The unconditional non-convex result requires no trajectory bound. Its star-convex last-iterate extension, stated next, assumes uniform mean-square boundedness of the iterates.

P.1 Fresh stochastic local calibration

The deterministic Scale-Calibrated method uses the exact current mismatch $\|\nabla f(x_k) - m_{k+1}\|_*$, which is not observable from one noisy gradient. The stochastic variant below replaces its dynamic smoothness component by a recursive scalar certificate r_k. Initialization lag and stochastic momentum noise are not treated as observable radius estimates. They are amortized in expectation. This yields an unconditional stationarity theorem for every smooth lower-bounded objective, without star-convexity or stochastic localization.

Let $\{\alpha_k\}_{k\ge 1} \subset (0,1]$ be deterministic coefficients and let $\beta \in (0,1]$ be a deterministic initialization coefficient. At round k define

$$a_k := (\|m_k\|_* - r_k)_+, \qquad \eta_k := \frac{\alpha_{k+1} a_k}{L}. \tag{168}$$

The coefficient α_{k+1} both damps the current step and mixes the next fresh gradient into the momentum. This indexing makes the coefficient predictable for the next oracle draw.

The first sample is reused by the first update, so T updates use exactly T stochastic-gradient evaluations. The output draw uses no oracle call and can be implemented online by weighted reservoir sampling. The default online initialization is $\beta = \alpha_1$, which is exactly the first update of a zero-initialized EMA. The ordinary warm start $\beta = 1$, hence $m_0 = G_0$, is also covered without a larger initial batch.

Algorithm 9 Stochastic variant of the local Scale-Calibrated method

Require: $x_0 \in \mathcal{E}$, $L > 0$, deterministic coefficients $\{\alpha_k\}_{k=1}^{T} \subset (0, 1]$, and $\beta \in (0, 1]$
Draw $G_0 \leftarrow g(x_0; \xi_0)$. Set $m_0 \leftarrow \beta G_0$ and $r_0 \leftarrow 0$
for $k = 0, 1, \dots, T-1$ **do**
  if $k \geq 1$ **then**
    Draw the fresh sample $G_k \leftarrow g(x_k; \xi_k)$
    $m_k \leftarrow (1 - \alpha_k) m_{k-1} + \alpha_k G_k$
  end if
  $u_k \leftarrow u(m_k)$ using the fixed measurable LMO selection
  $a_k \leftarrow (\|m_k\|_* - r_k)_+$ and $\eta_k \leftarrow \alpha_{k+1} a_k / L$
  $x_{k+1} \leftarrow x_k - \eta_k u_k$
  $r_{k+1} \leftarrow (1 - \alpha_{k+1})\big(r_k + L\|x_{k+1} - x_k\|\big)$
end for
Draw $K \in \{0, \dots, T-1\}$ independently with $\Pr(K = k) = \alpha_{k+1} / \sum_{j=0}^{T-1} \alpha_{j+1}$
return $\widehat{x}_T \leftarrow x_K$

In the common one-dimensional update notation,

$$c_k = x_k, \qquad R_k = a_k / L, \qquad \gamma_k = \alpha_{k+1}.$$

The momentum and the LMO over the unit norm ball are unchanged. The extra multiplier α_{k+1} is part of the proved method: its damping is what lets the EMA contraction absorb the stochastic momentum error.

For the result below assume $f^\star := \inf_x f(x) > -\infty$ and write $\Delta_0 := f(x_0) - f^\star$. Define

$$S_T := \sum_{k=0}^{T-1} \alpha_{k+1}, \qquad V_T := \sum_{k=0}^{T-1} \alpha_{k+1}^2. \tag{169}$$

Theorem P.1 (Unconditional stochastic Scale-Calibrated stationarity)**.** *Let Assumptions 2.1 and 2.2 hold, let x_0 be deterministic, and suppose $f^\star > -\infty$. Run Algorithm 9 for $T \geq 1$ updates with deterministic coefficients $\alpha_k \in (0, 1]$ and deterministic $\beta \in (0, 1]$. Define*

$$\mathfrak{B}_T(\beta) := \frac{62L\big(1 + 6(1-\beta)^2\big)\Delta_0 + 186\rho_{\mathrm{norm}}^2 \sigma^2 (\beta^2 + V_T)}{S_T}. \tag{170}$$

Then

$$\mathbb{E}\,\|\nabla f(\widehat{x}_T)\|_*^2 \leq \mathfrak{B}_T(\beta), \tag{171}$$

$$\min_{0 \leq k < T} \mathbb{E}\,\|\nabla f(x_k)\|_*^2 \leq \mathfrak{B}_T(\beta), \tag{172}$$

$$\mathbb{E} \min_{0 \leq k < T} \|\nabla f(x_k)\|_* \leq \sqrt{\mathfrak{B}_T(\beta)}. \tag{173}$$

For the online initialization $\beta = \alpha_1$,

$$\mathbb{E}\,\|\nabla f(\widehat{x}_T)\|_*^2 \leq \frac{434L\Delta_0 + 372\rho_{\mathrm{norm}}^2 \sigma^2 V_T}{S_T}. \tag{174}$$

No localization or bounded-trajectory assumption is required. The result also assumes neither star-convexity, a minimizer, bounded sublevel sets, bounded stochastic gradients, nor moments above order two. The full proof is in Appendix R.

Proof mechanism. Writing $h_k := \nabla f(x_k)$, the momentum decomposes as

$$m_k = h_k + p_k + d_k + n_k,$$

where p_k is the dynamic deterministic lag, d_k is the initialization transient, and n_k is the stochastic momentum noise. The scalar recursion certifies $\|p_k\|_* \leq r_k$ pathwise. The augmented potential

$$\Psi_k = f(x_k) - f^\star + \frac{r_k^2}{4L} + \frac{3}{L}\left(\|d_k\|_*^2 + \rho_{\mathrm{norm}}^2 \|n_k\|_2^2\right)$$

has a one-step expected drift controlling $a_k^2 + r_k^2$ and both error energies. A pathwise norm inequality then turns this controlled energy into $\|\nabla f(x_k)\|_*^2$. In particular, no factorization of an adaptive coefficient and a correlated noise state is used. The coefficients in the theorem are deterministic, and the actual noise state is kept inside the Lyapunov function.

P.1.1 Parameter choices and complexity

Theorem P.1 separates accumulated progress S_T, the squared-coefficient sum V_T, and the one-time initialization term β^2.

Corollary P.2 (Constant parameter choice)**.** *Let $\alpha_k \equiv \alpha \in (0, 1]$ and use the online initialization $\beta = \alpha$. Then*

$$\mathbb{E}\,\|\nabla f(\widehat{x}_T)\|_*^2 \le \frac{434L\Delta_0}{\alpha T} + 372\rho_{\mathrm{norm}}^2\sigma^2\alpha. \tag{175}$$

In the noiseless case, $\alpha = \beta = 1$ gives $\min_{k<T} \|\nabla f(x_k)\|_*^2 \le 62L\Delta_0/T$.

Corollary P.3 (Horizon-dependent and anytime parameter choices)**.** *For a prescribed horizon, choose $\alpha_k \equiv T^{-1/2}$ and $\beta = 1$. Then*

$$\mathbb{E}\,\|\nabla f(\widehat{x}_T)\|_*^2 \le \frac{62L\Delta_0 + 372\rho_{\mathrm{norm}}^2\sigma^2}{\sqrt{T}}. \tag{176}$$

Consequently $\mathbb{E}\,\|\nabla f(\widehat{x}_T)\|_* \le \varepsilon$ whenever

$$T \ge \frac{\left(62L\Delta_0 + 372\rho_{\mathrm{norm}}^2\sigma^2\right)^2}{\varepsilon^4}. \tag{177}$$

Alternatively, the anytime choice $\alpha_{k+1} = (k+1)^{-1/2}$ and $\beta = 1$ satisfies, for every $T \ge 1$,

$$\mathbb{E}\,\|\nabla f(\widehat{x}_T)\|_*^2 \le \frac{62L\Delta_0 + 186\rho_{\mathrm{norm}}^2\sigma^2(2 + \log T)}{\sqrt{T}}. \tag{178}$$

The horizon-dependent schedule uses T, but not σ, Δ_0, a target accuracy, or a distance parameter. The anytime schedule removes T as well, but is not a best-of-both-worlds method: when $\sigma = 0$ it gives $O(T^{-1/2})$ instead of the constant-parameter $O(T^{-1})$ squared-stationarity rate, and with noise it pays a logarithmic factor.

Corollary P.4 (Accuracy- and noise-dependent parameter choice)**.** *Let*

$$C_0 := 434L\Delta_0, \qquad C_1 := 372\rho_{\mathrm{norm}}^2\sigma^2.$$

If $C_1 > 0$, choose $\alpha_\varepsilon := \min\{1, \varepsilon^2/(2C_1)\}$. Otherwise choose $\alpha_\varepsilon = 1$. Set $\alpha_k \equiv \beta \equiv \alpha_\varepsilon$. Then $\mathbb{E}\,\|\nabla f(\widehat{x}_T)\|_* \le \varepsilon$ whenever

$$T \ge \max\left\{1, \frac{2C_0}{\alpha_\varepsilon\varepsilon^2}\right\}.$$

In particular, it is sufficient to use

$$T = O\left(1 + \frac{L\Delta_0}{\varepsilon^2} + \frac{L\Delta_0\rho_{\mathrm{norm}}^2\sigma^2}{\varepsilon^4}\right). \tag{179}$$

This benchmark assumes that σ and the target accuracy ε are known when the parameter is chosen.

P.1.2 Relation to deterministic Scale-Calibrated Muon and fixed-radius Muon

Deterministic Scale-Calibrated Muon in Algorithm 8 uses the exact mismatch

$$e_k = \|\nabla f(x_k) - m_{k+1}\|_*, \qquad \eta_k^{\mathrm{det}} = \frac{(\|m_{k+1}\|_* - e_k)_+}{L}.$$

That exact certificate gives pathwise descent and, under star-convexity and a bounded initial sublevel set, an $O(LD_{\mathrm{lev}}^2/T)$ last-iterate objective gap. Algorithm 9 instead certifies only the recursive dynamic lag, amortizes initialization and sampling noise, and damps its radius by α_{k+1}. Without additional geometry, this yields the non-convex stationarity guarantee above. Section P.2 shows that the constant-parameter trajectory also admits a last-iterate objective-gap bound under star-convexity and uniform mean-square boundedness. All three Scale-Calibrated variants share the current center and the norm-induced Muon direction, but the stochastic recursion and deterministic Scale-Calibrated Muon do not produce the same trajectory for a general momentum parameter. They coincide in the noiseless $\alpha = \beta = 1$ specialization.

For comparison, the unregularized non-convex result of Kovalev (2025, Corollary 2) has oracle complexity

$$O\left(\max\left\{\frac{\rho_{\mathrm{norm}}\sigma}{\varepsilon}, \frac{\rho_{\mathrm{norm}}^3\sigma^3}{\varepsilon^3}, \frac{L\Delta_0}{\varepsilon^2}, \frac{L\Delta_0\rho_{\mathrm{norm}}^2\sigma^2}{\varepsilon^4}\right\}\right). \tag{180}$$

Corollary P.4 has the same deterministic term and the same leading mixed stochastic term. Its displayed bound lacks the two standalone initialization-noise terms because the online initialization uses $m_0 = \alpha G_0$ rather than an unscaled noisy sample. This is a comparison of proven upper bounds for different adaptive stepsize rules and initializations, not a claim of a stronger minimax result. Kovalev's method also covers a convex composite regularizer. The present local method is proved only for the smooth objective.

Practical scope. The update analyzed in Theorem P.1 adds one scalar state and no extra gradient call, but it requires a valid smoothness constant L with respect to the chosen norm, an exact dual norm, and an exact LMO over the unit norm ball. In spectral–nuclear geometry this means an exact nuclear norm and polar direction. Experiments may replace the nuclear norm by the already available support value of an approximate polar direction, but that requires an approximate-LMO theorem to be fully covered. Moreover, a conventional EMA coefficient $\alpha \ll 1$ also multiplies the certified step by α. This coupling can make the analyzed method conservative. The result does not cover a general nonsmooth regularizer or decoupled weight decay. The unconditional theorem controls a random iterate sampled with probabilities proportional to α_{k+1}. The last-iterate variant in Section P.2 instead requires the explicit trajectory stability condition stated there. These limitations are algorithmic and geometric. No stronger stochastic moment or localization assumption is hidden in the stationarity theorem.

P.2 Last-iterate star-convex local calibration

The fresh local method also admits a last-iterate analysis under star-convexity. The recursion below uses constant coefficients in Algorithm 9, but it returns the final trajectory point rather than a random iterate sampled with probabilities proportional to α_{k+1}. This change in output requires a separate stability condition.

Fix $\alpha, \beta \in (0, 1]$, write $q := 1 - \alpha$, and use

$$a_k := \left(\|m_k\|_* - r_k\right)_+, \qquad \eta_k := \frac{\alpha a_k}{L}. \tag{181}$$

The algorithm uses exactly T fresh stochastic gradients for T updates, never evaluates the population objective, and contains no anchor or regularizer gradient. In the common one-dimensional update notation,

$$c_k = x_k, \qquad R_k = a_k/L, \qquad \gamma_k = \alpha.$$

Algorithms 8, 9, and 10 are, respectively, the deterministic, stochastic-stationarity, and stochastic last-iterate variants of the local Scale-Calibrated method. We display the last variant separately because it returns the last iterate at a fixed horizon.

Lemma R.1 applies to this constant-parameter specialization. Consequently, for every fixed finite horizon,

$$D_{\mathrm{ms},T}^2 := \max_{0\le k\le T} \mathbb{E}\|x_k - x^\star\|^2 < \infty. \tag{182}$$

This a posteriori quantity is finite for each run. It is not supplied to the algorithm, and its finite-horizon bound need not be uniform in T.

Theorem P.5 (Fixed-horizon stochastic local Scale-Calibrated last-iterate bound)**.** *Let Assumptions 2.1, 2.3, and 2.2 hold. Run Algorithm 10 for $T \ge 1$ with $\alpha, \beta \in (0, 1]$, let $s_\sigma := \rho_{\mathrm{norm}}\sigma$, and define*

$$\mathcal{J}_{0,T}(\beta) := f(x_0) - f(x^\star) + 3(1-\beta)^2 L D_{\mathrm{ms},T}^2 + \frac{3\beta^2 s_\sigma^2}{L}.$$

Algorithm 10 Last-iterate star-convex variant of stochastic Scale-Calibrated Muon

Require: $x_0 \in \mathcal{E}$, horizon $T \geq 1$, smoothness $L > 0$, parameters $\alpha, \beta \in (0, 1]$
Set $q \leftarrow 1 - \alpha$ and $r_0 \leftarrow 0$
Draw $G_0 \leftarrow g(x_0; \xi_0)$ and set $m_0 \leftarrow \beta G_0$
for $k = 0, 1, \ldots, T-1$ **do**
  if $k \geq 1$ **then**
    Draw the fresh sample $G_k \leftarrow g(x_k; \xi_k)$
    $m_k \leftarrow q m_{k-1} + \alpha G_k$
  end if
  $u_k \leftarrow u(m_k)$ using the fixed measurable LMO selection
  $a_k \leftarrow (\|m_k\|_* - r_k)_+$ and $\eta_k \leftarrow \alpha a_k / L$
  $x_{k+1} \leftarrow x_k - \eta_k u_k$
  $r_{k+1} \leftarrow q\big(r_k + L\|x_{k+1} - x_k\|\big)$
end for
return x_T

There is a universal numerical constant $C > 0$ such that

$$\mathbb{E}[f(x_T) - f(x^\star)] \leq C\Bigg[e^{-\alpha T/48}\mathcal{J}_{0,T}(\beta) + \frac{L D_{\mathrm{ms},T}^2}{\alpha(T+1)} + D_{\mathrm{ms},T} s_\sigma \sqrt{\alpha} + \frac{\alpha s_\sigma^2}{L}\Bigg]. \tag{183}$$

The algorithm does not use $D_{\mathrm{ms},T}$, σ, $f(x^\star)$, or a target accuracy. This is a fixed-horizon a posteriori inequality. Obtaining a convergence rate requires uniform control of the horizon-indexed trajectories.

The proof in Appendix R.5 starts from the same augmented stochastic drift used for Theorem P.1. Star-convexity closes the final-iterate gap against the mean-square trajectory radius, producing a noisy Riccati recursion instead of a summed stationarity bound.

Assumption P.6 (Uniform mean-square boundedness of horizon-tuned local Scale-Calibrated Muon)**.** For each $T \geq 1$, let $\{x_k^{[T]}\}_{k=0}^T$ be the trajectory of Algorithm 10 run with

$$\alpha_T = T^{-2/3}, \qquad \beta = 1.$$

There is a constant $\overline{D}_{\mathrm{ms}} < \infty$, independent of T, such that

$$\sup_{T \geq 1} \max_{0 \leq k \leq T} \mathbb{E}\|x_k^{[T]} - x^\star\|^2 \leq \overline{D}_{\mathrm{ms}}^2. \tag{184}$$

The constant $\overline{D}_{\mathrm{ms}}$ appears only in the guarantee and is not passed to the algorithm.

Corollary P.7 (Horizon-dependent stochastic local Scale-Calibrated last-iterate schedule)**.** *Under Assumptions 2.1, 2.3, 2.2, and P.6, Algorithm 10 with $\alpha_T = T^{-2/3}$ and $\beta = 1$ satisfies*

$$\mathbb{E}[f(x_T) - f(x^\star)] \leq C'\Bigg[\frac{L\overline{D}_{\mathrm{ms}}^2 + \overline{D}_{\mathrm{ms}} s_\sigma}{T^{1/3}} + \frac{s_\sigma^2}{L T^{2/3}} + e^{-T^{1/3}/48}\left(L\overline{D}_{\mathrm{ms}}^2 + \frac{s_\sigma^2}{L}\right)\Bigg] \tag{185}$$

for a universal constant $C' > 0$. Consequently,

$$\mathbb{E}[f(x_T) - f(x^\star)] = O\left(\frac{L\overline{D}_{\mathrm{ms}}^2 + \overline{D}_{\mathrm{ms}}\rho_{\mathrm{norm}}\sigma}{T^{1/3}} + \frac{\rho_{\mathrm{norm}}^2\sigma^2/L}{T^{2/3}}\right),$$

which gives $O(\varepsilon^{-3})$ stochastic iteration complexity for fixed problem parameters.

Proof. Assumption P.6 gives $D_{\mathrm{ms},T} \leq \overline{D}_{\mathrm{ms}}$. Smoothness at the global minimizer gives $f(x_0) - f(x^\star) \leq L\|x_0 - x^\star\|^2/2 \leq L\overline{D}_{\mathrm{ms}}^2/2$. Substitute $\alpha_T = T^{-2/3}$ and $\beta = 1$ in Theorem P.5. **Q.E.D.**

Corollary P.8 (Deterministic specialization of the stochastic local method)**.** *Let Assumptions 2.1, 2.3, and 2.2 hold with $\sigma = 0$. Use $\beta = 1$ and fix any $\alpha_0 \in (0, 1]$ independently of T. If Assumption O.1 holds, then*

$$f(x_T) - f(x^\star) \le C\left[e^{-\alpha_0 T/48}\big(f(x_0) - f(x^\star)\big) + \frac{LD_{\mathrm{lev}}^2}{\alpha_0(T+1)}\right].$$

Thus the damped stochastic variant retains deterministic $O(1/T)$ order for every fixed $\alpha_0 > 0$.

Proof. With exact gradients and $\beta = 1$, the initialization and noise components in Appendix R vanish. The certified descent is then pathwise, so every iterate remains in the initial sublevel set and $D_{\mathrm{ms},T} \le D_{\mathrm{lev}}$. Apply Theorem P.5. **Q.E.D.**

For one $m \times n$ matrix block in spectral–nuclear geometry, let $r := \min\{m, n\}$. Since $\rho_{\mathrm{norm}} = \sqrt{r}$, the stochastic terms in Corollary P.7 become $\overline{D}_{\mathrm{ms}}\sqrt{r}\,\sigma/T^{1/3}$ and $r\sigma^2/(LT^{2/3})$, where $\overline{D}_{\mathrm{ms}}$ is measured in spectral norm. No anchor-smoothness factor is introduced.

The main limitation is Assumption P.6. Finite-horizon integrability does not provide a constant uniform over the triangular family of horizon-tuned runs, and bounded initial population sublevels do not imply it for nonmonotone stochastic trajectories. The result is in expectation, is horizon-dependent rather than anytime, and assumes an exact LMO, exact dual norm, and valid global smoothness constant. The predictable recentered distance-free method in Appendix J removes trajectory localization at the cost of a different one-dimensional update and oracle timing.

Q Proofs for Scale-Calibrated Muon

This appendix proves Theorem O.2. We use the notation

$$\Delta_k := f(x_k) - f(x^\star), \qquad q := 1 - \alpha, \qquad \gamma := 1 - q^2 = \alpha(2 - \alpha).$$

Define

$$A := \frac{\gamma}{16q^2}, \qquad \chi_\alpha := \min\left\{\frac{3}{8}, \frac{\gamma^3}{16q^2}\right\} > 0,$$

and the Lyapunov function

$$\Phi_k := \Delta_k + \frac{A}{L}e_k^2.$$

Here χ_α is a Lyapunov decrease constant and should not be confused with a strong-convexity parameter.

Q.1 Basic descent and error estimates

Lemma Q.1 (Certified descent)**.** *For every $k \ge 0$, Algorithm 8 satisfies*

$$f(x_{k+1}) \le f(x_k) - \frac{a_k^2}{2L}.$$

Consequently, the objective values are nonincreasing.

Proof. If $a_k = 0$, then $\eta_k = 0$ and $x_{k+1} = x_k$. Suppose $a_k > 0$. Then $a_k = \|m_{k+1}\|_* - e_k$, and the choice of u_k gives

$$\begin{aligned}\langle g_k, u_k\rangle &= \langle m_{k+1}, u_k\rangle + \langle g_k - m_{k+1}, u_k\rangle \\ &\ge \|m_{k+1}\|_* - e_k = a_k.\end{aligned}$$

By L-smoothness and $\|u_k\| \le 1$,

$$\begin{aligned}f(x_{k+1}) &\le f(x_k) - \eta_k\langle g_k, u_k\rangle + \frac{L}{2}\eta_k^2\|u_k\|^2 \\ &\le f(x_k) - \eta_k a_k + \frac{L}{2}\eta_k^2.\end{aligned}$$

Substituting $\eta_k = a_k/L$ proves the claim. **Q.E.D.**

Lemma Q.2 (Momentum-error recursion)**.** *Let $q := 1 - \alpha$. Then*

$$e_{k+1} \le q(e_k + a_k) \qquad \forall k \ge 0.$$

Proof. Since $m_{k+2} = qm_{k+1} + \alpha g_{k+1}$, we have

$$g_{k+1} - m_{k+2} = q(g_{k+1} - m_{k+1}).$$

Therefore, using smoothness and $\|x_{k+1} - x_k\| \le \eta_k = a_k/L$,

$$\begin{aligned} e_{k+1} &= \|g_{k+1} - m_{k+2}\|_* \\ &\le q\|g_{k+1} - g_k\|_* + q\|g_k - m_{k+1}\|_* \\ &\le qL\|x_{k+1} - x_k\| + qe_k \le q(a_k + e_k). \end{aligned}$$

Q.E.D.

Lemma Q.3 (Gap-to-certificate relation)**.** *Let $\Delta_k := f(x_k) - f(x^\star)$. Under Assumptions 2.3 and O.1, the iterates of Algorithm 8 satisfy*

$$\Delta_k \le D_{\mathrm{lev}}(a_k + 2e_k) \qquad \forall k \ge 0.$$

Proof. By Lemma Q.1, $f(x_k) \le f(x_0)$ for all k. Hence Assumption O.1 gives

$$\|x_k - x^\star\| \le D_{\mathrm{lev}}.$$

Star-convexity implies

$$\Delta_k \le \langle g_k, x_k - x^\star \rangle \le D_{\mathrm{lev}}\|g_k\|_*.$$

If $a_k > 0$, then $a_k = \|m_{k+1}\|_* - e_k$, hence

$$\|m_{k+1}\|_* = a_k + e_k$$

and

$$\|g_k\|_* \le \|m_{k+1}\|_* + \|g_k - m_{k+1}\|_* = a_k + 2e_k.$$

If $a_k = 0$, then $\|m_{k+1}\|_* \le e_k$, so

$$\|g_k\|_* \le 2e_k = a_k + 2e_k.$$

Combining these inequalities proves the result. **Q.E.D.**

Q.2 Lyapunov analysis

The objective gap alone may not decrease rapidly when the momentum is stale. The Lyapunov function Φ_k combines the objective gap with the momentum-tracking error, allowing the descent generated by the certificate a_k to amortize the stale momentum error.

Lemma Q.4 (Lyapunov descent)**.** *For all $k \ge 0$,*

$$\Phi_{k+1} \le \Phi_k - \frac{\chi_\alpha}{L}(a_k^2 + e_k^2).$$

Proof. By Lemmas Q.1 and Q.2,

$$\begin{aligned} \Phi_k - \Phi_{k+1} &\ge \frac{a_k^2}{2L} - \frac{A}{L}\left(q^2(e_k + a_k)^2 - e_k^2\right) \\ &= \frac{1}{L}\left[\left(\frac{1}{2} - Aq^2\right)a_k^2 - 2Aq^2 a_k e_k + A(1 - q^2)e_k^2\right]. \end{aligned}$$

Using $A = \gamma/(16q^2)$ and $\gamma = 1 - q^2$, the bracket equals

$$\left(\frac{1}{2} - \frac{\gamma}{16}\right)a_k^2 - \frac{\gamma}{8}a_k e_k + \frac{\gamma^2}{16q^2}e_k^2.$$

Since $\gamma \le 1$, the first coefficient is at least $7/16$. By Young's inequality,

$$\frac{\gamma}{8}a_k e_k \le \frac{1}{16}a_k^2 + \frac{\gamma^2}{16}e_k^2.$$

Thus the bracket is at least

$$\frac{3}{8}a_k^2 + \frac{\gamma^3}{16q^2}e_k^2 \ge \chi_\alpha(a_k^2 + e_k^2),$$

which proves the lemma. **Q.E.D.**

Lemma Q.5 (Potential control)**.** *For all $k \geq 0$,*

$$\Phi_k \leq C_{\mathrm{sc},\alpha} D_{\mathrm{lev}} \sqrt{a_k^2 + e_k^2}, \qquad C_{\mathrm{sc},\alpha} := \sqrt{5} + 2A.$$

Proof. Lemma Q.3 gives

$$\Delta_k \leq D_{\mathrm{lev}}(a_k + 2e_k) \leq \sqrt{5}\, D_{\mathrm{lev}} \sqrt{a_k^2 + e_k^2}.$$

It remains to bound Ae_k^2/L. Since $f(x_k) \leq f(x_0)$, Assumption O.1 gives

$$\|x_k - x^\star\| \leq D_{\mathrm{lev}}.$$

Because we are in the unconstrained differentiable setting,

$$\nabla f(x^\star) = 0.$$

Smoothness gives

$$\|g_k\|_* \leq L D_{\mathrm{lev}}.$$

Moreover, m_{k+1} is a convex combination of past gradients and all past iterates remain in the same initial sublevel set, so

$$\|m_{k+1}\|_* \leq L D_{\mathrm{lev}}.$$

Therefore

$$e_k \leq 2L D_{\mathrm{lev}},$$

and hence

$$\frac{e_k^2}{L} \leq 2 D_{\mathrm{lev}} e_k \leq 2 D_{\mathrm{lev}} \sqrt{a_k^2 + e_k^2}.$$

Combining the two estimates yields the claim. **Q.E.D.**

Q.3 Refined Lyapunov guarantee

Theorem Q.6 (Refined Lyapunov guarantee for Scale-Calibrated Muon)**.** *Let Assumptions 2.1, 2.3, and O.1 hold, and consider Algorithm 8 with $\alpha \in (0, 1)$. Then*

$$\Delta_T \leq \Phi_T \qquad \textit{for all } T \geq 0.$$

Moreover, if $\Phi_0 = 0$, then $\Phi_T = 0$ for all $T \geq 0$. Otherwise, for all $T \geq 0$,

$$\Phi_T \leq \frac{1}{\Phi_0^{-1} + \dfrac{\chi_\alpha T}{C_{\mathrm{sc},\alpha}^2 L D_{\mathrm{lev}}^2}}, \qquad \Phi_0 = \Delta_0.$$

In particular, for all $T \geq 1$,

$$f(x_T) - f(x^\star) \leq \frac{C_{\mathrm{sc},\alpha}^2}{\chi_\alpha} \frac{L D_{\mathrm{lev}}^2}{T}.$$

Proof. By Lemma Q.5,

$$a_k^2 + e_k^2 \geq \frac{\Phi_k^2}{C_{\mathrm{sc},\alpha}^2 D_{\mathrm{lev}}^2}.$$

Combining this with Lemma Q.4,

$$\Phi_{k+1} \leq \Phi_k - \frac{\chi_\alpha}{C_{\mathrm{sc},\alpha}^2 L D_{\mathrm{lev}}^2} \Phi_k^2.$$

Let

$$\theta_\alpha := \frac{\chi_\alpha}{C_{\mathrm{sc},\alpha}^2 L D_{\mathrm{lev}}^2}.$$

If $\Phi_k = 0$, then the preceding recursion and nonnegativity of Φ_{k+1} imply $\Phi_{k+1} = 0$. Thus, if $\Phi_0 = 0$, then $\Phi_T = 0$ for all $T \geq 0$.

Now suppose $\Phi_0 > 0$. As long as $\Phi_k > 0$, the recursion gives

$$\Phi_{k+1} \leq \Phi_k (1 - \theta_\alpha \Phi_k).$$

Since $\Phi_{k+1} \geq 0$, we have $\theta_\alpha \Phi_k \leq 1$, and therefore

$$\frac{1}{\Phi_{k+1}} \geq \frac{1}{\Phi_k(1-\theta_\alpha \Phi_k)} \geq \frac{1}{\Phi_k} + \theta_\alpha.$$

Summing from $k = 0$ to $T-1$ yields

$$\frac{1}{\Phi_T} \geq \frac{1}{\Phi_0} + \theta_\alpha T.$$

Equivalently,

$$\Phi_T \leq \frac{1}{\Phi_0^{-1} + \dfrac{\chi_\alpha T}{C_{\mathrm{sc},\alpha}^2 L D_{\mathrm{lev}}^2}}.$$

Finally, since $m_0 = g_0$, we have

$$m_1 = (1-\alpha)m_0 + \alpha g_0 = g_0,$$

so $e_0 = 0$ and hence $\Phi_0 = \Delta_0$. Also,

$$\Delta_T \leq \Phi_T$$

by definition of Φ_T. For $T \geq 1$, the displayed bound implies

$$f(x_T) - f(x^\star) = \Delta_T \leq \frac{C_{\mathrm{sc},\alpha}^2}{\chi_\alpha} \frac{L D_{\mathrm{lev}}^2}{T}.$$

This proves the theorem. **Q.E.D.**

Proof of Theorem O.2. Theorem Q.6 gives, for all $T \geq 1$,

$$f(x_T) - f(x^\star) \leq \frac{C_{\mathrm{sc},\alpha}^2}{\chi_\alpha} \frac{L D_{\mathrm{lev}}^2}{T}.$$

Thus Theorem O.2 holds with

$$K_\alpha := \frac{C_{\mathrm{sc},\alpha}^2}{\chi_\alpha}.$$

The distance parameter D_{lev} appears only in the analysis through Assumption O.1. It is not used by Algorithm 8. **Q.E.D.**

R Proofs for the Stochastic Local Scale-Calibrated Method

We first record the integrability needed for all conditional expectations. No moment above order two is used.

Lemma R.1 (Finite-horizon second moments)**.** *Fix $T < \infty$ and run Algorithm 9 under Assumptions 2.1 and 2.2. Then, for every algorithmic $k \leq T$, the second moments of $\|x_k - x_0\|$, $\|m_k\|_*$, and r_k are finite. The momentum-noise states used below also have finite Euclidean second moments. The same statements hold after adding the single analysis-only sample at x_T.*

Proof. At x_0, the oracle assumption and $\|z\|_* \leq \rho_{\mathrm{norm}} \|z\|_2$ give a finite dual-norm second moment for G_0, hence for $m_0 = \beta G_0$. Suppose the claim holds through round k. Assumption 2.1 gives

$$\|\nabla f(x_k)\|_* \leq \|\nabla f(x_0)\|_* + L\|x_k - x_0\|,$$

and therefore $G_k = \nabla f(x_k) + \zeta_k$ has a finite dual-norm second moment. The momentum recursion gives the same for m_k. Since $a_k \leq \|m_k\|_*$,

$$\|x_{k+1} - x_k\| \leq \frac{\alpha_{k+1}}{L} \|m_k\|_*,$$

so x_{k+1} has a finite second moment. Finally,

$$r_{k+1} \leq (1-\alpha_{k+1})(r_k + \alpha_{k+1} a_k)$$

propagates the scalar second moment. The noise recursion is a deterministic linear filter of innovations with finite Euclidean second moment, and hence its Euclidean second moment is finite. Induction proves the claim. The same argument applies to the independent virtual sample at x_T. Finally, smoothness bounds $|f(x_k) - f(x_0)|$ by a constant times $\|x_k - x_0\| + \|x_k - x_0\|^2$, so every potential expectation used below is finite. **Q.E.D.**

Let $\mathcal{F}_k$ denote the sigma-field generated by $G_0, \dots, G_k$ and any trajectory randomization through round k. Then x_{k+1}, m_k, and r_k are $\mathcal{F}_k$-measurable, and the next oracle innovation has zero conditional mean and conditional Euclidean second moment at most σ^2.

Throughout the appendix, write

$$h_k := \nabla f(x_k).$$

Define the noiseless momentum driven by the same deterministic coefficients and initialization coefficient:

$$\bar{m}_0 := \beta h_0, \qquad \bar{m}_k := (1-\alpha_k)\bar{m}_{k-1} + \alpha_k h_k, \quad k \geq 1.$$

Separate its lag into the initialization transient

$$d_0 := -(1-\beta)h_0, \qquad d_{k+1} := (1-\alpha_{k+1})d_k, \tag{186}$$

and the dynamic lag

$$p_0 := 0, \qquad p_{k+1} := (1-\alpha_{k+1})(p_k + h_k - h_{k+1}). \tag{187}$$

With $n_k := m_k - \bar{m}_k$, induction gives

$$\bar{m}_k = h_k + p_k + d_k, \qquad m_k = h_k + p_k + d_k + n_k. \tag{188}$$

Set

$$D_k := \|d_k\|_*^2, \qquad Z_k := \rho_{\mathrm{norm}}^2 \|n_k\|_2^2, \qquad W_k := D_k + Z_k, \qquad H_k := \|d_k + n_k\|_*. \tag{189}$$

The norm comparison and $(a+b)^2 \leq 2a^2 + 2b^2$ imply

$$H_k^2 \leq 2W_k \qquad \text{pathwise.} \tag{190}$$

R.1 Initialization and lag certificate

Lemma R.2 (Self-bounding at initialization)**.** *Under Assumption 2.1 and $f^\star > -\infty$,*

$$\|h_0\|_*^2 \leq 2L\Delta_0.$$

Proof. Let $u_0 = u(h_0)$ and set

$$y = x_0 - \frac{\|h_0\|_*}{L} u_0.$$

Smoothness and the LMO identity give

$$\begin{aligned} f(y) &\leq f(x_0) - \frac{\|h_0\|_*}{L}\langle h_0, u_0\rangle + \frac{\|h_0\|_*^2}{2L}\|u_0\|^2 \\ &\leq f(x_0) - \frac{\|h_0\|_*^2}{2L}. \end{aligned}$$

Since $f^\star \leq f(y)$, the claim follows. **Q.E.D.**

Lemma R.3 (Validity of the lag-adjusted certificate)**.** *Every iterate of Algorithm 9 satisfies*

$$\|p_k\|_* \leq r_k.$$

Proof. At $k = 0$, $p_0 = 0 = r_0$. If $\|p_k\|_* \leq r_k$, then (187), Assumption 2.1, and the update rule give

$$\begin{aligned} \|p_{k+1}\|_* &\leq (1-\alpha_{k+1})\left(\|p_k\|_* + \|h_{k+1} - h_k\|_*\right) \\ &\leq (1-\alpha_{k+1})\left(r_k + L\|x_{k+1} - x_k\|\right) = r_{k+1}. \end{aligned}$$

The result follows by induction. **Q.E.D.**

R.2 Augmented Lyapunov drift

Define

$$\Phi_k := f(x_k) - f^\star + \frac{r_k^2}{4L}, \qquad \Psi_k := \Phi_k + \frac{3W_k}{L}, \qquad q_k := a_k^2 + r_k^2 + 12W_k. \tag{191}$$

All three quantities are nonnegative because $f^\star$ is a lower bound.

Lemma R.4 (Objective and base-potential decrease)**.** *For every $k \geq 0$, pathwise,*

$$f(x_{k+1}) \leq f(x_k) - \frac{\alpha_{k+1}}{4L} a_k^2 + \frac{\alpha_{k+1}}{L} H_k^2, \tag{192}$$

$$\Phi_{k+1} \leq \Phi_k - \frac{\alpha_{k+1}}{12L}(a_k^2 + r_k^2) + \frac{\alpha_{k+1}}{L} H_k^2. \tag{193}$$

Proof. Put $\gamma := \alpha_{k+1}$. If $a_k = 0$, then $x_{k+1} = x_k$ and (192) is immediate. Suppose $a_k > 0$. Lemma R.3 and the LMO identity imply

$$\begin{aligned} \langle h_k, u_k \rangle &= \langle m_k, u_k \rangle - \langle p_k + d_k + n_k, u_k \rangle \\ &\geq \|m_k\|_* - r_k - H_k = a_k - H_k. \end{aligned}$$

By smoothness and $x_{k+1} - x_k = -(\gamma a_k / L) u_k$,

$$\begin{aligned} f(x_{k+1}) - f(x_k) &\leq -\frac{\gamma a_k}{L}(a_k - H_k) + \frac{\gamma^2 a_k^2}{2L} \\ &\leq -\frac{\gamma}{2L} a_k^2 + \frac{\gamma}{L} a_k H_k \\ &\leq -\frac{\gamma}{4L} a_k^2 + \frac{\gamma}{L} H_k^2, \end{aligned}$$

where $\gamma \leq 1$ and $aH \leq a^2/4 + H^2$ were used. This proves (192).

Let $q := 1 - \gamma$. The certificate recursion, $\|u_k\| \leq 1$, and the step definition give

$$r_{k+1} \leq q(r_k + \gamma a_k).$$

Combining this inequality with (192),

$$\begin{aligned} 4L\left(\Phi_k - \Phi_{k+1} + \frac{\gamma}{L} H_k^2\right) &\geq \gamma a_k^2 + r_k^2 - q^2(r_k + \gamma a_k)^2 \\ &= \gamma \begin{bmatrix} a_k & r_k \end{bmatrix} M_\gamma \begin{bmatrix} a_k \\ r_k \end{bmatrix}, \end{aligned}$$

where

$$M_\gamma = \begin{bmatrix} 1 - q^2\gamma & -q^2 \\ -q^2 & (1 - q^2)/\gamma \end{bmatrix}.$$

We claim that $M_\gamma \succeq I/3$. Both diagonal entries of $M_\gamma - I/3$ are positive. Moreover,

$$\det(M_\gamma - I/3) = \frac{2(1 - q^2)}{3\gamma} - \frac{2}{9} - q^2 + \frac{q^2\gamma}{3}.$$

Since $\gamma = 1 - q$, the right-hand side is at least

$$\frac{4}{9} + \frac{2q}{3} - q^2 \geq \frac{1}{9}, \qquad q \in [0, 1).$$

Therefore

$$4L\left(\Phi_k - \Phi_{k+1} + \frac{\gamma}{L} H_k^2\right) \geq \frac{\gamma}{3}(a_k^2 + r_k^2),$$

which is equivalent to (193). **Q.E.D.**

For the final use of the next-noise recursion only, after x_T has been formed we may enlarge the probability space by a virtual sample ζ_T with the oracle law at x_T and define

$$G_T := h_T + \zeta_T, \qquad m_T := (1 - \alpha_T) m_{T-1} + \alpha_T G_T.$$

This ghost sample does not alter the trajectory, the output, or the count of T actual oracle evaluations.

Lemma R.5 (Unconditional augmented drift)**.** *Under Assumptions 2.1 and 2.2, Algorithm 9 satisfies, for $k = 0, \dots, T-1$,*

$$\mathbb{E}\Psi_{k+1} \le \mathbb{E}\Psi_k - \frac{\alpha_{k+1}}{12L}\mathbb{E}q_k + 3\frac{\rho_{\mathrm{norm}}^2\sigma^2}{L}\alpha_{k+1}^2. \tag{194}$$

Proof. The momentum-noise recursion is

$$n_{k+1} = (1-\alpha_{k+1})n_k + \alpha_{k+1}\zeta_{k+1}.$$

The coefficient α_{k+1} is deterministic and n_k is measurable before the next sample. Conditional unbiasedness eliminates the Euclidean cross term, and Assumption 2.2 gives

$$\mathbb{E}[Z_{k+1} \mid \mathcal{F}_k] \le (1-\alpha_{k+1})^2 Z_k + \rho_{\mathrm{norm}}^2\alpha_{k+1}^2\sigma^2,$$

where $\mathcal{F}_k$ is the history through the sample used at round k. Also, $D_{k+1} = (1-\alpha_{k+1})^2 D_k$. Hence

$$\mathbb{E}[W_{k+1} \mid \mathcal{F}_k] \le (1-\alpha_{k+1})^2 W_k + \rho_{\mathrm{norm}}^2\alpha_{k+1}^2\sigma^2. \tag{195}$$

Taking total expectation, adding three times (195) divided by L to (193), and using $H_k^2 \le 2W_k$, the coefficient of W_k/L relative to Ψ_k is

$$2\alpha_{k+1} + 3\left((1-\alpha_{k+1})^2 - 1\right) = -4\alpha_{k+1} + 3\alpha_{k+1}^2 \le -\alpha_{k+1}.$$

Thus the decrease in the W_k term is at least $-\alpha_{k+1}W_k/L = -\alpha_{k+1}(12W_k)/(12L)$, proving (194). **Q.E.D.**

Lemma R.6 (Certificate energy controls stationarity)**.** *For every $k \ge 0$, pathwise,*

$$\|\nabla f(x_k)\|_*^2 \le \frac{31}{6}q_k. \tag{196}$$

Proof. The definition of a_k implies $\|m_k\|_* \le a_k + r_k$. Therefore, by (188), Lemma R.3, and (190),

$$\begin{aligned}\|h_k\|_* &\le \|m_k\|_* + \|p_k\|_* + \|d_k + n_k\|_* \\ &\le a_k + 2r_k + \sqrt{2W_k}.\end{aligned}$$

Weighted Cauchy–Schwarz gives

$$\left(a_k + 2r_k + \sqrt{2W_k}\right)^2 \le \left(1 + 4 + \frac{2}{12}\right)\left(a_k^2 + r_k^2 + 12W_k\right) = \frac{31}{6}q_k.$$

Q.E.D.

R.3 Proof of Theorem P.1

Summing (194) over $k = 0, \dots, T-1$ and using $\Psi_T \ge 0$ yields

$$\sum_{k=0}^{T-1}\alpha_{k+1}\mathbb{E}q_k \le 12L\mathbb{E}\Psi_0 + 36\rho_{\mathrm{norm}}^2\sigma^2 V_T. \tag{197}$$

At initialization, $r_0 = 0$, $d_0 = -(1-\beta)h_0$, and $n_0 = \beta(G_0 - h_0)$. Lemma R.2 therefore gives

$$\begin{aligned}\mathbb{E}\Psi_0 &\le \Delta_0 + \frac{3}{L}\left((1-\beta)^2\|h_0\|_*^2 + \rho_{\mathrm{norm}}^2\beta^2\sigma^2\right) \\ &\le \left(1 + 6(1-\beta)^2\right)\Delta_0 + 3\frac{\rho_{\mathrm{norm}}^2\beta^2\sigma^2}{L}.\end{aligned}$$

By Lemma R.6 and (197),

$$\begin{aligned}\frac{1}{S_T}\sum_{k=0}^{T-1}\alpha_{k+1}\mathbb{E}\|\nabla f(x_k)\|_*^2 &\le \frac{31}{6S_T}\sum_{k=0}^{T-1}\alpha_{k+1}\mathbb{E}q_k \\ &\le \mathfrak{B}_T(\beta).\end{aligned}$$

The left-hand side is exactly $\mathbb{E}\|\nabla f(\widehat{x}_T)\|_*^2$, proving (171). It also upper-bounds the minimum in (172). Pointwise,

$$\min_{0\le k<T} \|\nabla f(x_k)\|_*^2 \le \frac{1}{S_T}\sum_{k=0}^{T-1} \alpha_{k+1}\|\nabla f(x_k)\|_*^2.$$

Taking expectation and applying Jensen's inequality proves (173).

Finally, for $\beta = \alpha_1$,

$$1 + 6(1-\alpha_1)^2 \le 7, \qquad \alpha_1^2 \le V_T.$$

Substitution in (170) proves (174). □

R.4 Proofs of the parameter-choice corollaries

Proof of Corollary P.2. The constant schedule gives $S_T = \alpha T$ and $V_T = \alpha^2 T$. With $\beta = \alpha$, use

$$1 + 6(1-\alpha)^2 \le 7, \qquad \frac{\alpha^2}{\alpha T} = \frac{\alpha}{T} \le \alpha$$

in (170) to obtain (175). The noiseless statement follows by setting $\alpha = \beta = 1$. □

Proof of Corollary P.3. For the horizon-dependent choice,

$$S_T = \sqrt{T}, \qquad V_T = 1, \qquad \beta = 1.$$

Theorem P.1 gives (176). Jensen's inequality and solving for T give (177).

For the anytime choice,

$$S_T = \sum_{j=1}^{T} j^{-1/2} \ge \sqrt{T}, \qquad V_T = \sum_{j=1}^{T} j^{-1} \le 1 + \log T.$$

Because $\alpha_1 = 1$, $\beta = 1$ is also the online initialization. Applying Theorem P.1 gives (178). □

Proof of Corollary P.4. The parameter choice gives

$$C_1\alpha_\varepsilon \le \frac{\varepsilon^2}{2},$$

whereas the lower bound on T gives

$$\frac{C_0}{\alpha_\varepsilon T} \le \frac{\varepsilon^2}{2}.$$

Thus (175) and Jensen's inequality prove the expected-norm statement. Finally,

$$\frac{1}{\alpha_\varepsilon} \le 1 + \frac{2C_1}{\varepsilon^2},$$

which yields (179). □

R.5 Last-Iterate Bound under Star-Convexity

This subsection proves Theorem P.5. It uses the same momentum decomposition, lag-adjusted certificate, and augmented drift as the unconditional stationarity theorem. Specialize the deterministic coefficients in Appendix R to $\alpha_k \equiv \alpha$ and consider the trajectory of Algorithm 10. For the analysis-only sample at x_T introduced before Lemma R.5, set

$$a_T := (\|m_T\|_* - r_T)_+, \qquad q_T := a_T^2 + r_T^2 + 12W_T.$$

This sample does not change the returned point or the count of T actual oracle calls.

If $D_{\mathrm{ms},T} = 0$, then $x_k = x^\star$ almost surely for every $k \le T$, and the theorem is immediate. Henceforth assume $D_{\mathrm{ms},T} > 0$. Write

$$\Delta_k := f(x_k) - f(x^\star), \qquad Y_k := \mathbb{E}\Psi_k, \qquad P_k := \mathbb{E}q_k, \qquad A := \sqrt{\frac{31}{6}}.$$

Under star-convexity, $f^\star = f(x^\star)$, so the potential Ψ_k defined in (191) contains exactly the gap Δ_k.

Lemma R.7 (Expected potential-to-certificate relation)**.** *For every $k = 0, \dots, T$,*

$$Y_k \le AD_{\mathrm{ms},T}\sqrt{P_k} + \frac{P_k}{4L}. \tag{198}$$

Consequently, whenever $Y_k > 0$,

$$P_k \ge \frac{Y_k^2}{A^2 D_{\mathrm{ms},T}^2 + Y_k/(2L)} \ge \frac{1}{2}\min\left\{\frac{Y_k^2}{A^2 D_{\mathrm{ms},T}^2}, 2LY_k\right\}. \tag{199}$$

Proof. Assumption 2.3, Cauchy–Schwarz, (182), and Lemma R.6 give

$$\begin{aligned}\mathbb{E}\Delta_k &\le \mathbb{E}\left[\|\nabla f(x_k)\|_* \, \|x_k - x^\star\|\right] \\ &\le \sqrt{\mathbb{E}\,\|\nabla f(x_k)\|_*^2}\sqrt{\mathbb{E}\|x_k - x^\star\|^2} \le AD_{\mathrm{ms},T}\sqrt{P_k}.\end{aligned}$$

Moreover, by the definition of q_k,

$$\frac{r_k^2}{4L} + \frac{3W_k}{L} \le \frac{q_k}{4L}.$$

Taking expectations proves (198).

For $Y_k > 0$, Young's inequality gives

$$AD_{\mathrm{ms},T}\sqrt{P_k} \le \frac{Y_k}{2} + \frac{A^2 D_{\mathrm{ms},T}^2 P_k}{2Y_k}.$$

Substitution in (198) and rearrangement give the first inequality in (199). The second follows from $u + v \le 2\max\{u, v\}$. **Q.E.D.**

Taking expectations in Lemma R.5, using the constant coefficient, and applying Lemma R.7 yield

$$Y_{k+1} \le Y_k - \min\left\{\frac{\alpha Y_k^2}{124LD_{\mathrm{ms},T}^2}, \frac{\alpha Y_k}{12}\right\} + \frac{3\alpha^2 s_\sigma^2}{L}. \tag{200}$$

When $Y_k = 0$, the same inequality follows directly from Lemma R.5 and $P_k \ge 0$.

Lemma R.8 (Noisy Riccati recursion)**.** *Suppose a nonnegative sequence satisfies*

$$z_{k+1} \le z_k - \min\{az_k^2, bz_k\} + c, \qquad a > 0, \quad 0 < b \le 1, \quad c \ge 0.$$

For every $T \ge 1$,

$$z_T \le 8\left[z_0 e^{-bT/4} + \frac{1}{a(T+1)} + \sqrt{\frac{c}{a}} + \frac{c}{b}\right]. \tag{201}$$

Proof. Let

$$\delta := 2\max\left\{\sqrt{\frac{c}{a}}, \frac{c}{b}\right\}.$$

If $z_k > \delta$, then $az_k^2 \ge 4c$ and $bz_k \ge 2c$, whence

$$z_{k+1} \le z_k - \frac{1}{2}\min\{az_k^2, bz_k\}. \tag{202}$$

Define

$$\tau := \inf\{k \ge 0 : z_k \le \delta\}, \qquad \inf \varnothing := +\infty.$$

For every $k < \tau$, (202) holds and the sequence is nonincreasing. Above the crossover b/a,

$$z_{k+1} \le (1 - b/2)z_k \le e^{-b/2} z_k.$$

Below the crossover, $az_k \le b \le 1$. Since $k < \tau$ implies $z_k > 0$, either $z_{k+1} = 0$ and the sequence has entered $[0, \delta]$, or

$$\frac{1}{z_{k+1}} \ge \frac{1}{z_k(1 - az_k/2)} \ge \frac{1}{z_k} + \frac{a}{2}.$$

Fix an integer $N \geq 1$ with $N < \tau$ and define

$$j := \inf\{0 \leq k \leq N : z_k \leq b/a\}, \qquad \inf \varnothing := +\infty.$$

If $j \leq N/2$, monotonicity keeps the sequence below b/a after time j, and summing the reciprocal estimate gives

$$z_N \leq \frac{2}{a(N-j)} \leq \frac{4}{aN}.$$

If $j > N/2$, the geometric regime lasts through time N, or it ends at $j \leq N$ and the sequence is nonincreasing thereafter. In both cases,

$$z_N \leq z_0 e^{-bN/4}.$$

Consequently, for every $N < \tau$,

$$z_N \leq z_0 e^{-bN/4} + \frac{4}{aN}. \tag{203}$$

It remains to control the sequence after entrance. Put $B := \delta + c$. If $z_k \leq \delta$, then $z_{k+1} \leq z_k + c \leq B$. If $\delta < z_k \leq B$, then $\min\{az_k^2, bz_k\} \geq 2c$, so $z_{k+1} \leq z_k - c \leq B$. Hence $[0, B]$ is forward invariant. Since $b \leq 1$ implies $c \leq c/b \leq \delta/2$,

$$B \leq \frac{3}{2}\delta \leq 3\left(\sqrt{\frac{c}{a}} + \frac{c}{b}\right).$$

If $\tau \leq T$, forward invariance gives $z_T \leq B$. If $\tau > T$, apply (203) with $N = T$. Thus

$$z_T \leq z_0 e^{-bT/4} + \frac{4}{aT} + 3\left(\sqrt{\frac{c}{a}} + \frac{c}{b}\right).$$

Finally, $4/(aT) \leq 8/(a(T+1))$ for $T \geq 1$, and enlarging every coefficient to 8 proves (201). **Q.E.D.**

R.5.1 Completion of Theorem P.5

In (200), take

$$a = \frac{\alpha}{124 L D_{\mathrm{ms},T}^2}, \qquad b = \frac{\alpha}{12}, \qquad c = \frac{3\alpha^2 s_\sigma^2}{L}.$$

Lemma R.8 gives

$$Y_T \leq C\left[Y_0 e^{-\alpha T/48} + \frac{L D_{\mathrm{ms},T}^2}{\alpha(T+1)} + D_{\mathrm{ms},T} s_\sigma \sqrt{\alpha} + \frac{\alpha s_\sigma^2}{L}\right] \tag{204}$$

after absorbing numerical factors into a universal C.

It remains to control the initialization. Since $\nabla f(x^\star) = 0$, smoothness and $\|x_0 - x^\star\| \leq D_{\mathrm{ms},T}$ imply

$$\|\nabla f(x_0)\|_* \leq L D_{\mathrm{ms},T}.$$

At $k = 0$, $r_0 = 0$, $d_0 = -(1-\beta)\nabla f(x_0)$, and $n_0 = \beta(G_0 - \nabla f(x_0))$. Therefore

$$\begin{aligned} Y_0 &\leq f(x_0) - f(x^\star) + 3(1-\beta)^2 L D_{\mathrm{ms},T}^2 + \frac{3\beta^2 s_\sigma^2}{L} \\ &= \mathcal{J}_{0,T}(\beta). \end{aligned}$$

Finally,

$$\mathbb{E}[f(x_T) - f(x^\star)] = \mathbb{E}\Delta_T \leq Y_T.$$

Substitution in (204) proves Theorem P.5.

S Additional Experiments

This appendix retains the separate experimental studies of DF-Muon-Proxy, Distance-Adaptive Muon, and Scale-Calibrated Muon. The GPT-124M three-seed comparison uses 5,000 optimizer steps, and the regularized ViT-Tiny comparison uses 100 epochs. These are distinct from both studies with 10,000 stochastic-gradient evaluations in Section 5.3. Numerical results, sweeps, and diagnostics are preserved under their original protocols.

S.1 Implementation Details for DF-Muon-Proxy

The deterministic DF-Muon analysis in Appendix F uses the exact minimizer of a one-dimensional majorized model for a general norm. Our neural-network experiments instead use a practical stochastic proxy, which we call *DF-Muon-Proxy*. The Muon direction itself is unchanged: for matrix-valued parameters we use the standard orthogonalized Muon direction associated with the spectral–nuclear geometry. The radius, however, is computed using product-Frobenius aggregate statistics over the Muon-updated matrix blocks, rather than by solving the exact one-dimensional problem in the spectral geometry.

Concretely, let $u_{t,j}$ denote the Muon direction for matrix block j, let $g_{t,j}$ be the corresponding stochastic gradient, and let $y_{t,j} = x_{t,j} - x_{0,j}$. At each optimizer step we aggregate

$$A_t = \sum_j \|u_{t,j}\|_F^2, \qquad B_t = \sum_j \langle y_{t,j}, u_{t,j}\rangle, \qquad C_t = \sum_j \|y_{t,j}\|_F^2, \qquad G_t = \sum_j \langle g_{t,j}, u_{t,j}\rangle.$$

These quantities define a cheap one-dimensional Frobenius-proxy majorized score for the radius. We minimize this score over the capped interval $[\eta_{\min}, \eta_{\max}]$ by a small grid search followed by local refinement, smooth the selected scale, and then multiply it by the common warmup–cosine schedule before applying the update. This procedure uses no additional forward or backward passes. It only reuses quantities already available during the optimizer step.

Thus, DF-Muon-Proxy is not the algorithm covered by Theorem F.1. It is also different from the predictable FTRL method in Algorithm 5: it uses a capped current-step proxy, whereas the predictable method fixes its radius from past feedback. The exact one-dimensional problem remains convex, but may be nonsmooth if written directly in the spectral norm. The Frobenius proxy avoids this cost while retaining the Muon orthogonalized direction.

Unless stated otherwise, the DF-Muon-Proxy experiments use a grid of 21 candidate scales and 6 local refinement steps. For GPT-124M/WikiText-103, we use $\eta_{\min} = 0.006$, initial scale 0.015, cap $\eta_{\max} = 0.03$, and smoothing coefficient 0.70. For the regularized ViT-Tiny/CIFAR-100 experiment, we use the same implementation with $\eta_{\max} = 0.01$, selected from the cap diagnostic in Appendix S.5.

S.2 Full Three-Seed Comparisons

Tables 16 and 17 compare DF-Muon-Proxy with the DA-Muon and Scale-Calibrated Muon rules described in the preceding supplementary sections. These tables are separate from the current FTRL study and should not be read as empirical tests of the predictable algorithms.

Table 16: GPT-124M/WikiText-103 results over three seeds. Runtime is relative to fixed Muon.

Method	Train loss	Val. loss	Base scale	Runtime
AdamW	3.9964 ± 0.0026	3.9353 ± 0.0143	0.001	0.65×
Fixed Muon	3.7662 ± 0.0029	3.7166 ± 0.0199	0.005	1.00×
DF-Muon-Proxy	$\mathbf{3.7192 \pm 0.0028}$	$\mathbf{3.6714 \pm 0.0237}$	adaptive	1.05×
DA-Muon	3.7794 ± 0.0017	3.7273 ± 0.0216	adaptive	1.03×
Scale-Calibrated Muon	3.8589 ± 0.0048	3.8029 ± 0.0180	adaptive	1.02×

Table 17: Regularized ViT-Tiny/CIFAR-100 results over three seeds. Runtime is relative to fixed Muon.

Method	Best val. CE	Top-1 at CE	Best top-1	Base scale	Runtime
AdamW	1.5561 ± 0.0196	60.04 ± 0.85	63.23 ± 0.45	0.0003	0.83×
Fixed Muon	$\mathbf{1.3851 \pm 0.0131}$	64.15 ± 1.28	67.69 ± 0.58	0.001	1.00×
DF-Muon-Proxy	1.3974 ± 0.0132	63.32 ± 0.67	$\mathbf{67.93 \pm 0.26}$	adaptive	1.05×
DA-Muon	1.5124 ± 0.0209	61.46 ± 0.55	63.38 ± 0.52	adaptive	1.02×
Scale-Calibrated Muon	1.4392 ± 0.0162	63.19 ± 0.51	65.28 ± 0.17	adaptive	1.02×

S.3 Additional GPT-124M/WikiText-103 Diagnostics

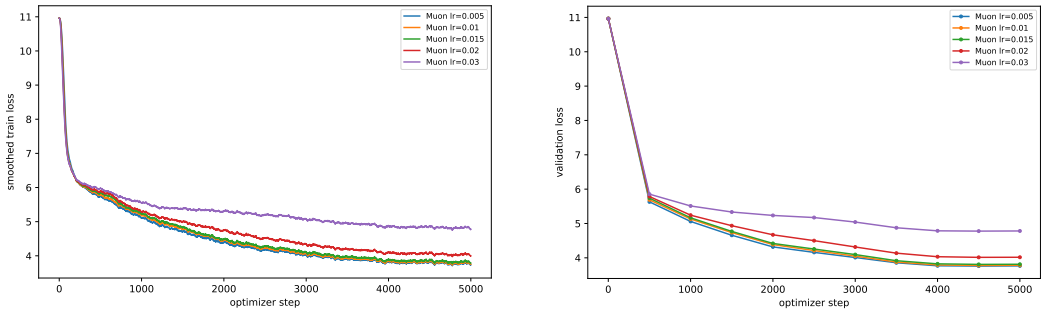
This section provides additional diagnostics for the earlier GPT-124M/WikiText-103 study summarized in Appendix S.2. We include the fixed-Muon learning-rate sweep used to select the tuned

baseline, the DA-Muon cap sweep used to select its cap, a one-seed DF-Muon-Proxy majorized-model diagnostic used to select the no-center variant, and the observed base scales selected by the adaptive variants.

Fixed-Muon learning-rate sweep. Table 18 reports the fixed-Muon learning-rate sweep on WikiText-103 for one representative seed. The sweep selects $\eta = 0.005$ as the best fixed-Muon learning rate. Larger fixed learning rates are substantially worse in this setting, indicating that the large Transformer benchmark is sensitive to the scalar scale.

Table 18: Fixed-Muon learning-rate sweep on the GPT-style Transformer/WikiText-103 benchmark for one representative seed. Runtime is reported relative to Muon with $\eta = 0.015$.

Method	Train loss	Val. loss	Mean η	Rel. time
Muon, $\eta = 0.005$	3.7558	**3.7650**	0.0050	1.01×
Muon, $\eta = 0.01$	3.7755	3.7834	0.0100	1.00×
Muon, $\eta = 0.015$	3.8081	3.8130	0.0150	1.00×
Muon, $\eta = 0.02$	4.0177	4.0166	0.0200	1.00×
Muon, $\eta = 0.03$	4.8027	4.7822	0.0300	1.00×



(a) Training loss

(b) Validation loss

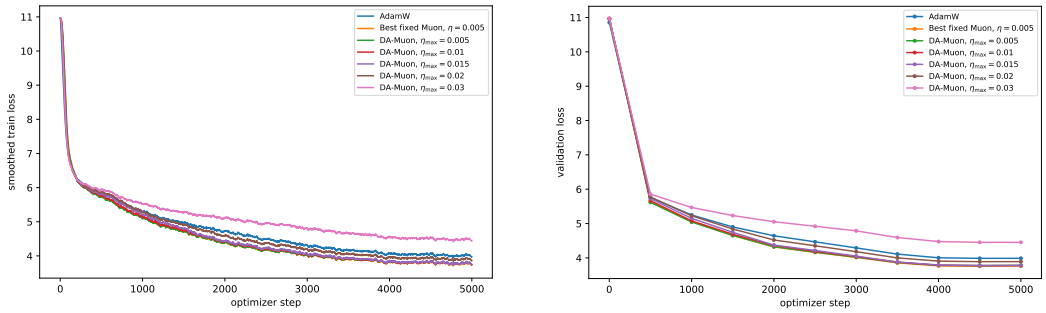
Figure 10: Fixed-Muon learning-rate sweep on GPT-124M/WikiText-103. The smaller learning rate $\eta = 0.005$ gives the best validation loss among the tested fixed scales.

DA-Muon cap sweep. The default DA-Muon cap $\eta_{\max} = 0.03$ is too aggressive on this benchmark. Table 19 reports a cap sweep. Tuning the cap makes DA-Muon competitive, with the best result at $\eta_{\max} = 0.01$, but it remains slightly worse than the tuned fixed-Muon baseline and DF-Muon-Proxy.

Table 19: DA-Muon cap sweep on GPT-124M/WikiText-103 for one representative seed. Runtime is reported relative to the tuned fixed-Muon baseline with $\eta = 0.005$.

Method	Train loss	Val. loss	Mean η	Rel. time
AdamW	3.9941	3.9874	0.0010	0.66×
Best fixed Muon, $\eta = 0.005$	3.7550	**3.7636**	0.0050	1.00×
DA-Muon, $\eta_{\max} = 0.005$	3.7728	3.7818	0.0050	1.03×
DA-Muon, $\eta_{\max} = 0.01$	3.7680	3.7779	0.0100	1.03×
DA-Muon, $\eta_{\max} = 0.015$	3.7724	3.7793	0.0150	1.03×
DA-Muon, $\eta_{\max} = 0.02$	3.8854	3.8908	0.0200	1.03×
DA-Muon, $\eta_{\max} = 0.03$	4.4655	4.4528	0.0300	1.02×

DF-Muon-Proxy majorized-model diagnostic. For DF-Muon-Proxy, we ran a one-seed diagnostic over five nearby variants of the practical one-dimensional majorized model. All variants use the same cap, floor, smoothing coefficient, grid size, and number of local refinement steps. The no-center variant, which sets the center-penalty coefficient to zero while keeping the step and proxy coefficients fixed, gives the best validation loss. We therefore use this variant in the final three-seed GPT-124M/WikiText-103 comparison.



(a) Training loss

(b) Validation loss

Figure 11: DA-Muon cap sweep on GPT-124M/WikiText-103. Smaller caps substantially improve DA-Muon over the default $\eta_{\max} = 0.03$ setting.

Table 20: One-seed DF-Muon-Proxy majorized-model diagnostic on GPT-124M/WikiText-103. All variants use $\eta_{\min} = 0.006$, initial scale 0.015, $\eta_{\max} = 0.03$, smoothing coefficient 0.70, 21 grid points, and 6 local refinement steps.

Variant	**Step coeff.**	**Center coeff.**	**Proxy coeff.**	**Val. loss**
Default	0.10	0.02	0.10	3.6947
Less conservative	0.05	0.02	0.10	3.6906
More conservative	0.15	0.02	0.10	3.6960
No-center	0.10	0.00	0.10	**3.6677**
Less proxy	0.10	0.02	0.05	3.6969

Observed adaptive scales. Because all methods use a warmup–cosine schedule, the effective learning rate approaches zero near the end of training. We therefore report the base scale η_t selected by each adaptive rule before multiplying by the common schedule. Table 21 reports the mean base scale over the final 20% of training, averaged over seeds.

Table 21: Observed base scales on GPT-124M/WikiText-103 over three seeds. For adaptive methods, the mean base scale is averaged over the final 20% of training and then averaged over seeds.

Method	**Selection rule**	**Mean base η**
AdamW	fixed	0.0010
Best fixed Muon	fixed sweep	0.0050
DF-Muon-Proxy	adaptive, no-center	0.0300
DA-Muon	adaptive cap sweep	0.0100
Scale-Calibrated Muon	adaptive	0.0263

Interpretation. The diagnostics show that the improvement of DF-Muon-Proxy is not due to comparing against an under-tuned fixed-Muon baseline. The fixed-Muon sweep selects $\eta = 0.005$, and the main three-seed comparison uses this tuned baseline. The DA-Muon sweep shows that the radius-based rule is sensitive to its cap in large Transformer training: the default $\eta_{\max} = 0.03$ is too aggressive, whereas $\eta_{\max} = 0.01$ is competitive. The DF-Muon-Proxy majorized-model diagnostic selects the no-center variant, which removes an overly conservative center penalty in this GPT-124M setting. With this variant, DF-Muon-Proxy remains the strongest method in the three-seed comparison, improving over the tuned fixed-Muon baseline with only a small runtime overhead.

S.4 NanoGPT/WikiText-2 Diagnostics

We include a smaller Transformer language-modeling experiment to complement the GPT-124M/WikiText-103 results in Appendix S.2. The goal is not to establish a strong language-modeling

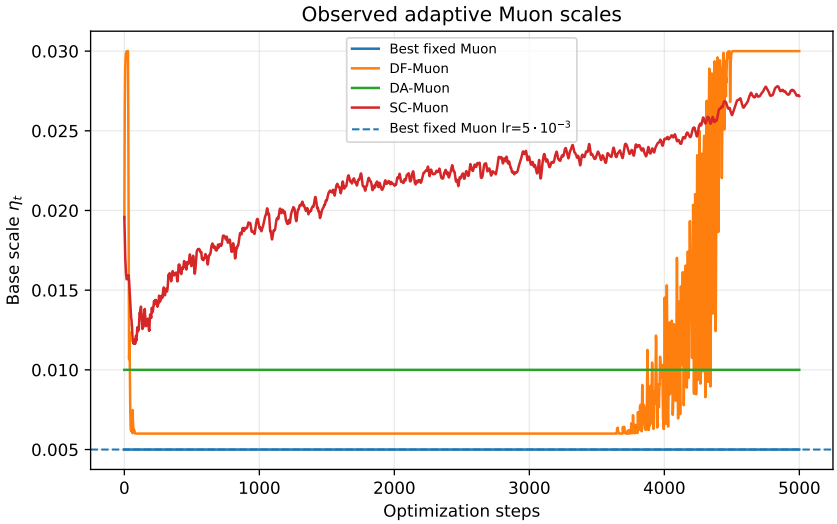


Figure 12: Observed base scales selected by the Muon variants on GPT-124M/WikiText-103 in a representative seed. The tuned fixed-Muon baseline uses $\eta = 0.005$, while DA-Muon uses the tuned cap $\eta_{\max} = 0.01$. DF-Muon-Proxy follows a data-dependent scale trajectory and achieves the best three-seed validation loss in Table 16.

benchmark, but to check whether the adaptive Muon scaling rules show the same qualitative behavior in a short-budget stochastic Transformer setting.

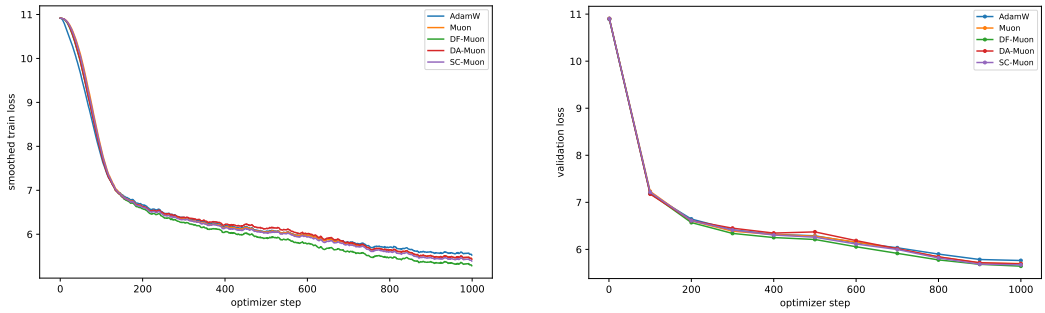
Experimental setup. We train a NanoGPT-style decoder-only Transformer on WikiText-2 using GPT-2 tokenization. The model has 6 layers, 6 attention heads, embedding dimension 384, context length 128, and approximately 30M parameters. All methods are run for $T = 1000$ optimizer steps. AdamW uses learning rate 10^{-3}. The fixed-Muon baseline uses $\eta = 0.015$, selected from a small fixed-Muon learning-rate grid. The adaptive Muon variants use the same maximum base scale $\eta_{\max} = 0.03$. As in the main experiment, Muon is applied to matrix-valued parameters, while biases, normalization parameters, embeddings, and other non-matrix parameters are optimized with AdamW.

Three-seed comparison. Table 22 reports the aggregate over three seeds. The updated majorized DF-Muon-Proxy implementation gives the best mean training and validation loss. DA-Muon and Scale-Calibrated Muon also improve over the tuned fixed-Muon baseline on validation loss, but the gains are smaller.

Table 22: NanoGPT/WikiText-2 results over three seeds for the updated majorized DF-Muon-Proxy implementation. We report mean $\pm$ standard deviation over seeds $\{42, 1337, 2024\}$. The fixed-Muon baseline uses $\eta = 0.015$. Adaptive Muon variants use $\eta_{\max} = 0.03$.

Method	Train loss	Val. loss	Mean base η
AdamW	5.5576 ± 0.0163	5.7926 ± 0.0339	0.0010
Best fixed Muon	5.4489 ± 0.0196	5.7162 ± 0.0256	0.0150
DF-Muon-Proxy	$\mathbf{5.3073 \pm 0.0156}$	$\mathbf{5.6630 \pm 0.0297}$	0.0202
DA-Muon	5.4445 ± 0.0205	5.7048 ± 0.0311	0.0300
Scale-Calibrated Muon	5.4122 ± 0.0176	5.6955 ± 0.0356	0.0202

Representative curves. Figure 13 shows representative training and validation curves. The validation curves confirm that the improvement is not restricted to the training objective: DF-Muon-Proxy gives the lowest validation loss on the representative seed, while DA-Muon and Scale-Calibrated Muon remain competitive with the fixed-Muon baseline.



(a) Training loss

(b) Validation loss

Figure 13: Representative NanoGPT/WikiText-2 curves for seed 42. We compare AdamW, the tuned fixed-Muon baseline, and the three adaptive Muon variants. Table 22 reports the aggregate over three seeds.

Observed adaptive scales. Because all methods use a warmup–cosine schedule, the effective learning rate approaches zero near the end of training. We therefore report the base scale η_t selected by each adaptive rule before multiplication by the common schedule. Table 23 reports the final-20% mean base scale averaged over seeds, and Figure 14 shows the representative scale trajectories.

DA-Muon saturates the cap $\eta_{\max} = 0.03$ on this small Transformer task. Scale-Calibrated Muon selects a moderate scale around 0.02. DF-Muon-Proxy follows a different trajectory: it first becomes conservative and then increases its scale late in training. This behavior is consistent with the validation results in Table 22, where DF-Muon-Proxy gives the strongest aggregate performance.

Table 23: Observed base scales on NanoGPT/WikiText-2 over three seeds. For adaptive methods, the mean base scale is averaged over the final 20% of training and then averaged over seeds.

Method	**Selection rule**	**Mean base η**
AdamW	fixed	0.0010
Best fixed Muon	fixed	0.0150
DF-Muon-Proxy	adaptive	0.0202
DA-Muon	adaptive	0.0300
Scale-Calibrated Muon	adaptive	0.0202

Interpretation. The NanoGPT experiment supports the same qualitative conclusion as the larger GPT-124M/WikiText-103 benchmark: adaptive scale selection can improve Muon even against a tuned fixed-Muon baseline. In this smaller setting, all three adaptive variants improve mean validation loss over fixed Muon. The updated majorized DF-Muon-Proxy is the strongest method, reducing mean validation loss from 5.7162 for fixed Muon to 5.6630. DA-Muon and Scale-Calibrated Muon also improve over fixed Muon, reaching validation losses 5.7048 and 5.6955, respectively. The observed-scale diagnostics show that these improvements arise from different scale-selection behavior rather than from a single shared effective learning rate.

S.5 Regularized ViT-Tiny/CIFAR-100 Diagnostics

This section provides additional diagnostics for the regularized ViT-Tiny/CIFAR-100 experiment in Appendix S.2. The goal is to document the fixed-Muon scale selection, the DF-Muon-Proxy cap selection, and the effective scales used by the adaptive methods.

Fixed-Muon learning-rate diagnostic. We first ran a one-seed fixed-Muon learning-rate diagnostic on ViT-Tiny/CIFAR-100. Table 24 reports the sweep. The smallest tested Muon scale, $\eta = 0.001$, gives the best validation cross-entropy and top-1 accuracy in this diagnostic, and is therefore used as the tuned fixed-Muon baseline in Table 17.

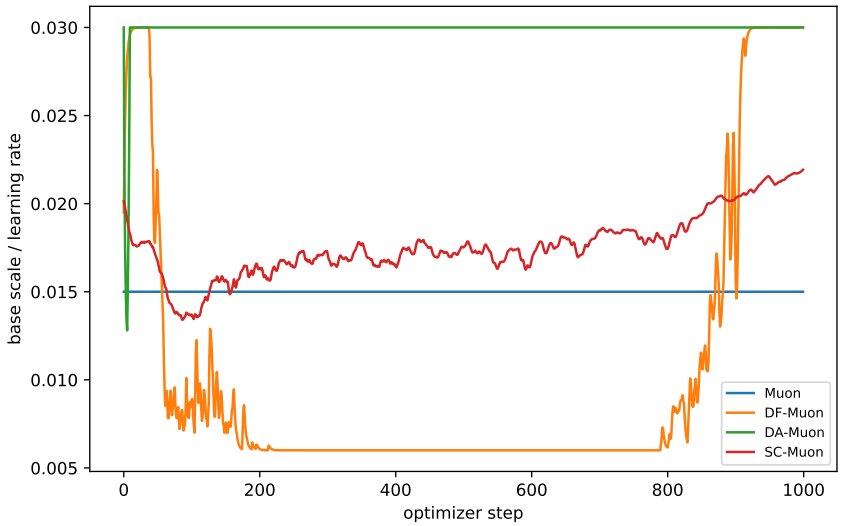


Figure 14: Observed base scales selected by the Muon variants on NanoGPT/WikiText-2 in representative seed 42. The fixed-Muon baseline uses $\eta = 0.015$, DA-Muon saturates the cap $\eta_{\max} = 0.03$, Scale-Calibrated Muon stabilizes around 0.02, and DF-Muon-Proxy follows a conservative-then-increasing scale trajectory.

Table 24: Fixed-Muon learning-rate diagnostic on ViT-Tiny/CIFAR-100 for one representative seed. CIFAR-100 images are resized to 224×224. Runtime is reported relative to the reference fixed-Muon run.

Method	**Train CE**	**Val. CE**	**Top-1**	**Mean η**	**Rel. time**
Muon, $\eta = 0.001$	0.3589	**1.6752**	**57.74**	0.0010	1.01×
Muon, $\eta = 0.002$	0.0775	2.0090	55.85	0.0020	0.99×
Muon, $\eta = 0.003$	0.0327	2.2269	53.85	0.0030	0.99×
Muon, $\eta = 0.005$	0.0295	2.3036	51.61	0.0050	1.00×
Muon, $\eta = 0.0075$	0.0561	2.3027	49.49	0.0075	1.01×
Muon, $\eta = 0.01$	0.1395	2.2052	48.98	0.0100	0.99×

DF-Muon-Proxy cap diagnostic. We then ran a matched regularized 100-epoch one-seed diagnostic for DF-Muon-Proxy, varying only the cap $\eta_{\max}$. Table 25 reports the result. The cap $\eta_{\max} = 0.01$ gives the best top-1 accuracy among the DF-Muon-Proxy variants, while remaining close to the tuned fixed-Muon baseline in best validation cross-entropy. We therefore use $\eta_{\max} = 0.01$ in the final three-seed comparison.

Table 25: Regularized ViT-Tiny/CIFAR-100 DF-Muon-Proxy cap diagnostic for one representative seed. CIFAR-100 images are resized to 224×224. Fixed Muon uses the tuned learning rate $\eta = 0.001$. DF-Muon-Proxy varies only the cap $\eta_{\max}$. We report best validation cross-entropy and best top-1 accuracy over epochs. Runtime is relative to fixed Muon. Since this is a one-seed diagnostic, no standard deviations are reported.

Variant	**Best Val. CE**	**Top-1 @ Best CE**	**Best Top-1**	**Mean η**	**Rel. time**
Fixed Muon, $\eta = 0.001$	**1.3508**	64.97	67.76	0.0010	1.00×
DF-Muon-Proxy, $\eta_{\max} = 0.005$	1.3721	64.11	68.25	0.0050	1.00×
DF-Muon-Proxy, $\eta_{\max} = 0.01$	1.3572	64.47	**68.57**	0.0100	1.00×
DF-Muon-Proxy, $\eta_{\max} = 0.015$	1.3744	63.94	68.42	0.0150	1.00×
DF-Muon-Proxy, $\eta_{\max} = 0.02$	1.3633	64.85	67.73	0.0200	1.04×
DF-Muon-Proxy, $\eta_{\max} = 0.03$	1.3543	64.38	67.56	0.0297	1.03×

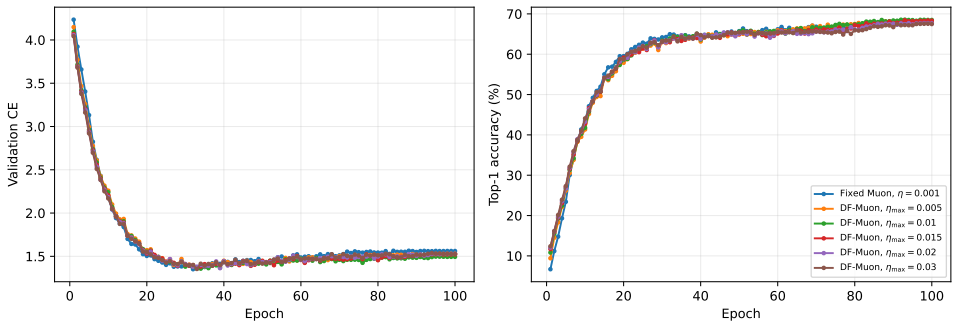


Figure 15: DF-Muon-Proxy cap diagnostic on regularized ViT-Tiny/CIFAR-100 in a representative seed. The cap $\eta_{\max} = 0.01$ gives the best top-1 accuracy among the tested DF-Muon-Proxy caps, and is used in the final three-seed comparison.

Observed effective scales. Because all ViT-Tiny/CIFAR-100 runs use a common warmup–cosine schedule, the schedule-multiplied effective step scale approaches zero near the end of training. We therefore report both the base scale selected by each method and the effective scale after multiplication by the common schedule. Table 26 summarizes the mean scales over the final 20% of training.

Table 26: Observed base and effective scales in the regularized ViT-Tiny/CIFAR-100 run. The effective scale equals the base scale selected by each method multiplied by the common warmup–cosine schedule.

Method	**Mean base η**	**Mean effective scale**
AdamW	0.0003	0.000011
Best fixed Muon	0.0010	0.000036
DF-Muon-Proxy	0.0100	0.000357
DA-Muon	0.0100	0.000357
Scale-Calibrated Muon	0.0168	0.000594

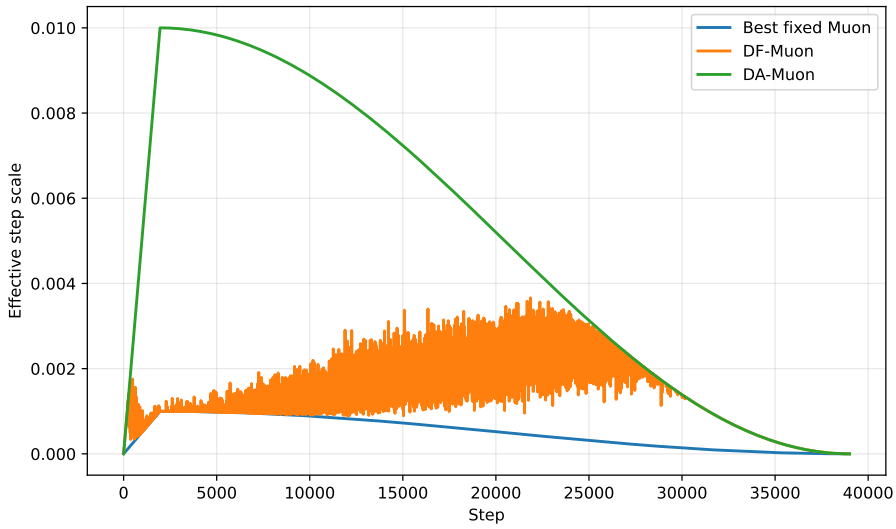


Figure 16: Effective step scales on regularized ViT-Tiny/CIFAR-100 for a representative seed, excluding Scale-Calibrated Muon for readability. The effective scale equals the base scale selected by each method multiplied by the common warmup–cosine schedule.

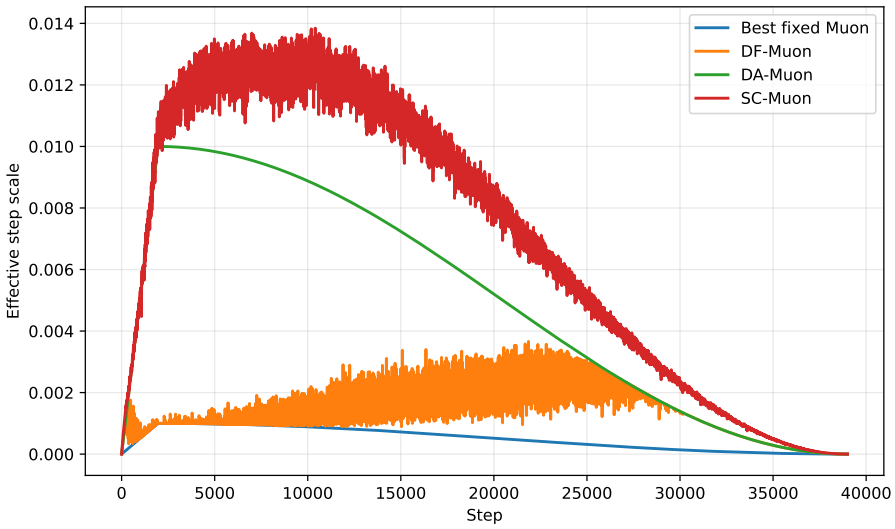


Figure 17: Effective step scales on regularized ViT-Tiny/CIFAR-100 for a representative seed, including Scale-Calibrated Muon. Scale-Calibrated Muon selects a larger and noisier effective scale, while DF-Muon-Proxy uses an intermediate scale between the tuned fixed-Muon baseline and the more aggressive adaptive variants.

Interpretation. The diagnostics show that the vision comparison is made against a tuned fixed-Muon baseline and that DF-Muon-Proxy is not evaluated with an arbitrary cap. The fixed-Muon diagnostic selects $\eta = 0.001$. The DF-Muon-Proxy cap diagnostic selects $\eta_{\max} = 0.01$, which gives the best top-1 accuracy among the tested caps. In the final three-seed run, fixed Muon obtains the lowest best validation cross-entropy, while DF-Muon-Proxy obtains the highest best top-1 accuracy with only a small runtime overhead. The scale diagnostics indicate that DF-Muon-Proxy chooses a more aggressive scale than the fixed-Muon baseline but remains less aggressive than DA-Muon and Scale-Calibrated Muon for much of training.

S.6 Additional CIFAR-100/ResNet-32 Diagnostics

This section reports a complementary image-classification diagnostic on CIFAR-100 with a CIFAR-style ResNet-32. The goal is to check whether the adaptive Muon scale rules remain competitive with a tuned fixed-Muon baseline outside Transformer language modeling. We train for $T = 2000$ optimizer steps with batch size 128. AdamW uses learning rate $3 \cdot 10^{-3}$. Fixed Muon is tuned by a learning-rate sweep, and the adaptive Muon variants use the same maximum base scale as the best fixed-Muon scale.

Fixed-Muon learning-rate sweep. Table 27 reports the fixed-Muon learning-rate sweep on one representative seed. The best fixed-Muon learning rate in the sweep is $\eta = 0.05$, which achieves the lowest training cross-entropy, lowest test cross-entropy, and highest test accuracy among the tested fixed scales. We therefore use $\eta = 0.05$ as the tuned fixed-Muon baseline and set $\eta_{\max} = 0.05$ for the adaptive variants.

Table 27: Fixed-Muon learning-rate sweep on CIFAR-100/ResNet-32 for one representative seed.

Method	Train CE	Test CE	Test acc.
Muon, $\eta = 0.01$	1.9777	3.8283	45.93%
Muon, $\eta = 0.015$	1.8756	3.7921	48.03%
Muon, $\eta = 0.02$	1.7952	3.7490	50.16%
Muon, $\eta = 0.03$	1.7495	3.7402	50.34%
Muon, $\eta = 0.05$	**1.6675**	**3.7171**	**52.20%**

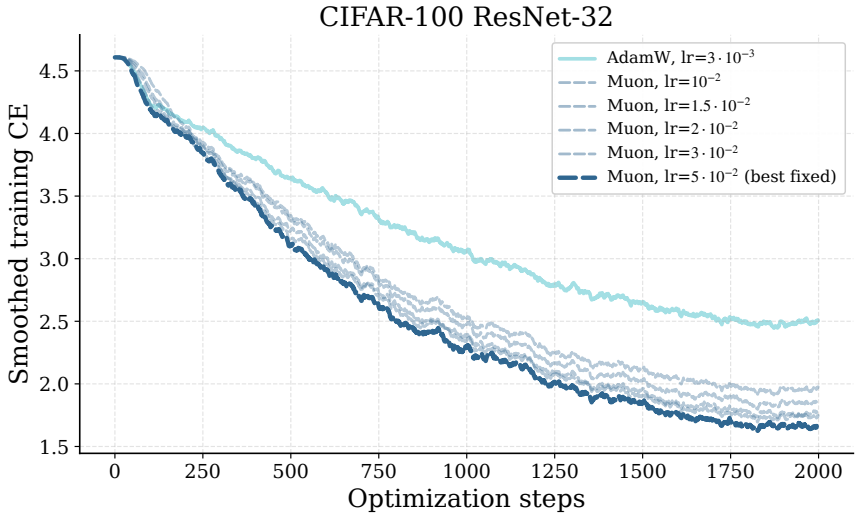


Figure 18: Fixed-Muon learning-rate sweep on CIFAR-100/ResNet-32. The largest tested scale, $\eta = 0.05$, gives the strongest final training curve and is used as the tuned fixed-Muon baseline.

Three-seed comparison. Table 28 reports the three-seed aggregate for AdamW, tuned fixed Muon, and the adaptive Muon variants. All Muon-based methods substantially outperform AdamW under this short training budget. The updated majorized DF-Muon-Proxy improves the tuned fixed-Muon baseline from 52.24% to 52.96% mean test accuracy and slightly improves mean test cross-entropy. DA-Muon gives the lowest mean training CE, while Scale-Calibrated Muon gives the lowest mean test CE.

Table 28: CIFAR-100/ResNet-32 results over three seeds. We report mean $\pm$ standard deviation over seeds $\{42, 1337, 2024\}$. The fixed-Muon baseline uses the best learning rate from the sweep, $\eta = 0.05$. Adaptive Muon variants use $\eta_{\max} = 0.05$.

Method	**Train CE**	**Test CE**	**Test acc.**	**Mean base η**
AdamW	2.5096 ± 0.0018	3.9469 ± 0.1759	$34.24 \pm 0.48\%$	0.0030
Best fixed Muon	1.6682 ± 0.0147	3.5589 ± 0.1477	$52.24 \pm 0.20\%$	0.0500
DF-Muon-Proxy	1.6271 ± 0.0115	3.5505 ± 0.1834	$\mathbf{52.96 \pm 0.54\%}$	0.0500
DA-Muon	$\mathbf{1.6241 \pm 0.0017}$	3.5430 ± 0.1754	$52.91 \pm 0.26\%$	0.0500
Scale-Calibrated Muon	1.6321 ± 0.0114	$\mathbf{3.5407 \pm 0.1894}$	$52.47 \pm 0.65\%$	0.0350

Representative training curve. Figure 19 shows the representative training curve for seed 1337. The Muon-based methods reduce training cross-entropy much faster than AdamW. Among the Muon variants, the adaptive methods remain close to the tuned fixed-Muon baseline throughout training.

Observed adaptive scales. Because all methods use a warmup–cosine schedule, the effective learning rate approaches zero near the end of training. We therefore report the base scale η_t selected by each adaptive rule before multiplying by the common schedule. Table 29 reports the mean base scale over the final 20% of training, averaged over seeds, and Figure 20 shows the representative scale trajectories.

DA-Muon reaches the cap $\eta_{\max} = 0.05$ on this task. Scale-Calibrated Muon selects a smaller stable scale around 0.035. DF-Muon-Proxy follows a data-dependent trajectory under the one-dimensional majorized rule: it is initially conservative and then increases toward the cap late in training.

Interpretation. The CIFAR-100/ResNet-32 experiment gives a complementary picture to the GPT-124M/WikiText-103 and NanoGPT/WikiText-2 language-modeling experiments. The fixed-Muon sweep identifies a strong baseline at $\eta = 0.05$, and the adaptive variants remain competitive with this

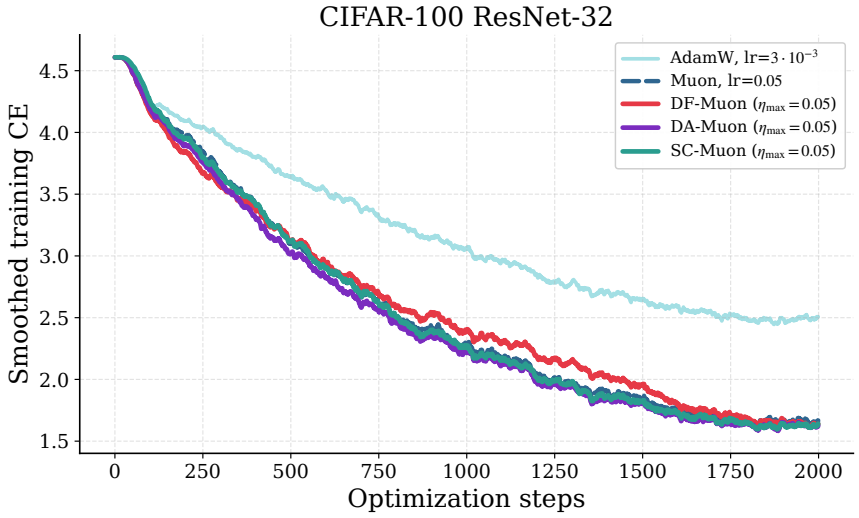


Figure 19: CIFAR-100/ResNet-32 training in a representative seed. We compare AdamW, the tuned fixed-Muon baseline, and the three adaptive Muon variants using the updated majorized DF-Muon-Proxy implementation.

Table 29: Observed base scales on CIFAR-100/ResNet-32 over three seeds. For adaptive methods, the mean base scale is averaged over the final 20% of training and then averaged over seeds.

Method	**Selection rule**	**Mean base η**
AdamW	fixed	0.0030
Best fixed Muon	fixed sweep	0.0500
DF-Muon-Proxy	majorized adaptive	0.0500
DA-Muon	trajectory radius	0.0500
Scale-Calibrated Muon	certificate	0.0350

tuned baseline. The updated majorized DF-Muon-Proxy improves mean test accuracy and slightly improves mean test cross-entropy relative to fixed Muon, while DA-Muon and Scale-Calibrated Muon are also competitive. The differences among the Muon variants are small relative to seed-to-seed variation, so we view this experiment as evidence that adaptive scale selection remains robust on a convolutional benchmark rather than as a decisive separation among the adaptive rules.

S.7 Existing Assets, Licenses, and Implementations

All experiments use standard public benchmark datasets and do not introduce or redistribute new datasets. For language modeling, we use WikiText-103 in the main GPT-124M experiment and WikiText-2 in the smaller NanoGPT diagnostic experiment. Both are standard WikiText benchmarks derived from Wikipedia text and distributed under Creative Commons Attribution–ShareAlike terms (Merity et al., 2016). For image classification, we use CIFAR-100 with its standard train/test split (Krizhevsky et al., 2009). In all cases, the datasets are used only as benchmarks and are not redistributed as part of this work.

The smaller language-modeling diagnostics use a NanoGPT-style GPT implementation based on the public NanoGPT codebase, which is released under the MIT License. The image-classification experiments use ViT-Tiny and CIFAR-style ResNet-32 implementations in PyTorch. Dataset loading and standard image transforms are based on TorchVision (Paszke et al., 2019). PyTorch and TorchVision are released under BSD-style licenses. We modified the training loops only to replace the optimizer on matrix-valued parameters by the Muon-based methods studied in this paper, while keeping the dataset splits, model families, and evaluation metrics standard.

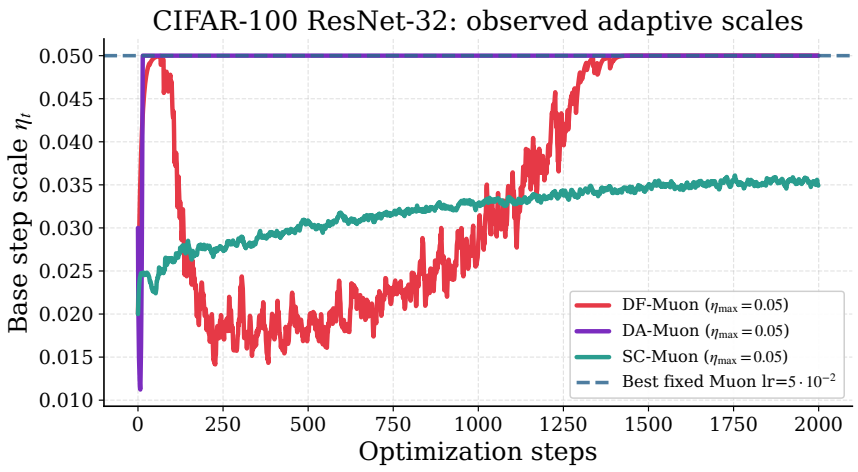


Figure 20: Observed base scales selected by the adaptive Muon variants on CIFAR-100/ResNet-32 in representative seed 1337. The dashed horizontal line marks the tuned fixed-Muon scale, $\eta = 0.05$. DA-Muon reaches the cap, Scale-Calibrated Muon stabilizes around 0.035, and DF-Muon-Proxy follows a conservative-then-increasing trajectory under the one-dimensional majorized rule.

T Additional Related Work

Broader Muon variants. Recent work has studied matrix-sign computations, variance reduction, error feedback, schedule-free spectral optimization, and broader non-Euclidean SGD frameworks (Amsel et al., 2025; Qian et al., 2025; Gruntkowska et al., 2025; Apte et al., 2026; Kovalev & Borodich, 2025). Most of these directions are complementary to adaptive scale selection. Momentum variance reduction is more directly related: Qian et al. (2025) improve the stochastic star-convex Muon rate, but use a prescribed radius proportional to a bound on the solution norm. Our focus is choosing the radius without knowing the initial distance to a minimizer, $D = \|x_0 - x^\star\|$.

Stochastic oracles. Our analysis uses unbiased stochastic oracles with bounded variance, a standard regime in stochastic non-convex optimization (Demidovich et al., 2023; Khaled & Richtárik, 2020).

Momentum and practical safeguards. Classical momentum originates in the two-step method of Polyak (1964), while Nesterov acceleration uses a related but distinct extrapolation mechanism (Nesterov, 1983). Momentum has also been studied in parameter-agnostic normalized error-feedback methods (Sadiev et al., 2025). Gradient clipping is a standard safeguard against exploding gradients (Pascanu et al., 2013). Its adaptive role has been analyzed under conventional and generalized smoothness assumptions (Zhang et al., 2019; Demidovich et al., 2025; Gorbunov et al., 2025). These practical mechanisms are complementary to the reference-radius problem studied here.