ACCELERATING STOCHASTIC SCHATTEN MOMENTUM DESCENT AGAINST UNBOUNDED GRADIENT NOISE

Anonymous authors
Paper under double-blind review

ABSTRACT

Matrix optimizers which exploit the spectral geometry are increasingly getting adopted for training large transformer based models. Inspired by these successes we seek to build foundations for these novel ML algorithms. Thus we are led to study a notion of stochastic Schatten momentum descent with Exponential Moving Average (EMA) and Momentum Variance Reduction (MVR). We introduce a generalized notion of smoothness and an affinely growing bound on the variance of stochastic gradient differences, both defined in the Schatten norm. Under these relaxed conditions of gradient smoothness and unbounded gradient noise, we prove (a) an optimal $O(T^{-1/3})$ convergence rate for the variance reduced form of our algorithm by exploiting the aforementioned control on stochastic gradient differences and (b) a $O(T^{-1/4})$ rate for the EMA variant. We further extend the EMA guarantee to Polyak's heavy-ball momentum, and thus getting new guarantees on the idealized stochastic Muon algorithm.

1 INTRODUCTION

MUON (Jordan et al., 2024b) has moved from a speedrunning curiosity to frontier-scale production faster than any comparable optimizer. It applies a Newton–Schulz iteration to the momentum buffer so that every singular value of the update is normalised to one. Liu et al. (2025a) showed that weight decay and a corrected per-parameter update scale make it work at scale, reporting roughly $2\times$ computational efficiency over AdamW under compute-optimal training — AdamW quality at about 52% of the FLOPs — and trained the 3B/16B Moonlight model on 5.7T tokens; Shah et al. (2025) independently found Muon expands the compute-time Pareto frontier. Kimi K2 (Kimi Team, 2025), a 1.04T-parameter mixture-of-experts model, was pre-trained on 15.5T tokens with MUONCLIP with no loss spikes.

What separates Muon from the Adam family is geometry, that it treats each weight matrix $\boldsymbol{W} \in \mathbb{R}^{m\times n}$ as a matrix rather than flattening it into a vector. Coordinate-wise methods are invariant to permutations of the entries of $\boldsymbol{W}$ and so cannot see its spectrum, whereas Muon's update is determined by the singular value decomposition of the momentum and equalises the influence of every singular direction. Bernstein & Newhouse (2024; 2025) made this precise by identifying the idealized polar update as steepest descent under the spectral norm. For $\boldsymbol{G} = \boldsymbol{U}\Sigma\boldsymbol{V}^\top$, the associated linear minimization oracle (LMO) admits the direction $-\boldsymbol{U}\boldsymbol{V}^\top$, which Newton–Schulz iterations approximate in practical implementations.

The spectral norm is one choice among many, and the Schatten family $\|\boldsymbol{X}\|_{S_p} = \|\boldsymbol{\sigma}(\boldsymbol{X})\|_{\ell_p}$ interpolates between normalized SGD ($p = 2$), momentum-free Muon ($p = \infty$) and a rank-one nuclear-ball update ($p = 1$). An informal Schatten-p treatment appeared in Cesista (2025); Fan et al. (2026) gave, to our knowledge, the first systematic theory across the family, characterising the implicit bias of normalized steepest descent under arbitrary entry-wise and Schatten norms. Concurrently Pethick et al. (2025a) introduced SCION, inverting the logic by taking the LMO as the design primitive with the norm a free parameter and they brings their insights to bear on pretraining and report hyperparameter transfer across scales. Riabinin et al. (2025) introduced GLUON, arguing that prior analyses describe a different algorithm from the one practitioners run — neglecting layer-wise structure — and they introduced a layer-wise generalized (L_i^0, L_i^1)-smoothness model whose prior form was seen in (Zhang et al., 2020a). In the released implementation (Pethick et al., 2025b), the algorithm is verified on NanoGPT and CNN.

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.

Convergence theory has developed rapidly. Convergence at a rate of $\mathcal{O}(\epsilon^{-4})$ appeared in Li & Hong (2025) and Kovalev (2025) for Muon and the latter via non-Euclidean trust regions in Pethick et al. (2025a) for the (un)constrained LMO variants. Shen et al. (2026) and Sato et al. (2025) analyzed for critical batch sizes for convergence and Sfyraki & Wang (2025) recast Muon with weight decay as stochastic Frank–Wolfe under heavy-tailed noise. Pethick et al. (2026) developed non-Euclidean (L_0, L_1)-smoothness in generality; Shulgin et al. (2026a) analysed the inexact LMO that a finite Newton–Schulz iteration actually delivers; and Shulgin et al. (2026b) replaced bounded variance altogether by the ABC (expected smoothness) condition of Khaled & Richtárik (2023), observing that bounded variance can fail even for a subsampled quadratic on an unbounded domain. *Our work probes much further in this direction to get optimal accelerated convergence for algorithms in the Schatten geometry while facing stochastic gradient noise with unbounded variance.*

In a seminal result Arjevani et al. (2023) have proven a convergence time lowerbound of $\Omega(\epsilon^{-4})$ for stochastic gradient based algorithms, that holds even when all derivatives are bounded. Its well-known that one can escape this restriction via variance reduction methods — SARAH (Nguyen et al., 2017), SPIDER (Fang et al., 2018) and SNVRG (Zhou et al., 2020) which attain $\widetilde{\mathcal{O}}(\epsilon^{-3})$, matching the corresponding lower bound of Arjevani et al. (2023) up to logarithmic factors. But these lowerbounds are unsatisfactory for modern deep learning because they use mega-batch checkpoints, SARAH's $\mathcal{O}(n + \sqrt{n}\epsilon^{-2})$ rate is pratically prohibitive for usual values of n = number of training-data. Further, these results are in the Euclidean geometry and hence not fully using the matrix structure of weights of modern ML models. Rate is moreover not the only axis along which the noise model matters: Shulgin et al. (2026b) show that natural sampling conditions are not compatible with the bounded-variance model of such results. A separate line of work escapes the ϵ^{-4} barrier by enlarging the oracle to stochastic Hessian-vector products, where Arjevani et al. (2020) show ϵ^{-3} is unimprovable under second-order smoothness even given pth-order derivatives for any $p \geq 2$.

Recent work has pursued $\mathcal{O}(\epsilon^{-3})$ guarantees for Muon-type methods through variance reduction (Huang et al., 2026; Liu et al., 2025b; Qian et al., 2025) and stochastic Hessian-vector corrections (Khirirat et al., 2025; Chayti, 2026). Table 1.1 compares their assumptions, oracle requirements, and convergence guarantees, with further details in Appendices E.5 and E.6. Our results here can be seen to add a novel aspect to this recently developing line of work.

Organization The paper is organized as follows. Section 2 introduces the background geometry of Schatten norms and the SSMD algorithm family. Section 3 presents our assumptions and convergence guarantees for SSMD-MVR, SSMD-EMA, and SSMD-Polyak's heavy-ball momentum. Section 4 reports the empirical evaluation, and Section 5 concludes. The appendices provide additional mathematical background (Appendix A), complete proofs (Appendices B and C), experimental details (Appendix D) and detailed comparisons with related methods (Appendix E).

1.1 Summary of Results

In the Schatten-p geometry our framework covers Momentum Variance Reduction (MVR) in Algorithm 1 and Exponential Moving Average (EMA) in Algorithm 2. For a fixed momentum coefficient and matched initialization, Polyak's heavy-ball momentum in Algorithm 3 differs from EMA only by a positive scaling, preserving the admissible descent directions. Our EMA guarantees therefore extend to heavy-ball momentum, with initialization differences contributing only a lower-order term. At $p = \infty$, Algorithm 3 reduces to idealized Muon (Jordan et al., 2024b). *We abbreviate the general family as Stochastic Schatten Momentum Descent (SSMD).*

Introducing a Relaxed Smoothness Assumption in the Schattten Norm We introduce a "clipped Schatten" formulation of (L_0, L_1)-smoothness. Specifically, Assumption 3 measures gradient and iterate differences in $S_{\bar{q}}$ and $S_{\bar{p}}$, respectively, where $\bar{p} = \min\{p, 2\}$ and $\bar{q} = \max\{q, 2\}$. For $p \leq 2$, this preserves the usual primal–dual Schatten geometry. For $p > 2$, clipping replaces both difference norms by the Frobenius norm, weakening the smoothness inequality at the same modulus compared with the unmodified (S_p, S_q) condition. The spectral-norm case $p = \infty$ requires Frobenius control of gradient variation instead of nuclear-norm control relative to spectral-norm displacements.

Algorithm 1 SSMD with MVR

Input: $\boldsymbol{W}_0, \eta > 0$, $a \in (0, 1]$
$\boldsymbol{M}_0 = \nabla f(\boldsymbol{W}_0; \xi_0)$
$\boldsymbol{W}_1 = \boldsymbol{W}_0 - \eta \mathcal{D}_p(\boldsymbol{M}_0)$
for $t = 1, 2, \ldots$ **do**
  Sample a fresh ξ_t.
  $\boldsymbol{G}_t^+ = \nabla f(\boldsymbol{W}_t; \xi_t)$
  $\boldsymbol{G}_t^- = \nabla f(\boldsymbol{W}_{t-1}; \xi_t)$
  $\boldsymbol{M}_t = \boldsymbol{G}_t^+ + (1-a)(\boldsymbol{M}_{t-1} - \boldsymbol{G}_t^-)$
  $\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \eta \mathcal{D}_p(\boldsymbol{M}_t)$
end for

Algorithm 2 SSMD with EMA

Input: $\boldsymbol{W}_0, \eta > 0$, $\beta \in [0, 1)$
$\boldsymbol{M}_0 = \nabla f(\boldsymbol{W}_0; \xi_0)$
$\boldsymbol{W}_1 = \boldsymbol{W}_0 - \eta \mathcal{D}_p(\boldsymbol{M}_0)$
for $t = 1, \ldots, T-1$ **do**
  Sample a fresh ξ_t.
  $\boldsymbol{G}_t = \nabla f(\boldsymbol{W}_t; \xi_t)$
  $\boldsymbol{M}_t = \beta \boldsymbol{M}_{t-1} + (1-\beta)\boldsymbol{G}_t$
  $\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \eta \mathcal{D}_p(\boldsymbol{M}_t)$
end for

Algorithm 3 SSMD with Heavy-Ball Momentum

Input: $\boldsymbol{W}_0, \eta > 0$, $\mu \in [0, 1)$
$\boldsymbol{M}_0 = \boldsymbol{0}$
for $t = 0, \ldots, T-1$ **do**
  Sample a fresh ξ_t.
  $\boldsymbol{G}_t = \nabla f(\boldsymbol{W}_t; \xi_t)$
  $\boldsymbol{M}_{t+1} = \mu \boldsymbol{M}_t + \boldsymbol{G}_t$
  $\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \eta \mathcal{D}_p(\boldsymbol{M}_{t+1})$
end for

Thus our clipped Schatten condition of smoothness always defaults to the weakest assumption in the given p. Meanwhile, the modulus of smoothness — the RHS of the smoothness condition — remains $L_0 + L_1 \sup_{\theta \in [0,1]} \|\nabla F(\theta W_1 + (1-\theta) W_2)\|_{S_q}$, allowing larger curvature in regions with larger gradients without requiring a uniform global smoothness constant.

For stochastic gradients, we use an affine noise variance condition.

Introducing a Relaxed Centered Gradient Difference Assumption in the Schattten Norm For analyzing SSMD-MVR type algorithms, standard mean-squared smoothness bounds the full stochastic gradient difference, whereas our analysis controls its centered fluctuation and the population gradient difference separately — and allows the former to be growing. Our assumption can be seen as a gradient-dependent relaxation of the constant-coefficient centered condition of Qian et al. (2025), allowing the coefficient in the variance bound to grow with the largest population-gradient norm along the segment between the two parameter points.

Informal Statement of Theorems. We consider the following three SSMD algorithms (Algorithm 1, 2 and 3) – in the pseudocodes we use the notation of $\mathcal{D}_p(\boldsymbol{M}_t) \in \arg\max_{\|\boldsymbol{D}\|_{S_p} \le 1} \langle \boldsymbol{M}, \boldsymbol{D} \rangle$ for $p, q > 0$ s.t $\frac{1}{p} + \frac{1}{q} = 1$. The MVR recursion uses the same-sample variance-reduction estimator of STORM (Cutkosky & Orabona, 2019), combined here with Schatten-p descent directions and analyzed under our relaxed assumptions.

Theorem 1 (Informal statement of Theorem 3)**.** *Suppose that the affine bound on the centered gradient-difference holds. Then SSMD with MVR (Algorithm 1) satisfies*

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla F(W_t)\|_{S_q} \le \mathcal{O}(T^{-1/3}).$$

Consequently, a uniformly sampled iterate has expected Schatten-q gradient norm at most ϵ within $\mathcal{O}(\epsilon^{-3})$ stochastic-gradient evaluations, using two evaluations of the same fresh sample per iteration.

Theorem 2 (Informal statement of Theorem 4 and Corollary 1)**.** *Under the common assumptions above, SSMD with EMA (Algorithm 2) and SSMD with Polyak's heavy-ball momentum (Algorithm 3) both satisfy,*

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \|\nabla F(W_t)\|_{S_q} \le \mathcal{O}(T^{-1/4}).$$

This yields $\mathcal{O}(\epsilon^{-4})$ stochastic-gradient complexity for a uniformly sampled output, using one gradient evaluation per iteration and requiring no gradient-difference variance condition.

We note that all guarantees above hold under our relaxed (L_0, L_1)-smoothness condition in clipped Schatten geometry introduced earlier.

Comparison to Related Works Table 1.1 compares convergence rates, smoothness and noise assumptions, oracle requirements, and optimization geometries. Detailed comparisons are provided in Appendices E.5 and E.6.

Method	Rate	GS	HS	Noise	Gradient-Difference Condition	Oracle	Geometry
STORM (Cutkosky & Orabona, 2019)	$\widetilde{\mathcal{O}}(T^{-1/3})$	L	–	BV	PS	SG	Euclidean
SOM (Tran & Cutkosky, 2022)	$\mathcal{O}(T^{-1/3})$	L	M	BV	–	SG + HVP	Euclidean
1-SFW (Zhang et al., 2020b)	$\mathcal{O}(T^{-1/3})^a$	L	–	BV	PS	SG	Euclidean
MARS (Yuan et al., 2025)	$\mathcal{O}(\log T\, T^{-2/3})^b$	L	–	BV	PS	SG	Preconditioned
HCM–V1 (Khirirat et al., 2025)	$\mathcal{O}(T^{-1/3})$	(L_0, L_1)	–	BV	–	SG + HVP	Arbitrary
HCM–MVR (Khirirat et al., 2025)	$\mathcal{O}(T^{-1/3})$	(L_0, L_1)	–	BV	GMSS	SG	Arbitrary
RanSOM (Chayti, 2026)	$\mathcal{O}(T^{-1/3})^c$	(L_0, L_1)	–	AV	–	SG + HVP	Arbitrary
Gluon-MVR (Qian et al., 2025)	$\mathcal{O}(T^{-1/3})$	(L_{0i}, L_{1i})	–	BV	CDV	SG	Layer-wise arbitrary
Ours: MVR	$\mathcal{O}(T^{-1/3})$	(L_0, L_1)	–	**AV**	**ACDV**	**SG**	Schatten-p
Ours: EMA / heavy-ball	$\mathcal{O}(T^{-1/4})$	(L_0, L_1)	–	**AV**	**–**	**SG**	Schatten-p

Notation, assumptions, and oracle models are detailed in Appendix E.5. Convergence criteria and theorem variants are given in Appendix E.6. [a]Frank–Wolfe gap. [b]Squared estimation errors and weighted squared steps. [c]Subject to batching and initialization conditions.

Table 1: Comparison of nonconvex convergence guarantees. GS and HS denote gradient and Hessian smoothness: L denotes Lipschitz gradients, M denotes Lipschitz Hessians, and (L_0, L_1) denotes generalized smoothness, with layer-specific constants where indicated. BV denotes uniformly bounded gradient-noise variance; AV denotes variance bounded by an affine function of the squared gradient norm. PS denotes per-sample gradient smoothness; GMSS denotes generalized mean-squared smoothness of stochastic gradients; CDV denotes centered gradient-difference variance with a constant coefficient; ACDV denotes affine centered gradient-difference variance. SG denotes stochastic-gradient evaluations and HVP denotes Hessian–vector products.

As summarized in Table 1.1, our MVR result achieves $\mathcal{O}(\epsilon^{-3})$ stochastic-gradient complexity under clipped Schatten smoothness, affine noise variance, and an affine centered gradient-difference condition — without requiring Hessian smoothness, Hessian–vector products, or uniformly bounded noise variance, one of which restrictions is used by each other result in Table 1.1. Our EMA and heavy-ball results also allow affine noise variance and achieve $\mathcal{O}(\epsilon^{-4})$ stochastic-gradient complexity without a gradient-difference condition.

Summary of Empirical Results. We compare SSMD with MVR and EMA using DeiT-Tiny (Touvron et al., 2021) on CIFAR-100 (Krizhevsky et al., 2009) and GPT-2 (Jordan et al., 2024a) pretraining on FineWeb (Henriksson et al., 2025). We consider theorem-scaled and empirical settings, with hyperparameters chosen according to the theoretical scalings and informed by practical hyperparameter tuning, respectively, as follows (a) Under theorem-scaled settings, MVR achieves lower GPT-2 test loss and initially reduces DeiT-Tiny test loss faster, while EMA reaches lower final DeiT-Tiny test loss. (b) Under empirical settings, EMA achieves lower final GPT-2 test loss across all three geometries and lower DeiT-Tiny test loss at $p = 2, \infty$, with similar results at $p = 1$. Among all evaluated configurations, EMA with $p = \infty$ under empirical settings achieves the lowest final test loss for both models: the peak learning rate and momentum coefficient are 0.004 and $\beta = 0.9$ for DeiT-Tiny, and 0.02 and $\beta = 0.95$ for GPT-2. Detailed configurations and results are provided in Section 4 and Appendix D.

2 Mathematical Preliminaries

For $\boldsymbol{G} \in \mathbb{R}^{m\times n}$, the Schatten-$p$ norm is $\|\boldsymbol{G}\|_{S_p} := \|\sigma(\boldsymbol{G})\|_{\ell_p}$, where $\sigma(\boldsymbol{G})$ is the vector of singular values. The cases $p = 1, 2, \infty$ give the nuclear, Frobenius, and spectral norms, respectively. Throughout, $p, q \in [1, \infty]$ satisfy $1/p + 1/q = 1$, $r := \min\{m, n\}$, and $\langle \boldsymbol{A}, \boldsymbol{B}\rangle := \mathrm{Tr}(\boldsymbol{A}^\top \boldsymbol{B})$.

Schatten Hölder inequality gives $|\langle \boldsymbol{A}, \boldsymbol{B}\rangle| \le \|\boldsymbol{A}\|_{S_q}\|\boldsymbol{B}\|_{S_p}$, while norm comparison gives $\|\boldsymbol{A}\|_{S_{\bar{q}}} \le \|\boldsymbol{A}\|_{S_q} \le \rho_q \|\boldsymbol{A}\|_{S_{\bar{q}}}$; see Lemmas 2 and 3 in Appendix A.1.

Definition 1 (Schatten Linear Minimization Oracle)**.** The linear minimization oracle over the Schatten-p unit ball is $\mathrm{LMO}_p(\boldsymbol{G}) := \arg\min_{\|\boldsymbol{D}\|_{S_p}\le 1}\langle \boldsymbol{G}, \boldsymbol{D}\rangle$. We write $\mathcal{D}_p(\boldsymbol{G})$ for a valid selection from $\arg\max_{\|\boldsymbol{D}\|_{S_p}\le 1}\langle \boldsymbol{G}, \boldsymbol{D}\rangle$, so $-\mathcal{D}_p(\boldsymbol{G}) \in \mathrm{LMO}_p(\boldsymbol{G})$.

Update direction	Primal norm	Dual norm	$\mathcal{D}_p(\boldsymbol{M})$
Muon-type polar	S_∞ (spectral)	S_1 (nuclear)	$\boldsymbol{U}\boldsymbol{V}^\top$
Normalized momentum	S_2 (Frobenius)	S_2 (Frobenius)	$\boldsymbol{M}/\|\boldsymbol{M}\|_F$
Rank-one	S_1 (nuclear)	S_∞ (spectral)	$\boldsymbol{u}_1\boldsymbol{v}_1^\top$

Table 2: Special cases of the Schatten direction map for $\boldsymbol{M} \neq \boldsymbol{0}$, with singular vectors from its compact SVD. The primal and dual indices satisfy $1/p + 1/q = 1$.

For $\boldsymbol{G} \neq \boldsymbol{0}$, let $\boldsymbol{G} = \boldsymbol{U}\,\mathrm{diag}(\sigma_1, \ldots, \sigma_k)\boldsymbol{V}^\top$ be a compact SVD, with $\sigma_1 \geq \cdots \geq \sigma_k > 0$ and columns $\boldsymbol{u}_i, \boldsymbol{v}_i$. Hence the aforementioned valid selection is explicitly given as,

$$\mathcal{D}_p(\boldsymbol{G}) = \begin{cases} \boldsymbol{U}\,\mathrm{diag}(\sigma_i^{q-1})\boldsymbol{V}^\top / \|\boldsymbol{G}\|_{S_q}^{q-1}, & 1 < q < \infty, \\ \boldsymbol{U}\boldsymbol{V}^\top, & q = 1 \ (p = \infty), \\ \boldsymbol{u}_1\boldsymbol{v}_1^\top, & q = \infty \ (p = 1). \end{cases} \tag{1}$$

We set $\mathcal{D}_p(\boldsymbol{0}) = \boldsymbol{0}$. By Lemma 4, $\|\mathcal{D}_p(\boldsymbol{G})\|_{S_p} \leq 1$ and $\langle \boldsymbol{G}, \mathcal{D}_p(\boldsymbol{G})\rangle = \|\boldsymbol{G}\|_{S_q}$. Thus, $p = \infty$ gives Muon-type polar directions, $p = 2$ gives Frobenius-normalized directions, and $p = 1$ gives rank-one directions.

For MVR in Algorithm 1 and EMA in Algorithm 2, the parameter update is

$$\boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \eta_t \mathcal{D}_p(\boldsymbol{M}_t), \tag{2}$$

with the momentum recursion specified by the corresponding algorithm. Additional properties of Schatten norms and their update directions are given in Appendix A.1.

The Schatten Momentum Algorithm Family. The Schatten momentum family combines a momentum estimator with the direction map $\mathcal{D}_p$. For example, with EMA momentum, $\beta \in [0, 1)$, and stepsize $\eta_t > 0$, the recursion takes the form

$$\boldsymbol{M}_t = \beta \boldsymbol{M}_{t-1} + (1 - \beta)\nabla f(\boldsymbol{W}_t; \xi_t), \qquad \boldsymbol{W}_{t+1} = \boldsymbol{W}_t - \eta_t \mathcal{D}_p(\boldsymbol{M}_t). \tag{3}$$

MVR modifies the momentum recursion using a same-sample gradient-difference correction, while the heavy-ball variant applies $\mathcal{D}_p$ to its updated momentum $\boldsymbol{M}_{t+1}$. These variants share the same Schatten geometry, with their recursions and initializations specified in the corresponding algorithms.

By Lemma 4, $\mathcal{D}_p(\boldsymbol{M}) \in \partial\|\boldsymbol{M}\|_{S_q}$. Thus, the primal index p determines the update constraint, while the conjugate index q determines the dual norm. Table 2 lists three special cases.

Wenote that Signum (Bernstein et al., 2018) instead uses entrywise ℓ_∞ geometry and its relation to Schatten geometry through a diagonal embedding is discussed in Remark 3 in Appendix A. Further properties of Schatten norms and their update directions are given in Appendix A.1.

3 Main Theoretical Results

We first introduce the smoothness and noise assumptions and derive a descent inequality in clipped Schatten geometry. We then establish convergence rates of $\mathcal{O}(T^{-1/3})$ for SSMD with MVR in Theorem 3 and $\mathcal{O}(T^{-1/4})$ for SSMD with EMA in Theorem 4, measured by the average expected Schatten-q norm of the gradient. Finally, we extend the SSMD with EMA guarantee to the direct generalization of Muon to the Schatten norm.

Assumption 1 (Objective Regularity and Lower Boundedness)**.** Let ξ be a random variable valued in a metric space E and $f : \mathbb{R}^{m\times n} \times E \to [0, \infty)$ be an at least once-differentiable function with well-defined expectation s.t we can define,

$$F(\boldsymbol{W}) := \mathbb{E}_\xi[f(\boldsymbol{W}; \xi)], \ \boldsymbol{W} \in \mathbb{R}^{m\times n}. \tag{4}$$

Suppose F is continuously differentiable on $\mathbb{R}^{m\times n}$. Further, suppose there exists $F^* > -\infty$ such that, $F(\boldsymbol{W}) \geq F^*$.

Assumption 2 (Unbiased Stochastic-Gradient Oracle)**.** For every $\boldsymbol{W} \in \mathbb{R}^{m\times n}$, the stochastic gradient satisfies,

$$\mathbb{E}_\xi \left[\nabla f(\boldsymbol{W};\xi)\right] = \nabla F(\boldsymbol{W}). \tag{5}$$

Assumption 3 (Generalized (L_0, L_1) Smoothness in Clipped Schatten Geometry)**.** Let $\bar{p} = \min(p, 2)$ and $\bar{q} = \max(q, 2)$ (and hence $\frac{1}{\bar{p}} + \frac{1}{\bar{q}} = 1$ when $\frac{1}{p} + \frac{1}{q} = 1$). There exist $L_0, L_1 \geq 0$ such that for all $\boldsymbol{W}_1, \boldsymbol{W}_2 \in \mathbb{R}^{m\times n}$,

$$\|\nabla F(\boldsymbol{W}_1) - \nabla F(\boldsymbol{W}_2)\|_{S_{\bar{q}}} \leq \left(L_0 + L_1 \sup_{\theta\in[0,1]} \|\nabla F(\theta\boldsymbol{W}_1 + (1-\theta)\boldsymbol{W}_2)\|_{S_q}\right) \|\boldsymbol{W}_1 - \boldsymbol{W}_2\|_{S_{\bar{p}}}. \tag{6}$$

Since Schatten-p norm is a non-increasing function of p, this is a weaker smoothness assumption that all usual options that are used.

Assumption 4 (Affine Gradient-Difference Variance in Schatten-p Geometry)**.** Let $\frac{1}{p} + \frac{1}{q} = 1$. There exist $\delta_0, \delta_1 \geq 0$ such that for all $\boldsymbol{W}_1, \boldsymbol{W}_2 \in \mathbb{R}^{m\times n}$,

$$\begin{aligned}&\mathbb{E}_\xi \left[\|\nabla f(\boldsymbol{W}_1;\xi) - \nabla f(\boldsymbol{W}_2;\xi) - (\nabla F(\boldsymbol{W}_1) - \nabla F(\boldsymbol{W}_2))\|_{S_q}^2\right] \\ &\leq \left(\delta_0 + \delta_1 \sup_{\theta\in[0,1]} \|\nabla F(\theta\boldsymbol{W}_1 + (1-\theta)\boldsymbol{W}_2)\|_{S_q}\right)^2 \|\boldsymbol{W}_1 - \boldsymbol{W}_2\|_{S_p}^2.\end{aligned} \tag{7}$$

We introduce this affine centered gradient-difference condition to control the stochastic correction error in MVR separately from population gradient variation. Building on the constant-coefficient centered condition of Qian et al. (2025), it allows the variance bound to grow with the gradient norm along the segment between the two points. Thus, $\delta_1 > 0$ permits greater variation in stochastic noise in regions with larger gradients.

Assumption 5 (Affine Variance for Noise)**.** Let $\frac{1}{p} + \frac{1}{q} = 1$. There exist $\sigma_0, \sigma_1 \geq 0$ such that for all $\boldsymbol{W} \in \mathbb{R}^{m\times n}$,

$$\mathbb{E}_\xi \left[\|\nabla f(\boldsymbol{W};\xi) - \nabla F(\boldsymbol{W})\|_F^2\right] \leq \sigma_0^2 + \sigma_1^2 \|\nabla F(\boldsymbol{W})\|_{S_q}^2. \tag{8}$$

Setting $\sigma_1 = 0$ the above recovers the standard bounded-variance assumption. Allowing $\sigma_1 > 0$ permits the variance bound to grow with $\|\nabla F(\boldsymbol{W})\|_{S_q}^2$.

To establish convergence, we first translate clipped Schatten smoothness in Assumption 3 into a bound on the objective change along an update. The segment form below provides the common descent estimate for our MVR and EMA analyses.

Lemma 1 (Descent Inequality in Clipped Schatten Smooothness)**.** *Let Assumptions 1 and 3 hold, let ρ_q be as in Lemma 3, and let $\boldsymbol{W}, \boldsymbol{Z} \in \mathbb{R}^{m\times n}$. Then:*

(i) (Segment Form.) *With $S := \sup_{\theta\in[0,1]} \|\nabla F(\boldsymbol{W} + \theta\boldsymbol{Z})\|_{S_q}$,*

$$F(\boldsymbol{W} + \boldsymbol{Z}) \leq F(\boldsymbol{W}) + \langle \nabla F(\boldsymbol{W}), \boldsymbol{Z}\rangle + \frac{L_0 + L_1 S}{2} \|\boldsymbol{Z}\|_{S_{\bar{p}}}^2. \tag{9}$$

(ii) (Endpoint Form.) *If $2\rho_q L_1 \|\boldsymbol{Z}\|_{S_{\bar{p}}} \leq 1$, then,*

$$F(\boldsymbol{W} + \boldsymbol{Z}) \leq F(\boldsymbol{W}) + \langle \nabla F(\boldsymbol{W}), \boldsymbol{Z}\rangle + \left(L_0 + L_1 \|\nabla F(\boldsymbol{W})\|_{S_q}\right) \|\boldsymbol{Z}\|_{S_{\bar{p}}}^2. \tag{10}$$

The proof of the above is given in Section A.

Theorem 3 (Convergence of SSMD with MVR)**.** *Consider Algorithm 1. Let $p, q \in [1, \infty]$ satisfy $1/p + 1/q = 1$, let $\boldsymbol{W}_0 \in \mathbb{R}^{m\times n}$ be a deterministic initial point, and let $r := \min\{m, n\}$. Suppose Assumptions 1, 3, 4, 5, and 2 hold. Fix any $c > 0$ and set $L_c := \max\{L_1, L_0/c\}$, $\sigma_c :=$*

$\max\{\sigma_1, \sigma_0/c\}$, *and* $\delta_c := \max\{\delta_1, \delta_0/c\}$. *Choose* $\eta_0 > 0$ *such that* $h_0 := \eta_0 r^{2\max\{\frac{1}{2}-\frac{1}{p},0\}} L_c < 2/3$. *For each horizon* T, *set* $a = a_T := T^{-2/3}$ *and* $\eta = \eta_T := \eta_0 T^{-2/3}$. *Then, for every integer* $T \geq T_0$,

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\,\|\nabla F(\boldsymbol{W}_t)\|_{S_q} \leq \frac{\mathcal{C}_1^{\mathrm{MVR}}}{T^{\frac{1}{3}}} + \frac{\mathcal{C}_2^{\mathrm{MVR}}}{T^{\frac{2}{3}}} = \mathcal{O}\left(T^{-\frac{1}{3}}\right), \tag{11}$$

where $\mathcal{C}_1^{\mathrm{MVR}} := 2[(F(\boldsymbol{W}_0)-F^*)/\eta_0 + 2r^{\max\{\frac{1}{q}-\frac{1}{2},0\}}\nu_0 + cb_0]/d_0$ *and* $\mathcal{C}_2^{\mathrm{MVR}} := ch_0/[d_0(1-h_0)]$, *with* $d_0 := (2-3h_0)/(2(1-h_0))$, $b_0 := 2r^{\max\{\frac{1}{q}-\frac{1}{2},0\}}(\sigma_c + r^{\max\{\frac{1}{2}-\frac{1}{q},0\}}\eta_0\delta_c/(1-h_0))/\sqrt{1-h_0}$, $\nu_0 := \sqrt{\sigma_0^2 + \sigma_1^2\|\nabla F(\boldsymbol{W}_0)\|_{S_q}^2}$, *and* $T_0 := \lceil \max\{1, (2b_0/d_0)^3\}\rceil$.

The proof is provided in Appendix B.

Theorem 3 establishes convergence to first-order stationarity, measured by the expected Schatten-q norm of the gradient, at rate $\mathcal{O}(T^{-1/3})$. For fixed problem parameters, returning an iterate sampled uniformly from $\{\boldsymbol{W}_0, \ldots, \boldsymbol{W}_{T-1}\}$ therefore yields an expected gradient norm at most ϵ within $\mathcal{O}(\epsilon^{-3})$ iterations. Each recursive iteration uses two stochastic-gradient evaluations with the same fresh sample, so the gradient-oracle complexity is also $\mathcal{O}(\epsilon^{-3})$.

The parameter $c > 0$ is an auxiliary reference scale for the gradient norm. It allows the smoothness and noise bounds to use a common factor $c + \|\nabla F(\boldsymbol{W})\|_{S_q}$, while retaining their additive contributions near stationary points. For example, $L_0 + L_1\|\nabla F(\boldsymbol{W})\|_{S_q} \leq L_c(c + \|\nabla F(\boldsymbol{W})\|_{S_q})$. The definitions of σ_c and δ_c provide corresponding noise bounds, and the same construction applies to the segment supremum of the gradient norm. Increasing c reduces the ratios defining L_c, σ_c, and δ_c, potentially relaxing the step-size restriction, but also increases the explicit factors of c in the convergence bound. Thus, c can be chosen to balance admissible step sizes and bound constants; any fixed choice independent of T preserves the stated convergence rates.

Increasing c reduces the ratios L_0/c, σ_0/c, and δ_0/c, while increasing the explicit factors of c in the convergence bound. Thus, c can be chosen jointly with η_0 to balance the admissible step size and the resulting constants, including the threshold T_0. Any fixed choice of c independent of T preserves the stated convergence rate.

Theorem 4 (Convergence of SSMD with EMA)**.** *Consider Algorithm 2. Let* $\boldsymbol{W}_0 \in \mathbb{R}^{m\times n}$ *be a deterministic initial point and let* $r := \min\{m, n\}$. *Suppose Assumptions 1, 2, 3 and 5 hold, and suppose each* ξ_t *is sampled independently of the preceding history. Fix any* $c > 0$ *and set* $L_c := \max\{L_1, L_0/c\}$ *and* $\sigma_c := \max\{\sigma_1, \sigma_0/c\}$. *For each horizon* T, *set* $\alpha = \alpha_T := T^{-1/2}$, $\beta = \beta_T := 1 - \alpha_T$, *and* $\eta = \eta_T := \eta_0 T^{-3/4}$. *Then, for every integer* $T \geq T_0$,

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\,\|\nabla F(\boldsymbol{W}_t)\|_{S_q} \leq \frac{\mathcal{C}_1^{\mathrm{EMA}}}{T^{\frac{1}{4}}} + \frac{\mathcal{C}_2^{\mathrm{EMA}}}{T^{\frac{1}{2}}} = \mathcal{O}\left(T^{-\frac{1}{4}}\right). \tag{12}$$

where $\mathcal{C}_1^{\mathrm{EMA}} := 2[(F(\boldsymbol{W}_0) - F^*)/\eta_0 + cb_0]/d_0$ *and* $\mathcal{C}_2^{\mathrm{EMA}} := 4r^{\max\{\frac{1}{q}-\frac{1}{2},0\}}\nu_0/d_0 + ch_0/[d_0(1-h_0)]$, *and* $\eta_0 > 0$ *has been set such that* $2/3 > h_0 := \eta_0 r^{2\max\{\frac{1}{2}-\frac{1}{p},0\}} L_c$, $d_0 := (2-3h_0)/(2(1-h_0))$, $b_0 := 2r^{\max\{\frac{1}{q}-\frac{1}{2},0\}}\sigma_c/\sqrt{1-h_0} + 2h_0/(1-h_0)$, $\nu_0 := \sqrt{\sigma_0^2 + \sigma_1^2\|\nabla F(\boldsymbol{W}_0)\|_{S_q}^2}$, *and* $T_0 := \lceil \max\{1, (2b_0/d_0)^4\}\rceil$.

The proof is provided in Appendix C

Theorem 4 establishes an $\mathcal{O}(T^{-1/4})$ convergence rate under generalized smoothness and affine noise varianc. For fixed problem parameters, an iterate sampled uniformly from $\{\boldsymbol{W}_0, \ldots, \boldsymbol{W}_{T-1}\}$ has expected Schatten-q gradient norm at most ϵ after $\mathcal{O}(\epsilon^{-4})$ iterations. Since EMA uses one stochastic-gradient evaluation per iteration, its gradient-oracle complexity is $\mathcal{O}(\epsilon^{-4})$.

Corollary 1 (Convergence of SSMD with Polyak's Heavy-Ball Momentum)**.** *Consider Algorithm 3. Let* $p, q \in [1, \infty]$ *satisfy* $1/p + 1/q = 1$, *let* $\boldsymbol{W}_0 \in \mathbb{R}^{m\times n}$ *be an arbitrary initial point, and let* $r := \min\{m, n\}$. *Suppose Assumptions 1, 2, 3 and 5, and hold. Fix* $c > 0$ *and choose* $\eta_0 > 0$ *satisfying the step-size condition in Theorem 4. Let* d_0, T_0, $\mathcal{C}_1^{\mathrm{EMA}}$, *and* $\mathcal{C}_2^{\mathrm{EMA}}$ *be the constants from that theorem, evaluated at the same* $\boldsymbol{W}_0$, c, *and* η_0. *For each horizon* T, *set* $\mu = \mu_T := 1 - T^{-1/2}$

and $\eta_t = \eta_T := \eta_0 T^{-3/4}$ *for* $t = 0, \dots, T-1$. *Then, for every integer* $T \geq T_0$,

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\|\nabla F(\boldsymbol{W}_t)\|_{S_q} \leq \frac{\mathcal{C}_1^{\mathrm{M}}}{T^{1/4}} + \frac{\mathcal{C}_2^{\mathrm{M}}}{T^{1/2}} = \mathcal{O}(T^{-1/4}), \tag{13}$$

where $\mathcal{C}_1^{\mathrm{M}} := \mathcal{C}_1^{\mathrm{EMA}}$ *and* $\mathcal{C}_2^{\mathrm{M}} := \mathcal{C}_2^{\mathrm{EMA}} + 4r^{\max\{\frac{1}{q}-\frac{1}{2},0\}}\|\nabla F(\boldsymbol{W}_0)\|_F/d_0$.

The proof of Corollary 1 is in Appendix C.3

Corollary 1 shows that Polyak's heavy-ball momentum achieves the same leading $\mathcal{O}(T^{-1/4})$ convergence rate as EMA under the stated assumptions. The connection follows from a key step in the proof, Lemma 11 in Appendix C: the scaled momentum $(1-\mu_T)\boldsymbol{M}_{t+1}$ satisfies the EMA recursion with $\beta_T = \mu_T$ and a correspondingly scaled initialization. Since positive scaling preserves the set of Schatten dual maximizers, this rescaling preserves the selected update directions.

4 Empirical Study of SSMD with MVR and EMA

We compare SSMD with MVR and EMA for $p \in \{1, 2, \infty\}$ on CIFAR-100 image classification (Krizhevsky et al., 2009) and GPT-2 pretraining (Jordan et al., 2024a). For CIFAR-100, we train DeiT-Tiny (Touvron et al., 2021) from scratch for 100 epochs, corresponding to 39,000 updates, with batch size 128. For GPT-2, we train a 124M-parameter model on tokenized FineWeb (Henriksson et al., 2025) for 6,000 updates, using sequence length 1024 and an effective batch size of 64 sequences. Detailed configurations and measurement procedures are provided in Appendix D.

Parameter Selection. We consider two experimental settings. The theorem-scaled settings use the horizon-dependent learning rates and momentum coefficients prescribed by the theorets proven. The empirical settings use larger learning rates, warmup, and cosine decay, with matched learning rates and coefficients $a = 1 - \beta$ for MVR and EMA at each fixed p.

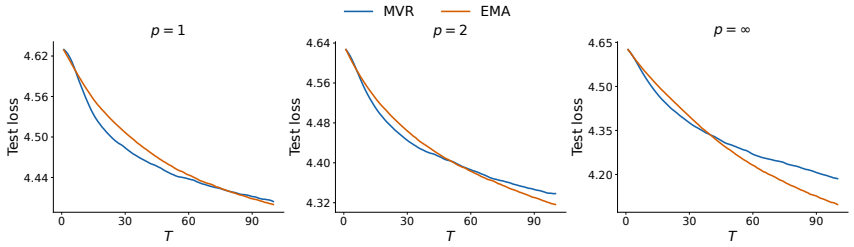


Figure 1: SSMD with MVR and EMA training DeiT-Tiny on CIFAR-100 with theorem-scaled parameters. From left to right, the panels correspond to $p = 1$, $p = 2$, and $p = \infty$.

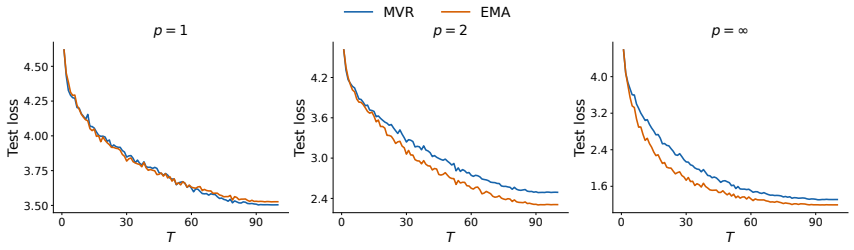


Figure 2: SSMD with MVR and EMA training DeiT-Tiny on CIFAR-100 under empirical parameter settings. From left to right, the panels correspond to $p = 1$, $p = 2$, and $p = \infty$.

Vision Transformer on CIFAR100 Under theorem-scaled parameters, Figure 1 shows that SSMD-MVR initially reduces test loss more quickly, while EMA reaches lower final test losses for all three geometries. Under empirical settings, Figure 2 shows similar final losses at $p = 1$, with a small advantage for MVR, whereas EMA achieves lower final losses at $p = 2$ and $p = \infty$. Thus, the early advantage of MVR does not consistently translate into a lower final test loss.

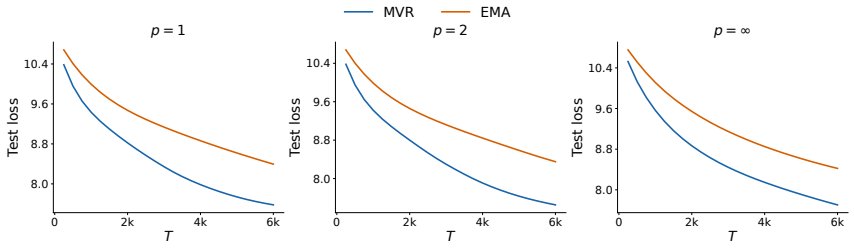


Figure 3: SSMD with MVR and EMA training GPT-2 on FineWeb with theorem-scaled parameters. From left to right, the panels correspond to $p = 1$, $p = 2$, and $p = \infty$.

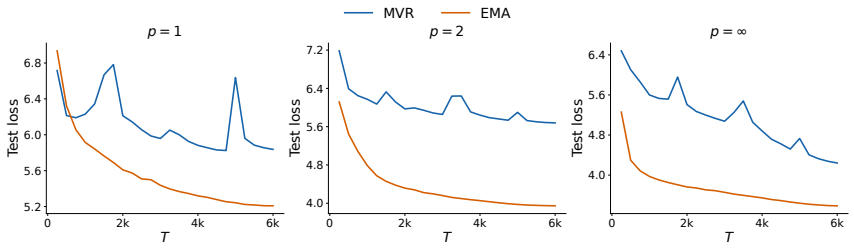


Figure 4: SSMD with MVR and EMA training GPT-2 on FineWeb under empirical settings. From left to right, the panels correspond to $p = 1$, $p = 2$, and $p = \infty$.

GPT-2 Pretraining With theorem-scaled parameters, Figure 3 shows lower test losses for MVR across all three geometries. Under empirical settings, Figure 4 instead shows lower final test losses for EMA, while MVR exhibits temporary increases in test loss, particularly at $p = 1$. Overall, both MVR and EMA achieve lower final test losses under empirical settings than under theorem-scaled settings on both tasks and across all three values of p. EMA shows a larger improvement in most cases, while MVR retains a small final-loss advantage on DeiT-Tiny at $p = 1$.

5 Conclusion

We established convergence guarantees for Stochastic Schatten Momentum Descent under relaxed clipped Schatten smoothness and gradient-dependent noise variance. With an additional affine bound on centered gradient-difference variance, the MVR variant is shown to achieve the optimal $\mathcal{O}(\epsilon^{-3})$ rate for first-order stationarity, while EMA and Polyak's heavy-ball momentum versions are shown to achieve the $\mathcal{O}(\epsilon^{-4})$ rate without this additional condition. These results cover the Schatten-p family, including idealized Muon at $p = \infty$, without requiring uniformly bounded noise variance. Our CIFAR-100 and GPT-2 experiments show that the relative performance of MVR and EMA depends on the training configuration, revealing a gap between stationarity guarantees and practical training performance. We posit that exciting future work is possible to understand training performance of these algorithms.

Next, it would be interesting to identify nonconvex objective classes that capture important features of LLM training and under which these methods provably converge to global minima. Secondly, the analysis here can possibly be extended to account for practical modifications such as Hyperball (Wen et al., 2026). For example, MuonH achieves 20–30% token-equivalent speedup over an AdamW scaling-law baseline at 1.2B parameters (Wen et al., 2026).

Understanding how its norm constraints and update rescaling interact with Schatten geometry could clarify when and how these modifications improve training stability or efficiency. Third, the explicit batch-size dependence in the convergence guarantees remains unknown for such algorithms as studied here – particularly about accounting for both gradient noise and the centered variance error. Such an analysis could guide the joint choice of batch size, learning rate, and momentum, and help explain how optimization efficiency changes as model size, data, and parallel computation increase (McCandlish et al., 2018; Shulgin et al., 2026c).

AI USE STATEMENT

We used generative AI tools to assist with language polishing, grammar and formatting checks, translation, code generation for implementing parts of our method, identifying related literature, and assisting in the writing of proofs. All AI-generated code used in this work was reviewed by the authors. We did not use generative AI tools for generating synthetic datasets, developing theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proofs, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, cleaning or reformatting datasets, supporting qualitative or thematic data analysis, or interpreting results. The authors take full responsibility for the final content of this work, including all text and code produced with AI assistance.

REFERENCES

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Ayush Sekhari, and Karthik Sridharan. Second-order information in non-convex stochastic optimization: Power and limitations. In Jacob Abernethy and Shivani Agarwal (eds.), Proceedings of Thirty Third Conference on Learning Theory, volume 125 of Proceedings of Machine Learning Research, pp. 242–299. PMLR, 09–12 Jul 2020. URL https://proceedings.mlr.press/v125/arjevani20a.html.

Yossi Arjevani, Yair Carmon, John C. Duchi, Dylan J. Foster, Nathan Srebro, and Blake Woodworth. Lower bounds for non-convex stochastic optimization. Mathematical Programming, 199(1):165–214, 2023. doi: 10.1007/s10107-022-01822-7. URL https://arxiv.org/abs/1912.02365.

Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. arXiv preprint arXiv:2409.20325, 2024. URL https://arxiv.org/abs/2409.20325.

Jeremy Bernstein and Laker Newhouse. Modular duality in deep learning. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), Proceedings of the 42nd International Conference on Machine Learning, volume 267 of Proceedings of Machine Learning Research, pp. 3920–3930. PMLR, 13–19 Jul 2025. URL https://proceedings.mlr.press/v267/bernstein25a.html.

Jeremy Bernstein, Yu-Xiang Wang, Kamyar Azizzadenesheli, and Animashree Anandkumar. signsgd: Compressed optimisation for non-convex problems. In International conference on machine learning, pp. 560–569. PMLR, 2018.

Franz Louis Cesista. Muon and a selective survey on steepest descent in Riemannian and non-Riemannian manifolds, 2025. URL https://leloykun.github.io/ponder/steepest-descent-non-riemannian/.

El Mahdi Chayti. RanSOM: Second-order momentum with randomized scaling for constrained and unconstrained optimization. arXiv preprint arXiv:2602.06824, 2026. URL https://arxiv.org/abs/2602.06824.

Frank H. Clarke. Optimization and Nonsmooth Analysis. Canadian Mathematical Society Series of Monographs and Advanced Texts. Wiley-Interscience, 1983.

Ashok Cutkosky and Francesco Orabona. Momentum-based variance reduction in non-convex SGD. In Advances in Neural Information Processing Systems, volume 32, 2019. URL https://arxiv.org/abs/1905.10018.

Chen Fan, Mark Schmidt, and Christos Thrampoulidis. Implicit bias of spectral descent and muon on multiclass separable data. Advances in Neural Information Processing Systems, 38:39622–39669, 2026.

Cong Fang, Chris Junchi Li, Zhouchen Lin, and Tong Zhang. Spider: Near-optimal non-convex optimization via stochastic path-integrated differential estimator. Advances in neural information processing systems, 31, 2018.

Erik Henriksson, Otto Tarkka, and Filip Ginter. Finerweb-10bt: Refining web data with llm-based line-level filtering. In Proceedings of the Joint 25th Nordic Conference on Computational Linguistics and 11th Baltic Conference on Human Language Technologies (NoDaLiDa/Baltic-HLT 2025), pp. 258–268, 2025.

Feihu Huang, Yuning Luo, and Songcan Chen. Limuon: Light and fast muon optimizer for large models. In Forty-third International Conference on Machine Learning, 2026. URL https://openreview.net/forum?id=mf8N8PTUAa.

Keller Jordan, Jeremy Bernstein, Brendan Rappazzo, @fernbear.bsky.social, Boza Vlado, You Jiacheng, Franz Cesista, Braden Koszarsky, and @Grad62304977. modded-nanogpt: Speedrunning the nanogpt baseline, 2024a. URL https://github.com/KellerJordan/modded-nanogpt.

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

Ahmed Khaled and Peter Richtárik. Better theory for SGD in the nonconvex world. Transactions on Machine Learning Research, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=AU4qHN2VkS. Survey Certification.

Sarit Khirirat, Abdurakhmon Sadiev, Yury Demidovich, and Peter Richtárik. Better LMO-based momentum methods with second-order information. arXiv preprint arXiv:2512.13227, 2025. URL https://arxiv.org/abs/2512.13227.

Kimi Team. Kimi K2: Open agentic intelligence. arXiv preprint arXiv:2507.20534, 2025. URL https://arxiv.org/abs/2507.20534.

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

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

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

Jiaxiang Li and Mingyi Hong. A note on the convergence of Muon and further. arXiv preprint arXiv:2502.02900, 2025. URL https://arxiv.org/abs/2502.02900.

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, 2025a. URL https://arxiv.org/abs/2502.16982.

Yifeng Liu, Angela Yuan, and Quanquan Gu. MARS-M: When variance reduction meets matrices. arXiv preprint arXiv:2510.21800, 2025b. URL https://arxiv.org/abs/2510.21800.

Sam McCandlish, Jared Kaplan, Dario Amodei, and OpenAI Dota Team. An empirical model of large-batch training, 2018. URL https://arxiv.org/abs/1812.06162.

Lam M Nguyen, Jie Liu, Katya Scheinberg, and Martin Takáč. Sarah: A novel method for machine learning problems using stochastic recursive gradient. In International conference on machine learning, pp. 2613–2621. PMLR, 2017.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Training deep learning models with norm-constrained LMOs. In Proceedings of the 42nd International Conference on Machine Learning, volume 267 of Proceedings of Machine Learning Research, pp. 49069–49104. PMLR, 2025a. URL https://arxiv.org/abs/2502.07529.

Thomas Pethick, Wanyun Xie, Kimon Antonakopoulos, Zhenyu Zhu, Antonio Silveti-Falls, and Volkan Cevher. Scion. GitHub repository, 2025b. URL https://github.com/LIONS-EPFL/scion.git.

Thomas Pethick, Wanyun Xie, Mete Erdogan, Kimon Antonakopoulos, Antonio Silveti-Falls, and Volkan Cevher. Generalized gradient norm clipping & non-euclidean (l_0, l_1)-smoothness. Advances in Neural Information Processing Systems, 38:21170–21208, 2026.

Xun Qian, Hussein Rammal, Dmitry Kovalev, and Peter Richtárik. Muon is provably faster with momentum variance reduction. arXiv preprint arXiv:2512.16598, 2025. URL https://arxiv.org/abs/2512.16598.

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. URL https://arxiv.org/abs/2505.13416.

Saber Salehkaleybar, Sadegh Khorasani, Negar Kiyavash, Niao He, and Patrick Thiran. Momentum-based policy gradient with second-order information. arXiv preprint arXiv:2205.08253, 2022. URL https://arxiv.org/abs/2205.08253.

Naoki Sato, Hiroki Naganuma, and Hideaki Iiduka. Analysis of Muon's convergence and critical batch size. arXiv preprint arXiv:2507.01598, 2025. URL https://arxiv.org/abs/2507.01598.

Maria-Eleni Sfyraki and Jun-Kun Wang. Lions and muons: Optimization via stochastic Frank–Wolfe under heavy-tailed noise. arXiv preprint arXiv:2506.04192, 2025. URL https://arxiv.org/abs/2506.04192.

Ishaan Shah, Anthony M. Polloreno, Karl Stratos, Philip Monk, Adarsh Chaluvaraju, Andrew Hojel, Andrew Ma, Anil Thomas, Ashish Tanwer, Darsh J. Shah, et al. Practical efficiency of Muon for pretraining. arXiv preprint arXiv:2505.02222, 2025. URL https://arxiv.org/abs/2505.02222.

Wei Shen, Ruichuan Huang, Minhui Huang, Cong Shen, and Jiawei Zhang. On the convergence analysis of muon. Transactions on Machine Learning Research, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=4nH4CulGaP. J2C Certification.

Egor Shulgin, Sultan AlRashed, Peter Richtárik, and Francesco Orabona. Beyond the ideal: Analyzing the inexact muon update. In The 29th International Conference on Artificial Intelligence and Statistics, 2026a. URL https://openreview.net/forum?id=IBRMWPBouf.

Egor Shulgin, Mohamed Awad, Peter Richtárik, and Eduard Gorbunov. General analysis of LMO-based optimizers: Beyond bounded variance. In Forty-third International Conference on Machine Learning, 2026b. URL https://openreview.net/forum?id=s0Q8nVsrMu.

Egor Shulgin, Dimitri von Rütte, Tianyue H. Zhang, Niccolò Ajroldi, Bernhard Schölkopf, and Antonio Orvieto. Deriving hyperparameter scaling laws via modern optimization theory, 2026c. URL https://arxiv.org/abs/2603.15958.

Hugo Touvron, Matthieu Cord, Matthijs Douze, Francisco Massa, Alexandre Sablayrolles, and Hervé Jégou. Training data-efficient image transformers & distillation through attention. In International conference on machine learning, pp. 10347–10357. PMLR, 2021.

Hoang Tran and Ashok Cutkosky. Better SGD using second-order momentum. In Advances in Neural Information Processing Systems, volume 35, 2022. URL https://arxiv.org/abs/2103.03265.

John von Neumann. Some matrix-inequalities and metrization of matric-space. Tomsk University Review, 1:286–300, 1937. Reprinted in Collected Works, Vol. IV, Pergamon, 1962, pp. 205–219.

G. Alistair Watson. Characterization of the subdifferential of some matrix norms. Linear Algebra and its Applications, 170:33–45, 1992. doi: 10.1016/0024-3795(92)90407-2.

Kaiyue Wen, Xingyu Dang, Kaifeng Lyu, Tengyu Ma, and Percy Liang. Fantastic pretraining optimizers and where to find them ii: Hyperball optimization, 2026. URL https://arxiv.org/abs/2606.16899.

Huizhuo Yuan, Yifeng Liu, Shuang Wu, Xun Zhou, and Quanquan Gu. MARS: Unleashing the power of variance reduction for training large models. In Proceedings of the 42nd International Conference on Machine Learning, volume 267 of Proceedings of Machine Learning Research. PMLR, 2025. URL https://arxiv.org/abs/2411.10438.

Jingzhao Zhang, Tianxing He, Suvrit Sra, and Ali Jadbabaie. Why gradient clipping accelerates training: A theoretical justification for adaptivity. In International Conference on Learning Representations, 2020a. URL https://arxiv.org/abs/1905.11881.

Mingrui Zhang, Zebang Shen, Aryan Mokhtari, Hamed Hassani, and Amin Karbasi. One sample stochastic Frank-Wolfe. In Proceedings of the 23rd International Conference on Artificial Intelligence and Statistics, volume 108 of Proceedings of Machine Learning Research, pp. 4012–4023. PMLR, 2020b. URL https://arxiv.org/abs/1910.04322.

Dongruo Zhou, Pan Xu, and Quanquan Gu. Stochastic nested variance reduction for nonconvex optimization. Journal of Machine Learning Research, 21(103):1–63, 2020. URL http://jmlr.org/papers/v21/18-447.html.

APPENDIX CONTENTS

A Background Mathematics **15**

A.1 Additional Properties of Schatten Geometry 15

A.2 Proofs of Properties of Schatten Geometry 16

A.3 Proof from Section 2 18

B Convergence of SSMD-MVR **21**

B.1 Proof of Theorem 3 21

B.2 Proof of the Intermediate Lemmas 23

C Convergence of SSMD-EMA **28**

C.1 Proof of Theorem 4 28

C.2 Proof of the Intermediate Lemma 30

C.3 Convergence of SSMD(Polyak's Heavy Ball) 32

D Experimental Details **34**

D.1 CIFAR-100 34

D.2 GPT-2 Pretraining 34

D.3 Theorem-Scaled Parameters 34

D.4 Empirical Parameters 35

D.5 Hyperparameter Search Ranges 35

D.6 Additional results 35

E Curvature-Corrected Momentum: A Unified Template and a Comparison of Seven Methods **38**

E.1 Motivation: what SARAH and SPIDER leave unsettled 38

E.2 Problem setting 38

E.3 A template covering all seven methods 39

E.4 Reduction of each published method to the template 39

E.5 Oracle models and assumptions 41

E.6 What each theorem actually bounds 43

E.7 Which results reach ϵ^{-3} without a Hessian-vector-product oracle or Hessian smoothness 44

E.8 Why every analysis passes through the same step 46

E.9 Empirical evidence, and a caveat about it 47

A Background Mathematics

A.1 Additional Properties of Schatten Geometry

Throughout this subsection, $p, q \in [1, \infty]$ are conjugate and $r := \min\{m, n\}$. We use the compact SVD $\boldsymbol{G} = \boldsymbol{U} \operatorname{diag}(\sigma_1, \ldots, \sigma_k)\boldsymbol{V}^\top$, where $k = \operatorname{rank}(\boldsymbol{G})$ and $\sigma_1 \geq \cdots \geq \sigma_k > 0$. A singular value is called simple if it has multiplicity one. For $\boldsymbol{G} = \boldsymbol{0}$, the compact SVD has no columns.

Lemma 2 (Schatten Hölder inequality). *For all $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m\times n}$, $|\langle \boldsymbol{A}, \boldsymbol{B}\rangle| \leq \|\boldsymbol{A}\|_{S_q}\|\boldsymbol{B}\|_{S_p}$.*

The proof is given in Section A.3.1.

Lemma 3 (Schatten Norm Comparison). *Let $\bar{q} := \max\{q, 2\}$ and $\rho_q := r^{\max\{1/2-1/p,0\}} = r^{\max\{1/q-1/2,0\}}$. For every $\boldsymbol{X} \in \mathbb{R}^{m\times n}$,*

$$\|\boldsymbol{X}\|_{S_{\bar{q}}} \leq \|\boldsymbol{X}\|_{S_q} \leq \rho_q \|\boldsymbol{X}\|_{S_{\bar{q}}}. \tag{14}$$

Both inequalities are sharp.

The proof is given in Section A.3.2

Lemma 4 (Duality of the Schatten LMO). *For conjugate $p, q \in [1, \infty]$, the following statements hold.*

(i) *The maximum $\max_{\|\boldsymbol{D}\|_{S_p}\leq 1}\langle \boldsymbol{G}, \boldsymbol{D}\rangle$ is attained and equals $\|\boldsymbol{G}\|_{S_q}$. If $\boldsymbol{G} \neq \boldsymbol{0}$, every maximizer has Schatten-p norm one.*

(ii) *For $\boldsymbol{G} \neq \boldsymbol{0}$, the selection in equation 1 is a maximizer.*

(iii) *For every $\boldsymbol{G}$, including $\boldsymbol{G} = \boldsymbol{0}$,*

$$\operatorname*{arg\,max}_{\|\boldsymbol{D}\|_{S_p}\leq 1} \langle \boldsymbol{G}, \boldsymbol{D}\rangle = \partial \|\boldsymbol{G}\|_{S_q}. \tag{15}$$

The maximizer is unique if and only if $\|\cdot\|_{S_q}$ is differentiable at $\boldsymbol{G}$, in which case $\mathcal{D}_p(\boldsymbol{G}) = \nabla_{\boldsymbol{G}}\|\boldsymbol{G}\|_{S_q}$.

The proof is given in Section A.3.3.

Remark 1 (Selections at nonsmooth points). The Schatten-p update direction is determined by the dual Schatten-q norm. The set identity equation 15 remains valid when that norm is not differentiable, including at $\boldsymbol{G} = \boldsymbol{0}$, at rank-deficient matrices when $q = 1$, and at matrices with a repeated largest singular value when $q = \infty$. A measurable selection makes the update well-defined in these cases.

Lemma 5 (Differentiability of Schatten norms). *Let $s \in (1, \infty)$ and $\boldsymbol{G} \neq \boldsymbol{0}$. Then $\|\cdot\|_{S_s}$ is differentiable at $\boldsymbol{G}$, without requiring distinct singular values or full rank, and*

$$\nabla_{\boldsymbol{G}}\|\boldsymbol{G}\|_{S_s} = \boldsymbol{U} \operatorname{diag}(s_1(s), \ldots, s_k(s))\boldsymbol{V}^\top, \qquad s_i(s) := \frac{\sigma_i^{s-1}}{\|\boldsymbol{G}\|_{S_s}^{s-1}}. \tag{16}$$

The nuclear norm is differentiable at $\boldsymbol{G}$ if and only if $\operatorname{rank}(\boldsymbol{G}) = r$. The spectral norm is differentiable at a nonzero $\boldsymbol{G}$ if and only if its largest singular value is simple. No Schatten norm is differentiable at $\boldsymbol{G} = \boldsymbol{0}$.

The proof is given in Section A.3.4.

Remark 2 (Perturbation interpretation). At a simple positive singular value, differentiating $\boldsymbol{G}^\top\boldsymbol{G}\boldsymbol{v}_i = \sigma_i^2\boldsymbol{v}_i$ gives $d\sigma_i = \boldsymbol{u}_i^\top(d\boldsymbol{G})\boldsymbol{v}_i = \langle \boldsymbol{u}_i\boldsymbol{v}_i^\top, d\boldsymbol{G}\rangle$. Applying the chain rule yields equation 16. This argument directly covers simple singular values; Lemma 5 also covers repeated singular values.

Lemma 6 (Subdifferential of the nuclear norm). *For every $\boldsymbol{G} \in \mathbb{R}^{m\times n}$,*

$$\partial\|\boldsymbol{G}\|_{S_1} = \left\{\boldsymbol{U}\boldsymbol{V}^\top + \boldsymbol{Q} : \boldsymbol{U}^\top\boldsymbol{Q} = \boldsymbol{0},\ \boldsymbol{Q}\boldsymbol{V} = \boldsymbol{0},\ \|\boldsymbol{Q}\|_{S_\infty} \leq 1\right\}. \tag{17}$$

If $\operatorname{rank}(\boldsymbol{G}) = r$, these constraints force $\boldsymbol{Q} = \boldsymbol{0}$, so $\nabla_{\boldsymbol{G}}\|\boldsymbol{G}\|_{S_1} = \boldsymbol{U}\boldsymbol{V}^\top$.

Direction	Primal norm	Dual norm	$\mathcal{D}_p(\boldsymbol{M})$
Muon-type polar	S_∞	S_1	$\boldsymbol{U}\boldsymbol{V}^\top$
Frobenius-normalized	S_2	S_2	$\boldsymbol{M}/\|\boldsymbol{M}\|_F$
Rank-one	S_1	S_∞	$\boldsymbol{u}_1\boldsymbol{v}_1^\top$

Table 3: Schatten update directions for a nonzero momentum matrix $\boldsymbol{M}$, with singular vectors defined by its compact SVD. The primal and dual indices are conjugate.

The proof is given in Section A.3.6.

Proposition 1 (Spectral concentration)**.** *Let $\boldsymbol{G} \neq \boldsymbol{0}$ and let $s_i(s)$ be defined in equation 16. For $s \in (1, \infty)$ and $i \leq k$,*

$$\frac{s_i(s)}{s_1(s)} = \left(\frac{\sigma_i}{\sigma_1}\right)^{s-1}. \tag{18}$$

If $\sigma_i < \sigma_1$, this ratio is strictly decreasing in s, with

$$\lim_{s\downarrow 1} \frac{s_i(s)}{s_1(s)} = 1, \qquad \lim_{s\to\infty} \frac{s_i(s)}{s_1(s)} = 0. \tag{19}$$

Moreover, if the largest singular value has multiplicity d,

$$\lim_{s\downarrow 1} \nabla_{\boldsymbol{G}} \|\boldsymbol{G}\|_{S_s} = \boldsymbol{U}\boldsymbol{V}^\top, \qquad \lim_{s\to\infty} \nabla_{\boldsymbol{G}} \|\boldsymbol{G}\|_{S_s} = \frac{1}{d}\sum_{i=1}^{d} \boldsymbol{u}_i\boldsymbol{v}_i^\top. \tag{20}$$

In particular, the second limit is $\boldsymbol{u}_1\boldsymbol{v}_1^\top$ when the largest singular value is simple.

Proof. The ratio identity follows directly from equation 16, and its monotonicity and limits follow from $0 < \sigma_i/\sigma_1 < 1$. As $s \downarrow 1$, every $s_i(s)$ tends to one. For the other limit, write

$$s_i(s) = \frac{(\sigma_i/\sigma_1)^{s-1}}{\left(d + \sum_{j=d+1}^{k} (\sigma_j/\sigma_1)^s\right)^{(s-1)/s}}. \tag{21}$$

The denominator tends to d, while the numerator equals one for $i \leq d$ and tends to zero for $i > d$. Substitution into equation 16 proves the result. □

Remark 3 (Relation to Signum). Signum uses the entrywise ℓ_∞ geometry, which generally differs from Schatten geometry on the original matrix space $\mathbb{R}^{m\times n}$. A diagonal embedding relates the two constructions: if $\boldsymbol{m} = \mathrm{vec}(\boldsymbol{M})$ and $\boldsymbol{Z} = \mathrm{diag}(\boldsymbol{m})$, then $\|\boldsymbol{Z}\|_{S_s} = \|\boldsymbol{m}\|_{\ell_s}$. Thus, the sign direction can be viewed as Schatten steepest descent restricted to a diagonal subspace of a larger matrix space. This embedding does not make the entrywise sign update a Schatten update on the original matrix space. At zero entries, the usual convention $\mathrm{sign}(0) = 0$ specifies a subgradient selection.

A.2 Proofs of Properties of Schatten Geometry

Lemma 7 (Smoothness Transfer between Schatten Norms)**.** *Let $p, s \in [1, \infty]$, let $r := \min\{m, n\}$, and let p^*, s^* denote their respective conjugate exponents, with $1/\infty := 0$. Suppose that $F : \mathbb{R}^{m\times n} \to \mathbb{R}$ is differentiable and L_p-smooth with respect to the Schatten-p norm, meaning that*

$$\|\nabla F(\boldsymbol{W}_1) - \nabla F(\boldsymbol{W}_2)\|_{S_{p^*}} \leq L_p \|\boldsymbol{W}_1 - \boldsymbol{W}_2\|_{S_p} \tag{22}$$

for all $\boldsymbol{W}_1, \boldsymbol{W}_2 \in \mathbb{R}^{m\times n}$. Then F is L_s-smooth with respect to the Schatten-s norm, where one may take

$$L_s := L_p r^{2\max\{1/p - 1/s, 0\}} = \begin{cases} L_p, & p \geq s, \\ L_p r^{2(1/p-1/s)}, & p < s. \end{cases} \tag{23}$$

In particular, for all $\boldsymbol{W}, \boldsymbol{X} \in \mathbb{R}^{m\times n}$,

$$F(\boldsymbol{W} + \boldsymbol{X}) \leq F(\boldsymbol{W}) + \langle \nabla F(\boldsymbol{W}), \boldsymbol{X}\rangle_F + \frac{L_s}{2}\|\boldsymbol{X}\|_{S_s}^2. \tag{24}$$

Proof of Lemma 7. Finite-dimensional norm comparison applied to the singular values gives

$$\|\boldsymbol{X}\|_{S_p} \le r^{\max\{1/p-1/s,0\}}\|\boldsymbol{X}\|_{S_s}, \tag{25}$$

$$\|\boldsymbol{Y}\|_{S_{s^*}} \le r^{\max\{1/p-1/s,0\}}\|\boldsymbol{Y}\|_{S_{p^*}}, \tag{26}$$

where the second inequality uses $1/s^* - 1/p^* = 1/p - 1/s$. Consequently,

$$\|\nabla F(\boldsymbol{W}_1) - \nabla F(\boldsymbol{W}_2)\|_{S_{s^*}} \le r^{\max\{1/p-1/s,0\}}\|\nabla F(\boldsymbol{W}_1) - \nabla F(\boldsymbol{W}_2)\|_{S_{p^*}} \tag{27}$$

$$\le L_p r^{\max\{1/p-1/s,0\}}\|\boldsymbol{W}_1 - \boldsymbol{W}_2\|_{S_p} \tag{28}$$

$$\le L_s\|\boldsymbol{W}_1 - \boldsymbol{W}_2\|_{S_s}. \tag{29}$$

This establishes the claimed gradient Lipschitz bound. Integrating along the segment from $\boldsymbol{W}$ to $\boldsymbol{W}+\boldsymbol{X}$ and applying Schatten Hölder inequality yields

$$\begin{aligned} &F(\boldsymbol{W}+\boldsymbol{X}) - F(\boldsymbol{W}) - \langle\nabla F(\boldsymbol{W}), \boldsymbol{X}\rangle_F \\ &= \int_0^1 \langle\nabla F(\boldsymbol{W}+t\boldsymbol{X}) - \nabla F(\boldsymbol{W}), \boldsymbol{X}\rangle_F\, dt \\ &\le \int_0^1 L_s t\|\boldsymbol{X}\|_{S_s}^2\, dt = \frac{L_s}{2}\|\boldsymbol{X}\|_{S_s}^2. \end{aligned} \tag{30}$$

□

Proof of Lemma 1. Assumption 3 holds throughout $\mathbb{R}^{m\times n}$ and hence along the segment $\{\boldsymbol{W}+\theta\boldsymbol{Z} : \theta\in[0,1]\}$. By Assumption 1, ∇F is continuous. Since the segment is compact, $S<\infty$.

(i) By the fundamental theorem of calculus along the segment,

$$F(\boldsymbol{W}+\boldsymbol{Z}) - F(\boldsymbol{W}) - \langle\nabla F(\boldsymbol{W}), \boldsymbol{Z}\rangle = \int_0^1 \langle\nabla F(\boldsymbol{W}+\theta\boldsymbol{Z}) - \nabla F(\boldsymbol{W}),\, \boldsymbol{Z}\rangle\, d\theta. \tag{31}$$

Since $\frac{1}{\bar p}+\frac{1}{\bar q}=1$, the von Neumann trace inequality followed by Hölder on the singular values gives $\langle\boldsymbol{A},\boldsymbol{B}\rangle \le \|\boldsymbol{A}\|_{S_{\bar q}}\|\boldsymbol{B}\|_{S_{\bar p}}$, whence

$$F(\boldsymbol{W}+\boldsymbol{Z}) - F(\boldsymbol{W}) - \langle\nabla F(\boldsymbol{W}), \boldsymbol{Z}\rangle \le \|\boldsymbol{Z}\|_{S_{\bar p}} \int_0^1 \|\nabla F(\boldsymbol{W}+\theta\boldsymbol{Z}) - \nabla F(\boldsymbol{W})\|_{S_{\bar q}}\, d\theta. \tag{32}$$

Apply Assumption 3 with $\boldsymbol{W}_1 = \boldsymbol{W}+\theta\boldsymbol{Z}$ and $\boldsymbol{W}_2 = \boldsymbol{W}$. The segment joining $\boldsymbol{W}_1$ and $\boldsymbol{W}_2$ is $\{\boldsymbol{W}+\theta'\theta\boldsymbol{Z} : \theta'\in[0,1]\}$, a subset of $\{\boldsymbol{W}+s\boldsymbol{Z} : s\in[0,1]\}$, so its supremum term is bounded by S; and $\|\boldsymbol{W}_1 - \boldsymbol{W}_2\|_{S_{\bar p}} = \theta\|\boldsymbol{Z}\|_{S_{\bar p}}$. Therefore

$$\|\nabla F(\boldsymbol{W}+\theta\boldsymbol{Z}) - \nabla F(\boldsymbol{W})\|_{S_{\bar q}} \le (L_0 + L_1 S)\,\theta\,\|\boldsymbol{Z}\|_{S_{\bar p}}. \tag{33}$$

Substituting equation 33 into equation 32 and using $\int_0^1 \theta\, d\theta = \frac{1}{2}$ yields equation 9.

(ii) Write $g(\theta) := \|\nabla F(\boldsymbol{W}+\theta\boldsymbol{Z})\|_{S_q}$, so that $S = \sup_{\theta\in[0,1]} g(\theta)$. By the triangle inequality, the right inequality of equation 14, and equation 33,

$$\begin{aligned} g(\theta) &\le g(0) + \|\nabla F(\boldsymbol{W}+\theta\boldsymbol{Z}) - \nabla F(\boldsymbol{W})\|_{S_q} \\ &\le g(0) + \rho_q\,\|\nabla F(\boldsymbol{W}+\theta\boldsymbol{Z}) - \nabla F(\boldsymbol{W})\|_{S_{\bar q}} \\ &\le g(0) + \rho_q\,(L_0 + L_1 S)\,\theta\,\|\boldsymbol{Z}\|_{S_{\bar p}} \;\le\; g(0) + \rho_q\,(L_0 + L_1 S)\,\|\boldsymbol{Z}\|_{S_{\bar p}}. \end{aligned} \tag{34}$$

This holds for every $\theta\in[0,1]$; taking the supremum on the left and using $S<\infty$ to rearrange,

$$S\left(1 - \rho_q L_1\,\|\boldsymbol{Z}\|_{S_{\bar p}}\right) \le g(0) + \rho_q L_0\,\|\boldsymbol{Z}\|_{S_{\bar p}}. \tag{35}$$

By the assumptions on $\boldsymbol{Z}$ we have $\rho_q L_1\|\boldsymbol{Z}\|_{S_{\bar p}} \le \frac{1}{2}$, so the bracket is at least $\frac{1}{2}$ and

$$S \le 2g(0) + 2\rho_q L_0\,\|\boldsymbol{Z}\|_{S_{\bar p}}. \tag{36}$$

Hence, using equation 36 and then $2\rho_q L_1\|\boldsymbol{Z}\|_{S_{\bar p}} \le 1$ once more,

$$L_0 + L_1 S \;\le\; L_0 + 2L_1 g(0) + 2\rho_q L_0 L_1\,\|\boldsymbol{Z}\|_{S_{\bar p}} \;\le\; 2L_0 + 2L_1 g(0). \tag{37}$$

Substituting this into equation 9 and recalling $g(0) = \|\nabla F(\boldsymbol{W})\|_{S_q}$ gives equation 10. □

A.3 Proof from Section 2

A.3.1 Proof of Lemma 2

By von Neumann's trace inequality (von Neumann, 1937), for any $\boldsymbol{A}, \boldsymbol{B} \in \mathbb{R}^{m\times n}$, $|\langle \boldsymbol{A}, \boldsymbol{B}\rangle| = |\mathrm{Tr}(\boldsymbol{A}^\top \boldsymbol{B})| \le \sum_{i=1}^{r} \sigma_i(\boldsymbol{A})\sigma_i(\boldsymbol{B})$. Applying the vector Hölder inequality to the singular-value vectors with conjugate exponents q and p,

$$\sum_{i=1}^{r} \sigma_i(\boldsymbol{A})\sigma_i(\boldsymbol{B}) \;\le\; \Big(\sum_i \sigma_i(\boldsymbol{A})^q\Big)^{1/q}\Big(\sum_i \sigma_i(\boldsymbol{B})^p\Big)^{1/p} \;=\; \|\boldsymbol{A}\|_{S_q}\|\boldsymbol{B}\|_{S_p}, \tag{38}$$

which is the claim. The cases $(p, q) = (1, \infty)$ and $(\infty, 1)$ follow from the same computation with the usual ℓ_∞ convention. □

A.3.2 Proof of Lemma 3

Write $\boldsymbol{\sigma}(\boldsymbol{X}) \in \mathbb{R}^r_{\ge 0}$ for the singular values, so that $\|\boldsymbol{X}\|_{S_a} = \|\boldsymbol{\sigma}(\boldsymbol{X})\|_{\ell_a}$. Since $a \mapsto \|\boldsymbol{v}\|_{\ell_a}$ is non-increasing and $\bar{q} = \max(q, 2) \ge q$, the left inequality in equation 14 is immediate.

For the right inequality, if $p \le 2$ then $q \ge 2$, so $\bar{q} = q$, $\rho_q = 1$, and there is nothing to prove. If $p > 2$ then $q < 2$ and $\bar{q} = 2$, and the standard ℓ_a comparison on $\mathbb{R}^r$ gives $\|\boldsymbol{v}\|_{\ell_q} \le r^{\frac{1}{q}-\frac{1}{2}}\|\boldsymbol{v}\|_{\ell_2}$. Since $\frac{1}{q} - \frac{1}{2} = \big(1 - \frac{1}{p}\big) - \frac{1}{2} = \frac{1}{2} - \frac{1}{p}$, this is the claim. Equality holds on the left when $q = \bar{q}$, and on the right at any $\boldsymbol{X}$ with flat spectrum $\sigma_1 = \cdots = \sigma_r$.

A.3.3 Proof of Lemma 4

(i) and (ii). Lemma 2 gives $\langle \boldsymbol{G}, \boldsymbol{D}\rangle \le \|\boldsymbol{G}\|_{S_q}\|\boldsymbol{D}\|_{S_p} \le \|\boldsymbol{G}\|_{S_q}$ for every feasible $\boldsymbol{D}$, so $\|\boldsymbol{G}\|_{S_q}$ is an upper bound. For attainment take $q \in (1, \infty)$ and $\boldsymbol{D}^\star$ as in equation 1. Since p and q are conjugate we have $p(q-1) = q$, whence

$$\|\boldsymbol{D}^\star\|_{S_p}^p = \sum_i \frac{\sigma_i^{(q-1)p}}{\|\boldsymbol{G}\|_{S_q}^{(q-1)p}} = \frac{\sum_i \sigma_i^q}{\|\boldsymbol{G}\|_{S_q}^q} = 1, \qquad \langle \boldsymbol{G}, \boldsymbol{D}^\star\rangle = \frac{\sum_i \sigma_i^q}{\|\boldsymbol{G}\|_{S_q}^{q-1}} = \|\boldsymbol{G}\|_{S_q}, \tag{39}$$

so $\boldsymbol{D}^\star$ is feasible and attains the bound. For $q = 1$ the choice $\boldsymbol{D}^\star = \boldsymbol{U}\boldsymbol{V}^\top$ gives $\|\boldsymbol{D}^\star\|_{S_\infty} = 1$ and $\langle \boldsymbol{G}, \boldsymbol{D}^\star\rangle = \sum_i \sigma_i = \|\boldsymbol{G}\|_{S_1}$; for $q = \infty$ the choice $\boldsymbol{D}^\star = \boldsymbol{u}_1\boldsymbol{v}_1^\top$ gives $\|\boldsymbol{D}^\star\|_{S_1} = 1$ and $\langle \boldsymbol{G}, \boldsymbol{D}^\star\rangle = \sigma_1 = \|\boldsymbol{G}\|_{S_\infty}$. Finally, if $\boldsymbol{G} \ne \boldsymbol{0}$ and a maximizer had $\|\boldsymbol{D}^\star\|_{S_p} < 1$, rescaling it to the boundary would strictly increase the objective, so every maximizer is on the boundary.

(iii). For $\boldsymbol{G} \ne \boldsymbol{0}$ this is the standard characterisation of the subdifferential of a norm: $\boldsymbol{D} \in \partial\|\boldsymbol{G}\|_{S_q}$ if and only if $\|\boldsymbol{D}\|_{S_p} \le 1$ and $\langle \boldsymbol{G}, \boldsymbol{D}\rangle = \|\boldsymbol{G}\|_{S_q}$ (Clarke, 1983, Cor. 2.4.16), and by (i) the latter pair of conditions is exactly the statement that $\boldsymbol{D}$ is a maximizer. For $\boldsymbol{G} = \boldsymbol{0}$ both sides equal the whole dual unit ball $\{\|\boldsymbol{D}\|_{S_p} \le 1\}$, so the identity persists — it is only the gradient description that fails there. The final sentence is the general fact that a convex function is differentiable at a point precisely when its subdifferential is a singleton, in which case that element is the gradient. □

A.3.4 Proof of Lemma 5

Write $\|\boldsymbol{G}\|_{S_s} = f(\sigma(\boldsymbol{G}))$ with $f(\boldsymbol{x}) = (\sum_i |x_i|^s)^{1/s}$, which is absolutely symmetric. By the theorem of Lewis (1995) on unitarily invariant matrix functions, the spectral function $f \circ \sigma$ is differentiable at $\boldsymbol{G}$ if and only if f is differentiable at $\sigma(\boldsymbol{G})$, and in that case $\nabla(f \circ \sigma)(\boldsymbol{G}) = \boldsymbol{U}\,\mathrm{diag}(\nabla f(\sigma(\boldsymbol{G})))\,\boldsymbol{V}^\top$. For $s \in (1, \infty)$ the map $x \mapsto |x|^s$ is continuously differentiable on all of $\mathbb{R}$, so f is differentiable at every $\sigma \ne \boldsymbol{0}$, whether or not its entries are distinct or some vanish; this is what removes the non-degeneracy and full-rank hypotheses. Differentiating,

$$\frac{\partial f}{\partial \sigma_i} = \frac{1}{s}\Big(\sum_j \sigma_j^s\Big)^{\frac{1}{s}-1} \cdot s\,\sigma_i^{s-1} = \frac{\sigma_i^{s-1}}{\big(\sum_j \sigma_j^s\big)^{(s-1)/s}}, \tag{40}$$

which gives equation 16.

For the endpoints, f fails to be differentiable exactly where the underlying ℓ_s norm does. At $s = 1$, $f(\boldsymbol{x}) = \sum_i |x_i|$ is non-differentiable whenever some $\sigma_i = 0$, i.e. whenever $\mathrm{rank}(\boldsymbol{G}) < r$; this is

also visible directly from Lemma 6, whose subdifferential is a singleton precisely in the full-rank case. At $s = \infty$, $f(\boldsymbol{x}) = \max_i |x_i|$ is non-differentiable whenever the maximum is attained more than once, i.e. whenever σ_1 is not simple. At $\boldsymbol{G} = \boldsymbol{0}$ every norm has subdifferential equal to its dual unit ball, which is not a singleton. □

A.3.5 Proof of Lemma 6

Write $\|\boldsymbol{G}\|_{S_s} = f(\sigma(\boldsymbol{G}))$, where $f(\boldsymbol{x}) = (\sum_i |x_i|^s)^{1/s}$ and $\sigma(\boldsymbol{G}) \in \mathbb{R}^r$ includes all singular values, including zeros. The function f is absolutely symmetric. By the theorem of Lewis (1995) on unitarily invariant matrix functions, $f \circ \sigma$ is differentiable at $\boldsymbol{G}$ if and only if f is differentiable at $\sigma(\boldsymbol{G})$, and its gradient is obtained by applying ∇f to the singular values in an SVD.

For $s \in (1, \infty)$, the function $x \mapsto |x|^s$ is continuously differentiable on $\mathbb{R}$. Hence f is differentiable at every nonzero vector, including vectors with repeated or zero entries. At a nonzero singular-value vector,

$$\frac{\partial f}{\partial \sigma_i} = \frac{\sigma_i^{s-1}}{\left(\sum_{j=1}^r \sigma_j^s\right)^{(s-1)/s}}. \tag{41}$$

The derivative is zero whenever $\sigma_i = 0$, so the matrix gradient can be written using the compact SVD of $\boldsymbol{G}$. This gives equation 16 without requiring distinct singular values or full rank.

At $s = 1$, the function $f(\boldsymbol{x}) = \sum_i |x_i|$ is differentiable at $\sigma(\boldsymbol{G})$ if and only if all r singular values are positive, equivalently $\operatorname{rank}(\boldsymbol{G}) = r$. At $s = \infty$ and $\boldsymbol{G} \neq \boldsymbol{0}$, the function $f(\boldsymbol{x}) = \max_i |x_i|$ is differentiable at $\sigma(\boldsymbol{G})$ if and only if its largest entry is unique, equivalently the largest singular value is simple. At $\boldsymbol{G} = \boldsymbol{0}$, the subdifferential of every norm is its dual unit ball, which is not a singleton. Thus no Schatten norm is differentiable at zero. □

A.3.6 Proof of Lemma 6

By the characterization used in the proof of Lemma 4(iii), $\boldsymbol{D} \in \partial\|\boldsymbol{G}\|_{S_1}$ if and only if $\|\boldsymbol{D}\|_{S_\infty} \leq 1$ and $\langle \boldsymbol{G}, \boldsymbol{D} \rangle = \|\boldsymbol{G}\|_{S_1}$. If $\boldsymbol{G} = \boldsymbol{0}$, this gives the spectral-norm unit ball, which is the stated formula with empty $\boldsymbol{U}$ and $\boldsymbol{V}$. Suppose henceforth that $\boldsymbol{G} \neq \boldsymbol{0}$, and write its compact SVD as $\boldsymbol{G} = \boldsymbol{U} \operatorname{diag}(\sigma_1, \ldots, \sigma_k) \boldsymbol{V}^\top$.

First let $\boldsymbol{D} \in \partial\|\boldsymbol{G}\|_{S_1}$. Since $\boldsymbol{u}_i^\top \boldsymbol{D} \boldsymbol{v}_i \leq \|\boldsymbol{D}\|_{S_\infty} \leq 1$ and $\sigma_i > 0$, the equality

$$\sum_{i=1}^k \sigma_i \boldsymbol{u}_i^\top \boldsymbol{D} \boldsymbol{v}_i = \langle \boldsymbol{G}, \boldsymbol{D} \rangle = \sum_{i=1}^k \sigma_i \tag{42}$$

forces $\boldsymbol{u}_i^\top \boldsymbol{D} \boldsymbol{v}_i = 1$ for every $i \leq k$. Consequently,

$$1 = \boldsymbol{u}_i^\top \boldsymbol{D} \boldsymbol{v}_i \leq \|\boldsymbol{D} \boldsymbol{v}_i\|_2 \leq 1. \tag{43}$$

Equality in Cauchy–Schwarz implies $\boldsymbol{D}\boldsymbol{v}_i = \boldsymbol{u}_i$. Applying the same argument to $\boldsymbol{D}^\top$ gives $\boldsymbol{D}^\top \boldsymbol{u}_i = \boldsymbol{v}_i$. Thus $\boldsymbol{D}\boldsymbol{V} = \boldsymbol{U}$ and $\boldsymbol{U}^\top \boldsymbol{D} = \boldsymbol{V}^\top$.

Define $\boldsymbol{Q} = \boldsymbol{D} - \boldsymbol{U}\boldsymbol{V}^\top$. These identities imply $\boldsymbol{U}^\top \boldsymbol{Q} = \boldsymbol{0}$ and $\boldsymbol{Q}\boldsymbol{V} = \boldsymbol{0}$, and also give

$$\boldsymbol{Q} = (\boldsymbol{I}_m - \boldsymbol{U}\boldsymbol{U}^\top) \boldsymbol{D} (\boldsymbol{I}_n - \boldsymbol{V}\boldsymbol{V}^\top). \tag{44}$$

Since both projection matrices have spectral norm at most one, $\|\boldsymbol{Q}\|_{S_\infty} \leq \|\boldsymbol{D}\|_{S_\infty} \leq 1$.

Conversely, suppose $\boldsymbol{D} = \boldsymbol{U}\boldsymbol{V}^\top + \boldsymbol{Q}$, where $\boldsymbol{U}^\top \boldsymbol{Q} = \boldsymbol{0}$, $\boldsymbol{Q}\boldsymbol{V} = \boldsymbol{0}$, and $\|\boldsymbol{Q}\|_{S_\infty} \leq 1$. For every $\boldsymbol{x} \in \mathbb{R}^n$, orthogonality gives

$$\|\boldsymbol{D}\boldsymbol{x}\|_2^2 = \|\boldsymbol{V}^\top \boldsymbol{x}\|_2^2 + \|\boldsymbol{Q}(\boldsymbol{I}_n - \boldsymbol{V}\boldsymbol{V}^\top)\boldsymbol{x}\|_2^2 \tag{45}$$

$$\leq \|\boldsymbol{V}^\top \boldsymbol{x}\|_2^2 + \|(\boldsymbol{I}_n - \boldsymbol{V}\boldsymbol{V}^\top)\boldsymbol{x}\|_2^2 = \|\boldsymbol{x}\|_2^2. \tag{46}$$

Hence $\|\boldsymbol{D}\|_{S_\infty} \leq 1$. Moreover, $\langle \boldsymbol{G}, \boldsymbol{Q} \rangle = 0$, so $\langle \boldsymbol{G}, \boldsymbol{D} \rangle = \sum_{i=1}^k \sigma_i = \|\boldsymbol{G}\|_{S_1}$. The subgradient characterization therefore gives $\boldsymbol{D} \in \partial\|\boldsymbol{G}\|_{S_1}$. This proves both inclusions; see also Watson (1992).

If $\operatorname{rank}(\boldsymbol{G}) = r = \min(m, n)$, then either $\boldsymbol{U}$ or $\boldsymbol{V}$ is square and orthogonal. The conditions $\boldsymbol{U}^\top \boldsymbol{Q} = \boldsymbol{0}$ and $\boldsymbol{Q}\boldsymbol{V} = \boldsymbol{0}$ therefore force $\boldsymbol{Q} = \boldsymbol{0}$. The subdifferential is then $\{\boldsymbol{U}\boldsymbol{V}^\top\}$, and

$\nabla_{\boldsymbol{G}}\|\boldsymbol{G}\|_{S_1} = \boldsymbol{U}\boldsymbol{V}^\top$. If $\operatorname{rank}(\boldsymbol{G}) < r$, both orthogonal complements contain unit vectors $\boldsymbol{u}$ and $\boldsymbol{v}$, and $\boldsymbol{Q} = \boldsymbol{u}\boldsymbol{v}^\top$ gives a nonzero admissible choice. Thus the subdifferential is not a singleton.

Finally, $\boldsymbol{U}\boldsymbol{V}^\top$ is the polar factor, as follows from

$$\boldsymbol{G} = (\boldsymbol{U}\boldsymbol{V}^\top)\big(\boldsymbol{V}\operatorname{diag}(\sigma_1, \ldots, \sigma_k)\boldsymbol{V}^\top\big). \tag{47}$$

□

A.3.7 Proof of Proposition 1

From equation 16 the normalising denominator $(\sum_j \sigma_j^s)^{(s-1)/s}$ is common to all entries, so it cancels in the ratio and $s_i(s)/s_1(s) = \sigma_i^{s-1}/\sigma_1^{s-1} = (\sigma_i/\sigma_1)^{s-1}$, which lies in $(0, 1]$ because $0 < \sigma_i \leq \sigma_1$; this is equation 18. Fix i with $\sigma_i < \sigma_1$ and set $\rho_i := \sigma_i/\sigma_1 \in (0, 1)$. Then $s \mapsto \rho_i^{s-1} = \exp\big((s-1)\log\rho_i\big)$ has derivative $\rho_i^{s-1}\log\rho_i < 0$, so it is strictly decreasing, and the two limits in equation 19 follow from $\rho_i^0 = 1$ and $\rho_i^{s-1} \to 0$. Since the ratio tends to 1 for every $i \leq k$ as $s \downarrow 1$, the rescaling in equation 16 tends to a multiple of $\boldsymbol{U}\boldsymbol{V}^\top$, matching the $q = 1$ case of equation 1; since it tends to 0 for every i with $\sigma_i < \sigma_1$ as $s \to \infty$, only the top singular direction survives, matching the $q = \infty$ case.

B CONVERGENCE OF SSMD-MVR

Under Assumption 1, ∇F is continuous on $\mathbb{R}^{m\times n}$. Consequently, its norm has a finite supremum on every compact update segment.

Theorem 5. *Consider Algorithm 1. Let $p, q \in [1,\infty]$ satisfy $1/p + 1/q = 1$, let $\boldsymbol{W}_0 \in \mathbb{R}^{m\times n}$ be a deterministic initial point, and let $r := \min\{m,n\}$. Suppose Assumptions 1, 3, 4, 5, and 2 hold. Fix any $c > 0$ and set $L_c := \max\{L_1, L_0/c\}$, $\sigma_c := \max\{\sigma_1, \sigma_0/c\}$, and $\delta_c := \max\{\delta_1, \delta_0/c\}$. Choose $\eta_0 > 0$ such that $h_0 := \eta_0 r^{2\max\{\frac{1}{2}-\frac{1}{p},0\}} L_c < 2/3$. For each horizon T, set $a = a_T := T^{-2/3}$ and $\eta = \eta_T := \eta_0 T^{-2/3}$. Then, for every integer $T \geq T_0$,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left\|\nabla F(\boldsymbol{W}_t)\right\|_{S_q} \leq \frac{\mathcal{C}_1^{\mathrm{AVR}}}{T^{\frac{1}{3}}} + \frac{\mathcal{C}_2^{\mathrm{AVR}}}{T^{\frac{2}{3}}} = \mathcal{O}\left(T^{-\frac{1}{3}}\right), \tag{48}$$

where $\mathcal{C}_1^{\mathrm{AVR}} := 2[(F(\boldsymbol{W}_0) - F^*)/\eta_0 + 2r^{\max\{\frac{1}{q}-\frac{1}{2},0\}}\nu_0 + cb_0]/d_0$ and $\mathcal{C}_2^{\mathrm{AVR}} := ch_0/[d_0(1-h_0)]$, with $d_0 := (2-3h_0)/(2(1-h_0))$, $b_0 := 2r^{\max\{\frac{1}{q}-\frac{1}{2},0\}}(\sigma_c + r^{\max\{\frac{1}{2}-\frac{1}{q},0\}}\eta_0\delta_c/(1-h_0))/\sqrt{1-h_0}$, $\nu_0 := \sqrt{\sigma_0^2 + \sigma_1^2\|\nabla F(\boldsymbol{W}_0)\|_{S_q}^2}$, and $T_0 := \lceil \max\{1, (2b_0/d_0)^3\}\rceil$.

The proof is given in Section B.1. It uses the following two intermediate lemmas.

Lemma 8 (General descent under generalized smoothness)**.** *Suppose Assumptions 1 and 3 hold. For an integer $T \geq 1$, recall $a = a_T := T^{-2/3}$, $\eta = \eta_T := \eta_0 T^{-2/3}$, and r, c, L_c, h_0 from Theorem 3. Define the alignment error $\mathcal{A}_t^{\mathrm{AVR}} := \|\nabla F(\boldsymbol{W}_t)\|_{S_q} - \langle \nabla F(\boldsymbol{W}_t), \mathcal{D}_p(\boldsymbol{M}_t)\rangle$. Define $h_T := \eta_T r^{2\max\{\frac{1}{2}-\frac{1}{p},0\}} L_c = h_0 T^{-2/3}$. Then, for every $t = 0,\dots,T-1$,*

$$F(\boldsymbol{W}_{t+1}) \leq F(\boldsymbol{W}_t) - \eta_T \frac{2-3h_T}{2(1-h_T)}\left\|\nabla F(\boldsymbol{W}_t)\right\|_{S_q} + \eta_T \mathcal{A}_t^{\mathrm{AVR}} + \frac{c\eta_T h_T}{2(1-h_T)}. \tag{49}$$

Moreover,

$$0 \leq \mathcal{A}_t^{\mathrm{AVR}} \leq 2\left\|\boldsymbol{M}_t - \nabla F(\boldsymbol{W}_t)\right\|_{S_q}. \tag{50}$$

Lemma 9 (Alignment error of recursive variance reduction under affine noise)**.** *Suppose Assumptions 1, 2, 3, 4, and 5 hold. For an integer $T \geq 1$, recall $a = a_T := T^{-2/3}$ and $\eta = \eta_T := \eta_0 T^{-2/3}$, together with $r, c, L_c, \sigma_c, \delta_c, \nu_0$ from Theorem 3. Recall $\mathcal{A}_t^{\mathrm{AVR}}$ and h_T from Lemma 8, and define*

$$\varrho_{\mathrm{AVR},T} := \frac{(1-a_T)^2}{1-h_T}, \quad \mathsf{K}_{\mathrm{AVR},T} := a_T\sigma_c + \frac{(1-a_T)r^{\max\{\frac{1}{q}-\frac{1}{2},0\}}\eta_T\delta_c}{1-h_T}, \quad b_T := \frac{2r^{\max\{\frac{1}{q}-\frac{1}{2},0\}}\mathsf{K}_{\mathrm{AVR},T}}{\sqrt{1-\varrho_{\mathrm{AVR},T}}}. \tag{51}$$

The parameters a_T and η_T are constant across iterations for each fixed horizon T. Then

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\mathcal{A}_t^{\mathrm{AVR}} \leq \frac{2r^{\max\{\frac{1}{q}-\frac{1}{2},0\}}\nu_0}{a_T T} + b_T\left[c + \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla F(\boldsymbol{W}_t)\right\|_{S_q}\right]. \tag{52}$$

B.1 PROOF OF THEOREM 3

Proof. Fix an integer $T \geq T_0$. The parameters $a_T = T^{-2/3}$ and $\eta_T = \eta_0 T^{-2/3}$ are fixed throughout this run. By Lemma 8, taking expectations, summing over $t = 0,\dots,T-1$, using $F(\boldsymbol{W}_T) \geq F^*$, and dividing by $\eta_T T$, we obtain

$$\frac{2-3h_T}{2(1-h_T)}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\nabla F(\boldsymbol{W}_t)\right\|_{S_q} \leq \frac{F(\boldsymbol{W}_0) - F^*}{\eta_T T} + \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\mathcal{A}_t^{\mathrm{AVR}} + \frac{ch_T}{2(1-h_T)}, \tag{53}$$

where

$$h_T = \eta_T r^{2\max\{\frac{1}{2}-\frac{1}{p},0\}} L_c = h_0 T^{-\frac{2}{3}}. \tag{54}$$

Since $T \geq 1$ and $0 \leq h_0 < 2/3 < 1$, the definitions of h_T and a_T give

$$h_T = h_0 T^{-2/3} < \frac{2}{3} T^{-2/3} < T^{-2/3} \leq T^{-2/3}(2 - T^{-2/3}) = a_T(2 - a_T). \tag{55}$$

Thus $1 - \varrho_{\mathrm{AVR},T} = [a_T(2-a_T) - h_T]/(1-h_T) > 0$, so the constant b_T in Lemma 9 is well defined and the lemma applies. Substituting equation 52 into equation 53 and collecting the average-gradient terms gives

$$\left[\frac{2 - 3h_T}{2(1-h_T)} - b_T\right] \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left\|\nabla F(\boldsymbol{W}_t)\right\|_{S_q} \leq \frac{F(\boldsymbol{W}_0) - F^*}{\eta_T T} + \frac{2 r^{\max\left\{\frac{1}{q} - \frac{1}{2}, 0\right\}} \nu_0}{a_T T} + c\left[b_T + \frac{h_T}{2(1-h_T)}\right]. \tag{56}$$

We next bound b_T. From equation 51 and the choices $a_T = T^{-2/3}$ and $\eta_T = \eta_0 T^{-2/3}$,

$$b_T = 2 r^{\max\left\{\frac{1}{q} - \frac{1}{2}, 0\right\}} T^{-\frac{1}{3}} \left[\sigma_c + \frac{(1 - T^{-\frac{2}{3}}) r^{\max\left\{\frac{1}{2} - \frac{1}{q}, 0\right\}} \eta_0 \delta_c}{1 - h_0 T^{-\frac{2}{3}}}\right] \sqrt{\frac{1 - h_0 T^{-\frac{2}{3}}}{2 - h_0 - T^{-\frac{2}{3}}}}. \tag{57}$$

Because $0 \leq h_0 < 2/3$ and $T \geq 1$, we have $0 \leq 1 - T^{-2/3} \leq 1$, $1 \geq 1 - h_0 T^{-\frac{2}{3}} \geq 1 - h_0 > 0$ and $2 - h_0 - T^{-\frac{2}{3}} \geq 1 - h_0 > 0$, consequently,

$$\frac{1 - T^{-\frac{2}{3}}}{1 - h_0 T^{-\frac{2}{3}}} \leq \frac{1}{1 - h_0 T^{-2/3}} \leq \frac{1}{1 - h_0}, \quad \sqrt{\frac{1 - h_0 T^{-\frac{2}{3}}}{2 - h_0 - T^{-\frac{2}{3}}}} \leq \sqrt{\frac{1}{2 - h_0 - T^{-2/3}}} \leq \frac{1}{\sqrt{1 - h_0}}. \tag{58}$$

Using the exact expression for b_T,

$$b_T = 2 r^{\max\left\{\frac{1}{q} - \frac{1}{2}, 0\right\}} T^{-\frac{1}{3}} \left[\sigma_c + \frac{\left(1 - T^{-\frac{2}{3}}\right) r^{\max\left\{\frac{1}{2} - \frac{1}{q}, 0\right\}} \eta_0 \delta_c}{1 - h_0 T^{-\frac{2}{3}}}\right] \sqrt{\frac{1 - h_0 T^{-\frac{2}{3}}}{2 - h_0 - T^{-\frac{2}{3}}}} \tag{59}$$

$$\leq \frac{2 r^{\max\left\{\frac{1}{q} - \frac{1}{2}, 0\right\}}}{\sqrt{1 - h_0}} \left[\sigma_c + \frac{r^{\max\left\{\frac{1}{2} - \frac{1}{q}, 0\right\}} \eta_0 \delta_c}{1 - h_0}\right] T^{-\frac{1}{3}} = b_0 T^{-\frac{1}{3}}. \tag{60}$$

Recall that $h_T = h_0 T^{-\frac{2}{3}} \leq h_0 < \frac{2}{3} < 1$. Therefore, both $1 - h_T$ and $1 - h_0$ are positive. Recall $d_0 = (2 - 3h_0)/(2(1 - h_0))$ from Theorem 3, we obtain

$$\frac{2 - 3h_T}{2(1-h_T)} - d_0 = \frac{2 - 3h_T}{2(1-h_T)} - \frac{2 - 3h_0}{2(1-h_0)} = \frac{h_0 - h_T}{2(1-h_T)(1-h_0)} \geq 0. \tag{61}$$

Hence, keeping the intermediate comparison explicit,

$$b_T \leq b_0 T^{-\frac{1}{3}}, \qquad \frac{2 - 3h_T}{2(1-h_T)} = d_0 + \frac{h_0 - h_T}{2(1-h_T)(1-h_0)} \geq d_0. \tag{62}$$

Since $L_0, L_1 \geq 0, c > 0, \eta_0 > 0$ and $r = \min\{m, n\} \geq 1$, then $L_c = \max\left\{L_1, \frac{L_0}{c}\right\} \geq 0, h_0 = \eta_0 r^{2\max\left\{\frac{1}{2} - \frac{1}{p}, 0\right\}} L_c \geq 0$. Since $h_0 < 2/3$,

$$d_0 = \frac{2 - 3h_0}{2(1-h_0)} > 0. \tag{63}$$

Since $\sigma_0, \sigma_1, \delta_0, \delta_1 \geq 0$, $c > 0$, $r \geq 1$, $\eta_0 > 0$ and $1 - h_0 > 0$

$$b_0 = \frac{2 r^{\max\left\{\frac{1}{q} - \frac{1}{2}, 0\right\}}}{\sqrt{1 - h_0}} \left[\sigma_c + \frac{r^{\max\left\{\frac{1}{2} - \frac{1}{q}, 0\right\}} \eta_0 \delta_c}{1 - h_0}\right] \geq 0. \tag{64}$$

For $T \geq T_0$, recall $T_0 := \lceil \max\{1, (2b_0/d_0)^3\} \rceil$ from Theorem 3, where $\lceil x \rceil$ denotes the smallest integer not less than x. Then $T \geq T_0 \geq \left(\frac{2b_0}{d_0}\right)^3$. Combining equation 63 and equation 64 gives $T^{\frac{1}{3}} \geq \frac{2b_0}{d_0}$, and hence $b_0 \leq (d_0/2) T^{1/3}$. Consequently,

$$b_0 T^{-\frac{1}{3}} = \frac{b_0}{T^{\frac{1}{3}}} \leq \frac{(d_0/2) T^{1/3}}{T^{1/3}} = \frac{d_0}{2}. \tag{65}$$

This argument also covers $b_0 = 0$, since it divides only by $d_0 > 0$ and $T^{1/3} > 0$. Combining equation 65 with equation 62,

$$b_T \le b_0 T^{-1/3} = \frac{b_0}{T^{1/3}} \le \frac{(d_0/2)T^{1/3}}{T^{1/3}} = \frac{d_0}{2}. \tag{66}$$

Combining equation 62 and equation 66,

$$\frac{2-3h_T}{2(1-h_T)} - b_T \ge d_0 - b_T \ge d_0 - \frac{d_0}{2} = \frac{d_0}{2}. \tag{67}$$

The remaining terms in equation 56 satisfy

$$\frac{F(\boldsymbol{W}_0) - F^*}{\eta_T T} = \frac{F(\boldsymbol{W}_0) - F^*}{\eta_0} T^{-\frac{1}{3}}, \tag{68}$$

$$\frac{2r^{\max\left\{\frac{1}{q}-\frac{1}{2},0\right\}}\nu_0}{a_T T} = 2r^{\max\left\{\frac{1}{q}-\frac{1}{2},0\right\}}\nu_0 T^{-\frac{1}{3}}, \tag{69}$$

$$cb_T \le cb_0 T^{-\frac{1}{3}}, \tag{70}$$

$$\frac{ch_T}{2(1-h_T)} = \frac{ch_0 T^{-\frac{2}{3}}}{2\left(1-h_0 T^{-\frac{2}{3}}\right)} \le \frac{ch_0 T^{-\frac{2}{3}}}{2(1-h_0)} = \frac{ch_0}{2(1-h_0)} T^{-\frac{2}{3}}, \tag{71}$$

Combining equation 56, equation 67, and equation 71, we obtain

$$\begin{aligned}
\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\left\|\nabla F(\boldsymbol{W}_t)\right\|_{S_q} &\le \frac{2}{d_0}\left[\frac{F(\boldsymbol{W}_0) - F^*}{\eta_T T} + \frac{2r^{\max\left\{\frac{1}{q}-\frac{1}{2},0\right\}}\nu_0}{a_T T} + cb_T + \frac{ch_T}{2(1-h_T)}\right] \\
&\le \frac{2}{d_0}\left[\frac{F(\boldsymbol{W}_0) - F^*}{\eta_0} + 2r^{\max\left\{\frac{1}{q}-\frac{1}{2},0\right\}}\nu_0 + cb_0\right] T^{-\frac{1}{3}} + \frac{ch_0}{d_0(1-h_0)} T^{-\frac{2}{3}} \\
&= \frac{\mathcal{C}_1^{\mathrm{AVR}}}{T^{1/3}} + \frac{\mathcal{C}_2^{\mathrm{AVR}}}{T^{2/3}}.
\end{aligned} \tag{72}$$

This proves equation 48. □

B.2 Proof of the Intermediate Lemmas

Proof of Lemma 8. Fix a horizon $T \ge 1$ and use $a_T = T^{-2/3}$ and $\eta_T = \eta_0 T^{-2/3}$ throughout. The freedom to choose $c > 0$ follows from the elementary inequality

$$L_0 + L_1 g \le cL_c + L_c g = L_c(c+g), \qquad g \ge 0, \tag{73}$$

because $L_c = \max\{L_1, L_0/c\}$ satisfies $L_0 \le cL_c$ and $L_1 \le L_c$. Let

$$\boldsymbol{W}_t(\theta) := \boldsymbol{W}_t - \theta\eta_T \mathcal{D}_p(\boldsymbol{M}_t), \qquad \theta \in [0,1], \tag{74}$$

and define the segment supremum

$$\mathcal{S}_t := \sup_{\theta\in[0,1]} \left\|\nabla F(\boldsymbol{W}_t(\theta))\right\|_{S_q}. \tag{75}$$

By Assumption 1, ∇F is continuous on the compact update segment, so $\mathcal{S}_t < \infty$. The definitions $\bar{p} = \min\{p,2\}$ and $\bar{q} = \max\{q,2\}$ imply $\frac{1}{\bar{p}} - \frac{1}{p} = \max\{\frac{1}{2} - \frac{1}{p}, 0\} = \max\{\frac{1}{q} - \frac{1}{2}, 0\}$ and $\frac{1}{q} - \frac{1}{\bar{q}} = \max\{\frac{1}{q} - \frac{1}{2}, 0\}$. Therefore, the Schatten norm-comparison inequalities and Lemma 4, which gives $\|\mathcal{D}_p(\boldsymbol{M}_t)\|_{S_p} \le 1$ and $\langle \boldsymbol{M}_t, \mathcal{D}_p(\boldsymbol{M}_t)\rangle = \|\boldsymbol{M}_t\|_{S_q}$, give

$$\begin{aligned}
\left\|\mathcal{D}_p(\boldsymbol{M}_t)\right\|_{S_{\bar{p}}} &\le r^{\frac{1}{\bar{p}}-\frac{1}{p}} \left\|\mathcal{D}_p(\boldsymbol{M}_t)\right\|_{S_p} = r^{\max\left\{\frac{1}{q}-\frac{1}{2},0\right\}} \left\|\mathcal{D}_p(\boldsymbol{M}_t)\right\|_{S_p} \le r^{\max\left\{\frac{1}{q}-\frac{1}{2},0\right\}}, \\
\left\|\boldsymbol{A}\right\|_{S_q} &\le r^{\frac{1}{q}-\frac{1}{\bar{q}}} \left\|\boldsymbol{A}\right\|_{S_{\bar{q}}} = r^{\max\left\{\frac{1}{q}-\frac{1}{2},0\right\}} \left\|\boldsymbol{A}\right\|_{S_{\bar{q}}}.
\end{aligned} \tag{76}$$

For $\lambda, \theta \in [0,1]$, the point $(1-\lambda)\boldsymbol{W}_t + \lambda \boldsymbol{W}_t(\theta) = \boldsymbol{W}_t(\lambda\theta)$ lies on the same update segment. Hence the gradient supremum over the subsegment from $\boldsymbol{W}_t$ to $\boldsymbol{W}_t(\theta)$ is at most $\mathcal{S}_t$. Therefore, for every

$\theta \in [0,1]$, Lemma 3, Assumption 3, and equation 76 imply

$$\begin{aligned}
\|\nabla F(\boldsymbol{W}_t(\theta))\|_{S_q} &\le \|\nabla F(\boldsymbol{W}_t)\|_{S_q} + \|\nabla F(\boldsymbol{W}_t(\theta)) - \nabla F(\boldsymbol{W}_t)\|_{S_q} \\
&\le \|\nabla F(\boldsymbol{W}_t)\|_{S_q} + r^{\max\left\{\frac{1}{q}-\frac{1}{2},0\right\}} \|\nabla F(\boldsymbol{W}_t(\theta)) - \nabla F(\boldsymbol{W}_t)\|_{S_{\bar{q}}} \\
&\le \|\nabla F(\boldsymbol{W}_t)\|_{S_q} + r^{\max\left\{\frac{1}{q}-\frac{1}{2},0\right\}} (L_0 + L_1 \mathcal{S}_t) \|\boldsymbol{W}_t(\theta) - \boldsymbol{W}_t\|_{S_{\bar{p}}} \\
&= \|\nabla F(\boldsymbol{W}_t)\|_{S_q} + \theta\eta_T r^{\max\left\{\frac{1}{q}-\frac{1}{2},0\right\}} (L_0 + L_1 \mathcal{S}_t) \|\mathcal{D}_p(\boldsymbol{M}_t)\|_{S_{\bar{p}}} \\
&\le \|\nabla F(\boldsymbol{W}_t)\|_{S_q} + \theta\eta_T r^{2\max\left\{\frac{1}{q}-\frac{1}{2},0\right\}} (L_0 + L_1 \mathcal{S}_t) \\
&\le \|\nabla F(\boldsymbol{W}_t)\|_{S_q} + \eta_T r^{2\max\left\{\frac{1}{2}-\frac{1}{p},0\right\}} (L_0 + L_1 \mathcal{S}_t) . \qquad (77)
\end{aligned}$$

Taking the supremum over θ and using

$$L_0 + L_1 \mathcal{S}_t \le c L_c + L_c \mathcal{S}_t = L_c (c + \mathcal{S}_t) \qquad (78)$$

gives

$$c + \mathcal{S}_t \le c + \|\nabla F(\boldsymbol{W}_t)\|_{S_q} + h_T (c + \mathcal{S}_t). \qquad (79)$$

Since $0 \le h_T = h_0 T^{-2/3} \le h_0 < 2/3 < 1$ and $\mathcal{S}_t < \infty$, subtracting $h_T(c+\mathcal{S}_t)$ gives $(1-h_T)(c+\mathcal{S}_t) \le c + \|\nabla F(\boldsymbol{W}_t)\|_{S_q}$, and therefore

$$c + \mathcal{S}_t \le \frac{c + \|\nabla F(\boldsymbol{W}_t)\|_{S_q}}{1 - h_T}. \qquad (80)$$

Applying Lemma 1(i), specifically equation 9, with $\boldsymbol{W} = \boldsymbol{W}_t$ and $\boldsymbol{Z} = -\eta_T \mathcal{D}_p(\boldsymbol{M}_t)$, and then using equation 76, $L_0 + L_1 \mathcal{S}_t \le L_c (c + \mathcal{S}_t)$, and equation 80, gives

$$\begin{aligned}
F(\boldsymbol{W}_{t+1}) &\le F(\boldsymbol{W}_t) - \eta_T \langle \nabla F(\boldsymbol{W}_t), \mathcal{D}_p(\boldsymbol{M}_t) \rangle + \frac{\eta_T^2}{2} (L_0 + L_1 \mathcal{S}_t) \|\mathcal{D}_p(\boldsymbol{M}_t)\|_{S_{\bar{p}}}^2 \\
&\le F(\boldsymbol{W}_t) - \eta_T \langle \nabla F(\boldsymbol{W}_t), \mathcal{D}_p(\boldsymbol{M}_t) \rangle + \frac{\eta_T^2 r^{2\max\left\{\frac{1}{2}-\frac{1}{p},0\right\}} L_c}{2(1-h_T)} \left[c + \|\nabla F(\boldsymbol{W}_t)\|_{S_q} \right] \\
&= F(\boldsymbol{W}_t) - \eta_T \langle \nabla F(\boldsymbol{W}_t), \mathcal{D}_p(\boldsymbol{M}_t) \rangle + \frac{\eta_T h_T}{2(1-h_T)} \left[c + \|\nabla F(\boldsymbol{W}_t)\|_{S_q} \right]. \qquad (81)
\end{aligned}$$

Substituting $\langle \nabla F(\boldsymbol{W}_t), \mathcal{D}_p(\boldsymbol{M}_t) \rangle = \|\nabla F(\boldsymbol{W}_t)\|_{S_q} - \mathcal{A}_t^{\mathrm{AVR}}$ into equation 81 and using $1 - h_T/[2(1-h_T)] = (2-3h_T)/[2(1-h_T)]$ yields equation 49.

Finally, Schatten Hölder inequality and Lemma 4 give

$$\langle \nabla F(\boldsymbol{W}_t), \mathcal{D}_p(\boldsymbol{M}_t) \rangle \le \|\nabla F(\boldsymbol{W}_t)\|_{S_q} \|\mathcal{D}_p(\boldsymbol{M}_t)\|_{S_p} \le \|\nabla F(\boldsymbol{W}_t)\|_{S_q}, \qquad (82)$$

which proves $\mathcal{A}_t^{\mathrm{AVR}} \ge 0$. Moreover,

$$\begin{aligned}
\langle \nabla F(\boldsymbol{W}_t), \mathcal{D}_p(\boldsymbol{M}_t) \rangle &= \langle \boldsymbol{M}_t, \mathcal{D}_p(\boldsymbol{M}_t) \rangle - \langle \boldsymbol{M}_t - \nabla F(\boldsymbol{W}_t), \mathcal{D}_p(\boldsymbol{M}_t) \rangle \\
&\ge \|\boldsymbol{M}_t\|_{S_q} - \|\boldsymbol{M}_t - \nabla F(\boldsymbol{W}_t)\|_{S_q} \|\mathcal{D}_p(\boldsymbol{M}_t)\|_{S_p} \\
&\ge \|\boldsymbol{M}_t\|_{S_q} - \|\boldsymbol{M}_t - \nabla F(\boldsymbol{W}_t)\|_{S_q} \\
&\ge \|\nabla F(\boldsymbol{W}_t)\|_{S_q} - 2\|\boldsymbol{M}_t - \nabla F(\boldsymbol{W}_t)\|_{S_q}. \qquad (83)
\end{aligned}$$

This proves equation 50. □

Proof of Lemma 9. Fix a horizon $T \ge 1$. Throughout the proof, $a_T = T^{-2/3}$, $\eta_T = \eta_0 T^{-2/3}$, and $h_T = h_0 T^{-2/3}$ are fixed with respect to t. As in equation 55, $h_T = h_0 T^{-2/3} < T^{-2/3} \le a_T(2 - a_T)$, so $0 \le \varrho_{\mathrm{AVR},T} < 1$. The definitions of σ_c and δ_c imply, for every $g \ge 0$,

$$\begin{aligned}
\sqrt{\sigma_0^2 + \sigma_1^2 g^2} \le \sigma_0 + \sigma_1 g \le c\sigma_c + \sigma_c g = \sigma_c (c + g), \\
\delta_0 + \delta_1 g \le c\delta_c + \delta_c g = \delta_c (c + g). \qquad (84)
\end{aligned}$$

For $t \geq 1$, let

$$\mathcal{F}_{t-1} := \sigma(\boldsymbol{W}_0, \xi_0, \ldots, \xi_{t-1}). \tag{85}$$

Then $\boldsymbol{W}_t$, $\boldsymbol{W}_{t-1}$, and $\boldsymbol{M}_{t-1}$ are $\mathcal{F}_{t-1}$-measurable, while the fresh sample ξ_t is independent of $\mathcal{F}_{t-1}$. Under the standard stochastic-oracle convention, the unbiased-oracle and affine second-moment assumptions therefore hold conditionally at these query points. Define

$$\boldsymbol{\Xi}_t := \nabla f(\boldsymbol{W}_t; \xi_t) - \nabla F(\boldsymbol{W}_t), \tag{86}$$

$$\boldsymbol{Z}_t := \nabla f(\boldsymbol{W}_t; \xi_t) - \nabla f(\boldsymbol{W}_{t-1}; \xi_t) - [\nabla F(\boldsymbol{W}_t) - \nabla F(\boldsymbol{W}_{t-1})], \tag{87}$$

$$\boldsymbol{U}_t := a_T \boldsymbol{\Xi}_t + (1 - a_T)\boldsymbol{Z}_t. \tag{88}$$

By Assumption 2, both $\mathbb{E}[\boldsymbol{\Xi}_t \mid \mathcal{F}_{t-1}]$ and $\mathbb{E}[\boldsymbol{Z}_t \mid \mathcal{F}_{t-1}]$ are zero. Consequently,

$$\mathbb{E}[\boldsymbol{U}_t \mid \mathcal{F}_{t-1}] = a_T \mathbb{E}[\boldsymbol{\Xi}_t \mid \mathcal{F}_{t-1}] + (1 - a_T)\mathbb{E}[\boldsymbol{Z}_t \mid \mathcal{F}_{t-1}] = \mathbf{0}. \tag{89}$$

Let

$$\mathcal{S}_t^- := \sup_{\theta \in [0,1]} \|\nabla F(\theta \boldsymbol{W}_t + (1-\theta)\boldsymbol{W}_{t-1})\|_{S_q}. \tag{90}$$

The same argument used to prove equation 80, together with

$$\|\boldsymbol{W}_t - \boldsymbol{W}_{t-1}\|_{S_p} = \eta_T \|\mathcal{D}_p(\boldsymbol{M}_{t-1})\|_{S_p} \leq \eta_T, \tag{91}$$

gives $c + \mathcal{S}_t^- \leq c + \|\nabla F(\boldsymbol{W}_t)\|_{S_q} + h_T(c + \mathcal{S}_t^-)$ when the segment is anchored at $\boldsymbol{W}_t$. Since $h_T < 1$, this yields

$$c + \mathcal{S}_t^- \leq \frac{c + \|\nabla F(\boldsymbol{W}_t)\|_{S_q}}{1 - h_T}. \tag{92}$$

Since $\boldsymbol{W}_{t-1}$ belongs to this segment, equation 92 also gives

$$c + \|\nabla F(\boldsymbol{W}_t)\|_{S_q} \geq (1 - h_T)(c + \mathcal{S}_t^-) \geq (1 - h_T)\left[c + \|\nabla F(\boldsymbol{W}_{t-1})\|_{S_q}\right]. \tag{93}$$

Applying equation 80 to the forward segment from $\boldsymbol{W}_{t-1}$ also gives

$$c + \|\nabla F(\boldsymbol{W}_t)\|_{S_q} \leq \frac{c + \|\nabla F(\boldsymbol{W}_{t-1})\|_{S_q}}{1 - h_T} \leq \frac{c + \|\nabla F(\boldsymbol{W}_0)\|_{S_q}}{(1 - h_T)^t}. \tag{94}$$

Thus the gradient norms are bounded by deterministic constants for fixed T. Together with the affine second-moment assumptions and the recursions below, this ensures that the expectations used in the proof are finite.

By the conditional Minkowski inequality in L_2, Assumptions 4 and 5, the norm comparison in Lemma 3, and equation 92,

$$\begin{aligned}
\left(\mathbb{E}\left[\|\boldsymbol{U}_t\|_F^2 \mid \mathcal{F}_{t-1}\right]\right)^{\frac{1}{2}} &\leq a_T \left(\mathbb{E}[\|\boldsymbol{\Xi}_t\|_F^2 \mid \mathcal{F}_{t-1}]\right)^{1/2} + (1 - a_T)\left(\mathbb{E}[\|\boldsymbol{Z}_t\|_F^2 \mid \mathcal{F}_{t-1}]\right)^{1/2} \\
&\leq a_T \left(\mathbb{E}[\|\boldsymbol{\Xi}_t\|_F^2 \mid \mathcal{F}_{t-1}]\right)^{1/2} + (1 - a_T) r^{\max\{\frac{1}{2} - \frac{1}{q}, 0\}} \left(\mathbb{E}[\|\boldsymbol{Z}_t\|_{S_q}^2 \mid \mathcal{F}_{t-1}]\right)^{1/2} \\
&\leq a_T \sqrt{\sigma_0^2 + \sigma_1^2 \|\nabla F(\boldsymbol{W}_t)\|_{S_q}^2} + (1 - a_T) r^{\max\{\frac{1}{2} - \frac{1}{q}, 0\}} \eta_T \left(\delta_0 + \delta_1 \mathcal{S}_t^-\right) \\
&\leq \left[a_T \sigma_c + \frac{(1 - a_T) r^{\max\{\frac{1}{2} - \frac{1}{q}, 0\}} \eta_T \delta_c}{1 - h_T}\right] \left[c + \|\nabla F(\boldsymbol{W}_t)\|_{S_q}\right] \\
&= \mathsf{K}_{\mathrm{AVR},T} \left[c + \|\nabla F(\boldsymbol{W}_t)\|_{S_q}\right].
\end{aligned} \tag{95}$$

Subtracting $\nabla F(\boldsymbol{W}_t)$ from the momentum update in Algorithm 1 gives

$$\boldsymbol{M}_t - \nabla F(\boldsymbol{W}_t) = (1 - a_T)\left[\boldsymbol{M}_{t-1} - \nabla F(\boldsymbol{W}_{t-1})\right] + \boldsymbol{U}_t. \tag{96}$$

Separate the initialization error by writing

$$\boldsymbol{M}_t - \nabla F(\boldsymbol{W}_t) = (1 - a_T)^t \left[\boldsymbol{M}_0 - \nabla F(\boldsymbol{W}_0)\right] + \widetilde{\boldsymbol{E}}_t, \tag{97}$$

where $\widetilde{\boldsymbol{E}}_0 = \mathbf{0}$ and

$$\widetilde{\boldsymbol{E}}_t = (1 - a_T)\widetilde{\boldsymbol{E}}_{t-1} + \boldsymbol{U}_t. \tag{98}$$

Define the weighted second-moment potential

$$\mathcal{Q}_t := \mathbb{E}\left[\frac{\|\widetilde{\boldsymbol{E}}_t\|_F^2}{c + \|\nabla F(\boldsymbol{W}_t)\|_{S_q}}\right]. \tag{99}$$

Since $\widetilde{\boldsymbol{E}}_{t-1}$ and $c + \|\nabla F(\boldsymbol{W}_t)\|_{S_q}$ are $\mathcal{F}_{t-1}$-measurable, equation 89 gives

$$\begin{aligned}\mathbb{E}\Big[\Big\langle \widetilde{\boldsymbol{E}}_{t-1}, \boldsymbol{U}_t\Big\rangle_F \mid \mathcal{F}_{t-1}\Big] &= \Big\langle \widetilde{\boldsymbol{E}}_{t-1}, \mathbb{E}[\boldsymbol{U}_t \mid \mathcal{F}_{t-1}]\Big\rangle_F = 0,\\ \mathbb{E}\Big[\|\widetilde{\boldsymbol{E}}_t\|_F^2 \mid \mathcal{F}_{t-1}\Big] &= (1 - a_T)^2\|\widetilde{\boldsymbol{E}}_{t-1}\|_F^2 + \mathbb{E}[\|\boldsymbol{U}_t\|_F^2 \mid \mathcal{F}_{t-1}].\end{aligned} \tag{100}$$

Using equation 89, equation 95, and equation 93, we obtain

$$\begin{aligned}\mathcal{Q}_t &= \mathbb{E}\left[\frac{\mathbb{E}[\|\widetilde{\boldsymbol{E}}_t\|_F^2 \mid \mathcal{F}_{t-1}]}{c + \|\nabla F(\boldsymbol{W}_t)\|_{S_q}}\right]\\ &= (1 - a_T)^2\mathbb{E}\left[\frac{\|\widetilde{\boldsymbol{E}}_{t-1}\|_F^2}{c + \|\nabla F(\boldsymbol{W}_t)\|_{S_q}}\right] + \mathbb{E}\left[\frac{\mathbb{E}[\|\boldsymbol{U}_t\|_F^2 \mid \mathcal{F}_{t-1}]}{c + \|\nabla F(\boldsymbol{W}_t)\|_{S_q}}\right]\\ &\le \frac{(1 - a_T)^2}{1 - h_T}\mathcal{Q}_{t-1} + \mathsf{K}_{\mathrm{AVR},T}^2\mathbb{E}\left[c + \|\nabla F(\boldsymbol{W}_t)\|_{S_q}\right]\\ &= \varrho_{\mathrm{AVR},T}\mathcal{Q}_{t-1} + \mathsf{K}_{\mathrm{AVR},T}^2\mathbb{E}\left[c + \|\nabla F(\boldsymbol{W}_t)\|_{S_q}\right].\end{aligned} \tag{101}$$

The identity $1 - \varrho_{\mathrm{AVR},T} = [a_T(2 - a_T) - h_T]/(1 - h_T) > 0$ confirms $\varrho_{\mathrm{AVR},T} < 1$. Since $\mathcal{Q}_0 = 0$, unrolling equation 101 gives

$$\begin{aligned}\mathcal{Q}_t &\le \varrho_{\mathrm{AVR},T}^t\mathcal{Q}_0 + \mathsf{K}_{\mathrm{AVR},T}^2\sum_{s=1}^{t}\varrho_{\mathrm{AVR},T}^{t-s}\mathbb{E}[c + \|\nabla F(\boldsymbol{W}_s)\|_{S_q}]\\ &= \mathsf{K}_{\mathrm{AVR},T}^2\sum_{s=1}^{t}\varrho_{\mathrm{AVR},T}^{t-s}\mathbb{E}\left[c + \|\nabla F(\boldsymbol{W}_s)\|_{S_q}\right].\end{aligned} \tag{102}$$

Define

$$\mathcal{P}_t := (1 - \varrho_{\mathrm{AVR},T})\sum_{s=1}^{t}\varrho_{\mathrm{AVR},T}^{t-s}\mathbb{E}\left[c + \|\nabla F(\boldsymbol{W}_s)\|_{S_q}\right]. \tag{103}$$

By Cauchy–Schwarz inequality, equation 102, and $2\sqrt{uv} \le u + v$,

$$\begin{aligned}\mathbb{E}\|\widetilde{\boldsymbol{E}}_t\|_F &= \mathbb{E}\left[\frac{\|\widetilde{\boldsymbol{E}}_t\|_F}{\sqrt{c + \|\nabla F(\boldsymbol{W}_t)\|_{S_q}}}\sqrt{c + \|\nabla F(\boldsymbol{W}_t)\|_{S_q}}\right]\\ &\le \sqrt{\mathcal{Q}_t\mathbb{E}\left[c + \|\nabla F(\boldsymbol{W}_t)\|_{S_q}\right]}\\ &\le \frac{\mathsf{K}_{\mathrm{AVR},T}}{\sqrt{1 - \varrho_{\mathrm{AVR},T}}}\sqrt{\mathcal{P}_t\mathbb{E}[c + \|\nabla F(\boldsymbol{W}_t)\|_{S_q}]}\\ &\le \frac{\mathsf{K}_{\mathrm{AVR},T}}{2\sqrt{1 - \varrho_{\mathrm{AVR},T}}}\left\{\mathcal{P}_t + \mathbb{E}\left[c + \|\nabla F(\boldsymbol{W}_t)\|_{S_q}\right]\right\}.\end{aligned} \tag{104}$$

Moreover,

$$\begin{aligned}\sum_{t=0}^{T-1}\mathcal{P}_t &= (1-\varrho_{\mathrm{AVR},T})\sum_{t=1}^{T-1}\sum_{s=1}^{t}\varrho_{\mathrm{AVR},T}^{t-s}\mathbb{E}[c+\|\nabla F(\boldsymbol{W}_s)\|_{S_q}]\\ &= \sum_{s=1}^{T-1}\mathbb{E}[c+\|\nabla F(\boldsymbol{W}_s)\|_{S_q}](1-\varrho_{\mathrm{AVR},T})\sum_{t=s}^{T-1}\varrho_{\mathrm{AVR},T}^{t-s}\\ &= \sum_{s=1}^{T-1}(1-\varrho_{\mathrm{AVR},T}^{T-s})\mathbb{E}[c+\|\nabla F(\boldsymbol{W}_s)\|_{S_q}]\\ &\le \sum_{s=1}^{T-1}\mathbb{E}[c+\|\nabla F(\boldsymbol{W}_s)\|_{S_q}] \le \sum_{t=0}^{T-1}\mathbb{E}\left[c+\|\nabla F(\boldsymbol{W}_t)\|_{S_q}\right]. \end{aligned} \tag{105}$$

For the initialization error, $\boldsymbol{M}_0 = \nabla f(\boldsymbol{W}_0;\xi_0)$, Cauchy–Schwarz inequality, and Assumption 5 give

$$\begin{aligned}\mathbb{E}\,\|\boldsymbol{M}_0-\nabla F(\boldsymbol{W}_0)\|_F &\le \left(\mathbb{E}\,\|\nabla f(\boldsymbol{W}_0;\xi_0)-\nabla F(\boldsymbol{W}_0)\|_F^2\right)^{1/2}\\ &\le \sqrt{\sigma_0^2+\sigma_1^2\|\nabla F(\boldsymbol{W}_0)\|_{S_q}^2} = \nu_0. \end{aligned} \tag{106}$$

Combining Lemma 3, equation 50, equation 97, equation 104, equation 105, and

$$\sum_{t=0}^{T-1}(1-a_T)^t = \frac{1-(1-a_T)^T}{1-(1-a_T)} = \frac{1-(1-a_T)^T}{a_T} \le \frac{1}{a_T}, \tag{107}$$

we obtain

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\mathcal{A}_t^{\mathrm{AVR}} &\le \frac{2r^{\max\{\frac{1}{q}-\frac{1}{2},0\}}}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\boldsymbol{M}_t-\nabla F(\boldsymbol{W}_t)\|_F\\ &\le \frac{2r^{\max\{\frac{1}{q}-\frac{1}{2},0\}}\nu_0}{T}\sum_{t=0}^{T-1}(1-a_T)^t + \frac{2r^{\max\{\frac{1}{q}-\frac{1}{2},0\}}}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\widetilde{\boldsymbol{E}}_t\|_F\\ &\le \frac{2r^{\max\{\frac{1}{q}-\frac{1}{2},0\}}\nu_0}{a_T T} + \frac{r^{\max\{\frac{1}{q}-\frac{1}{2},0\}}\mathsf{K}_{\mathrm{AVR},T}}{T\sqrt{1-\varrho_{\mathrm{AVR},T}}}\sum_{t=0}^{T-1}\left\{\mathcal{P}_t+\mathbb{E}[c+\|\nabla F(\boldsymbol{W}_t)\|_{S_q}]\right\}\\ &\le \frac{2r^{\max\left\{\frac{1}{q}-\frac{1}{2},0\right\}}\nu_0}{a_T T} + \frac{2r^{\max\left\{\frac{1}{q}-\frac{1}{2},0\right\}}\mathsf{K}_{\mathrm{AVR},T}}{\sqrt{1-\varrho_{\mathrm{AVR},T}}}\left[c+\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\,\|\nabla F(\boldsymbol{W}_t)\|_{S_q}\right]. \end{aligned} \tag{108}$$

By the definition of b_T in equation 51, this is equation 52. □

C CONVERGENCE OF SSMD-EMA

Under Assumption 1, ∇F is continuous on $\mathbb{R}^{m\times n}$. Consequently, its norm has a finite supremum on every compact update segment.

Theorem 6. *Consider Algorithm 2. Let $\boldsymbol{W}_0 \in \mathbb{R}^{m\times n}$ be a deterministic initial point and let $r := \min\{m,n\}$. Suppose Assumptions 1, 2, 3 and 5 hold, and suppose each ξ_t is sampled independently of the preceding history. Fix any $c>0$ and set $L_c := \max\{L_1, L_0/c\}$ and $\sigma_c := \max\{\sigma_1, \sigma_0/c\}$. For each horizon T, set $\alpha = \alpha_T := T^{-1/2}$, $\beta = \beta_T := 1-\alpha_T$, and $\eta = \eta_T := \eta_0 T^{-3/4}$. Then, for every integer $T \geq T_0$,*

$$\frac{1}{T}\sum_{t=0}^{T-1} \mathbb{E}\,\|\nabla F(\boldsymbol{W}_t)\|_{S_q} \leq \frac{\mathcal{C}_1^{\mathrm{EMA}}}{T^{\frac{1}{4}}} + \frac{\mathcal{C}_2^{\mathrm{EMA}}}{T^{\frac{1}{2}}} = \mathcal{O}\left(T^{-\frac{1}{4}}\right). \tag{109}$$

where $\mathcal{C}_1^{\mathrm{EMA}} := 2[(F(\boldsymbol{W}_0)-F^*)/\eta_0 + cb_0]/d_0$ and $\mathcal{C}_2^{\mathrm{EMA}} := 4r^{\max\{\frac{1}{q}-\frac{1}{2},0\}}\nu_0/d_0 + ch_0/[d_0(1-h_0)]$, and $\eta_0 > 0$ has been set such that $2/3 > h_0 := \eta_0 r^{2\max\{\frac{1}{2}-\frac{1}{p},0\}} L_c$, $d_0 := (2-3h_0)/(2(1-h_0))$, $b_0 := 2r^{\max\{\frac{1}{q}-\frac{1}{2},0\}}\sigma_c/\sqrt{1-h_0} + 2h_0/(1-h_0)$, $\nu_0 := \sqrt{\sigma_0^2 + \sigma_1^2\|\nabla F(\boldsymbol{W}_0)\|_{S_q}^2}$, and $T_0 := \lceil \max\{1, (2b_0/d_0)^4\}\rceil$.

The proof is given in Section C.1.

Lemma 10 (Alignment error of EMA momentum under affine noise)**.** *Suppose Assumptions 1, 2, 3, and 5 hold. For an integer $T \geq 1$, recall $\alpha_T = T^{-1/2}$, $\beta_T = 1-\alpha_T$, $\eta_T = \eta_0 T^{-3/4}$, and $r, c, L_c, \sigma_c, h_0, \nu_0$ from Theorem 4. Define $\mathcal{A}_t^{\mathrm{EMA}} := \|\nabla F(\boldsymbol{W}_t)\|_{S_q} - \langle \nabla F(\boldsymbol{W}_t), \mathcal{D}_p(\boldsymbol{M}_t)\rangle$, and define $h_T := \eta_T r^{2\max\{\frac{1}{2}-\frac{1}{p},0\}} L_c = h_0 T^{-\frac{3}{4}}$, $\varrho_{\mathrm{EMA},T} := \frac{\beta_T^2}{1-h_T}$, $b_T := \frac{2r^{\max\{\frac{1}{q}-\frac{1}{2},0\}}\alpha_T\sigma_c}{\sqrt{1-\varrho_{\mathrm{EMA},T}}} + \frac{2\beta_T h_T}{\alpha_T(1-h_T)}$. If $h_T < \alpha_T(2-\alpha_T)$, then*

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\mathcal{A}_t^{\mathrm{EMA}} \leq \frac{2r^{\max\{\frac{1}{q}-\frac{1}{2},0\}}\nu_0}{\alpha_T T} + b_T\left[c + \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\,\|\nabla F(\boldsymbol{W}_t)\|_{S_q}\right]. \tag{110}$$

C.1 PROOF OF THEOREM 4

Proof. Recall $\alpha_T = T^{-1/2}$, $\beta_T = 1-\alpha_T$, $\eta_T = \eta_0 T^{-3/4}$, and $h_T = \eta_T r^{2\max\{\frac{1}{2}-\frac{1}{p},0\}} L_c = h_0 T^{-3/4}$. Since $T \geq 1$ and $h_0 < 2/3 < 1$, we have

$$h_T = h_0 T^{-\frac{3}{4}} < T^{-\frac{3}{4}} \leq T^{-\frac{1}{2}} \leq T^{-\frac{1}{2}}\left(2 - T^{-\frac{1}{2}}\right) = \alpha_T(2-\alpha_T), \tag{111}$$

where both middle inequalities follow from $T \geq 1$. Thus, Lemma 10 applies.

Since $\|\mathcal{D}_p(\boldsymbol{M}_t)\|_{S_p} \leq 1$, Lemma 3 gives $\|\mathcal{D}_p(\boldsymbol{M}_t)\|_{S_{\bar{p}}} \leq r^{\max\{\frac{1}{2}-\frac{1}{p},0\}}$. Thus, Assumption 3, the same norm comparison, and $L_0 + L_1 g \leq L_c(c+g)$ for $g \geq 0$ imply, for every $\theta \in [0,1]$,

$$\left\|\nabla F\left(\boldsymbol{W}_t - \theta\eta_T \mathcal{D}_p(\boldsymbol{M}_t)\right)\right\|_{S_q} \leq \|\nabla F(\boldsymbol{W}_t)\|_{S_q} + h_T\left[c + \sup_{u\in[0,1]}\left\|\nabla F\left(\boldsymbol{W}_t - u\eta_T\mathcal{D}_p(\boldsymbol{M}_t)\right)\right\|_{S_q}\right].$$

Taking the supremum over θ and using its finiteness and $h_T < 1$, we obtain

$$c + \sup_{\theta\in[0,1]}\left\|\nabla F\left(\boldsymbol{W}_t - \theta\eta_T\mathcal{D}_p(\boldsymbol{M}_t)\right)\right\|_{S_q} \leq \frac{c + \|\nabla F(\boldsymbol{W}_t)\|_{S_q}}{1-h_T}.$$

Applying Lemma 1(i), specifically equation 9, with $\boldsymbol{W} = \boldsymbol{W}_t$ and $\boldsymbol{Z} = -\eta_T\mathcal{D}_p(\boldsymbol{M}_t)$ therefore gives

$$\begin{aligned} F(\boldsymbol{W}_{t+1}) &\leq F(\boldsymbol{W}_t) - \eta_T\langle\nabla F(\boldsymbol{W}_t), \mathcal{D}_p(\boldsymbol{M}_t)\rangle + \frac{\eta_T^2 r^{2\max\{\frac{1}{2}-\frac{1}{p},0\}} L_c}{2(1-h_T)}\left[c + \|\nabla F(\boldsymbol{W}_t)\|_{S_q}\right] \\ &= F(\boldsymbol{W}_t) - \eta_T\frac{2-3h_T}{2(1-h_T)}\|\nabla F(\boldsymbol{W}_t)\|_{S_q} + \eta_T\mathcal{A}_t^{\mathrm{EMA}} + \frac{c\eta_T h_T}{2(1-h_T)}, \end{aligned}$$

where the equality uses the definitions of h_T and $\mathcal{A}_t^{\mathrm{EMA}}$. Taking expectations, summing over $t = 0, \dots, T-1$, using $F(\boldsymbol{W}_T) \geq F^*$, and dividing by $\eta_T T$, we obtain

$$\frac{2-3h_T}{2(1-h_T)} \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left\| \nabla F(\boldsymbol{W}_t) \right\|_{S_q} \leq \frac{F(\boldsymbol{W}_0) - F^*}{\eta_T T} + \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \mathcal{A}_t^{\mathrm{EMA}} + \frac{c h_T}{2(1-h_T)}. \tag{112}$$

Substituting equation 110 into equation 112 and collecting the average-gradient terms gives

$$\left[\frac{2-3h_T}{2(1-h_T)} - b_T \right] \frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left\| \nabla F(\boldsymbol{W}_t) \right\|_{S_q} \leq \frac{F(\boldsymbol{W}_0) - F^*}{\eta_T T} + \frac{2 r^{\max\left\{\frac{1}{q} - \frac{1}{2}, 0\right\}} \nu_0}{\alpha_T T} + c b_T + \frac{c h_T}{2(1-h_T)}. \tag{113}$$

Using $\beta_T = 1 - T^{-1/2}$, $\alpha_T = T^{-1/2}$, and $h_T = h_0 T^{-3/4}$, we obtain

$$\begin{aligned} b_T &= 2T^{-\frac{1}{4}} \left[r^{\max\left\{\frac{1}{q} - \frac{1}{2}, 0\right\}} \sigma_c \sqrt{\frac{1 - h_0 T^{-\frac{3}{4}}}{2 - T^{-\frac{1}{2}} - h_0 T^{-\frac{1}{4}}}} + \frac{h_0 (1 - T^{-\frac{1}{2}})}{1 - h_0 T^{-\frac{3}{4}}} \right] \\ &\leq 2 \left[\frac{r^{\max\left\{\frac{1}{q} - \frac{1}{2}, 0\right\}} \sigma_c}{\sqrt{1-h_0}} + \frac{h_0}{1-h_0} \right] T^{-\frac{1}{4}} = b_0 T^{-\frac{1}{4}}. \end{aligned} \tag{114}$$

Indeed, $1 - h_0 T^{-3/4} \leq 1$ and

$$2 - T^{-\frac{1}{2}} - h_0 T^{-\frac{1}{4}} = 1 - h_0 + 1 - T^{-\frac{1}{2}} + h_0 (1 - T^{-\frac{1}{4}}) \geq 1 - h_0, \tag{115}$$

so $\sqrt{(1 - h_0 T^{-3/4})/(2 - T^{-1/2} - h_0 T^{-1/4})} \leq 1/\sqrt{1-h_0}$. The second term follows from $1 - T^{-1/2} \leq 1$ and $1 - h_0 T^{-3/4} \geq 1 - h_0$.

Since $L_c \geq 0$, $\eta_0 > 0$, and $r \geq 1$, we have $0 \leq h_T \leq h_0 < 2/3$. Hence $1 - h_T > 0$ and $d_0 = (2 - 3h_0)/(2(1-h_0)) > 0$. Moreover, $\sigma_c \geq 0$ implies $b_0 \geq 0$. We also have

$$\frac{2-3h_T}{2(1-h_T)} - d_0 = \frac{2-3h_T}{2(1-h_T)} - \frac{2-3h_0}{2(1-h_0)} = \frac{h_0 - h_T}{2(1-h_T)(1-h_0)} \geq 0. \tag{116}$$

Moreover, if $b_0 = 0$, then $b_T = 0 \leq d_0/2$. If $b_0 > 0$, the definition of T_0 gives $T^{1/4} \geq 2b_0/d_0$, and hence

$$b_T \leq b_0 T^{-\frac{1}{4}} = \frac{b_0}{T^{\frac{1}{4}}} \leq \frac{b_0}{2b_0/d_0} = \frac{d_0}{2}. \tag{117}$$

Combining equation 116 and equation 117 gives

$$\frac{2-3h_T}{2(1-h_T)} - b_T \geq \frac{d_0}{2}. \tag{118}$$

The terms on the right-hand side of equation 113 satisfy

$$\frac{F(\boldsymbol{W}_0) - F^*}{\eta_T T} = \frac{F(\boldsymbol{W}_0) - F^*}{\eta_0} T^{-\frac{1}{4}}, \tag{119}$$

$$\frac{2 r^{\max\left\{\frac{1}{q} - \frac{1}{2}, 0\right\}} \nu_0}{\alpha_T T} = 2 r^{\max\left\{\frac{1}{q} - \frac{1}{2}, 0\right\}} \nu_0 T^{-\frac{1}{2}}, \tag{120}$$

$$c b_T \leq c b_0 T^{-\frac{1}{4}}, \tag{121}$$

$$\frac{c h_T}{2(1-h_T)} = \frac{c h_0 T^{-\frac{3}{4}}}{2(1 - h_0 T^{-\frac{3}{4}})} \leq \frac{c h_0}{2(1-h_0)} T^{-\frac{3}{4}}. \tag{122}$$

Combining equation 113, equation 118, and equation 122, and using $T^{-3/4} \leq T^{-1/2}$, we obtain

$$\frac{1}{T} \sum_{t=0}^{T-1} \mathbb{E} \left\| \nabla F(\boldsymbol{W}_t) \right\|_{S_q} \leq \frac{2}{d_0} \left[\frac{F(\boldsymbol{W}_0) - F^*}{\eta_0} + c b_0 \right] T^{-\frac{1}{4}} + \left[\frac{4 r^{\max\left\{\frac{1}{q} - \frac{1}{2}, 0\right\}} \nu_0}{d_0} + \frac{c h_0}{d_0 (1-h_0)} \right] T^{-\frac{1}{2}}. \tag{123}$$

This proves equation 109. □

C.2 Proof of the Intermediate Lemma

Proof of Lemma 10. Fix a horizon T satisfying $h_T < \alpha_T(2-\alpha_T)$. The parameters α_T, β_T, and η_T are fixed with respect to the iteration index t. Since $\alpha_T \in (0,1]$, the condition also gives $h_T < \alpha_T(2-\alpha_T) \le 1$. For any $g \ge 0$, the definitions of L_c and σ_c give

$$\begin{aligned} L_0 + L_1 g &\le \max\left\{L_1, \frac{L_0}{c}\right\}(c+g) = L_c(c+g), \\ \sqrt{\sigma_0^2 + \sigma_1^2 g^2} &\le \sigma_0 + \sigma_1 g \le \max\left\{\sigma_1, \frac{\sigma_0}{c}\right\}(c+g) = \sigma_c(c+g). \end{aligned} \tag{124}$$

For $t \ge 1$, let $\mathcal{F}_{t-1} := \sigma(\boldsymbol{W}_0, \xi_0, \dots, \xi_{t-1})$. Then $\boldsymbol{W}_t$ and $\boldsymbol{M}_{t-1}$ are $\mathcal{F}_{t-1}$-measurable, while ξ_t is independent of $\mathcal{F}_{t-1}$. Define

$$\boldsymbol{\Xi}_t := \nabla f(\boldsymbol{W}_t; \xi_t) - \nabla F(\boldsymbol{W}_t). \tag{125}$$

Assumptions 2 and 5 give

$$\begin{aligned} &\mathbb{E}[\boldsymbol{\Xi}_t \mid \mathcal{F}_{t-1}] = \mathbf{0}, \\ &\mathbb{E}[\|\boldsymbol{\Xi}_t\|_F^2 \mid \mathcal{F}_{t-1}] \le \sigma_0^2 + \sigma_1^2 \|\nabla F(\boldsymbol{W}_t)\|_{S_q}^2 \le \left[\sigma_0 + \sigma_1 \|\nabla F(\boldsymbol{W}_t)\|_{S_q}\right]^2 \le \sigma_c^2 \left[c + \|\nabla F(\boldsymbol{W}_t)\|_{S_q}\right]^2. \end{aligned} \tag{126}$$

For $t \ge 1$, define the previous update segment and its gradient supremum by

$$\boldsymbol{W}_t^-(\theta) := (1-\theta)\boldsymbol{W}_t + \theta \boldsymbol{W}_{t-1}, \qquad \mathcal{S}_t^- := \sup_{\theta \in [0,1]} \left\|\nabla F(\boldsymbol{W}_t^-(\theta))\right\|_{S_q}. \tag{127}$$

The update rule, and Lemma 3 give

$$\|\boldsymbol{W}_t - \boldsymbol{W}_{t-1}\|_{S_p} = \eta_T \|\mathcal{D}_p(\boldsymbol{M}_{t-1})\|_{S_p} \le \eta_T, \tag{128}$$

$$\|\boldsymbol{W}_t - \boldsymbol{W}_{t-1}\|_{S_{\bar{p}}} \le r^{\max\left\{\frac{1}{\bar{p}} - \frac{1}{p}, 0\right\}} \|\boldsymbol{W}_t - \boldsymbol{W}_{t-1}\|_{S_p} \le \eta_T r^{\max\left\{\frac{1}{q} - \frac{1}{2}, 0\right\}}. \tag{129}$$

Therefore, for every $\theta \in [0,1]$, the triangle inequality, the Schatten norm-comparison inequalities, and Assumption 3 imply

$$\begin{aligned} \left\|\nabla F(\boldsymbol{W}_t^-(\theta))\right\|_{S_q} &\le \|\nabla F(\boldsymbol{W}_t)\|_{S_q} + \left\|\nabla F(\boldsymbol{W}_t^-(\theta)) - \nabla F(\boldsymbol{W}_t)\right\|_{S_q} \\ &\le \|\nabla F(\boldsymbol{W}_t)\|_{S_q} + r^{\max\left\{\frac{1}{q} - \frac{1}{2}, 0\right\}} \left\|\nabla F(\boldsymbol{W}_t^-(\theta)) - \nabla F(\boldsymbol{W}_t)\right\|_{S_{\bar{q}}} \\ &\le \|\nabla F(\boldsymbol{W}_t)\|_{S_q} + r^{\max\left\{\frac{1}{q} - \frac{1}{2}, 0\right\}} (L_0 + L_1 \mathcal{S}_t^-) \|\boldsymbol{W}_t^-(\theta) - \boldsymbol{W}_t\|_{S_{\bar{p}}} \\ &= \|\nabla F(\boldsymbol{W}_t)\|_{S_q} + \theta r^{\max\left\{\frac{1}{q} - \frac{1}{2}, 0\right\}} (L_0 + L_1 \mathcal{S}_t^-) \|\boldsymbol{W}_t - \boldsymbol{W}_{t-1}\|_{S_{\bar{p}}} \\ &\le \|\nabla F(\boldsymbol{W}_t)\|_{S_q} + \eta_T r^{2\max\left\{\frac{1}{q} - \frac{1}{2}, 0\right\}} (L_0 + L_1 \mathcal{S}_t^-). \end{aligned} \tag{130}$$

Taking the supremum over θ, using equation 124, and recalling $\max\{\frac{1}{q} - \frac{1}{2}, 0\} = \max\{\frac{1}{2} - \frac{1}{p}, 0\}$ give

$$\begin{aligned} c + \mathcal{S}_t^- &\le c + \|\nabla F(\boldsymbol{W}_t)\|_{S_q} + h_T(c + \mathcal{S}_t^-), \\ c + \mathcal{S}_t^- &\le \frac{c + \|\nabla F(\boldsymbol{W}_t)\|_{S_q}}{1 - h_T}. \end{aligned} \tag{131}$$

Since $\boldsymbol{W}_{t-1} = \boldsymbol{W}_t^-(1)$, equation 131 also yields

$$c + \|\nabla F(\boldsymbol{W}_t)\|_{S_q} \ge (1 - h_T)\left[c + \|\nabla F(\boldsymbol{W}_{t-1})\|_{S_q}\right]. \tag{132}$$

The definitions $\bar{p} = \min\{p, 2\}$ and $\bar{q} = \max\{q, 2\}$, the Schatten norm-comparison inequalities, equation 129, and Assumption 3 imply

$$\begin{aligned} \|\nabla F(\boldsymbol{W}_t) - \nabla F(\boldsymbol{W}_{t-1})\|_{S_q} &\le r^{\max\left\{\frac{1}{q} - \frac{1}{2}, 0\right\}} \|\nabla F(\boldsymbol{W}_t) - \nabla F(\boldsymbol{W}_{t-1})\|_{S_{\bar{q}}} \\ &\le r^{\max\left\{\frac{1}{q} - \frac{1}{2}, 0\right\}} (L_0 + L_1 \mathcal{S}_t^-) \|\boldsymbol{W}_t - \boldsymbol{W}_{t-1}\|_{S_{\bar{p}}} \\ &\le \eta_T r^{2\max\left\{\frac{1}{q} - \frac{1}{2}, 0\right\}} L_c(c + \mathcal{S}_t^-) \\ &\le \frac{h_T}{1 - h_T}\left[c + \|\nabla F(\boldsymbol{W}_t)\|_{S_q}\right], \end{aligned} \tag{133}$$

where $\max\{\frac{1}{q}-\frac{1}{2},0\}=\max\{\frac{1}{2}-\frac{1}{p},0\}$ was used in the last two lines.

Define $\boldsymbol{N}_0=\mathbf{0}$ and $\boldsymbol{R}_0=\mathbf{0}$, and, for $t\geq 1$, define

$$\boldsymbol{N}_t := \beta_T \boldsymbol{N}_{t-1} + \alpha_T \boldsymbol{\Xi}_t, \tag{134}$$

$$\boldsymbol{R}_t := \beta_T \boldsymbol{R}_{t-1} + \beta_T \left[\nabla F(\boldsymbol{W}_{t-1}) - \nabla F(\boldsymbol{W}_t)\right]. \tag{135}$$

Subtracting $\nabla F(\boldsymbol{W}_t)$ from the EMA update and iterating the resulting recursion gives

$$\boldsymbol{M}_t - \nabla F(\boldsymbol{W}_t) = \beta_T^t \left[\boldsymbol{M}_0 - \nabla F(\boldsymbol{W}_0)\right] + \boldsymbol{N}_t + \boldsymbol{R}_t. \tag{136}$$

We first control the centered noise term. Define

$$\mathcal{Q}_t := \mathbb{E}\left[\frac{\|\boldsymbol{N}_t\|_F^2}{c+\|\nabla F(\boldsymbol{W}_t)\|_{S_q}}\right]. \tag{137}$$

Using the conditional centering in equation 126, the Frobenius inner product, and the $\mathcal{F}_{t-1}$-measurability of $\boldsymbol{N}_{t-1}$, the cross term satisfies

$$\mathbb{E}\left[\langle \boldsymbol{N}_{t-1}, \boldsymbol{\Xi}_t\rangle_F \mid \mathcal{F}_{t-1}\right] = \langle \boldsymbol{N}_{t-1}, \mathbb{E}[\boldsymbol{\Xi}_t \mid \mathcal{F}_{t-1}]\rangle_F = 0. \tag{138}$$

Therefore, equation 132 gives

$$\begin{aligned}\mathcal{Q}_t &= \mathbb{E}\left[\frac{\beta_T^2\|\boldsymbol{N}_{t-1}\|_F^2}{c+\|\nabla F(\boldsymbol{W}_t)\|_{S_q}}\right] + \alpha_T^2\mathbb{E}\left[\frac{\mathbb{E}[\|\boldsymbol{\Xi}_t\|_F^2 \mid \mathcal{F}_{t-1}]}{c+\|\nabla F(\boldsymbol{W}_t)\|_{S_q}}\right] \\ &\leq \frac{\beta_T^2}{1-h_T}\mathcal{Q}_{t-1} + \alpha_T^2\sigma_c^2\mathbb{E}\left[c+\|\nabla F(\boldsymbol{W}_t)\|_{S_q}\right] = \varrho_{\mathrm{EMA},T}\mathcal{Q}_{t-1} + \alpha_T^2\sigma_c^2\mathbb{E}\left[c+\|\nabla F(\boldsymbol{W}_t)\|_{S_q}\right].\end{aligned} \tag{139}$$

The condition $h_T<\alpha_T(2-\alpha_T)$ gives $\beta_T^2=(1-\alpha_T)^2<1-h_T$ and hence $\varrho_{\mathrm{EMA},T}<1$. Since $\mathcal{Q}_0=0$, unrolling equation 139 gives

$$\mathcal{Q}_t \leq \alpha_T^2\sigma_c^2\sum_{s=1}^{t}\varrho_{\mathrm{EMA},T}^{t-s}\mathbb{E}\left[c+\|\nabla F(\boldsymbol{W}_s)\|_{S_q}\right]. \tag{140}$$

Define

$$\mathcal{P}_t := (1-\varrho_{\mathrm{EMA},T})\sum_{s=1}^{t}\varrho_{\mathrm{EMA},T}^{t-s}\mathbb{E}\left[c+\|\nabla F(\boldsymbol{W}_s)\|_{S_q}\right]. \tag{141}$$

Cauchy–Schwarz inequality, equation 140, and $2\sqrt{uv}\leq u+v$ give

$$\begin{aligned}\mathbb{E}\|\boldsymbol{N}_t\|_F &\leq \sqrt{\mathcal{Q}_t\mathbb{E}\left[c+\|\nabla F(\boldsymbol{W}_t)\|_{S_q}\right]} \leq \frac{\alpha_T\sigma_c}{\sqrt{1-\varrho_{\mathrm{EMA},T}}}\sqrt{\mathcal{P}_t\mathbb{E}\left[c+\|\nabla F(\boldsymbol{W}_t)\|_{S_q}\right]} \\ &\leq \frac{\alpha_T\sigma_c}{2\sqrt{1-\varrho_{\mathrm{EMA},T}}}\left\{\mathcal{P}_t+\mathbb{E}\left[c+\|\nabla F(\boldsymbol{W}_t)\|_{S_q}\right]\right\},\end{aligned} \tag{142}$$

Moreover, summing the geometric weights gives

$$\sum_{t=0}^{T-1}\mathcal{P}_t = (1-\varrho_{\mathrm{EMA},T})\sum_{s=1}^{T-1}\mathbb{E}\left[c+\|\nabla F(\boldsymbol{W}_s)\|_{S_q}\right]\sum_{t=s}^{T-1}\varrho_{\mathrm{EMA},T}^{t-s} \leq \sum_{t=0}^{T-1}\mathbb{E}\left[c+\|\nabla F(\boldsymbol{W}_t)\|_{S_q}\right], \tag{143}$$

and hence

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\boldsymbol{N}_t\|_F \leq \frac{\alpha_T\sigma_c}{\sqrt{1-\varrho_{\mathrm{EMA},T}}}\left[c+\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(\boldsymbol{W}_t)\|_{S_q}\right]. \tag{144}$$

For the gradient-drift term, equation 135 and equation 133 give

$$\|\boldsymbol{R}_t\|_{S_q} \leq \frac{\beta_T h_T}{1-h_T}\sum_{s=1}^{t}\beta_T^{t-s}\left[c+\|\nabla F(\boldsymbol{W}_s)\|_{S_q}\right], \tag{145}$$

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\boldsymbol{R}_t\|_{S_q} &\leq \frac{\beta_T h_T}{T(1-h_T)}\sum_{s=1}^{T-1}\mathbb{E}\left[c+\|\nabla F(\boldsymbol{W}_s)\|_{S_q}\right]\sum_{t=s}^{T-1}\beta_T^{t-s} \\ &\leq \frac{\beta_T h_T}{\alpha_T(1-h_T)}\left[c+\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(\boldsymbol{W}_t)\|_{S_q}\right].\end{aligned} \tag{146}$$

Assumption 5 also gives

$$\mathbb{E}\|\boldsymbol{M}_0 - \nabla F(\boldsymbol{W}_0)\|_F \le \sqrt{\mathbb{E}\|\boldsymbol{M}_0 - \nabla F(\boldsymbol{W}_0)\|_F^2} \le \nu_0. \tag{147}$$

Finally, the Schatten norm-comparison inequality and equation 50 give

$$\mathbb{E}\mathcal{A}_t^{\mathrm{EMA}} \le 2r^{\max\left\{\frac{1}{q}-\frac{1}{2},0\right\}}\beta_T^t\mathbb{E}\|\boldsymbol{M}_0 - \nabla F(\boldsymbol{W}_0)\|_F + 2r^{\max\left\{\frac{1}{q}-\frac{1}{2},0\right\}}\mathbb{E}\|\boldsymbol{N}_t\|_F + 2\mathbb{E}\|\boldsymbol{R}_t\|_{S_q}. \tag{148}$$

Combining equation 144, equation 146, equation 147, and $\sum_{t=0}^{T-1}\beta_T^t \le 1/\alpha_T$, we obtain

$$\begin{aligned}\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\mathcal{A}_t^{\mathrm{EMA}} &\le \frac{2r^{\max\left\{\frac{1}{q}-\frac{1}{2},0\right\}}\nu_0}{\alpha_T T} + \left[\frac{2r^{\max\left\{\frac{1}{q}-\frac{1}{2},0\right\}}\alpha_T\sigma_c}{\sqrt{1-\varrho_{\mathrm{EMA},T}}} + \frac{2\beta_T h_T}{\alpha_T(1-h_T)}\right]\left[c + \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(\boldsymbol{W}_t)\|_{S_q}\right] \\ &= \frac{2r^{\max\left\{\frac{1}{q}-\frac{1}{2},0\right\}}\nu_0}{\alpha_T T} + b_T\left[c + \frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(\boldsymbol{W}_t)\|_{S_q}\right].\end{aligned} \tag{149}$$

This proves equation 110. □

C.3 Convergence of SSMD(Polyak's Heavy Ball)

We consider SSMD with Polyak's heavy-ball momentum, as specified in Algorithm 3. The proof of Corollary 1 reuses the argument of Lemma 10. By Lemma 11, the scaled momentum $(1-\mu_T)\boldsymbol{M}_{t+1}$ satisfies the EMA recursion at $\boldsymbol{W}_t$ with $\beta_T = \mu_T$ and a correspondingly scaled initialization. Positive scaling preserves the set of Schatten dual maximizers, so the EMA argument applies along the Polyak iterates with the same selected update directions. Specifically, the error decomposition equation 136 and the centered-noise and gradient-drift bounds equation 144 and equation 146 remain valid.

The only adjustment is to replace the initialization bound equation 147 by equation 151 below. Consequently, the average alignment bound equation 110 holds with ν_0 replaced by $\nu_0 + \|\nabla F(\boldsymbol{W}_0)\|_F$. Combining this bound with the generalized descent inequality equation 112 and the coefficient bound equation 118 yields the corollary. This replacement changes only $\mathcal{C}_2^{\mathrm{EMA}}$, while $\mathcal{C}_1^{\mathrm{EMA}}$, b_0, d_0, and T_0 remain unchanged.

Lemma 11. *Let $\mu \in [0,1)$ and let $\{\boldsymbol{G}_t\}_{t\ge 0} \subset \mathbb{R}^{m\times n}$ be any matrix sequence. Define Polyak's heavy-ball momentum by $\boldsymbol{B}_0 = \boldsymbol{0}$ and $\boldsymbol{B}_{t+1} = \mu\boldsymbol{B}_t + \boldsymbol{G}_t$, and define the EMA momentum by $\boldsymbol{M}_0 = \boldsymbol{0}$ and $\boldsymbol{M}_{t+1} = \mu\boldsymbol{M}_t + (1-\mu)\boldsymbol{G}_t$. Then, for every integer $t \ge 0$,*

$$\boldsymbol{M}_t = (1-\mu)\boldsymbol{B}_t. \tag{150}$$

If the matrices $\boldsymbol{G}_t$ are random, the identity holds for every realized sequence.

The proof of Lemma 11 is in Appendix C.3.

Proof of Corollary 1. Fix $T \ge T_0$ and recall $\alpha_T = 1-\mu_T = T^{-1/2}$ and $\beta_T = \mu_T$. Since $\boldsymbol{M}_1 = \boldsymbol{G}_0$, the sequence $\boldsymbol{0}, \boldsymbol{M}_1, \boldsymbol{M}_2, \dots$ satisfies the zero-initialized Polyak recursion driven by $\boldsymbol{G}_t = \nabla f(\boldsymbol{W}_t;\xi_t)$. Applying Lemma 11 to each realized gradient sequence identifies $\alpha_T\boldsymbol{M}_{t+1}$ as the corresponding EMA momentum at $\boldsymbol{W}_t$. Positive scaling preserves the set of Schatten dual maximizers, so the proof of Lemma 10 applies with the same selected directions and learning rate, subject to the initialization estimate below.

Since $\boldsymbol{M}_1 = \boldsymbol{G}_0$, the triangle inequality and Assumption 5 give

$$\begin{aligned}\mathbb{E}\|\alpha_T\boldsymbol{M}_1 - \nabla F(\boldsymbol{W}_0)\|_F &\le \alpha_T\sqrt{\sigma_0^2 + \sigma_1^2\|\nabla F(\boldsymbol{W}_0)\|_{S_q}^2} + \mu_T\|\nabla F(\boldsymbol{W}_0)\|_F \\ &\le \nu_0 + \|\nabla F(\boldsymbol{W}_0)\|_F.\end{aligned} \tag{151}$$

The initialization enters the EMA alignment proof only through equation 147. Substituting the bound above in that step leaves all centered-noise and gradient-drift estimates unchanged. Therefore, the proof of Theorem 4 applies with ν_0 replaced by $\nu_0 + \|\nabla F(\boldsymbol{W}_0)\|_F$. Since b_0, d_0, T_0 are

unchanged, it gives

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\|\nabla F(\boldsymbol{W}_t)\|_{S_q} \le \frac{\mathcal{C}_1^{\mathrm{EMA}}}{T^{1/4}} + \left[\mathcal{C}_2^{\mathrm{EMA}} + \frac{4r^{\max\{\frac{1}{q}-\frac{1}{2},0\}}\|\nabla F(\boldsymbol{W}_0)\|_F}{d_0}\right] T^{-1/2}$$
$$= \frac{\mathcal{C}_1^{\mathrm{M}}}{T^{1/4}} + \frac{\mathcal{C}_2^{\mathrm{M}}}{T^{1/2}}. \tag{152}$$

This proves equation 13. □

Proof of Lemma 11. Since $\boldsymbol{M}_0 = \boldsymbol{0} = (1-\mu)\boldsymbol{B}_0$, the identity holds for $t = 0$. Suppose $\boldsymbol{M}_t = (1-\mu)\boldsymbol{B}_t$ for some $t \ge 0$. Then

$$\boldsymbol{M}_{t+1} = \mu\boldsymbol{M}_t + (1-\mu)\boldsymbol{G}_t = \mu(1-\mu)\boldsymbol{B}_t + (1-\mu)\boldsymbol{G}_t$$
$$= (1-\mu)(\mu\boldsymbol{B}_t + \boldsymbol{G}_t) = (1-\mu)\boldsymbol{B}_{t+1}. \tag{153}$$

The conclusion follows by induction. Since the argument is purely algebraic, it also applies to every realization of a random matrix sequence. □

D EXPERIMENTAL DETAILS

D.1 CIFAR-100

We use CIFAR-100 with 50,000 training images and 10,000 test images. The model is DeiT-Tiny, with 12 Transformer blocks, hidden dimension 192, three attention heads, and patch size 16. Inputs have resolution 224×224, and the classification head has 100 outputs. The model is randomly initialized without pretrained weights, and all parameters are trained.

Each configuration is trained for 100 epochs, with 390 optimizer updates per epoch and terminal horizon $H = 39{,}000$. The batch size is 128 without gradient accumulation. We use training seed 42 for each combination of momentum method and Schatten geometry. Training uses BF16 mixed precision and zero weight decay.

For the theorem-scaled runs, momentum is initialized using the first mini-batch gradient, and gradient clipping is disabled. Learning rates and momentum coefficients remain fixed throughout training.

For the empirical runs, all parameter blocks receive Schatten updates, with directions computed using exact SVDs. The learning rate starts at 10^{-5} and warms up over five epochs, followed by cosine decay to 10^{-6}. The peak learning rates and momentum coefficients are listed in Table 5.

D.2 GPT-2 PRETRAINING

We train GPT-2 on tokenized FineWeb data using separate training and validation shards. The model contains 124,439,808 unique parameters and has 12 Transformer blocks, hidden dimension 768, 12 attention heads, and MLP dimension 3072. The sequence length is 1024, and the vocabulary size is 50,257. The token embedding and output projection share weights. The model is randomly initialized, and all parameters are trained. Both settings use BF16 mixed precision and $H = 6{,}000$ optimizer updates.

Theorem-scaled setting. We use a micro-batch size of eight sequences and accumulate gradients over eight micro-batches. The effective batch size is 64 sequences, corresponding to 65,536 tokens per update and 393,216,000 sampled training tokens per run. We use training seeds 1337, 2027, and 2029, with all three runs sharing initialization seed 1337. Learning rates and momentum coefficients remain fixed throughout training, with no warmup, no learning-rate decay, and zero weight decay.

Empirical setting. We use a micro-batch size of eight sequences and accumulate gradients over 64 micro-batches. The effective batch size is 512 sequences, corresponding to 524,288 tokens per update and 3,145,728,000 sampled training tokens per run. These runs use seed 1337.

Transformer two-dimensional weight matrices receive Schatten updates. Directions are computed using 20 power iterations for $p = 1$, Frobenius normalization for $p = 2$, and 12 Newton–Schulz iterations for $p = \infty$. Thus, the endpoint directions are computed approximately in this setting. The remaining parameters are optimized with AdamW, using learning rate 6×10^{-4} and momentum coefficients $(0.9, 0.95)$.

The empirical configuration uses weight decay 0.1. The learning-rate schedule consists of 250 updates of linear warmup followed by cosine decay, with a minimum learning rate equal to 0.1 times the peak value. The Schatten optimizer parameters are listed in Table 5.

D.3 THEOREM-SCALED PARAMETERS

For both tasks, the theorem-scaled runs follow

$$\text{MVR:} \qquad a = H^{-2/3}, \qquad \eta = \eta_0 H^{-2/3}, \tag{154}$$

$$\text{EMA:} \qquad 1 - \beta = H^{-1/2}, \qquad \eta = \eta_0 H^{-3/4}. \tag{155}$$

Here, $H = 39{,}000$ for CIFAR-100 and $H = 6{,}000$ for GPT-2. We use $\eta_0 = 0.008$ for $p \in \{1, 2\}$ and $\eta_0 = 0.004$ for $p = \infty$. The MVR correction retains the previous estimation error with coefficient $1 - a$, while EMA retains the previous momentum with coefficient β.

Parameter	CIFAR-100 MVR	CIFAR-100 EMA	GPT-2 MVR	GPT-2 EMA
Terminal horizon H	39,000	39,000	6,000	6,000
Learning rate, $p \in \{1, 2\}$	6.95633×10^{-6}	2.88265×10^{-6}	2.42283×10^{-5}	1.17348×10^{-5}
Learning rate, $p = \infty$	3.47817×10^{-6}	1.44132×10^{-6}	1.21141×10^{-5}	5.86741×10^{-6}
Momentum parameter	$a = 8.69541 \times 10^{-4}$	$\beta = 0.99493630$	$a = 0.00302853$	$\beta = 0.98709006$

Table 4: Parameters for the theorem-scaled runs. Learning rates are the actual constant values used during training, rather than the base scale η_0. Values are rounded for presentation.

Dataset	p	MVR peak LR	MVR a	EMA peak LR	EMA β
CIFAR-100	$1, 2$	0.008	0.1	0.008	0.9
CIFAR-100	∞	0.004	0.1	0.004	0.9
GPT-2	$1, 2, \infty$	0.02	0.05	0.02	0.95

Table 5: Schatten optimizer parameters for the empirical runs. Learning rates are peak values reached after warmup. The auxiliary AdamW optimizer in GPT-2 uses the separate settings described above.

The resulting values are listed in Table 4. These parameters are computed from the terminal horizon before training and remain fixed throughout each run.

These choices follow the horizon exponents in the convergence theorems. Using these exponents alone does not certify the remaining assumptions or step-size conditions of the theorems, nor does it establish the best task-specific hyperparameters.

D.4 Empirical Parameters

Table 5 lists the parameters used in the empirical figures. At each fixed p, MVR and EMA use the same peak learning rate and satisfy $a = 1 - \beta$. Their learning-rate schedules are specified in the task descriptions above.

D.5 Hyperparameter Search Ranges

CIFAR-100. For MVR, we searched learning rates in $\{0.00025, 0.0005, 0.001, 0.002, 0.004, 0.008\}$ for each $p \in \{1, 2, \infty\}$, with $a = 0.1$ fixed.

GPT-2. For MVR, the search considered learning rates in $\{0.01, 0.03, 0.05, 0.07, 0.1\}$ and $a \in \{0.01, 0.05, 0.1, 1.0\}$.

For the empirical comparisons, EMA uses the same learning rates as MVR and momentum coefficient $\beta = 1 - a$, as listed in Table 5.

D.6 Additional results

We provide training and test loss curves under both theorem-scaled and empirical settings for CIFAR-100 (Figures 5 and 7) and GPT-2 (Figures 6 and 8). These figures supplement the main-text comparisons using the configurations described above.

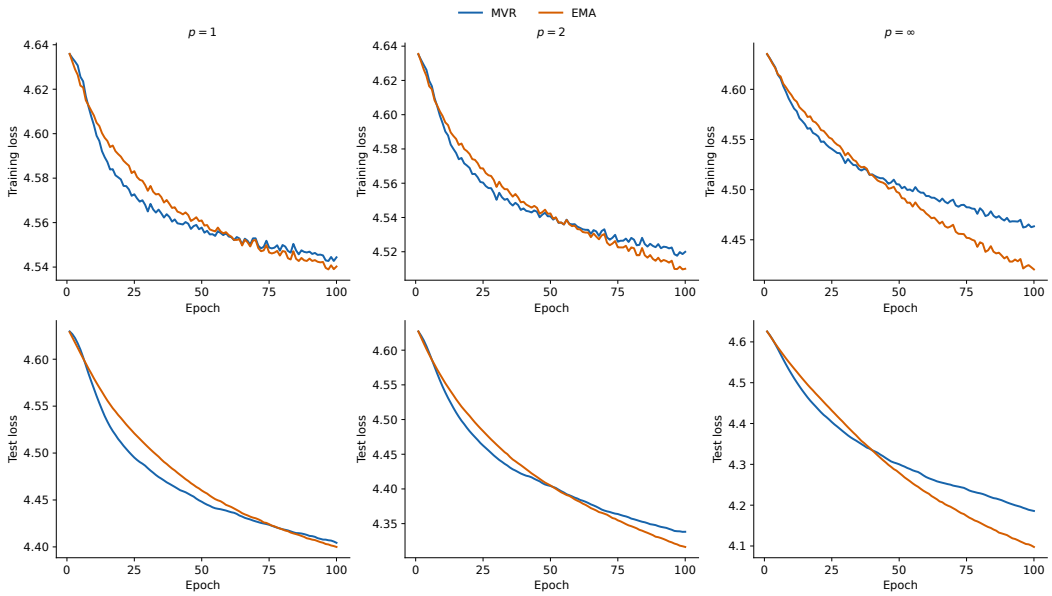


Figure 5: Training and test loss on CIFAR-100 with theorem-scaled parameters. The top and bottom rows show training and test loss, respectively. From left to right, the columns correspond to $p = 1$, $p = 2$, and $p = \infty$. The horizontal axis shows training epochs, and the vertical axis shows the corresponding loss. Blue and orange curves represent MVR and EMA, respectively.

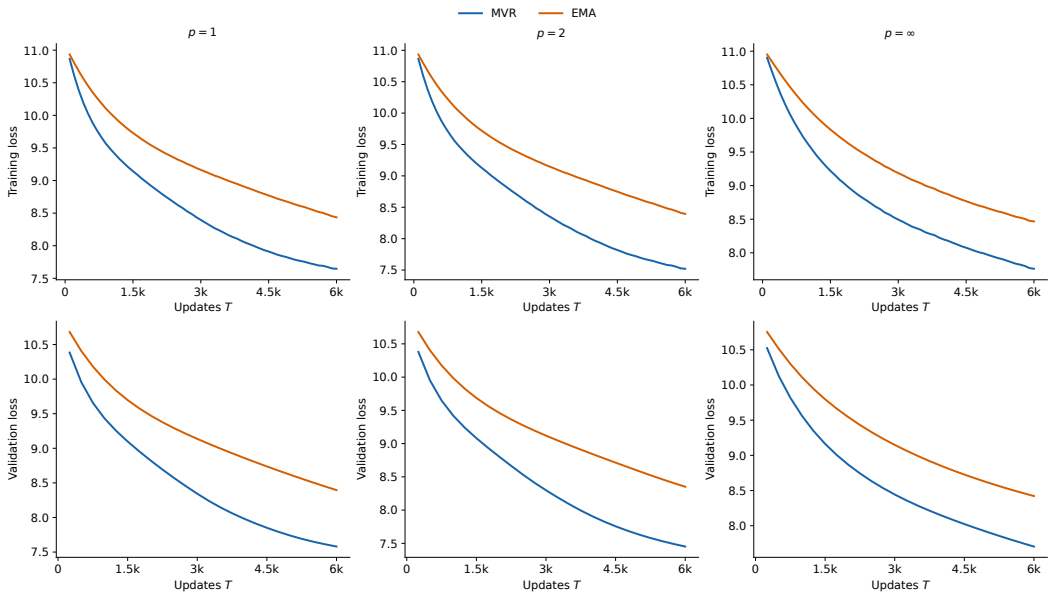


Figure 6: Training and test loss on GPT-2 with theorem-scaled parameters. The top and bottom rows show training and test loss, respectively. From left to right, the columns correspond to $p = 1$, $p = 2$, and $p = \infty$. The horizontal axis shows optimizer updates T, and the vertical axis shows the corresponding loss. Blue and orange curves represent MVR and EMA, respectively.

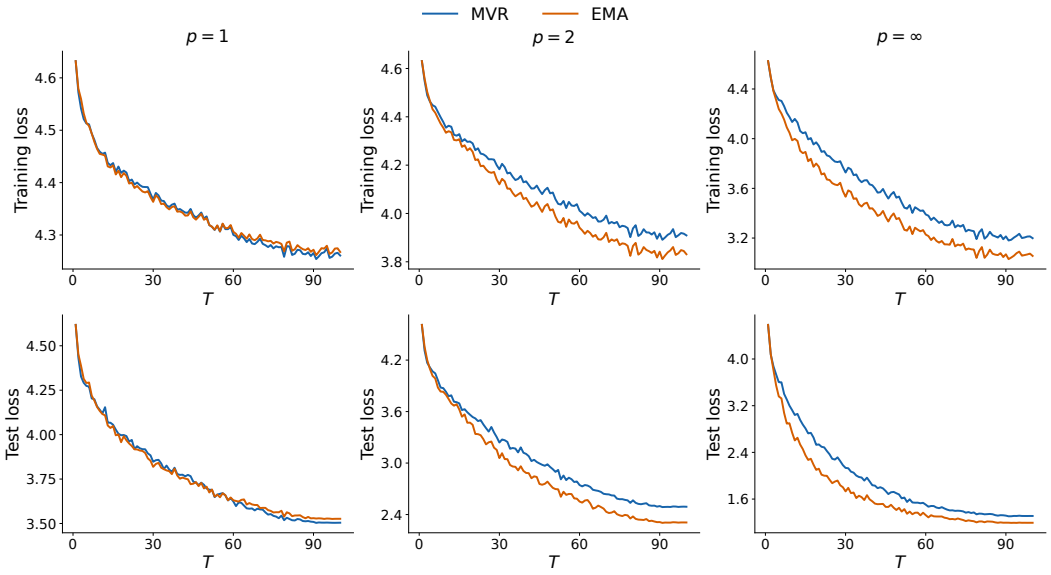


Figure 7: Training and test loss on CIFAR-100 under empirical settings. The top and bottom rows show training and test loss, respectively. From left to right, the columns correspond to $p = 1$, $p = 2$, and $p = \infty$. The horizontal axis shows training epochs, and the vertical axis shows the corresponding loss. Blue and orange curves represent MVR and EMA, respectively.

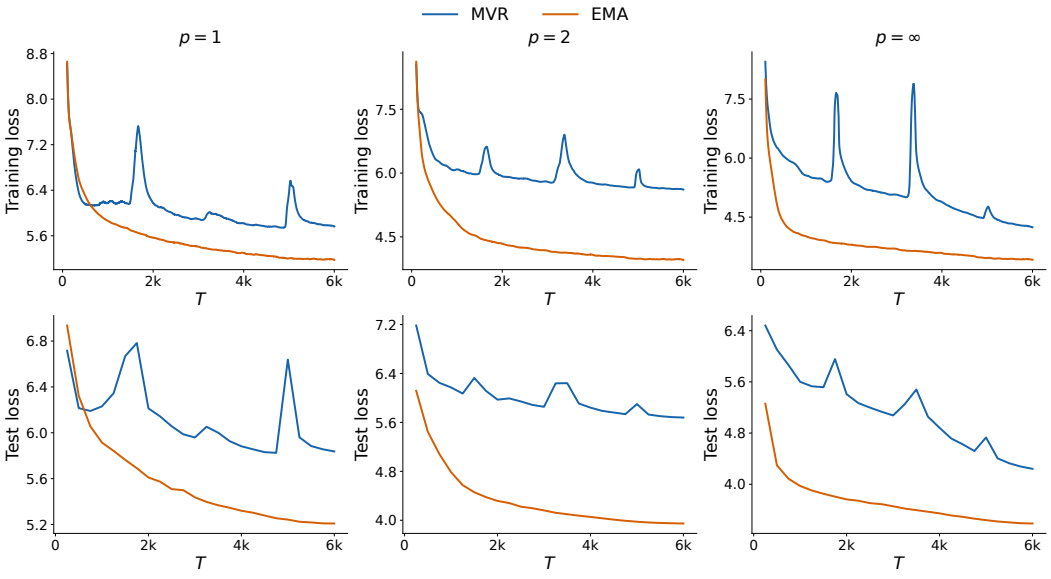


Figure 8: Training and test loss on GPT-2 under empirical settings. The top and bottom rows show training and test loss, respectively. From left to right, the columns correspond to $p = 1$, $p = 2$, and $p = \infty$. The horizontal axis shows optimizer updates T, and the vertical axis shows the corresponding loss. Blue and orange curves represent MVR and EMA, respectively.

E CURVATURE-CORRECTED MOMENTUM: A UNIFIED TEMPLATE AND A COMPARISON OF SEVEN METHODS

E.1 MOTIVATION: WHAT SARAH AND SPIDER LEAVE UNSETTLED

The optimal $\mathcal{O}(\epsilon^{-3})$ oracle complexity for finding an ϵ-stationary point of a smooth non-convex stochastic objective has been available for some years, through SPIDER and SNVRG and ultimately through SARAH's recursive gradient estimator. It is therefore fair to ask what the methods surveyed here are for. The answer is that the rate alone is not the object of interest: what matters is the pair (rate, admissible assumption and oracle class), and on the second coordinate the SARAH–SPIDER line is expensive in four distinct ways.

(i) The rate is bought with a per-sample smoothness condition, and that purchase is provably necessary. SPIDER-type analyses require mean-squared smoothness, a condition on the individual realisations f_ξ rather than on f. Arjevani et al. (2023) show this is not an artifact of the analysis: in the stochastic first-order oracle model under bounded variance alone, $\Omega(\epsilon^{-4})$ is unavoidable, so no algorithm attains ϵ^{-3} without either strengthening the assumption class or enlarging the oracle. Mean-squared smoothness is moreover substantively restrictive in the settings where these methods are deployed, since it fails for networks with non-smooth activations.

(ii) The oracle is non-standard. The recursive correction $\nabla f_\xi(\boldsymbol{x}_{t+1}) - \nabla f_\xi(\boldsymbol{x}_t)$ requires evaluating the same drawn sample at two parameter values. In the terminology of Arjevani et al. (2023) this is a $K = 2$ oracle. Every known algorithm attaining ϵ^{-3} under mean-squared smoothness requires $K = 2$; with $K = 1$ the best known guarantee remains SGD's ϵ^{-4}, and whether ϵ^{-3} is attainable at $K = 1$ is an open problem. A comparison that reports only a rate conceals this difference in what the algorithm is permitted to ask for.

(iii) The loop structure is impractical at scale. SARAH and SPIDER are epoch-based, resetting the estimator to a full or mega-batch gradient at checkpoints. SARAH's headline finite-sum complexity $\mathcal{O}(n + \sqrt{n}\epsilon^{-2})$ is not commensurable with a streaming rate at all.

(iv) The geometry is Euclidean. Contemporary practice increasingly uses non-Euclidean steepest descent — sign-SGD in ℓ_∞, Muon and Scion in the spectral norm — for which a guarantee stated in $\|\cdot\|_2$ is simply the wrong statement.

Each method compared below relaxes one or more of these while retaining ϵ^{-3}. STORM removes the loop and the mega-batches but keeps (i) and (ii). The Hessian-vector-product line — Tran & Cutkosky (2022), Zhang et al. (2020b), Khirirat et al. (2025), Chayti (2026) — removes (i) outright, assuming nothing about f_ξ beyond unbiasedness and a variance bound. It escapes the $\Omega(\epsilon^{-4})$ barrier not through a stronger assumption on f_ξ but by enlarging the oracle to include stochastic Hessian-vector products, which places it outside the model in which that lower bound is proved; the applicable optimality reference for this regime is Arjevani et al. (2020), which shows ϵ^{-3} is unimprovable for such oracles under second-order smoothness. Khirirat et al. (2025) and Qian et al. (2025) remove (iv). Yuan et al. (2025) address a deficiency of a different kind: before it, variance reduction had not produced convincing gains at language-model scale.

The purpose of this appendix is to make these trade-offs comparable. We exhibit all seven methods as instances of a single template, so that they differ in exactly one vector, and then tabulate what each pays for it.

E.2 PROBLEM SETTING

We minimise $f(\boldsymbol{x}) = \mathbb{E}_{\xi\sim\mathcal{D}}\big[f_\xi(\boldsymbol{x})\big]$ over a closed convex set $\mathcal{K} \subseteq \mathbb{R}^d$, with $\mathcal{K} = \mathbb{R}^d$ in the unconstrained case. Write $f^{\inf} = \inf_{\boldsymbol{x}\in\mathcal{K}} f(\boldsymbol{x})$ and $\Delta_0 = f(\boldsymbol{x}_0) - f^{\inf}$.

The space $\mathbb{R}^d$ carries a norm $\|\cdot\|$, not necessarily Euclidean; its dual $\|\boldsymbol{g}\|_\star = \sup_{\|\boldsymbol{x}\|\le 1}\langle \boldsymbol{g}, \boldsymbol{x}\rangle$ is the norm in which gradients are measured. We write $\|\cdot\|_2$ for the Euclidean (equivalently Frobenius) norm and $\|\cdot\|_{\mathrm{op}}$ for the operator norm induced by $\|\cdot\| \to \|\cdot\|_\star$. When $\|\cdot\| = \|\cdot\|_2$ all three

coincide. In the layer-wise setting of Qian et al. (2025), $\boldsymbol{x} = [\boldsymbol{x}_1, \dots, \boldsymbol{x}_p]$ and each block carries its own norm $\|\cdot\|_{(i)}$.

E.3 A TEMPLATE COVERING ALL SEVEN METHODS

Fix a horizon T, an initial point $\boldsymbol{x}_0 \in \mathcal{K}$, and an initial momentum $\boldsymbol{m}_0$ formed from a batch of size $B_{\text{init}} \geq 1$. A step map is a function $\mathcal{D} : \mathbb{R}^d \times \mathcal{K} \to \mathbb{R}^d$ producing a search direction from the current momentum and the current iterate. For $t = 0, 1, \dots, T-1$:

(T1) Direction. $\boldsymbol{d}_t = \mathcal{D}(\boldsymbol{m}_t, \boldsymbol{x}_t)$.

(T2) Step. Draw or set a scalar step scale $s_t > 0$ with $\mathbb{E}[s_t] = \eta_t$, and move:

$$\boldsymbol{x}_{t+1} \;=\; \boldsymbol{x}_t + s_t \boldsymbol{d}_t, \qquad \text{so that} \qquad \boldsymbol{x}_{t+1} - \boldsymbol{x}_t = s_t \boldsymbol{d}_t. \tag{156}$$

(T3) Sample and query. Draw $\xi_{t+1} \sim \mathcal{D}$ and query the oracle. Write $\boldsymbol{g}_{t+1}$ for the resulting stochastic gradient at $\boldsymbol{x}_{t+1}$.

(T4) Correction. Form a vector $\boldsymbol{c}_{t+1}$ from the same query, intended as an estimator of the gradient variation

$$\nabla f(\boldsymbol{x}_{t+1}) - \nabla f(\boldsymbol{x}_t) \;=\; \int_0^1 \nabla^2 f\big(\boldsymbol{x}_t + \tau(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t)\big)\,(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t)\,d\tau. \tag{157}$$

(T5) Momentum. With a mixing weight $\alpha_t \in (0, 1]$,

$$\boldsymbol{m}_{t+1} \;=\; (1 - \alpha_t)\big(\boldsymbol{m}_t + \boldsymbol{c}_{t+1}\big) \;+\; \alpha_t \boldsymbol{g}_{t+1}. \tag{158}$$

Setting $\boldsymbol{c}_{t+1} = \mathbf{0}$ recovers Polyak momentum. The role of $\boldsymbol{c}_{t+1}$ is visible from equation 158: the buffer $\boldsymbol{m}_t$ is an average of gradients evaluated at past iterates, so once the iterate moves it is stale by exactly the amount equation 157, and adding $\boldsymbol{c}_{t+1}$ before mixing transports $\boldsymbol{m}_t$ from $\boldsymbol{x}_t$ to $\boldsymbol{x}_{t+1}$. Define the correction residual and the momentum error

$$\boldsymbol{w}_{t+1} \;\coloneqq\; \boldsymbol{c}_{t+1} - \big(\nabla f(\boldsymbol{x}_{t+1}) - \nabla f(\boldsymbol{x}_t)\big), \qquad \boldsymbol{e}_t \;\coloneqq\; \boldsymbol{m}_t - \nabla f(\boldsymbol{x}_t). \tag{159}$$

Three step maps suffice for all seven methods:

$$\mathcal{D}_{\text{sgd}}(\boldsymbol{m}, \boldsymbol{x}) = -\boldsymbol{m}, \qquad \text{plain stochastic gradient step;} \tag{160}$$

$$\mathcal{D}_{\text{lmo}}(\boldsymbol{m}, \boldsymbol{x}) = \operatorname*{arg\,min}_{\|\boldsymbol{v}\| \leq 1} \langle \boldsymbol{m}, \boldsymbol{v} \rangle, \qquad \text{linear minimization oracle on the unit ball;} \tag{161}$$

$$\mathcal{D}_{\text{fw}}(\boldsymbol{m}, \boldsymbol{x}) = \operatorname*{arg\,min}_{\boldsymbol{v} \in \mathcal{K}} \langle \boldsymbol{m}, \boldsymbol{v} \rangle - \boldsymbol{x}, \qquad \text{Frank–Wolfe direction.} \tag{162}$$

Here $\mathcal{D}_{\text{lmo}}$ gives normalized SGD for $\|\cdot\|_2$, sign-SGD for $\|\cdot\|_\infty$ and Muon-type updates for the spectral norm, while $\mathcal{D}_{\text{fw}}$ keeps iterates in $\mathcal{K}$ whenever $s_t \in [0, 1]$.

Two observations make the reduction in §E.4 possible. First, $\mathcal{D}_{\text{fw}}$ genuinely depends on $\boldsymbol{x}$, which is why the step map is not a function of $\boldsymbol{m}$ alone. Second, the linear minimization oracle over a ball of radius η is positively homogeneous in that radius,

$$\operatorname*{arg\,min}_{\|\boldsymbol{v}\| \leq \eta} \langle \boldsymbol{m}, \boldsymbol{v} \rangle \;=\; \eta \cdot \operatorname*{arg\,min}_{\|\boldsymbol{v}\| \leq 1} \langle \boldsymbol{m}, \boldsymbol{v} \rangle \;=\; \eta\, \mathcal{D}_{\text{lmo}}(\boldsymbol{m}, \boldsymbol{x}), \tag{163}$$

which is what allows a method that folds its stepsize into the LMO radius to be written in the form equation 156, with a unit-ball direction and a separate scale $s_t = \eta$.

E.4 REDUCTION OF EACH PUBLISHED METHOD TO THE TEMPLATE

We now verify that each of the seven methods is an instance of **(T1)**–**(T5)**. In each case we state the published update in the source's own notation and then identify $\mathcal{D}$, s_t, $\boldsymbol{c}_{t+1}$ and α_t. Indexing differs between sources by a shift; everything is aligned here to the template's convention, in which $\boldsymbol{c}_{t+1}$ and $\boldsymbol{g}_{t+1}$ are formed after the move to $\boldsymbol{x}_{t+1}$.

STORM (Cutkosky & Orabona, 2019). The published recursion and update are $\boldsymbol{d}_{t+1} = \nabla f_{\xi_{t+1}}(\boldsymbol{x}_{t+1}) + (1-a_{t+1})\big(\boldsymbol{d}_t - \nabla f_{\xi_{t+1}}(\boldsymbol{x}_t)\big)$ and $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_t \boldsymbol{d}_t$. Identifying $\boldsymbol{m}_t$ with $\boldsymbol{d}_t$ and regrouping,

$$\begin{aligned}\boldsymbol{d}_{t+1} &= (1-a_{t+1})\boldsymbol{d}_t - (1-a_{t+1})\nabla f_{\xi_{t+1}}(\boldsymbol{x}_t) + \nabla f_{\xi_{t+1}}(\boldsymbol{x}_{t+1}) \\ &= (1-a_{t+1})\Big(\boldsymbol{d}_t + \big[\nabla f_{\xi_{t+1}}(\boldsymbol{x}_{t+1}) - \nabla f_{\xi_{t+1}}(\boldsymbol{x}_t)\big]\Big) + a_{t+1}\,\nabla f_{\xi_{t+1}}(\boldsymbol{x}_{t+1}),\end{aligned} \tag{164}$$

which is exactly equation 158 with $\boldsymbol{c}_{t+1} = \nabla f_{\xi_{t+1}}(\boldsymbol{x}_{t+1}) - \nabla f_{\xi_{t+1}}(\boldsymbol{x}_t)$, $\boldsymbol{g}_{t+1} = \nabla f_{\xi_{t+1}}(\boldsymbol{x}_{t+1})$ and $\alpha_t = a_{t+1}$. The update gives $\mathcal{D} = \mathcal{D}_{\mathrm{sgd}}$ and $s_t = \eta_t$.

SOM (Tran & Cutkosky, 2022). Algorithm 1 uses $\hat{\boldsymbol{g}}_t = (1-\alpha_{t-1})\big(\hat{\boldsymbol{g}}_{t-1}^{\mathrm{clip}} + \nabla^2 f_{z_t}(\boldsymbol{x}_t)(\boldsymbol{x}_t - \boldsymbol{x}_{t-1})\big) + \alpha_{t-1}\nabla f_{z_t}(\boldsymbol{x}_t)$ with $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_t \hat{\boldsymbol{g}}_t^{\mathrm{clip}}$. After the index shift, this is a clipped version of equation 158, with $\boldsymbol{m}_t$ the clipped buffer, $\boldsymbol{c}_{t+1} = \nabla^2 f_{\xi_{t+1}}(\boldsymbol{x}_{t+1})(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t)$, and clipping applied to the new buffer before it is stored. The step map is $\mathcal{D}(\boldsymbol{m}, \boldsymbol{x}) = -\boldsymbol{m}$ with $s_t = \eta_t$, where $\mathrm{clip}_G(\boldsymbol{m}) = \boldsymbol{m}\min\{1, G/\|\boldsymbol{m}\|_2\}$ for $\boldsymbol{m} \neq \boldsymbol{0}$ and $\mathrm{clip}_G(\boldsymbol{0}) = \boldsymbol{0}$. Algorithm 2 uses the unclipped momentum recursion and the normalized update $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta\,\boldsymbol{m}_t/\|\boldsymbol{m}_t\|_2$, that is $\mathcal{D} = \mathcal{D}_{\mathrm{lmo}}$ in $\|\cdot\|_2$ with $s_t = \eta$; Algorithm 3 retains clipping and makes η_t adaptive.

1-SFW (Zhang et al., 2020b). The published estimator $\boldsymbol{d}_t = (1-\rho_t)(\boldsymbol{d}_{t-1} + \tilde{\Delta}_t) + \rho_t \nabla f_{z_t}(\boldsymbol{x}_t)$ is already in the form equation 158, with $\alpha_t = \rho_t$ and $\boldsymbol{c}_{t+1} = \tilde{\Delta}_{t+1} = \nabla^2 f_{\xi_{t+1}}(\hat{\boldsymbol{x}}_{t+1})(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t)$, where $\hat{\boldsymbol{x}}_{t+1} = a\boldsymbol{x}_{t+1} + (1-a)\boldsymbol{x}_t$ with $a \sim \mathcal{U}[0,1]$. The update $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t + \eta_t(\boldsymbol{v}_t - \boldsymbol{x}_t)$ with $\boldsymbol{v}_t = \arg\min_{\boldsymbol{v}\in\mathcal{K}}\langle \boldsymbol{d}_t, \boldsymbol{v}\rangle$ is $\mathcal{D} = \mathcal{D}_{\mathrm{fw}}$ with $s_t = \eta_t \in [0,1]$.

MARS (Yuan et al., 2025). Write $\Delta_t = \nabla f_{\xi_t}(\boldsymbol{x}_t) - \nabla f_{\xi_t}(\boldsymbol{x}_{t-1})$. The published estimator is $\boldsymbol{c}_t^{\mathrm{MARS}} = \nabla f_{\xi_t}(\boldsymbol{x}_t) + \gamma_t \frac{\beta_1}{1-\beta_1}\Delta_t$, clipped by norm to $\tilde{\boldsymbol{c}}_t^{\mathrm{MARS}}$, followed by $\boldsymbol{m}_t = \beta_1 \boldsymbol{m}_{t-1} + (1-\beta_1)\tilde{\boldsymbol{c}}_t^{\mathrm{MARS}}$. Setting the clip aside and expanding,

$$\boldsymbol{m}_t = \beta_1\boldsymbol{m}_{t-1} + (1-\beta_1)\nabla f_{\xi_t}(\boldsymbol{x}_t) + \beta_1\gamma_t\Delta_t = (1-\alpha)\big(\boldsymbol{m}_{t-1} + \gamma_t\Delta_t\big) + \alpha\,\nabla f_{\xi_t}(\boldsymbol{x}_t), \qquad \alpha = 1-\beta_1, \tag{165}$$

so the template's correction is $\boldsymbol{c}_{t+1} = \gamma_{t+1}\big(\nabla f_{\xi_{t+1}}(\boldsymbol{x}_{t+1}) - \nabla f_{\xi_{t+1}}(\boldsymbol{x}_t)\big)$, which is STORM's at $\gamma \equiv 1$. The update $\boldsymbol{x}_{t+1} = \arg\min_{\boldsymbol{x}}\{\eta_t\langle \boldsymbol{m}_t, \boldsymbol{x}\rangle + \frac{1}{2}\|\boldsymbol{x} - \boldsymbol{x}_t\|^2_{\boldsymbol{H}_t}\}$ has the closed form $\boldsymbol{x}_{t+1} = \boldsymbol{x}_t - \eta_t \boldsymbol{H}_t^{-1}\boldsymbol{m}_t$, so $\mathcal{D}(\boldsymbol{m}, \boldsymbol{x}) = -\boldsymbol{H}_t^{-1}\boldsymbol{m}$ with $s_t = \eta_t$.

HCM (Khirirat et al., 2025). The published update is $\boldsymbol{x}_{k+1} = \boldsymbol{x}_k + \mathrm{lmo}(\boldsymbol{m}_k)$ with $\mathrm{lmo}(\boldsymbol{m}) = \arg\min_{\|\boldsymbol{x}\|\le\eta_k}\langle \boldsymbol{m}, \boldsymbol{x}\rangle$, so the stepsize appears as the LMO radius rather than as a separate multiplier. By the homogeneity identity equation 163 this is $\mathcal{D} = \mathcal{D}_{\mathrm{lmo}}$ with $s_t = \eta_k$. The momentum recursion $\boldsymbol{m}_{k+1} = (1-\alpha_k)\big(\boldsymbol{m}_k + \frac{\beta_k}{1-\alpha_k}\nabla^2 f_{\xi_{k+1}}(\hat{\boldsymbol{x}}_{k+1})(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)\big) + \alpha_k \nabla f_{\xi_{k+1}}(\boldsymbol{x}_{k+1})$ is equation 158 with $\boldsymbol{c}_{k+1} = \frac{\beta_k}{1-\alpha_k}\nabla^2 f_{\xi_{k+1}}(\hat{\boldsymbol{x}}_{k+1})(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)$, reducing to the bare Hessian-vector product at $\beta_k = 1-\alpha_k$, the only case the theory covers. Variant 1 takes $\hat{\boldsymbol{x}}_{k+1} = b\boldsymbol{x}_{k+1} + (1-b)\boldsymbol{x}_k$ with $b \sim \mathcal{U}(0,1)$; Variant 2 takes $\hat{\boldsymbol{x}}_{k+1} = \boldsymbol{x}_{k+1}$. Theorem 4 (Appendix F) instead uses the STORM correction, i.e. $\boldsymbol{c}_{k+1} = \nabla f_{\xi_{k+1}}(\boldsymbol{x}_{k+1}) - \nabla f_{\xi_{k+1}}(\boldsymbol{x}_k)$, with the same LMO step.

RanSOM (Chayti, 2026). RanSOM-E uses $\mathcal{D} = \mathcal{D}_{\mathrm{lmo}}$ rescaled to radius ρ and RanSOM-B uses $\mathcal{D} = \mathcal{D}_{\mathrm{fw}}$. What is new is that s_t is random: $s_t \sim \mathrm{Exp}(1/\eta_t)$ for the former and $s_t \sim \mathrm{Beta}(1, K_t)$ with $K_t = \eta_t^{-1} - 1$ for the latter, in both cases with $\mathbb{E}[s_t] = \eta_t$ as **(T2)** requires. The correction is $\boldsymbol{c}_{t+1} = \eta_t \nabla^2 f_{\xi_{t+1}}(\boldsymbol{x}_{t+1})\boldsymbol{d}_t$ for RanSOM-E and $\boldsymbol{c}_{t+1} = \frac{1-s_t}{K_t}\nabla^2 f_{\xi_{t+1}}(\boldsymbol{x}_{t+1})\boldsymbol{d}_t$ for RanSOM-B. Since $\boldsymbol{x}_{t+1} - \boldsymbol{x}_t = s_t\boldsymbol{d}_t$ by equation 156, the former equals $(\eta_t/s_t)\,\nabla^2 f_{\xi_{t+1}}(\boldsymbol{x}_{t+1})(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t)$: the classic endpoint Hessian-vector product, reweighted by the ratio of the mean step to the realised step.

Gluon-MVR (Qian et al., 2025). Three variants, all layer-wise: for each block i the update is an LMO over $\{\boldsymbol{x}_i : \|\boldsymbol{x}_i - \boldsymbol{x}_i^k\|_{(i)} \le t_i\eta\}$, i.e. $\mathcal{D} = \mathcal{D}_{\mathrm{lmo}}$ in $\|\cdot\|_{(i)}$ with $s_t = t_i\eta$ by equation 163. **Gluon-MVR-1** sets $\boldsymbol{m}_i^k = \nabla_i f_{\xi^k}(\boldsymbol{x}^k) + \beta(\boldsymbol{m}_i^{k-1} - \nabla_i f_{\xi^k}(\boldsymbol{x}^{k-1}))$, which regroups exactly as in equation 164 into equation 158 with $\alpha = 1-\beta$ and $\boldsymbol{c}_i^k = \nabla_i f_{\xi^k}(\boldsymbol{x}^k) - \nabla_i f_{\xi^k}(\boldsymbol{x}^{k-1})$: it is STORM's correction verbatim. **Gluon-MVR-2** introduces an inner buffer $\boldsymbol{g}_i^k = \nabla_i f_{\xi^k}(\boldsymbol{x}^k) + (1-$

$q)(\boldsymbol{g}_i^{k-1} - \nabla_i f_{\xi^k}(\boldsymbol{x}^{k-1}))$, which is itself equation 158 with mixing weight q, and then applies a second smoothing $\boldsymbol{m}_i^k = \beta \boldsymbol{m}_i^{k-1} + (1-\beta)\boldsymbol{g}_i^k$; it is therefore not a single instance of the template but a composition of one with an outer exponential average. **Gluon-MVR-3** adds the correction at the outer level as well, $\boldsymbol{m}_i^k = \beta \boldsymbol{m}_i^{k-1} + (1-\beta)\boldsymbol{g}_i^k + \beta(\nabla_i f_{\xi^k}(\boldsymbol{x}^k) - \nabla_i f_{\xi^k}(\boldsymbol{x}^{k-1}))$. The authors note that Gluon-MVR-1 with the spectral norm coincides with LiMuon (option 1) and with MVR2 at $\gamma = 1$ of Chang et al.

Table 6 collects these identifications.

Method	Step map $\mathcal{D}(\boldsymbol{m}, \boldsymbol{x})$ and scale s_t	Correction $\boldsymbol{c}_{t+1}$	α_t
STORM Cutkosky & Orabona (2019)	$\mathcal{D}_{\mathrm{sgd}} = -\boldsymbol{m}$; $s_t = \eta_t \propto \left(w + \sum_{i \le t} G_i^2\right)^{-1/3}$, $G_i = \|\nabla f_{\xi_i}(\boldsymbol{x}_i)\|_2$	$\nabla f_{\xi_{t+1}}(\boldsymbol{x}_{t+1}) - \nabla f_{\xi_{t+1}}(\boldsymbol{x}_t)$ same sample at both endpoints	$c\eta_t^2$
SOM Tran & Cutkosky (2022)	$-\mathrm{clip}_G(\boldsymbol{m})$ (Alg. 1, 3) or $\mathcal{D}_{\mathrm{lmo}}$ in $\|\cdot\|_2$ (Alg. 2); $s_t = \eta_t \propto t^{-1/3}$	$\nabla^2 f_{\xi_{t+1}}(\boldsymbol{x}_{t+1})(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t)$ HVP at the current iterate	$2K\eta_t\eta_{t+1}$
1-SFW Zhang et al. (2020b)	$\mathcal{D}_{\mathrm{fw}}$; $s_t = \eta_t$ deterministic	$\nabla^2 f_{\xi_{t+1}}(\hat{\boldsymbol{x}}_{t+1})(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t)$ with $\hat{\boldsymbol{x}}_{t+1} = a\boldsymbol{x}_{t+1} + (1-a)\boldsymbol{x}_t$, $a \sim \mathcal{U}[0,1]$ HVP at a uniform midpoint	$(t-1)^{-2/3}$
MARS Yuan et al. (2025)	$-\boldsymbol{H}_t^{-1}\boldsymbol{m}$; $s_t = \eta_t = (s+t)^{-1/3}$	$\gamma_{t+1}\big(\nabla f_{\xi_{t+1}}(\boldsymbol{x}_{t+1}) - \nabla f_{\xi_{t+1}}(\boldsymbol{x}_t)\big)$ $\gamma \equiv 1$ is STORM; $\gamma = 0.025$ in practice	$1-\beta_1$
HCM Khirirat et al. (2025)	$\mathcal{D}_{\mathrm{lmo}}$ in an arbitrary $\|\cdot\|$; $s_t = \eta_k = \hat{\eta}(T+1)^{-2/3}$, via equation 163	Thms. 1–2: $\frac{\beta_k}{1-\alpha_k}\nabla^2 f_{\xi_{k+1}}(\hat{\boldsymbol{x}}_{k+1})(\boldsymbol{x}_{k+1} - \boldsymbol{x}_k)$; V1: $\hat{\boldsymbol{x}}_{k+1} = b\boldsymbol{x}_{k+1} + (1-b)\boldsymbol{x}_k$; V2: $\hat{\boldsymbol{x}}_{k+1} = \boldsymbol{x}_{k+1}$. Thm. 4: the STORM correction	$(T+1)^{-2/3}$
RanSOM Chayti (2026)	$\mathcal{D}_{\mathrm{lmo}}$ (-E) or $\mathcal{D}_{\mathrm{fw}}$ (-B); $s_t \sim \mathrm{Exp}(1/\eta_t)$ (-E), $s_t \sim \mathrm{Beta}(1, K_t)$ (-B)	(-E): $\eta_t \nabla^2 f_{\xi_{t+1}}(\boldsymbol{x}_{t+1})\boldsymbol{d}_t$; (-B): $\frac{1-s_t}{K_t}\nabla^2 f_{\xi_{t+1}}(\boldsymbol{x}_{t+1})\boldsymbol{d}_t$ HVP at the next iterate	β
Gluon-MVR Qian et al. (2025)	$\mathcal{D}_{\mathrm{lmo}}$ in the layer norm $\|\cdot\|_{(i)}$; $s_t = t_i\eta$ (Thm. 2, 4, 5) or $t_i(k+1)^{-2/3}$ (Thm. 3)	MVR-1: $\nabla_i f_{\xi^k}(\boldsymbol{x}^k) - \nabla_i f_{\xi^k}(\boldsymbol{x}^{k-1})$, i.e. STORM's, per layer. MVR-2/3 apply the same correction to an inner buffer with weight q, then smooth again	$1-\beta$

Table 6: Instantiation of the template **(T1)–(T5)**, as derived in §E.4. All seven methods estimate the same object equation 157; they differ only in $\boldsymbol{c}_{t+1}$ and in the step map.

E.5 Oracle models and assumptions

The labels below are used in the final column of Table 7, which maps every numbered assumption of every source onto exactly one of them. The organising distinction is whether a condition constrains the averaged objective f or every realisation f_ξ.

Lower bound and noise.

(LB) $f^{\inf} > -\infty$, so that $\Delta_0 < \infty$. Required by every result in this appendix without exception.

(V) Unbiasedness and bounded variance, $\mathbb{E}_\xi[\nabla f_\xi(\boldsymbol{x})] = \nabla f(\boldsymbol{x})$ and $\mathbb{E}_\xi\|\nabla f_\xi(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|_2^2 \le \sigma_g^2$.

(V_p) Heavy-tailed noise, $\mathbb{E}_\xi\|\nabla f_\xi(\boldsymbol{x}) - \nabla f(\boldsymbol{x})\|_\star^p \le \sigma^p$ with $p \in (1,2]$; at $p = 2$ this is (V), for $p < 2$ the variance is infinite.

(V_H) Bounded Hessian variance, $\mathbb{E}_\xi\|(\nabla^2 f_\xi(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{x}))\boldsymbol{w}\|_2^2 \le \sigma_H^2\|\boldsymbol{w}\|_2^2$.

(V_Δ) Bounded gradient-difference variance,

$$\mathbb{E}_\xi\big\|\big(\nabla f_\xi(\boldsymbol{x}) - \nabla f_\xi(\boldsymbol{v})\big) - \big(\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{v})\big)\big\|_2^2 \;\le\; \delta^2\|\boldsymbol{x} - \boldsymbol{v}\|^2. \tag{166}$$

This is precisely a bound on $\mathbb{E}\|\boldsymbol{w}_{t+1}\|_2^2$ for the same-sample correction of equation 159.

Smoothness.

(S_f) Smoothness of f only: $\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{v})\|_\star \le L\|\boldsymbol{x} - \boldsymbol{v}\|$.

(MSS) Mean-squared smoothness, $\mathbb{E}_\xi\|\nabla f_\xi(\boldsymbol{x}) - \nabla f_\xi(\boldsymbol{v})\|^2 \le \bar{L}^2\|\boldsymbol{x} - \boldsymbol{v}\|^2$ (Arjevani et al., 2023).

(S_ξ) Per-sample smoothness: $\|\nabla f_\xi(\boldsymbol{x}) - \nabla f_\xi(\boldsymbol{v})\| \le L\|\boldsymbol{x} - \boldsymbol{v}\|$ for every ξ, almost surely.

(H) Lipschitz Hessian of f: $\|\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{v})\|_{\mathrm{op}} \le \rho\|\boldsymbol{x} - \boldsymbol{v}\|$.

(H_ξ) Lipschitz Hessian of every f_ξ, almost surely.

Generalized (relaxed) smoothness. Each of the following is stated in the symmetric Lipschitz form used by Khirirat et al. (2025), and reduces to its classical counterpart when the second constant vanishes. In each case the modulus involves the gradient norm, taken as a supremum along the segment joining the two points.

(L_0, L_1) Symmetric generalized gradient smoothness: for all $\boldsymbol{x}, \boldsymbol{v} \in \mathbb{R}^d$,

$$\|\nabla f(\boldsymbol{x}) - \nabla f(\boldsymbol{v})\|_\star \;\le\; \Big(L_0 + L_1 \sup_{\theta\in[0,1]} \|\nabla f(\theta\boldsymbol{x} + (1-\theta)\boldsymbol{v})\|_\star\Big)\|\boldsymbol{x} - \boldsymbol{v}\|. \tag{167}$$

This is the relaxation of (S_f), recovered at $L_1 = 0$ (Zhang et al., 2020a). It implies the Hessian bound $\|\nabla^2 f(\boldsymbol{x})\|_{\mathrm{op}} \le L_0 + L_1\|\nabla f(\boldsymbol{x})\|_\star$, but that bound is a consequence (their Lemma 1, from Chen et al., 2023) and not the assumption. In Qian et al. (2025) the condition is imposed blockwise, with constants (L_i^0, L_i^1) per layer.

(M_0, M_1) Symmetric generalized Hessian smoothness:

$$\|\nabla^2 f(\boldsymbol{x}) - \nabla^2 f(\boldsymbol{v})\|_\star \;\le\; \Big(M_0 + M_1 \sup_{\theta\in[0,1]} \|\nabla f(\theta\boldsymbol{x} + (1-\theta)\boldsymbol{v})\|_\star\Big)\|\boldsymbol{x} - \boldsymbol{v}\|. \tag{168}$$

The relaxation of (H), recovered at $M_1 = 0$. Note that the modulus is the gradient norm, not the Hessian norm.

($\mathcal{L}_0, \mathcal{L}_1$) Symmetric generalized mean-squared smoothness:

$$\mathbb{E}_\xi\Big[\|\nabla f_\xi(\boldsymbol{x}) - \nabla f_\xi(\boldsymbol{v})\|_\star^2\Big] \;\le\; \mathbb{E}_\xi\Big[\big(\mathcal{L}_0 + \mathcal{L}_1 \sup_{\theta\in[0,1]} \|\nabla f_\xi(\theta\boldsymbol{x} + (1-\theta)\boldsymbol{v})\|_\star\big)^2\Big]\|\boldsymbol{x} - \boldsymbol{v}\|^2. \tag{169}$$

The relaxation of (MSS), recovered at $\mathcal{L}_1 = 0$. Two structural points distinguish equation 169 from equation 167: the modulus involves the stochastic gradient ∇f_ξ rather than ∇f, and the entire squared bracket sits inside the expectation. Khirirat et al. (2025) attribute the condition to Chen et al. (2023) and record that it is strictly stronger than equation 167; consequently their Assumption 5 subsumes their Assumption 3, and listing both for their Theorem 4 is redundant.

Boundedness and geometry.

(G) Bounded stochastic gradient, $\|\nabla f_\xi(\boldsymbol{x})\| \le G$ almost surely.

(B_f) Bounded stochastic function value, $|f_\xi(\boldsymbol{x})| \le B$ almost surely.

(B_H) Bounded stochastic Hessian, $\|\nabla^2 f_\xi(\boldsymbol{x})\|_{\mathrm{op}} \le L_f$ almost surely.

(D) Bounded domain or bounded iterates: either $\mathcal{K}$ compact of diameter D, or $\|\boldsymbol{x}_t\| \le D$ along the trajectory.

(NE) Norm equivalence between the dual norm and the Euclidean norm, e.g. $\|\boldsymbol{x}\|_\star \le \rho\|\boldsymbol{x}\|_2$. Needed whenever the criterion is measured in $\|\cdot\|_\star$ but the variance in (V) is measured in $\|\cdot\|_2$.

(P) Preconditioner lower bound, $\boldsymbol{H}_t \succ \rho\boldsymbol{I}$.

Ordering. These conditions are not independent. Directly from the definitions,

$$(B_H) \implies (S_\xi) \implies (\mathrm{MSS}) \implies (V_\Delta), \qquad (\mathrm{MSS}) \implies (S_f), \tag{170}$$

the third implication because $\mathbb{E}\|\zeta - \mathbb{E}\zeta\|_2^2 \le \mathbb{E}\|\zeta\|_2^2$ applied to $\zeta = \nabla f_\xi(\boldsymbol{x}) - \nabla f_\xi(\boldsymbol{v})$. Thus ($\mathrm{V}_\Delta$) is the weakest condition in the chain and (B_H) the strongest. The generalized conditions sit above their classical counterparts in the same order: equation 169 implies equation 167, and setting the second constant to zero in equation 167, equation 168 and equation 169 returns (S_f), (H) and (MSS) respectively. Note also that (H) constrains f alone, so it is not comparable to (MSS): neither implies the other.

Method	Oracle query per iteration; batching	Passes	Source assumption → local label
STORM	one sample at **two points** $\boldsymbol{x}_t, \boldsymbol{x}_{t+1}$; no batches	$2(F_c{+}B_c)$	§3: $F^\star > -\infty$ → (LB); unbiased, variance σ^2 → (V); $f(\cdot,\xi)$ L-smooth a.s. → (S_ξ); $f(\cdot,\xi)$ G-Lipschitz → (G)
SOM	one sample at **one point**: gradient and HVP, both at $\boldsymbol{x}_{t+1}$; no batches	$\approx$ $F_c{+}2B_c$	eq. (1) → (LB); eq. (2) → (S_f); eq. (3) → (V); eq. (4) → (V_H); eq. (5) → (H); eq. (6) → (G), for Alg. 1 and 3 only
1-SFW	one sample; HVP at the auxiliary midpoint $\hat{\boldsymbol{x}}_{t+1}$; batch size 1	$3(F_c{+}B_c)$	A1 → (D); A2 → (B_f), which with A1 gives (LB); A3 → (G); A4 → (B_H), which by equation 170 gives (S_ξ) and hence (S_f); A5 → (H_ξ), for the finite-difference variant only. Unbiasedness and variance follow from A3–A4
MARS	one sample at **two points** $\boldsymbol{x}_t, \boldsymbol{x}_{t+1}$; APPROX variant reuses ξ_t; no batches	$2(F_c{+}B_c)$	$\min_{\boldsymbol{x}} f(\boldsymbol{x})$ finite in Thm. B.5 → (LB); A B.1 → (V); A B.2 → (S_ξ); A B.3 → (P); $\|\boldsymbol{x}_t\|_2 \le D$ in Thm. B.6 → (D)
HCM	one sample; gradient at $\boldsymbol{x}_{t+1}$, plus one HVP at a midpoint (V1) or at $\boldsymbol{x}_{t+1}$ (V2); Thm. 4 uses two points, no HVP; no batches	3 / 2 / 2 $(F_c{+}B_c)$	A2 → (LB); A1 → (V) and (V_H); eq. (2) → (NE); A3 → (L_0, L_1); A4 → (M_0, M_1), Thm. 2 only; A5 → ($\mathcal{L}_0, \mathcal{L}_1$), Thm. 4 only, and subsumes A3
RanSOM	one sample; gradient and HVP computed jointly at $\boldsymbol{x}_{t+1}$; large initial batch B_{init}	$2(F_c{+}B_c)$	A4.1 → (LB); A4.2 → (L_0, L_1); A4.4 → (V_p); norm-equivalence constant → (NE)
Gluon-MVR	one sample at **two points** $\boldsymbol{x}^k, \boldsymbol{x}^{k-1}$; no batches	$2(F_c{+}B_c)$	§1 → (LB); A5 → (L_0, L_1), layer-wise; A6 → (V); A7 → (NE); A8 → (V_Δ)

Table 7: Oracle model and assumptions. F_c and B_c denote the cost of one forward and one backward pass. "Two points" means the same drawn sample is evaluated at two parameter values — the oracle SPIDER/SARAH-type estimators require, and which a single-query stochastic first-order oracle forbids. The final column lists every assumption the cited results actually invoke; nothing is omitted and nothing not invoked is listed. Assumption numbering is that of the source ("A" = Assumption). Qian et al. (2025) also prove a star-convex result (their Thm. 1) which additionally assumes star-convexity and replaces A8 by (MSS); the row above covers their non-convex Thms. 2–5.

E.6 What each theorem actually bounds

The headline "$\mathcal{O}(\epsilon^{-3})$" conceals real differences in the criterion. Before reading Table 8, note the conventions for the reported point $\boldsymbol{x}_{\text{out}}$ and the criterion:

- **Uniformly random iterate.** $\boldsymbol{x}_{\text{out}}$ is drawn uniformly from $\{\boldsymbol{x}_1, \dots, \boldsymbol{x}_T\}$, so that $\mathbb{E}\|\nabla f(\boldsymbol{x}_{\text{out}})\| = \frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{x}_t)\|$. Used by STORM, SOM and 1-SFW.
- **Best iterate.** $\min_{t\le T}\mathbb{E}\|\nabla f(\boldsymbol{x}_t)\|_\star$, a weaker statement than the average and not achievable without evaluating all iterates. Used by HCM and Gluon-MVR, the latter with the criterion aggregated across layers as $\sum_i t_i \mathbb{E}\|\nabla_i f(\boldsymbol{x}^k)\|_{(i)\star}$.

- **Frank–Wolfe gap.** $\mathcal{G}(\boldsymbol{x}) = \max_{\boldsymbol{v}\in\mathcal{K}}\langle \boldsymbol{v}-\boldsymbol{x}, -\nabla f(\boldsymbol{x})\rangle$, the correct stationarity measure on a constrained domain, where the gradient norm need not vanish at a solution. Used by 1-SFW and RanSOM-B.
- **Momentum error and step size.** MARS bounds neither a gradient norm nor a gap, but the averaged squared momentum error together with a normalized-step quantity; see the entry below.

Squared versus unsquared norms also differ across rows and are stated explicitly.

E.7 Which results reach ϵ^{-3} without a Hessian-vector-product oracle or Hessian smoothness

Tables 7 and 8 together isolate a distinguished subset. The following results attain $\mathcal{O}(\epsilon^{-3})$ while querying only stochastic gradients — no $\nabla^2 f_\xi$ in the algorithm — and while assuming neither (H) nor (H_ξ) nor (M_0, M_1) in the analysis:

1. **STORM**, Theorem 1 of Cutkosky & Orabona (2019). Euclidean, plain SGD step. Assumes (S_ξ) and additionally (G).
2. **MARS**, Theorems B.5 and B.6 of Yuan et al. (2025). Euclidean with a preconditioner. Assumes (S_ξ), plus (P) and, for B.6, (D).
3. **HCM**, Theorem 4 (Appendix F) of Khirirat et al. (2025). Arbitrary norm. Assumes $(\mathcal{L}_0, \mathcal{L}_1)$ as displayed in equation 169, i.e. generalized (MSS), plus (NE).
4. **Gluon-MVR-1**, Theorems 2 and 3 of Qian et al. (2025); **Gluon-MVR-2**, Theorem 4 (for $q < \alpha$); **Gluon-MVR-3**, Theorem 5. Layer-wise arbitrary norms. Assume only (V_Δ), plus (NE).

Four remarks follow: on why this subset is exactly the same-sample methods, on their relative assumption strength, on the Schatten-p instantiation, and on which of them generalizes STORM.

These are exactly the same-sample methods, and that is not a coincidence. If a Hessian oracle is refused, the only available estimator of equation 157 is a difference of gradients. Evaluating that difference at the same ξ makes $\boldsymbol{w}_{t+1}$ mean-zero for free, but leaves $\mathbb{E}\|\boldsymbol{w}_{t+1}\|^2$ to be controlled, which is what (S_ξ), (MSS), $(\mathcal{L}_0, \mathcal{L}_1)$ or (V_Δ) supply. Evaluating it at different samples leaves a bias that only (H) or (M_0, M_1) can absorb — and that route is Theorem 3 of Khirirat et al. (2025), which reaches only $\epsilon^{-7/2}$. So every ϵ^{-3} result in this appendix invokes at least one of {HVP oracle, Hessian smoothness, a condition on f_ξ}; stripping all three returns SGD's ϵ^{-4}, as Arjevani et al. (2023) require. None of the four entries above escapes the $K = 2$ oracle: the same-sample requirement is what makes the residual mean-zero, so "no HVP" here does not mean "standard oracle".

Assumption strength within the set. By the ordering equation 170, entry 4 is strictly the weakest: (V_Δ) is implied by $(\mathcal{L}_0, \mathcal{L}_1)$ at $\mathcal{L}_1 = 0$, which is implied by (S_ξ). Qian et al. (2025) make this explicit, noting that their δ_i is bounded both by the expected Hessian variance and by the expected smoothness constant, since the left side of equation 166 is itself bounded by $\mathbb{E}\|\nabla_i f_\xi(\boldsymbol{x}) - \nabla_i f_\xi(\boldsymbol{v})\|_2^2$. Assuming ($\mathrm{V}_\Delta$) directly is the sharp move: it is precisely the quantity the proof needs, and nothing more.

The Schatten-p subset: entries 3 and 4. Entries 1 and 2 are Euclidean, so for matrix-parameterised problems — where the natural geometry is a Schatten norm S_p on $\mathbb{R}^{m\times n}$, giving the nuclear ball at $p = 1$, normalized SGD at $p = 2$, and Muon at $p = \infty$ — only entries 3 and 4 apply. This pair is worth isolating, because together they are the only results in the literature surveyed here that simultaneously

(a) attain $\mathcal{O}(\epsilon^{-3})$ in a genuinely non-Euclidean norm;

(b) do not assume a bounded stochastic gradient (G) — Khirirat et al. (2025) name this explicitly as an assumption of Tran & Cutkosky (2022) that they dispense with;

(c) query no Hessian-vector product; and

(d) assume no Hessian smoothness, in either the (H) or the (M_0, M_1) form.

Moreover the analyses are valid for every $p \in [1, \infty]$, which is not automatic and deserves comment. The natural obstruction is that a martingale argument in a Banach space ordinarily requires 2-uniform smoothness of the dual norm, and S_q is 2-uniformly smooth only for $q \geq 2$, i.e. only for $p \leq 2$ — which would exclude the spectral-norm case that motivates Muon. Neither proof incurs this restriction, because neither works in the Banach space at all: both embed into the Frobenius norm, apply orthogonality there, and convert back. Concretely, the load-bearing step of Khirirat et al. (2025), Appendix F is

$$\mathbb{E}_{k+1}\Big\| \sum_{t\leq k-1} c_t \hat{S}_{t+1} + (1-\alpha_k)\hat{S}_{k+1}\Big\|_2^2 = \Big\| \sum_{t\leq k-1} c_t \hat{S}_{t+1}\Big\|_2^2 + (1-\alpha_k)^2\, \mathbb{E}_{k+1}\big\|\hat{S}_{k+1}\big\|_2^2, \quad (171)$$

an identity in S_2 that holds for any p. The linear minimization oracle is likewise well defined throughout, returning a rank-one matrix at $p = 1$, a rescaling at $p = 2$, and the orthogonal polar factor at $p = \infty$.

What does vary with p is the constant, through (NE). Writing $r = \min(m, n)$,

$$\frac{\bar{\rho}}{\underline{\rho}} = r^{|1/2-1/p|} \quad \text{for Khirirat et al. (2025)}, \qquad \rho = r^{(1/2-1/p)_+} \quad \text{for Qian et al. (2025)}. \quad (172)$$

The asymmetry is structural rather than incidental. Qian et al. (2025) require only the one-sided bound $\|X\|_{(i)\star} \leq \rho \|X\|_2$ of their Assumption 7, because their Assumption 8 is already stated with a Frobenius left-hand side and a primal-norm right-hand side, so no back-conversion arises; Khirirat et al. (2025) convert in both directions and therefore pay both $\bar{\rho}$ and $\underline{\rho}$. Consequently at $p = 1$ the Gluon-MVR constant is dimension-free while HCM's is $\sqrt{r}$; at $p = \infty$ both are $\sqrt{r}$; at $p = 2$ both vanish.

Which of the two to instantiate. Qian et al. (2025), Theorem 2, on three grounds. Its assumption (V_Δ) is strictly implied by $(\mathcal{L}_0, \mathcal{L}_1)$ and is exactly the quantity the proof consumes; its norm-equivalence penalty vanishes for all $p \leq 2$ rather than only at $p = 2$, by equation 172; and, decisively, it treats $L_i^1 = 0$ as an explicit case, whereas the stepsize $\hat{\eta} = \frac{1}{3}\min\{1/L_1, \frac{1}{11\mathcal{L}_1}(\bar{\rho}/\underline{\rho})^{-1}\}$ of Khirirat et al. (2025), Theorem 4 diverges as $L_1, \mathcal{L}_1 \to 0$ and so does not cover the standard-smooth Schatten-p case — which is the case one would most often want to instantiate.

The condition this subset does not remove. Neither result escapes the $K = 2$ oracle: both evaluate the same ξ at $\boldsymbol{x}^k$ and $\boldsymbol{x}^{k-1}$. This is not incidental to the construction but constitutive of it, since the same-sample property is what makes $\boldsymbol{w}_{t+1}$ mean-zero; with two independent samples the residual acquires a bias that only (H) or (M_0, M_1) can absorb. Hence (c) above should be read as "no second-order oracle", not as "a single-query stochastic first-order oracle". If one additionally demands $K = 1$, then no result in this appendix qualifies, and to our knowledge none in the literature does: the best available is the $\epsilon^{-7/2}$ of Khirirat et al. (2025), Theorem 3, which purchases it with (M_0, M_1) and therefore violates (d) in any case, and whether ϵ^{-3} is attainable at $K = 1$ remains open (Arjevani et al., 2023).

Which of these are arbitrary-norm generalizations of STORM? Only entries 3 and 4, and among those only some. The test is whether $\boldsymbol{c}_{t+1}$ and equation 158 coincide with STORM's after the regrouping equation 164, with the step map replaced by an LMO.

- **Gluon-MVR-1** (Qian et al., 2025, Thms. 2–3) is the faithful generalization. Its recursion $\boldsymbol{m}_i^k = \nabla_i f_{\xi^k}(\boldsymbol{x}^k) + \beta(\boldsymbol{m}_i^{k-1} - \nabla_i f_{\xi^k}(\boldsymbol{x}^{k-1}))$ is STORM's, verbatim and blockwise, with $\mathcal{D}_{\mathrm{lmo}}$ in $\|\cdot\|_{(i)}$ replacing $\mathcal{D}_{\mathrm{sgd}}$. The authors record that with the spectral norm it coincides with LiMuon (option 1) and with MVR2 at $\gamma = 1$ of Chang et al.
- **HCM Theorem 4** (Khirirat et al., 2025) is the same generalization for a single global norm rather than layer-wise, and under $(\mathcal{L}_0, \mathcal{L}_1)$ rather than (V_Δ). It is therefore strictly weaker than Gluon-MVR-1 in assumptions and less general in geometry, which is why its novelty is thin: the authors themselves position it as matching a known rate.

- **Gluon-MVR-2 and -3** are not generalizations of STORM. They apply the STORM recursion to an inner buffer with weight q and then smooth a second time; MVR-3 additionally re-injects the correction at the outer level. These are new estimators, and the q they introduce is what produces the phase boundary $q < \alpha$ recorded in Table 8.
- **MARS** is not an arbitrary-norm generalization either. Its γ_t scaling generalizes STORM in a different direction — towards control variates — while the geometry stays Euclidean up to the preconditioner $\boldsymbol{H}_t$.

The $q < \alpha$ boundary is worth flagging against Khirirat et al. (2025), who leave the optimal scaling β_k in their equation (6) explicitly open. Qian et al. (2025) do not answer that question for MARS's estimator, but they do answer the structurally analogous one for theirs, and the answer is a strict inequality rather than an optimum. Their empirical selection, $q = 0.7$ against $\alpha = 0.8$, satisfies it.

E.8 Why every analysis passes through the same step

The descent argument is common to all seven papers and is indifferent to the contents of $\boldsymbol{m}_t$. For an LMO step of radius η_t one has the duality identity $\langle \boldsymbol{m}_t, \eta_t \mathcal{D}_{\mathrm{lmo}}(\boldsymbol{m}_t, \boldsymbol{x}_t)\rangle = -\eta_t \|\boldsymbol{m}_t\|_\star$, and combining it with smoothness gives

$$f(\boldsymbol{x}_{t+1}) \;\le\; f(\boldsymbol{x}_t) - \eta_t \|\nabla f(\boldsymbol{x}_t)\|_\star + 2\eta_t \|\boldsymbol{e}_t\|_\star + \tfrac{L}{2}\eta_t^2, \tag{173}$$

so that, telescoping,

$$\frac{1}{T}\sum_t \mathbb{E}\|\nabla f(\boldsymbol{x}_t)\|_\star \;\lesssim\; \underbrace{\frac{\Delta_0}{\eta T}}_{\text{progress}} \;+\; \underbrace{\frac{2}{T}\sum_t \mathbb{E}\|\boldsymbol{e}_t\|_\star}_{\text{the only nontrivial term}} \;+\; \underbrace{\frac{L\eta}{2}}_{\text{discretization}}. \tag{174}$$

The outer terms depend only on the stepsize. The entire design space therefore reduces to the decay of $\mathbb{E}\|\boldsymbol{e}_t\|_\star$, which is why these papers share a skeleton. Substituting equation 158 and equation 159,

$$\boldsymbol{e}_{t+1} = (1-\alpha_t)\boldsymbol{e}_t + (1-\alpha_t)\boldsymbol{w}_{t+1} + \alpha_t\big(\boldsymbol{g}_{t+1} - \nabla f(\boldsymbol{x}_{t+1})\big), \tag{175}$$

whence

$$\mathbb{E}\|\boldsymbol{e}_{t+1}\|_\star \;\le\; (1-\alpha)^{t+1}\mathbb{E}\|\boldsymbol{e}_0\|_\star \;+\; \mathbb{E}\Big\|\sum_{j=0}^{t}(1-\alpha)^{t-j}\boldsymbol{w}_{j+1}\Big\|_\star \;+\; \rho\sigma_g\sqrt{\alpha}. \tag{176}$$

Here ρ satisfies $\|\boldsymbol{v}\|_\star \le \rho\|\boldsymbol{v}\|_2$ for every $\boldsymbol{v}$, and $\alpha \in (0,1]$ is constant.

Everything hinges on the middle term of equation 176. If $\mathbb{E}_t[\boldsymbol{w}_{t+1}] = \boldsymbol{0}$ and $\mathbb{E}_t\|\boldsymbol{w}_{t+1}\|_2^2 = \mathcal{O}(\eta^2)$ uniformly, where $\mathbb{E}_t$ conditions on the history before sampling at step $t+1$, the residuals form a martingale difference sequence. Orthogonality in the Euclidean norm, followed by norm conversion, bounds the weighted sum by $\mathcal{O}(\rho\eta/\sqrt{\alpha})$. A conditional bias of size $\mathcal{O}(\eta)$ in the dual norm can instead contribute $\mathcal{O}(\eta/\alpha)$. With $\eta = \hat{\eta}\alpha$ and $\alpha = T^{-2/3}$, the martingale case yields $\mathcal{O}(\rho\hat{\eta}T^{-1/3})$, matching the noise term $\rho\sigma_g\sqrt{\alpha}$ and the progress term in equation 174. For fixed problem parameters, this gives the rate $T^{-1/3}$, corresponding to $\mathcal{O}(\epsilon^{-3})$.

With no correction ($\boldsymbol{c}_{t+1} \equiv \boldsymbol{0}$), $\boldsymbol{w}_{t+1} = -(\nabla f(\boldsymbol{x}_{t+1}) - \nabla f(\boldsymbol{x}_t))$ is generally not conditionally mean-zero. Under standard smoothness and updates of norm at most η, its dual norm is $\mathcal{O}(\eta)$. Balancing $\Delta_0 T^{-1}\eta^{-1}$, η/α, and $\rho\sigma_g\sqrt{\alpha}$ with $\eta = T^{-a}$ and $\alpha = T^{-b}$ gives $a = 3/4$ and $b = 1/2$, yielding $T^{-1/4}$ and $\mathcal{O}(\epsilon^{-4})$ complexity.

The improved rate therefore requires quantitative control of both the residual variance and its conditional bias. Retaining the same second-moment bound, a sufficient alternative to exact conditional centering is $\|\mathbb{E}_t[\boldsymbol{w}_{t+1}]\|_\star = \mathcal{O}(\eta\sqrt{\alpha})$, whose accumulated contribution is $\mathcal{O}(\eta/\sqrt{\alpha})$. A bias of merely $o(\eta)$ does not ensure the claimed rate.

The seven methods differ precisely in how they enforce this.

STORM, MARS and Gluon-MVR. Reusing the same ξ_{t+1} at both endpoints makes $\boldsymbol{w}_{t+1}$ mean-zero by construction; the work lies in bounding $\mathbb{E}\|\boldsymbol{w}_{t+1}\|_2^2 \lesssim \eta^2$. STORM and MARS

obtain this from (S_ξ), HCM's Theorem 4 from ($\mathcal{L}_0, \mathcal{L}_1$), and Qian et al. (2025) by assuming it directly as (V_Δ) — which, by equation 170, is the weakest of the three. MARS additionally rescales by γ_{t+1}, which reintroduces bias in exchange for variance: their Lemma C.2 exhibits a subtracted term $M_{t+1} \geq 0$ that vanishes exactly at $\gamma_{t+1} = 1$, so STORM is the worst admissible choice within that family.

SOM. Does not achieve mean-zero: $\mathbb{E}_\xi \boldsymbol{c}_{t+1} = \nabla^2 f(\boldsymbol{x}_{t+1})(\boldsymbol{x}_{t+1} - \boldsymbol{x}_t)$, which differs from equation 157 by the Taylor remainder. Assumption (H) bounds that remainder by $\mathcal{O}(\rho\eta^2) = o(\eta)$, small enough to absorb.

1-SFW and HCM V1. Genuinely mean-zero, because $\int_0^1 \nabla^2 f(a\boldsymbol{x}_{t+1} + (1-a)\boldsymbol{x}_t)\, da$ equals the gradient variation exactly; the uniform a evaluates equation 157 by Monte Carlo. The price is that $\hat{\boldsymbol{x}}_{t+1}$ is neither endpoint, so the HVP needs its own forward–backward pass.

HCM. For Theorems 1–2 the estimator is one of the two above; what is new is the norm in which $\|\boldsymbol{e}_t\|$ is measured, embedding into ℓ_2 to run the orthogonality argument and converting back. The price is (NE), whose constant for a Schatten-p LMO on $\mathbb{R}^{m\times n}$ scales as $r^{|1/2-1/p|}$ with $r = \min(m, n)$.

RanSOM. Makes the endpoint HVP mean-zero by randomising the step length rather than the evaluation point. Writing $g(s) = \nabla f(\boldsymbol{x}_t + s\boldsymbol{d}_t)$, integration by parts gives $\mathbb{E}[g(s) - g(0)] = \lambda^{-1}\mathbb{E}[g'(s)]$ for $s \sim \mathrm{Exp}(\lambda)$, and $\mathbb{E}[g(s) - g(0)] = \mathbb{E}[\frac{1-s}{K} g'(s)]$ for $s \sim \mathrm{Beta}(1, K)$, so a single HVP at $\boldsymbol{x}_{t+1}$ is exactly unbiased. The randomness in s_t performs the integration for which 1-SFW paid a third pass.

Reading the prices in order: STORM and MARS pay in assumptions on f_ξ; SOM pays in third derivatives of f; 1-SFW pays in compute and in per-sample almost-sure bounds; HCM pays in the dimension-dependent constant of (NE); Gluon-MVR pays only (V_Δ) and (NE); and RanSOM pays in the tail of the step distribution, since $\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\| = s_t\rho$ is unbounded under $\mathrm{Exp}(1/\eta_t)$, which removes the deterministic trust-region property $\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\| \leq \eta_t$ on which the LMO analyses otherwise rely, and in the initial batch B_{init}.

E.9 Empirical evidence, and a caveat about it

Cutkosky & Orabona (2019) report CIFAR-10 with a ResNet-32, concluding that STORM is only marginally better than AdaGrad on test accuracy. Tran & Cutkosky (2022) report CIFAR image classification and neural machine translation against Adam, SGD and AdaHessian. Zhang et al. (2020b) is essentially theoretical; the policy-gradient adaptation (Salehkaleybar et al., 2022) reports MuJoCo control benchmarks. Khirirat et al. (2025) report an MLP on the `splice` dataset (1000 samples, 60 features) and a two-layer LSTM on Penn Treebank. Chayti (2026) report `splice`, MNIST-1D, and constrained matrix completion on a reduced MovieLens instance under a nuclear-norm ball.

Two entries are at a different scale. Yuan et al. (2025) train GPT-2 at 125M, 355M and 770M parameters on OpenWebText, where MARS-AdamW reaches validation loss 2.511 against 2.568 for AdamW and 2.606 for Muon on GPT-2 large, and attains a target loss of 2.58 in 28B tokens where AdamW needs 50B. Qian et al. (2025) train NanoGPT with 124M parameters on FineWeb-10B for 5000 steps: against a Gluon baseline at validation loss 3.832, Gluon-MVR-2 reaches 3.682 and Gluon-MVR-3 reaches 3.742, while Gluon-MVR-1 is essentially tied with the baseline at $B = 512$ (3.829) and helps only at $B = 128$ (4.107 against 4.256).

Two observations. First, the empirical ordering MVR-2 > MVR-3 > MVR-1 inverts the theoretical one: Gluon-MVR-1 is the faithful STORM generalization and the cleanest to analyse, yet it is the weakest in practice, while the two estimators with no interpretation as a norm generalization of STORM are the ones that work. Second, the only entry with language-model-scale evidence and a matching theorem is Qian et al. (2025): Yuan et al. (2025) ship MARS-APPROX, which replaces $\nabla f_{\xi_{t+1}}(\boldsymbol{x}_t)$ by $\nabla f_{\xi_t}(\boldsymbol{x}_t)$ and thereby breaks the same-sample property that makes $\boldsymbol{w}_{t+1}$ mean-zero, so the shipped variant is not covered by Theorems B.5 or B.6.

Method	Result	Quantity bounded, and the bound	Complexity
STORM	Thm. 1	$\mathbb{E}\|\nabla f(\boldsymbol{x}_{\text{out}})\|_2 = \mathbb{E}\big[\frac{1}{T}\sum_{t=1}^{T}\|\nabla f(\boldsymbol{x}_t)\|_2\big] \le \frac{w^{1/6}\sqrt{2M}+2M^{3/4}}{\sqrt{T}} + \frac{2\sigma_g^{1/3}}{T^{1/3}}$, where $M = \frac{8}{k}\Delta_0 + \frac{w^{1/3}\sigma_g^2}{4L^2k^2} + \frac{k^2c^2}{2L^2}\ln(T{+}2)$ is $\mathcal{O}(\log T)$. Unsquared, Euclidean, uniformly random iterate.	$\mathcal{O}(\epsilon^{-3})$, adaptive to σ_g
SOM	Thm. 1 (Alg. 1)	$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\|\nabla f(\boldsymbol{x}_t)\|_2^2 \le \frac{20C\Delta_0+96C^2G^2/K}{T^{2/3}} + \frac{20G^2D(1+\log T)}{C^2T^{2/3}}$. **Squared** norm.	
	Thm. 5 (Alg. 2)	$\mathbb{E}\|\nabla f(\boldsymbol{x}_{\text{out}})\|_2 \le \frac{6\sigma_g+54^{1/3}(\Delta_0\sigma_H)^{1/3}\sigma_g^{1/3}}{T^{1/3}} + \frac{6\sigma_g^{2/5}\Delta_0^{2/5}\rho^{1/5}}{T^{2/5}} + \frac{\sqrt{9\Delta_0 L}+\sqrt{72\Delta_0\sigma_H}}{\sqrt{T}}$; normalized step, no (G) needed.	
	Thm. 6 (Alg. 3)	Adaptive AdaGrad-type version of Thm. 5, matching STORM's form.	$\mathcal{O}(\epsilon^{-3})$
1-SFW	Thm. 2	$\mathbb{E}[\mathcal{G}(\boldsymbol{x}_{\text{out}})] \le \frac{2B+3\sqrt{C}D/2}{T^{1/3}} + \frac{\bar{L}D^2}{2T^{2/3}}$, **Frank–Wolfe gap**, $\boldsymbol{x}_{\text{out}}$ uniform over the iterates. Thm. 1 gives $\mathbb{E}[f(\boldsymbol{x}_{T+1}) - f(\boldsymbol{x}^\star)]$ in the convex case.	$\mathcal{O}(\epsilon^{-3})$, one sample/iter
MARS	B.5, B.6	**Not a gradient norm.** Two quantities are bounded simultaneously: $\frac{1}{T}\sum_t \mathbb{E}\|\nabla f(\boldsymbol{x}_t) - \boldsymbol{m}_t\|_2^2 \le \big(2\rho G' + \frac{\rho c^2\sigma_g^2}{4L^2}\log(s{+}T)\big)T^{-2/3} - \frac{\rho^2}{8L^2T^{1/3}}\sum_t M_{t+1}$, and $\frac{1}{T}\sum_t \eta_t^{-1}\mathbb{E}\|\boldsymbol{x}_{t+1} - \boldsymbol{x}_t\|_2^2 \le \big(\frac{16G'}{3\rho} + \frac{2c^2\sigma_g^2}{3L^2}\log(s{+}T)\big)T^{-2/3} - \frac{1}{6L^2T^{1/3}}\sum_t M_{t+1}$, with $G' = \Delta_0 + \rho s^{1/3}\sigma_g^2/(16L^2)$ and $M_{t+1} \ge 0$ from Lemma C.2. Thm. B.6 adds $\lambda D^2\log(s{+}T)$ terms.	stated in prose as $\mathcal{O}(\log T \cdot T^{-1/3})$
HCM	Thm. 1, 2	$\min_{k\le K}\mathbb{E}\|\nabla f(\boldsymbol{x}_k)\|_\star \le \frac{2\Delta_0}{\hat{\eta}(K+1)^{1/3}} + \frac{4(\mathbb{E}\|\boldsymbol{e}_0\|_\star+\bar{\rho}\sigma_g)}{(K+1)^{1/3}} + \frac{\hat{\eta}L_0e^{L_1\hat{\eta}}}{(K+1)^{2/3}} + \cdots + \frac{c\,(\bar{\rho}/\underline{\theta})\hat{\eta}\sigma_H}{(K+1)^{1/3}}$, in the **dual norm**, best iterate, unsquared. Thm. 1 (V1) takes $\hat{\eta} = \frac{1}{80L_1}(\bar{\rho}/\underline{\rho})^{-1}$; Thm. 2 (V2) takes $\hat{\eta} = \frac{1}{3}\min\{1/L_1,\, 1/\sqrt{M_1}\}$.	
	Thm. 3	Same criterion, rate $\mathcal{O}(K^{-2/7})$ with $\alpha = (K{+}1)^{-4/7}$, $\eta = \hat{\eta}(K{+}1)^{-5/7}$ and $\hat{\eta} = \min\{\frac{1}{3L_1}, \frac{1}{3\sqrt{M_1}}\}$, using extrapolated momentum: assumes (M_0, M_1) but **queries no Hessian**.	
	Thm. 4	Same criterion, using the STORM correction, with $\hat{\eta} = \frac{1}{3}\min\big\{\frac{1}{L_1}, \frac{1}{11\mathcal{L}_1}(\bar{\rho}/\underline{\rho})^{-1}\big\}$: $\min_{k\le K}\mathbb{E}\|\nabla f(\boldsymbol{x}_k)\|_\star \le \frac{2\Delta_0}{\hat{\eta}(K+1)^{1/3}} + \frac{4(\mathbb{E}\|\boldsymbol{e}_0\|_\star+\bar{\rho}\sigma_g)}{(K+1)^{1/3}} + \frac{\hat{\eta}L_0e^{L_1\hat{\eta}}}{(K+1)^{2/3}} + \frac{8\bar{\rho}}{\underline{\rho}}(\mathcal{L}_0 + 2\mathcal{L}_1\bar{\rho}\sigma_g)e^{6\mathcal{L}_1^2\hat{\eta}^2}\frac{\hat{\eta}}{(K+1)^{1/3}}$, i.e. $\mathcal{O}(K^{-1/3})$.	$\mathcal{O}(\epsilon^{-3})$; $\epsilon^{-7/2}$ for Thm. 3
RanSOM	4.5, 4.7; -B thm.	$\mathbb{E}\|\nabla f(\boldsymbol{x}_{\text{out}})\|_\star$ for RanSOM-E and the **Frank–Wolfe gap** for RanSOM-B, both at the optimal $\mathcal{O}(T^{-1/3})$ under light-tailed noise. Under (V_p) with $p \in (1, 2]$ the rate becomes $\mathcal{O}\big(T^{-(p-1)/(2p-1)}\big)$, attained without gradient or Hessian clipping.	$\mathcal{O}(\epsilon^{-3})$
Gluon-MVR	Thm. 2 (MVR-1)	$\min_{k<K}\sum_{i=1}^{p} t_i\mathbb{E}\|\nabla_i f(\boldsymbol{x}^k)\|_{(i)\star} \le \frac{\Delta_0}{\eta K} + \frac{2\sum_i t_i\rho\sigma}{\alpha K} + 2\sqrt{\alpha}\sum_i t_i\rho\sigma + \frac{2\sum_i \delta_i t_i^2\rho\eta}{\sqrt{\alpha}} + \frac{\sum_i L_i^0 t_i^2\eta}{2}$ at $L_i^1 = 0$; layer-weighted sum of **dual norms**, best iterate. With $\eta = \alpha = K^{-2/3}$ this is $\mathcal{O}(K^{-1/3})$.	
	Thm. 3	Same criterion with decreasing t_i^k, giving $\tilde{\mathcal{O}}(K^{-1/3})$ up to a $\log K$ factor.	
	Thm. 4 (MVR-2)	Same criterion. **A phase boundary at $q = \alpha$:** for $q \ge \alpha$ the bound contains $\frac{1}{\eta K} + \frac{\eta}{\alpha} + \sqrt{\alpha} \ge 2\sqrt{2}K^{-1/4}$, capping the rate at $\mathcal{O}(K^{-1/4})$; only for $q < \alpha$, with $\eta = q = K^{-2/3}$ and $\alpha \ge K^{-1/3}$, is $\mathcal{O}(K^{-1/3})$ attained.	
	Thm. 5 (MVR-3)	Same criterion; $\mathcal{O}(K^{-1/3})$ for $\eta = \alpha = K^{-2/3}$, $q \ge K^{-2/3}$, with no constraint of the form η/α required.	$\mathcal{O}(\epsilon^{-3})$

Table 8: Guarantees, with left-hand sides written out. Constants are those printed in the sources. RanSOM's closed-form right-hand-side constants could not be transcribed reliably and are omitted; the criteria, rates and heavy-tailed exponent are as stated.