IS MUON GEOMETRICALLY UNIQUE? ORTHONORMAL RANK-ONE SYSTEMS FOR NORMALIZED OPTIMIZERS

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent normalized optimizers such as Lion and Muon highlight the importance of geometry in modern optimizer design. Yet it remains unclear whether Muon's two-sided singular-vector orthogonality is essential to its effectiveness or instead represents one realization of a broader geometric structure. We extend the Lion-$\mathcal{K}$ perspective by representing the momentum variable in a possibly time-varying orthonormal rank-one system and defining $\mathcal{K}_t$ as the corresponding coefficient-wise ℓ_1-type potential. Different selection rules over orthonormal rank-one systems recover SGD, Lion, and Muon within a single geometric framework. Within this framework, one-sided orthogonality defines a broader family that contains Muon's two-sided orthogonality as a structured special case. To probe whether this additional two-sided structure is necessary, we introduce Random Orthogonal Basis Normalization (ROBN) and Split Polar Update (SPU) as realizations of one-sided orthogonality. ROBN shows that one-sided orthogonality alone can retain a meaningful portion of Muon's empirical performance, while SPU shows that a realization based on one-sided orthogonality can match Muon across both vision and language modeling. These results suggest that Muon is best understood as a distinguished realization of two-sided orthogonality within the broader geometric principle of one-sided orthogonality, rather than as a uniquely privileged geometric structure.

1 INTRODUCTION

AdamW (Loshchilov & Hutter, 2017) remains a widely used optimizer for large-scale language-model training (Liu et al., 2025), while Muon has recently emerged as a strong alternative based on matrix orthogonalization of momentum updates (Jordan et al., 2024; Liu et al., 2025). Muon and related normalized optimizers have attracted rapidly growing theoretical, algorithmic, and computational interest (Chen et al., 2024; 2026; Pethick et al., 2025; Sfyraki & Wang, 2025; Riabinin et al., 2026; Shen et al., 2025; Ahn et al., 2025; Si et al., 2025; Liu et al., 2026; Li et al., 2025; Zhang et al., 2026; Amsel et al., 2026). For a momentum matrix $m = U\Sigma V^\top$, Muon's normalized update is $UV^\top$, effectively replacing the nonzero singular values by one while retaining orthogonal left and right singular-vector directions. Yet it remains unclear whether Muon's use of orthogonality on both sides is essential to its effectiveness or instead represents a realization within a broader geometric structure.

This motivates a central question: *Is Muon's orthogonality on both sides of the matrix update necessary for strong optimization performance, or can a method that enforces orthogonality on only one side achieve comparable performance?*

To answer this question, we show that orthonormal rank-one systems recover the update directions of SGD, Lion, and Muon within a single geometric framework.

This framework reveals *one-sided orthogonality* as a broader geometric family containing the balanced SGD realization and Muon's two-sided singular-vector orthogonality as a structured special case. To probe whether this additional two-sided structure is necessary, we introduce Random Orthogonal Basis Normalization (ROBN) and Split Polar Update (SPU) as realizations of one-sided orthogonality. ROBN retains a meaningful portion of Muon's empirical performance while enforcing only one-sided orthogonality, while SPU matches or outperforms Muon across language modeling

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.

and vision. Together, these results support viewing Muon through the broader geometric principle of one-sided orthogonality rather than through a uniquely superior or independent principle of two-sided orthogonality.

Our main contributions are therefore threefold:

- We introduce a possibly time-varying orthonormal rank-one framework in which a coefficient-wise ℓ_1-type potential $\mathcal{K}_t$ and its selected subgradient recover the update directions of SGD, Lion, and Muon. We further establish convergence of the resulting momentum formulation under the assumptions stated in Appendix D.
- We identify one-sided orthogonality as a broader geometric family that includes the balanced SGD realization and contains Muon's two-sided singular-vector orthogonality as a structured special case. ROBN empirically shows that one-sided orthogonality can retain a meaningful portion of Muon's performance without its two-sided singular-vector structure.
- We construct SPU as a complementary realization of one-sided orthogonality. Across language modeling and vision, SPU matches or outperforms Muon. Together, the ROBN and SPU results support the view that Muon's effectiveness is better understood through the broader principle of one-sided orthogonality than through a uniquely superior or independent principle of two-sided orthogonality.

2 Related Work

First-order and structure-aware optimization. Deep-learning optimization has traditionally been dominated by first-order methods based on gradients and momentum. Classical momentum methods accumulate information from previous iterations (Polyak, 1964; Sutskever et al., 2013), while adaptive methods modify the update through coordinate-wise statistics. AdaGrad uses accumulated squared gradients (Duchi et al., 2011), RMSProp uses an exponential moving average of squared gradients (Tieleman, 2012), and Adam combines adaptive first- and second-moment estimates (Kingma & Ba, 2014), while AdamW decouples weight decay from the loss-gradient update (Loshchilov & Hutter, 2017). Beyond coordinate-wise adaptation, structure-aware methods exploit richer curvature or parameter structure. K-FAC uses a Kronecker-factored approximation of the Fisher information matrix (Martens & Grosse, 2015), Sophia incorporates lightweight diagonal-Hessian information (Liu et al., 2024), Shampoo maintains tensor-structured preconditioners (Gupta et al., 2018), and SOAP applies Adam in the eigenbasis of Shampoo's preconditioner (Vyas et al., 2025).

Normalized optimization and geometric interpretations. Normalized update rules provide another route for imposing geometry on optimization. Lion was discovered through symbolic program search and uses a sign-based momentum update (Chen et al., 2023), whereas Muon uses Newton–Schulz iterations to approximately orthogonalize a momentum update (Jordan et al., 2024). These methods have received several complementary theoretical interpretations. The Lion-$\mathcal{K}$ framework interprets Lion with decoupled weight decay through an ℓ_∞-constrained optimization problem (Chen et al., 2024). This perspective was extended to Muon by identifying the nuclear norm as the corresponding convex map $\mathcal{K}$, yielding a spectral-norm-constrained formulation on the weight matrices under decoupled weight decay (Chen et al., 2026). Our framework extends this potential-based perspective by allowing the selection rule over orthonormal rank-one systems to vary explicitly with time, thereby inducing a time-varying coefficient-wise potential $\mathcal{K}_t$. A separate line of work studies normalized updates through linear minimization oracles (LMOs) over norm balls, where different norm choices recover normalized SGD, sign-based updates, and the non-Nesterov variant of Muon (Pethick et al., 2025). Stochastic Frank–Wolfe analysis likewise places Lion and Muon with weight decay within a common Frank–Wolfe formulation (Sfyraki & Wang, 2025). Building on the LMO viewpoint, Gluon develops a layer-wise LMO framework for large-language-model training (Riabinin et al., 2026), while direct convergence analysis relates Muon's matrix-valued updates to structural properties of the Hessian (Shen et al., 2025). Recent work has further examined which aspects of Muon's geometry are essential. Alternative singular-value transformations, including randomized and inverted spectra, have been studied (Shumaylov et al., 2026), while Muse (Chang et al., 2026) studies how the matrix representation used before polar normalization changes the induced optimization geometry.

Muon extensions. Muon has also been extended in several algorithmic and computational directions. A large-scale study demonstrates its applicability to language-model pre-training (Liu et al., 2025), Dion develops distributed orthonormalized updates (Ahn et al., 2025), and MuonBP combines blockwise orthogonalization of matrix shards with periodic full orthogonalization to reduce model-parallel communication overhead (Khaled et al., 2026). Adaptive variants include AdaMuon (Si et al., 2025), Muon^2 (Liu et al., 2026), and NorMuon (Li et al., 2025). TEON extends orthonormalization to tensorized parameter structures (Zhang et al., 2026), while Polar Express develops efficient matrix-sign methods for computing Muon's polar factor (Amsel et al., 2026). These works broaden Muon's theoretical and practical scope, but do not directly isolate whether Muon's two-sided singular-vector orthogonality itself is necessary, which is the geometric question studied in this work.

3 Orthonormal Rank-One Systems

We first introduce a unified notation for normalized update rules. Let m_t denote the momentum or update variable along the optimization trajectory, and let m denote a generic argument in the same ambient space.

For each pair (t, m), a selection rule specifies an admissible orthonormal rank-one system representing m. Depending on the update geometry, the rule may constrain the coefficients (as in balanced SGD), restrict the admissible orthonormal rank-one systems through additional structural conditions (as in Muon, SPU, and ROBN), or select a fixed admissible orthonormal rank-one system (as in Lion).

$$\mathcal{B}_t(m) = \{M_i(t, m)\}_i,$$

where the elements are mutually orthonormal and rank one:

$$\langle M_i(t, m), M_j(t, m)\rangle = \delta_{ij}, \qquad \operatorname{rank}(M_i(t, m)) = 1. \tag{1}$$

The corresponding coefficients $a_i(t, m) \in \mathbb{R}$ are defined by

$$a_i(t, m) = \langle m, M_i(t, m)\rangle, \qquad m = \sum_i a_i(t, m) M_i(t, m). \tag{2}$$

Here, $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product for vectors and the Frobenius inner product for matrices. Along the optimization trajectory, we write $a_i(t) := a_i(t, m_t)$, so that

$$m_t = \sum_i a_i(t) M_i(t, m_t).$$

This orthonormal rank-one representation provides the common structure of the framework, while the selection rule determines the specific update geometry within this structure.

With the orthonormal rank-one system selected according to the rule, we define the corresponding coefficient-wise potential

$$\mathcal{K}_t(m) = \sum_i |\langle m, M_i(t, m)\rangle| = \sum_i |a_i(t, m)|. \tag{3}$$

We assume that, for each t, $\mathcal{K}_t$ is proper, closed, convex, and positively homogeneous on $\mathcal{X}$, where positive homogeneity means

$$\mathcal{K}_t(\lambda m) = \lambda \mathcal{K}_t(m), \qquad \lambda \geq 0, \quad m \in \mathcal{X}. \tag{4}$$

We consider selection rules for which this coefficient-wise construction satisfies these properties and for which the signed combination below is a valid subgradient of $\mathcal{K}_t$. Along the optimization trajectory,

$$\mathcal{K}_t(m_t) = \sum_i |a_i(t)|. \tag{5}$$

We use the sign function with the convention $\operatorname{sign}(0) = 0$, corresponding to the zero subgradient selection at zero coefficients. The selected momentum subgradient is

$$\partial_m \mathcal{K}_t(m_t) := \sum_i \operatorname{sign}(a_i(t)) M_i(t, m_t) \in \partial \mathcal{K}_t(m_t). \tag{6}$$

The corresponding normalized update direction is determined by this selected subgradient.

3.1 One-Sided and Two-Sided Orthogonality

We next distinguish two structural forms of orthogonality within the orthonormal rank-one system. Since each $M_i(t,m)$ has rank one and unit Frobenius norm, it can be written as

$$M_i(t,m) = p_i(t,m)q_i(t,m)^\top, \qquad \|p_i(t,m)\|_2 = \|q_i(t,m)\|_2 = 1. \tag{7}$$

Collecting the left and right factors, define

$$P_t(m) = [p_1(t,m) \quad \cdots \quad p_r(t,m)], \qquad Q_t(m) = [q_1(t,m) \quad \cdots \quad q_r(t,m)]. \tag{8}$$

For two rank-one elements,

$$\langle M_i(t,m), M_j(t,m)\rangle_F = \langle p_i(t,m), p_j(t,m)\rangle\langle q_i(t,m), q_j(t,m)\rangle. \tag{9}$$

Therefore, orthogonality of the orthonormal rank-one system can be ensured by imposing orthogonality on only one side of the factorization.

We call the system *one-sided orthogonal* when orthogonality is imposed on at least one factor side, that is,

$$P_t(m)^\top P_t(m) = I \qquad \text{or} \qquad Q_t(m)^\top Q_t(m) = I. \tag{10}$$

Because the factors on the other side have unit norm, either condition is sufficient to preserve

$$\langle M_i(t,m), M_j(t,m)\rangle_F = \delta_{ij}. \tag{11}$$

We call the stronger case *two-sided orthogonal* when both factor systems are orthonormal:

$$P_t(m)^\top P_t(m) = I, \qquad Q_t(m)^\top Q_t(m) = I. \tag{12}$$

Thus, two-sided orthogonality is a structured special case within the broader one-sided orthogonal family: orthogonality of both factor sides is sufficient, but not necessary, for constructing an orthonormal rank-one system.

3.2 Recovering SGD, Lion, and Muon

We now show how SGD, Lion, and Muon arise from different selection rules within the orthonormal rank-one framework.

SGD. For SGD, let $m_t \in \mathbb{R}^{d_1\times d_2}$, fix an integer n independently of t and m_t such that $\min(d_1,d_2) \le n \le \max(d_1,d_2)$, and consider a balanced orthonormal rank-one representation of m_t, that is, one with equal coefficients:

$$a_i(t) = \frac{\|m_t\|_F}{\sqrt{n}}, \qquad i = 1,\dots,n. \tag{13}$$

For $m_t \neq 0$, the selected subgradient becomes

$$\partial_m \mathcal{K}_t(m_t) = \sum_{i=1}^{n} M_i(t,m_t) = \frac{\sqrt{n}}{\|m_t\|_F} m_t. \tag{14}$$

Thus the induced direction is aligned with m_t, recovering the update direction of SGD up to normalization. The existence of such a balanced orthonormal rank-one representation is shown in Appendix A.1. The construction given there is also one-sided orthogonal: depending on the construction, either the left or the right factor system is orthonormal. Hence, the balanced SGD representation considered here belongs to the one-sided orthogonal family while satisfying the equal-coefficient condition above.

Lion. Lion is obtained by selecting the fixed coordinate rank-one basis

$$M_{ij} = e_i e_j^\top, \qquad a_{ij}(t) = \langle m_t, M_{ij}\rangle = (m_t)_{ij}, \tag{15}$$

where e_i and e_j are standard basis vectors. The corresponding coefficient-wise potential is

$$\mathcal{K}_t(m_t) = \sum_{i,j} |a_{ij}(t)| = \sum_{i,j} |(m_t)_{ij}| = \|m_t\|_1. \tag{16}$$

Consequently, under the zero-selection convention at zero entries,

$$\partial_m \mathcal{K}_t(m_t) = \operatorname{sign}(m_t), \tag{17}$$

which recovers the sign update used by Lion.

Muon. Muon is obtained by imposing two-sided orthogonality on the rank-one factors. For $m_t \neq 0$, we retain only the active terms with $a_i(t) \neq 0$ and relabel them as $i = 1, \ldots, r$. For $m_t = 0$, the potential and the selected update are both zero. Let

$$M_i(t, m_t) = p_i(t, m_t) q_i(t, m_t)^\top, \tag{18}$$

and collect the factors as

$$P_t(m_t) = [p_1(t, m_t) \quad \cdots \quad p_r(t, m_t)], \qquad Q_t(m_t) = [q_1(t, m_t) \quad \cdots \quad q_r(t, m_t)]. \tag{19}$$

Under two-sided orthogonality,

$$P_t(m_t)^\top P_t(m_t) = I_r, \qquad Q_t(m_t)^\top Q_t(m_t) = I_r. \tag{20}$$

The momentum representation can therefore be written as

$$m_t = \sum_{i=1}^{r} a_i(t) p_i(t, m_t) q_i(t, m_t)^\top = P_t(m_t) A_t Q_t(m_t)^\top, \tag{21}$$

where

$$A_t = \operatorname{diag}(a_1(t), \ldots, a_r(t)). \tag{22}$$

Since both factor systems are orthonormal, the signs of the coefficients can be absorbed into either factor system, yielding

$$m_t = \widetilde{P}_t(m_t) |A_t| Q_t(m_t)^\top, \tag{23}$$

where

$$\widetilde{P}_t(m_t) = P_t(m_t) \operatorname{diag}\big(\operatorname{sign}(a_1(t)), \ldots, \operatorname{sign}(a_r(t))\big), \tag{24}$$

which also has orthonormal columns. Hence this representation is a singular-vector decomposition of m_t, up to the usual ordering and sign ambiguities, with singular values $|a_i(t)|$.

It follows that the coefficient-wise potential becomes

$$\mathcal{K}_t(m_t) = \sum_{i=1}^{r} |a_i(t)| = \|m_t\|_*. \tag{25}$$

Thus, imposing two-sided orthogonality within orthonormal rank-one systems recovers the nuclear-norm geometry underlying Muon. Moreover, Appendix A.2 proves the converse: any orthonormal rank-one representation satisfying two-sided orthogonality is necessarily a singular-vector decomposition. Hence, within this framework, two-sided orthogonality characterizes Muon's singular-vector geometry.

3.3 Time-Dependent Selection Rules

The selected orthonormal rank-one system may change over time along the optimization trajectory as its argument m_t evolves. However, this argument-induced temporal variation does not by itself imply that the induced potential $\mathcal{K}_t$ is explicitly time-dependent. For example, in Muon, the selected singular-vector orthonormal rank-one system changes with m_t, while the induced potential remains the time-independent nuclear norm.

More generally, the selection rule may also change with time even when its argument m is held fixed. In this case, the selected orthonormal rank-one system, and consequently the induced potential $\mathcal{K}_t(m)$, may depend explicitly on t, as in ROBN.

Assuming that $t \mapsto m_t$ and $t \mapsto M_i(t, m_t)$ are locally absolutely continuous, differentiating $\mathcal{K}_t(m_t) = \sum_i |a_i(t)|$ yields, for almost every t,

$$\frac{d}{dt}\mathcal{K}_t(m_t) = \langle \partial_m \mathcal{K}_t(m_t), \dot{m}_t \rangle + \partial_t \mathcal{K}_t(m_t). \tag{26}$$

For almost every t, we denote the remaining time-variation term by

$$\partial_t \mathcal{K}_t(m_t) := \left\langle m_t, \sum_i \operatorname{sign}(a_i(t)) \frac{d}{dt} M_i(t, m_t) \right\rangle. \tag{27}$$

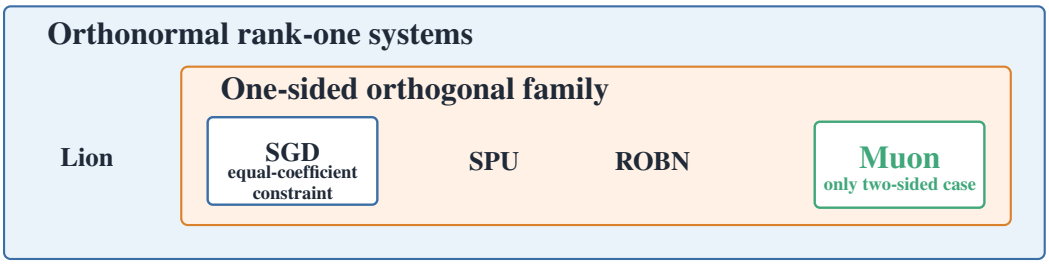


Figure 1: Geometric organization of the orthonormal rank-one framework. Balanced SGD, ROBN, SPU, and Muon belong to the one-sided orthogonal family, while Lion belongs to the broader orthonormal rank-one framework. Balanced SGD is distinguished by its equal-coefficient condition, whereas the two-sided case corresponds to Muon's singular-vector geometry.

Although the selected orthonormal rank-one system is fixed over time for Lion, the systems selected for balanced SGD and Muon may change over time as the momentum argument m_t evolves along the optimization trajectory. Nevertheless, the induced potentials in all three cases are time-independent. Consequently, the time-variation term defined above vanishes along their optimization trajectories:

$$\partial_t \mathcal{K}_t(m_t) = 0 \qquad \text{for SGD, Lion, and Muon.} \tag{28}$$

This property is verified in Appendix A.3. In contrast, the time-dependent extension considered in this work allows the term to be nonzero.

4 One-Sided Orthogonality Instantiations

We now introduce two concrete realizations of the one-sided orthogonality structure defined in Section 3.1: Random Orthogonal Basis Normalization (ROBN) and the Split Polar Update (SPU). Both construct orthonormal rank-one systems by enforcing orthogonality on only one factor side, but they realize this principle in different ways.

4.1 Random Orthogonal Basis Normalization

ROBN is designed to test whether one-sided orthogonality can retain a meaningful portion of Muon's empirical performance when the orthogonal matrix A_t is randomly resampled independently of the current momentum at every step. Given a momentum matrix $m_t \in \mathbb{R}^{r\times c}$, ROBN first applies a randomly sampled orthogonal matrix to m_t along its larger dimension. It then normalizes each row of the resulting matrix when $r \geq c$, and each column when $r < c$.

Throughout this subsection, $D_t^\dagger$ denotes the Moore–Penrose inverse of the diagonal matrix D_t; thus, positive diagonal entries are inverted, while zero diagonal entries remain zero.

When $r \geq c$, let $A_t \in \mathbb{R}^{r\times r}$ be a randomly sampled orthogonal matrix and define

$$\widetilde{m}_t = A_t^\top m_t, \qquad D_t = \operatorname{diag}\left(\|\widetilde{m}_{t,1:}\|_2, \ldots, \|\widetilde{m}_{t,r:}\|_2\right).$$

Then

$$\Delta_t^{\mathrm{ROBN}} = A_t D_t^\dagger A_t^\top m_t. \tag{29}$$

Define the active row index set

$$\mathcal{I}_t^{\mathrm{row}} = \{i \in \{1,\ldots,r\} : \|\widetilde{m}_{t,i:}\|_2 > 0\}.$$

Writing

$$A_t = [p_1(t), \ldots, p_r(t)], \qquad q_i(t)^\top = \frac{\widetilde{m}_{t,i:}}{\|\widetilde{m}_{t,i:}\|_2}, \qquad i \in \mathcal{I}_t^{\mathrm{row}},$$

gives

$$\Delta_t^{\mathrm{ROBN}} = \sum_{i\in\mathcal{I}_t^{\mathrm{row}}} p_i(t) q_i(t)^\top.$$

Hence, for all $i, j \in \mathcal{I}_t^{\mathrm{row}}$,

$$\langle p_i(t), p_j(t) \rangle = \delta_{ij}, \qquad \|q_i(t)\|_2 = 1.$$

Thus, the active ROBN orthonormal rank-one system is left-side one-sided orthogonal.

When $r < c$, ROBN is defined symmetrically along the column dimension. Let $A_t \in \mathbb{R}^{c \times c}$ be a randomly sampled orthogonal matrix and define

$$\widetilde{m}_t = m_t A_t^\top, \qquad D_t = \operatorname{diag}\left(\|\widetilde{m}_{t,:1}\|_2, \ldots, \|\widetilde{m}_{t,:c}\|_2\right).$$

Then

$$\Delta_t^{\mathrm{ROBN}} = m_t A_t^\top D_t^\dagger A_t. \tag{30}$$

The resulting active orthonormal rank-one system is right-side one-sided orthogonal; a derivation establishing this property is provided in Appendix F.1.

In implementation, a small $\varepsilon > 0$ is added to the normalization denominators for numerical stability.

We consider two variants of ROBN that differ in how the orthogonal matrix A_t is generated: ROBN-Haar uses a Haar-like random orthogonal matrix, intended to approximate uniform sampling over the orthogonal group, whereas ROBN-τ samples orthogonal matrices concentrated near the identity. Their sampling procedures and full algorithms are provided in Appendix F.

4.2 Split Polar Update

Split Polar Update (SPU) is designed to test whether a structured realization based on one-sided orthogonality can match or outperform Muon without requiring its two-sided singular-vector structure. To this end, given $m_t \in \mathbb{R}^{r \times c}$, SPU partitions m_t along its larger dimension, splitting it into two row blocks when $r \geq c$ and into two column blocks when $r < c$, and applies polar normalization independently to each block. For a matrix M with compact singular-value decomposition $M = U\Sigma V^\top$, we denote its polar factor by $\mathrm{Polar}(M) := UV^\top$. The resulting ideal two-block polar update coincides with a two-block special case of the BlockMuon construction analyzed in MuonBP (Khaled et al., 2026). In this work, however, the SPU construction arises from the one-sided orthogonality framework and is used to investigate the role of Muon's two-sided singular-vector structure.

The geometric characterization below is stated for this ideal blockwise polar normalization. In practice, each blockwise polar factor is approximated using a finite-step Newton–Schulz iteration.

When $r \geq c$ and r is even, split m_t along the row dimension:

$$m_t = \begin{bmatrix} m_t^{(1)} \\ m_t^{(2)} \end{bmatrix}, \qquad m_t^{(1)}, m_t^{(2)} \in \mathbb{R}^{(r/2) \times c}.$$

Apply polar normalization independently to the two blocks and concatenate the results:

$$\Delta_t^{\mathrm{SPU}} = \begin{bmatrix} \mathrm{Polar}(m_t^{(1)}) \\ \mathrm{Polar}(m_t^{(2)}) \end{bmatrix}.$$

When $r < c$ and c is even, we instead split along the column dimension:

$$m_t = \begin{bmatrix} m_t^{(1)} & m_t^{(2)} \end{bmatrix}, \qquad m_t^{(1)}, m_t^{(2)} \in \mathbb{R}^{r \times (c/2)}.$$

The ideal polar-normalized update is

$$\Delta_t^{\mathrm{SPU}} = \begin{bmatrix} \mathrm{Polar}(m_t^{(1)}) & \mathrm{Polar}(m_t^{(2)}) \end{bmatrix}.$$

Under row splitting, the blockwise left singular vectors are embedded into the full row space by zero-padding opposite halves of the rows and form the left factors of the combined system. Under column splitting, the blockwise right singular vectors are embedded into the full column space by zero-padding opposite halves of the columns and form the right factors of the combined system. Consequently, these embedded factors from different blocks are mutually orthogonal. Together with the orthonormality of the singular-vector factors within each block, the combined system forms an orthonormal rank-one system. The factors on the other side need not be mutually orthogonal. Therefore, row splitting yields left-sided one-sided orthogonality, whereas column splitting yields right-sided one-sided orthogonality.

The full SPU algorithm and an explicit derivation of the one-sided orthogonality of its ideal polar-normalization form are provided in Appendix G.

Table 1: Final validation performance over three random seeds (mean ± standard deviation). AdamW is included as a widely used reference point; the primary comparison for ROBN and SPU is Muon.

Dataset	Metric	AdamW	ROBN-Haar	ROBN-τ	Muon	SPU
ImageNet-1K	Acc. (%) ↑	79.217 ± 0.125	79.652 ± 0.171	80.187 ± 0.140	80.408 ± 0.057	$\mathbf{80.561 \pm 0.084}$
	Loss ↓	0.9429 ± 0.0021	0.9228 ± 0.0092	$\mathbf{0.8964 \pm 0.0059}$	0.9160 ± 0.0077	0.9078 ± 0.0038
FineWeb-Edu	Loss ↓	2.8248 ± 0.0031	2.8281 ± 0.0017	2.8104 ± 0.0014	2.7931 ± 0.0009	$\mathbf{2.7823 \pm 0.0014}$
	PPL ↓	16.858 ± 0.052	16.913 ± 0.029	16.616 ± 0.024	16.331 ± 0.014	$\mathbf{16.157 \pm 0.022}$
OpenWebText	Loss ↓	3.0224 ± 0.0022	3.0361 ± 0.0023	3.0208 ± 0.0017	2.9890 ± 0.0015	$\mathbf{2.9813 \pm 0.0033}$
	PPL ↓	20.541 ± 0.046	20.823 ± 0.048	20.507 ± 0.036	19.865 ± 0.029	$\mathbf{19.713 \pm 0.066}$

5 Experiments

Our experiments empirically examine the broader one-sided orthogonal geometry developed above. ROBN tests whether competitive performance can be retained without Muon's two-sided singular-vector structure, while SPU tests whether a different realization of one-sided orthogonality can even surpass Muon. Although the balanced SGD realization also belongs to the one-sided orthogonal family, it additionally imposes the equal-coefficient constraint described in Section 3.2; our empirical comparison of alternative realizations of one-sided orthogonality therefore focuses on ROBN and SPU, which do not impose this SGD-specific coefficient-balancing constraint.

We compare AdamW, Muon, ROBN-Haar, ROBN-τ, and SPU[1] across all experiments. For vision, we evaluate these methods on ImageNet-1K (Russakovsky et al., 2015) using a ViT-B/16-style model (Dosovitskiy et al., 2021; Touvron et al., 2021). For language modeling, we consider two settings. On FineWeb-Edu (Penedo et al., 2024), we use a GPT-style 210M model (Radford et al., 2019) trained for 10B tokens. On OpenWebText (Gokaslan & Cohen, 2019), we use GPT-2 Small (Radford et al., 2019) trained for 8B tokens. All final comparisons are reported over three random seeds.

The results show that ROBN retains a meaningful portion of Muon's empirical performance using only one-sided orthogonality, while SPU demonstrates that a realization of one-sided orthogonality can in some settings outperform Muon. These results place Muon as a strong member of the broader one-sided orthogonal family, rather than a uniquely special or best realization of that geometry, and do not support an independent or uniquely superior principle of two-sided orthogonality. Detailed experimental settings and hyperparameter-selection procedures are provided in Appendix H.

5.1 Results

Table 1 summarizes the final validation performance across ImageNet-1K, FineWeb-Edu, and OpenWebText over three random seeds.

Across the two language-modeling settings, the ordering is identical for both validation loss and perplexity: SPU performs best, followed by Muon, ROBN-τ, AdamW, and ROBN-Haar. On ImageNet-1K, final validation accuracy follows the ordering SPU, Muon, ROBN-τ, ROBN-Haar, and AdamW, while ROBN-τ achieves the lowest final validation loss. SPU also attains a lower final validation loss than Muon.

Taken together, the ROBN results indicate that one-sided orthogonality can retain a meaningful portion of Muon's empirical performance. In particular, ROBN-Haar remains competitive despite randomly resampling the orthogonal matrix A_t independently of the current momentum at every step, while the difference between ROBN-Haar and ROBN-τ shows that performance also depends on how a realization within the one-sided orthogonal family is constructed. SPU provides stronger evidence: it achieves better mean final validation performance than Muon on every reported metric across ImageNet-1K, FineWeb-Edu, and OpenWebText. These results support the view that a realization of one-sided orthogonality can match or outperform Muon without requiring Muon's two-sided singular-vector structure.

[1]ROBN-Haar, ROBN-τ, and SPU inherit Muon's hyperparameters without separate tuning; these hyperparameters are selected from the corresponding Muon sweep.

Additional best-validation results for ImageNet-1K and late-stage validation trajectories for all three settings are provided in Appendix H.4.

6 DISCUSSION

The orthonormal rank-one framework suggests a concrete design principle for normalized optimizers: rather than fixing a particular matrix geometry in advance, one can design an update by specifying how an admissible orthonormal rank-one system is selected and which additional structural constraints it satisfies. This perspective differs from norm- or LMO-based viewpoints, which often characterize an update after a feasible geometry or norm has been specified. In our framework, the selection rule itself is a design variable. Within the resulting one-sided orthogonal family, the balanced SGD realization and Muon's two-sided singular-vector geometry appear as structured special cases, while other realizations arise by changing how the orthonormal rank-one system is selected. ROBN illustrates this distinction most directly: its randomly resampled, momentum-independent orthogonal matrices are not naturally obtained by choosing a standard norm and taking its LMO. SPU can be interpreted retrospectively through a blockwise norm, but its split construction arises naturally from designing a different realization of one-sided orthogonality. Thus, the framework is not only a common interpretation of existing normalized updates, but also a constructive principle for designing new ones.

The ROBN experiments address whether Muon's momentum-adaptive singular-vector selection is necessary for effective normalized optimization. ROBN-Haar randomly resamples the orthogonal matrix A_t independently of the current momentum at every step, yet still retains a meaningful portion of Muon's empirical performance. This provides evidence that a meaningful portion of Muon's empirical performance is already reflected at the level of the broader one-sided orthogonality structure, rather than arising only from Muon's particular two-sided singular-vector realization.

ROBN-τ addresses a different question: whether the particular realization within the one-sided orthogonal family is irrelevant. The comparison with ROBN-Haar indicates otherwise. Both methods satisfy the same one-sided orthogonality condition, but ROBN-τ consistently outperforms ROBN-Haar across ImageNet-1K, FineWeb-Edu, and OpenWebText. Thus, performance also depends on how a realization within the one-sided orthogonal family is constructed. In ROBN, changing the distribution from which the orthogonal matrices are sampled changes optimization behavior even though the defining one-sided orthogonality property is preserved. This leaves the selection rule, including the distribution from which the orthogonal matrices are sampled, as an additional degree of freedom for optimizer design.

SPU addresses the final and more consequential question: whether this broader family should be viewed merely as a weaker relaxation of Muon's two-sided geometry. The experiments do not support such an interpretation. SPU is a different structured realization of one-sided orthogonality, yet it matches or outperforms Muon across the vision and language-modeling settings considered here, and achieves better mean final validation performance than Muon on every reported metric. Hence, the one-sided orthogonal family contains not only momentum-independent random realizations that remain competitive, but also structured realizations that can surpass Muon's two-sided construction.

Taken together, these results support one-sided orthogonality as a broader geometric principle for normalized optimizer design. ROBN shows that one-sided orthogonality can retain a meaningful portion of Muon's empirical performance without its two-sided singular-vector structure, while SPU shows that a different realization of one-sided orthogonality can match or outperform Muon. Accordingly, Muon's effectiveness is better understood through the broader principle of one-sided orthogonality than through a uniquely superior or independent principle of two-sided orthogonality.

Future work. The framework developed here suggests a broader direction for designing new normalized optimizers by constructing alternative selection rules over orthonormal rank-one systems. Beyond the specific ROBN and SPU realizations studied in this work, exploring new structured selection rules may lead to additional optimizer designs within the one-sided orthogonal family. The performance gap between ROBN-Haar and ROBN-τ also suggests that the distribution of the selected orthogonal matrices matters in practice. A natural next step is therefore to study how performance varies with τ, and more generally with how strongly the sampling distribution is concentrated around the identity.

AI USE STATEMENT

In this work, we used generative AI tools to assist in formulating mathematical claims, provide critical ingredients for proving mathematical claims, assist in the writing of proofs, provide feedback on research methodology and experimental design, and assist with translation. We did not use generative AI tools to help develop theoretical models or conceptual frameworks, propose or refine hypotheses, implement methods, or interpret results. The remaining required disclosure tasks, including synthetic dataset generation, dataset cleaning or reformatting, and qualitative or thematic data analysis, are not applicable to this work.

Additionally, we used generative AI tools to identify relevant literature, summarize or analyse existing literature, create or edit software code (for the vision and language-modeling experimental setups; the ROBN and SPU optimizer implementations were written by the authors), draft parts of the research paper, and edit the research paper to improve readability. We have reviewed all AI-assisted work. In particular, all AI-assisted mathematical claims, proof arguments, and proofs were independently reviewed and checked for correctness by the authors, and all AI-assisted code was manually reviewed and tested for correctness. We take responsibility for the final content of this work, including text, claims, and artifacts produced with the aid of generative AI.

REPRODUCIBILITY STATEMENT

To support reproducibility, the assumptions underlying the theoretical results and the corresponding proofs are provided in the appendices. Section 4 describes the constructions of ROBN and SPU, while Appendices F and G provide their full algorithms and implementation details. Appendix H provides the experimental configurations, model architectures, data processing and augmentation procedures, optimizer parameter groups, training schedules, hyperparameter sweeps and selection procedures, random seeds, and evaluation protocols used for the vision and language-modeling experiments.

REFERENCES

Kwangjun Ahn, Byron Xu, Natalie Abreu, Ying Fan, Gagik Magakyan, Pratyusha Sharma, Zheng Zhan, and John Langford. Dion: Distributed orthonormalized updates. *arXiv preprint arXiv:2504.05295*, 2025.

Noah Amsel, David Persson, Christopher Musco, and Robert M Gower. The polar express: Optimal matrix sign methods and their application to the muon algorithm. In *International Conference on Learning Representations*, volume 2026, pp. 138323–138360, 2026.

Da Chang, Qiankun Shi, Lvgang Zhang, Di He, Yaoshuai Ma, Ganzhao Yuan, and Yongxiang Liu. Muse: Representation geometry of muon beyond normalized momentum. *arXiv preprint arXiv:2607.14536*, 2026.

Lizhang Chen, Bo Liu, Kaizhao Liang, and Qiang Liu. Lion secretly solves a constrained optimization: As lyapunov predicts. In *International Conference on Learning Representations*, volume 2024, pp. 35404–35439, 2024.

Lizhang Chen, Jonathan Li, and qiang liu. Muon optimizes under spectral norm constraints. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=Blz4hjxLwU.

Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, et al. Symbolic discovery of optimization algorithms. *Advances in neural information processing systems*, 36:49205–49233, 2023.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=YicbFdNTTy.

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

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

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

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

Ahmed Khaled, Kaan Ozkara, Tao Yu, Mingyi Hong, and Youngsuk Park. Muonbp: Faster muon via block-periodic orthogonalization. In *International Conference on Learning Representations*, volume 2026, pp. 17389–17416, 2026.

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

Zichong Li, Liming Liu, Chen Liang, Weizhu Chen, and Tuo Zhao. Normuon: Making muon more efficient and scalable. *arXiv preprint arXiv:2510.05491*, 2025.

Hong Liu, Zhiyuan Li, David Hall, Percy Liang, and Tengyu Ma. Sophia: A scalable stochastic second-order optimizer for language model pre-training. In *International conference on learning representations*, volume 2024, pp. 1621–1650, 2024.

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

Ziyue Liu, Ruijie Zhang, Zhengyang Wang, Yequan Zhao, Yupeng Su, Zi Yang, and Zheng Zhang. Muon ^{2}: Boosting muon via adaptive second-moment preconditioning. *arXiv preprint arXiv:2604.09967*, 2026.

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

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

Guilherme Penedo, Hynek Kydlíček, Loubna Ben allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 30811–30849. Curran Associates, Inc., 2024. doi: 10.52202/079017-0970. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/370df50ccfdf8bde18f8f9c2d9151bda-Paper-Datasets_and_Benchmarks_Track.pdf.

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

B.T. Polyak. Some methods of speeding up the convergence of iteration methods. *USSR Computational Mathematics and Mathematical Physics*, 4(5):1–17, 1964. ISSN 0041-5553. doi: https://doi.org/10.1016/0041-5553(64)90137-5. URL https://www.sciencedirect.com/science/article/pii/0041555364901375.

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

Artem Riabinin, Egor Shulgin, Kaja Gruntkowska, and Peter Richtárik. From muon to gluon: Bridging theory and practice of LMO-based optimizers for LLMs. In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=IelAHU5MVz.

Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, Alexander C. Berg, and Li Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision (IJCV)*, 115(3):211–252, 2015. doi: 10.1007/s11263-015-0816-y.

Maria-Eleni Sfyraki and Jun-Kun Wang. Lions and muons: Optimization via stochastic frank-wolfe. *arXiv preprint arXiv:2506.04192*, 2025.

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

Zakhar Shumaylov, Nathaël Da Costa, Peter Zaika, Bálint Mucsányi, Alex Massucco, Yoav Gelberg, Carola-Bibiane Schönlieb, Yarin Gal, and Philipp Hennig. Muon is not that special: Random or inverted spectra work just as well. *arXiv preprint arXiv:2605.11181*, 2026.

Chongjie Si, Debing Zhang, and Wei Shen. Adamuon: Adaptive muon optimizer. *arXiv preprint arXiv:2507.11005*, 2025.

Ilya Sutskever, James Martens, George Dahl, and Geoffrey Hinton. On the importance of initialization and momentum in deep learning. In Sanjoy Dasgupta and David McAllester (eds.), *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pp. 1139–1147, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/sutskever13.html.

Tijmen Tieleman. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude. *COURSERA: Neural networks for machine learning*, 4(2):26, 2012.

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.

Nikhil Vyas, Depen Morwani, Rosie Zhao, Itai Shapira, David Brandfonbrener, Lucas Janson, and Sham Kakade. Soap: Improving and stabilizing shampoo using adam for language modeling. In *International Conference on Learning Representations*, volume 2025, pp. 93423–93444, 2025.

Ruijie Zhang, Yequan Zhao, Ziyue Liu, Zhengyang Wang, Dongyang Li, Yupeng Su, Sijia Liu, and Zheng Zhang. Teon: Tensorized orthonormalization beyond layer-wise muon for large language model pre-training. *arXiv preprint arXiv:2601.23261*, 2026.

A Static Examples and Balanced Rank-One Representations

A.1 Balanced Orthonormal Rank-One Representation of SGD with One-Sided Orthogonality

Let

$$0 \neq m \in \mathbb{R}^{p\times q}. \tag{31}$$

We show that, under a suitable choice of an orthonormal basis, the matrix m can be represented as an equal-coefficient sum of orthonormal rank-one matrices, and that the resulting representation is one-sided orthogonal.

Let

$$B = m^\top m \quad \text{or} \quad B = mm^\top. \tag{32}$$

Then B is symmetric positive semidefinite. Let

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_N \geq 0 \tag{33}$$

be the eigenvalues of B, counted with multiplicity, where N is the dimension of B. Suppose that

$$\operatorname{rank}(m) = \operatorname{rank}(B) \leq n \leq N. \tag{34}$$

Define

$$d = \left(\frac{\operatorname{tr}(B)}{n}, \ldots, \frac{\operatorname{tr}(B)}{n}, 0, \ldots, 0\right) \in \mathbb{R}^N, \tag{35}$$

where the first n entries are equal to $\operatorname{tr}(B)/n$.

Let

$$r = \operatorname{rank}(B) \leq n. \tag{36}$$

Since $B \succeq 0$, we have

$$\lambda_1 \geq \cdots \geq \lambda_r > 0, \qquad \lambda_{r+1} = \cdots = \lambda_N = 0. \tag{37}$$

For $1 \leq k \leq r$,

$$\sum_{i=1}^{k} \lambda_i \geq \frac{k}{r}\operatorname{tr}(B) \geq \frac{k}{n}\operatorname{tr}(B) = \sum_{i=1}^{k} d_i. \tag{38}$$

For $r < k \leq n$,

$$\sum_{i=1}^{k} \lambda_i = \operatorname{tr}(B) \geq \frac{k}{n}\operatorname{tr}(B) = \sum_{i=1}^{k} d_i, \tag{39}$$

and for $n < k \leq N$,

$$\sum_{i=1}^{k} \lambda_i = \sum_{i=1}^{k} d_i = \operatorname{tr}(B). \tag{40}$$

Therefore,

$$d \prec \lambda. \tag{41}$$

By the Schur–Horn theorem, there exists an orthogonal matrix Q such that

$$\operatorname{diag}(Q^\top B Q) = d. \tag{42}$$

Let

$$Q = \begin{bmatrix} u_1 & u_2 & \cdots & u_N \end{bmatrix}. \tag{43}$$

Then

$$u_i^\top B u_i = d_i. \tag{44}$$

Moreover,

$$\frac{\operatorname{tr}(B)}{n} = \frac{\|m\|_F^2}{n}. \tag{45}$$

Define

$$\alpha = \frac{\|m\|_F}{\sqrt{n}}. \tag{46}$$

Then, for every index i with $d_i \neq 0$, we have

$$\sqrt{d_i} = \alpha. \tag{47}$$

Let

$$I = \{i : d_i \neq 0\}. \tag{48}$$

Then $|I| = n$.

Case 1: $B = m^\top m$**.** For $i \in I$, we have

$$u_i^\top m^\top m u_i = d_i. \tag{49}$$

Hence

$$\|mu_i\|_2 = \sqrt{d_i} = \alpha. \tag{50}$$

Define

$$v_i = \frac{mu_i}{\alpha}. \tag{51}$$

Then

$$\|v_i\|_2 = 1. \tag{52}$$

Moreover,

$$mu_i = \alpha v_i. \tag{53}$$

Define

$$M_i = v_i u_i^\top, \qquad i \in I. \tag{54}$$

Then each M_i is rank one. Also,

$$\langle M_i, M_j \rangle_F = \mathrm{tr}(M_i^\top M_j) = \mathrm{tr}(u_i v_i^\top v_j u_j^\top) = (v_i^\top v_j)(u_j^\top u_i). \tag{55}$$

Since the vectors u_i are orthonormal, we obtain

$$\langle M_i, M_j \rangle_F = \delta_{ij}. \tag{56}$$

Therefore, $\{M_i\}_{i \in I}$ is an orthonormal rank-one system.

Furthermore,

$$\alpha \sum_{i \in I} M_i = \alpha \sum_{i \in I} v_i u_i^\top = \sum_{i \in I} m u_i u_i^\top. \tag{57}$$

If $i \notin I$, then $d_i = 0$, and hence

$$\|mu_i\|_2^2 = u_i^\top m^\top m u_i = u_i^\top B u_i = d_i = 0. \tag{58}$$

Thus

$$mu_i = 0 \qquad \text{for } i \notin I. \tag{59}$$

Therefore,

$$\sum_{i \in I} m u_i u_i^\top = \sum_{i=1}^{N} m u_i u_i^\top. \tag{60}$$

Since Q is orthogonal,

$$\sum_{i=1}^{N} u_i u_i^\top = QQ^\top = I. \tag{61}$$

Hence

$$\alpha \sum_{i \in I} M_i = m \sum_{i=1}^{N} u_i u_i^\top = mQQ^\top = m. \tag{62}$$

Thus,

$$m = \alpha \sum_{i \in I} M_i. \tag{63}$$

Moreover, the right factors $\{u_i\}_{i \in I}$ are orthonormal, while the left factors $\{v_i\}_{i \in I}$ have unit norm. Hence, this construction is right-sided one-sided orthogonal in the sense of Section 3.1.

Case 2: $B = mm^\top$**.** For $i \in I$, we have

$$u_i^\top m m^\top u_i = d_i. \tag{64}$$

Hence

$$\|m^\top u_i\|_2 = \sqrt{d_i} = \alpha. \tag{65}$$

Define

$$v_i = \frac{m^\top u_i}{\alpha}. \tag{66}$$

Then

$$\|v_i\|_2 = 1. \tag{67}$$

Moreover,

$$m^\top u_i = \alpha v_i. \tag{68}$$

Define

$$M_i = u_i v_i^\top, \qquad i \in I. \tag{69}$$

Then each M_i is rank one. Also,

$$\langle M_i, M_j \rangle_F = \mathrm{tr}(M_i^\top M_j) = \mathrm{tr}(v_i u_i^\top u_j v_j^\top) = (u_i^\top u_j)(v_j^\top v_i). \tag{70}$$

Since the vectors u_i are orthonormal, we obtain

$$\langle M_i, M_j \rangle_F = \delta_{ij}. \tag{71}$$

Therefore, $\{M_i\}_{i \in I}$ is an orthonormal rank-one system.

Furthermore,

$$\alpha \sum_{i \in I} M_i = \alpha \sum_{i \in I} u_i v_i^\top = \sum_{i \in I} u_i u_i^\top m. \tag{72}$$

If $i \notin I$, then $d_i = 0$, and hence

$$\|m^\top u_i\|_2^2 = u_i^\top m m^\top u_i = u_i^\top B u_i = d_i = 0. \tag{73}$$

Thus

$$m^\top u_i = 0 \qquad \text{for } i \notin I. \tag{74}$$

Therefore,

$$\sum_{i \in I} u_i u_i^\top m = \sum_{i=1}^{N} u_i u_i^\top m. \tag{75}$$

Since Q is orthogonal,

$$\sum_{i=1}^{N} u_i u_i^\top = QQ^\top = I. \tag{76}$$

Hence

$$\alpha \sum_{i \in I} M_i = \left(\sum_{i=1}^{N} u_i u_i^\top \right) m = QQ^\top m = m. \tag{77}$$

Thus,

$$m = \alpha \sum_{i \in I} M_i. \tag{78}$$

Moreover, the left factors $\{u_i\}_{i \in I}$ are orthonormal, while the right factors $\{v_i\}_{i \in I}$ have unit norm. Hence, this construction is left-sided one-sided orthogonal in the sense of Section 3.1.

Combining the two cases, we obtain an equal-coefficient orthonormal rank-one representation of the matrix-valued momentum:

$$m = \alpha \sum_{i \in I} M_i, \qquad \alpha = \frac{\|m\|_F}{\sqrt{n}}, \tag{79}$$

where

$$\langle M_i, M_j \rangle_F = \delta_{ij}, \qquad \mathrm{rank}(M_i) = 1. \tag{80}$$

Moreover, the right factor system is orthonormal in Case 1, whereas the left factor system is orthonormal in Case 2. Hence, m admits a balanced orthonormal rank-one representation that is one-sided orthogonal.

A.2 Two-Sided Orthogonality Characterizes SVD Geometry

We first show that every nonzero matrix admits an orthonormal rank-one system satisfying two-sided orthogonality; the zero matrix is trivial. Let

$$m = U\Sigma V^\top = \sum_{i=1}^{r} \sigma_i u_i v_i^\top \tag{81}$$

be a compact singular-value decomposition of m, where $r = \operatorname{rank}(m)$, $\sigma_i > 0$, and

$$\langle u_i, u_j \rangle = \delta_{ij}, \qquad \langle v_i, v_j \rangle = \delta_{ij}. \tag{82}$$

Defining

$$M_i := u_i v_i^\top, \qquad i = 1, \dots, r, \tag{83}$$

we have

$$\langle M_i, M_j \rangle_F = \left\langle u_i v_i^\top, u_j v_j^\top \right\rangle_F \tag{84}$$

$$= \langle u_i, u_j \rangle \langle v_i, v_j \rangle \tag{85}$$

$$= \delta_{ij}. \tag{86}$$

Hence every nonzero matrix admits an orthonormal rank-one system satisfying two-sided orthogonality, as provided by its singular-value decomposition, while the zero matrix is trivial.

We next show the converse: any orthonormal rank-one representation of m satisfying two-sided orthogonality is necessarily an SVD-type representation.

Suppose

$$m = \sum_i a_i p_i q_i^\top. \tag{87}$$

Let

$$I := \{i : a_i \neq 0\} \tag{88}$$

be the active index set, and assume that, for all $i, j \in I$,

$$\langle p_i, p_j \rangle = \delta_{ij}, \qquad \langle q_i, q_j \rangle = \delta_{ij}. \tag{89}$$

Define

$$M_i := p_i q_i^\top, \qquad i \in I. \tag{90}$$

Then

$$\langle M_i, M_j \rangle_F = \left\langle p_i q_i^\top, p_j q_j^\top \right\rangle_F \tag{91}$$

$$= \langle p_i, p_j \rangle \langle q_i, q_j \rangle \tag{92}$$

$$= \delta_{ij}, \tag{93}$$

so $\{M_i\}_{i \in I}$ is an orthonormal rank-one system.

For each $i \in I$, absorb the sign of a_i into p_i by defining

$$\widetilde{p}_i := \operatorname{sign}(a_i) p_i. \tag{94}$$

Then $\{\widetilde{p}_i\}_{i \in I}$ remains orthonormal. Define

$$\widetilde{P} := [\, \widetilde{p}_i \,]_{i \in I}, \qquad Q := [\, q_i \,]_{i \in I}. \tag{95}$$

Then

$$m = \sum_{i \in I} |a_i| \, \widetilde{p}_i q_i^\top. \tag{96}$$

For each $j \in I$,

$$m q_j = \sum_{i \in I} |a_i| \, \widetilde{p}_i q_i^\top q_j \tag{97}$$

$$= |a_j| \, \widetilde{p}_j, \tag{98}$$

and similarly,

$$m^\top \widetilde{p}_j = \sum_{i\in I} |a_i|\, q_i \widetilde{p}_i^\top \widetilde{p}_j \tag{99}$$

$$= |a_j|\, q_j. \tag{100}$$

Therefore, $\widetilde{p}_j$ and q_j form a left-right singular-vector pair of m with singular value $|a_j|$.

Moreover, since $\widetilde{P}$ and Q have orthonormal columns,

$$\widetilde{P}^\top m Q = \mathrm{diag}(|a_i|)_{i\in I}. \tag{101}$$

The diagonal matrix on the right is invertible, so

$$\mathrm{rank}(m) \geq |I|. \tag{102}$$

On the other hand, the factorization

$$m = \widetilde{P}\,\mathrm{diag}(|a_i|)_{i\in I} Q^\top \tag{103}$$

implies

$$\mathrm{rank}(m) \leq |I|. \tag{104}$$

Therefore,

$$\mathrm{rank}(m) = |I|. \tag{105}$$

Hence

$$m = \widetilde{P}\,\mathrm{diag}(|a_i|)_{i\in I} Q^\top \tag{106}$$

is a compact singular-value decomposition of m, up to the standard nonuniqueness associated with permutations, signs, and rotations within repeated singular-value subspaces.

Consequently, the nonzero singular values of m, counted with multiplicity, are precisely $|a_i|$ for $i \in I$, and therefore

$$\mathcal{K}(m) = \sum_i |a_i| = \|m\|_*. \tag{107}$$

Since

$$m = \widetilde{P}\,\mathrm{diag}(|a_i|)_{i\in I} Q^\top \tag{108}$$

is a compact singular-value decomposition, $\widetilde{P}Q^\top$ is a valid subgradient of the nuclear norm at m.

Moreover, the selected coefficient-wise subgradient is

$$\partial_m \mathcal{K}(m) := \sum_{i\in I} \mathrm{sign}(a_i) p_i q_i^\top \tag{109}$$

$$= \sum_{i\in I} \widetilde{p}_i q_i^\top \tag{110}$$

$$= \widetilde{P} Q^\top. \tag{111}$$

Thus, two-sided orthogonality necessarily induces the nuclear-norm geometry and the corresponding singular-vector normalized direction used by Muon.

A.3 Zero Time-Variation Contribution of Standard Examples

Throughout this subsection, $\partial_t \mathcal{K}_t(m)$ denotes the time-variation term associated with an arbitrary differentiable path through m, and

$$\frac{d}{dt} V_i(t, m)$$

denotes the corresponding total derivative of a selected orthonormal rank-one element, with m varying with velocity $\dot{m}$.

We first record a simple consequence of differentiating an orthonormal system when the selected system depends on both time and its argument. Let $\{V_i(t,m)\}_i$ be continuously differentiable in (t,m) on a neighborhood of the local variation considered below and satisfy

$$\langle V_i(t,m), V_j(t,m)\rangle = \delta_{ij}. \tag{112}$$

Consider simultaneous increments

$$t \mapsto t + \Delta t, \qquad m \mapsto m + \Delta m.$$

We decompose

$$\begin{aligned} V_i(t + \Delta t, m + \Delta m) = V_i(t, m) \\ + \Delta t \frac{V_i(t + \Delta t, m + \Delta m) - V_i(t, m + \Delta m)}{\Delta t} \\ + \Big(V_i(t, m + \Delta m) - V_i(t, m) \Big). \end{aligned} \tag{113}$$

Hence,

$$V_i(t + \Delta t, m + \Delta m) = V_i(t, m) + \Delta t \, R_i(\Delta t, \Delta m), \tag{114}$$

where

$$\begin{aligned} R_i(\Delta t, \Delta m) := \frac{V_i(t + \Delta t, m + \Delta m) - V_i(t, m + \Delta m)}{\Delta t} \\ + \frac{V_i(t, m + \Delta m) - V_i(t, m)}{\Delta t}. \end{aligned} \tag{115}$$

Since the system is orthonormal at every (t, m),

$$\langle V_i(t + \Delta t, m + \Delta m), V_j(t + \Delta t, m + \Delta m) \rangle = \delta_{ij}. \tag{116}$$

Substituting the expansion above gives

$$\begin{aligned} \delta_{ij} &= \langle V_i(t, m) + \Delta t \, R_i(\Delta t, \Delta m), V_j(t, m) + \Delta t \, R_j(\Delta t, \Delta m) \rangle \\ &= \langle V_i(t, m), V_j(t, m) \rangle \\ &\quad + \Delta t \langle R_i(\Delta t, \Delta m), V_j(t, m) \rangle \\ &\quad + \Delta t \langle V_i(t, m), R_j(\Delta t, \Delta m) \rangle \\ &\quad + (\Delta t)^2 \langle R_i(\Delta t, \Delta m), R_j(\Delta t, \Delta m) \rangle. \end{aligned} \tag{117}$$

Since

$$\langle V_i(t, m), V_j(t, m) \rangle = \delta_{ij}, \tag{118}$$

subtracting this term and dividing by Δt yields

$$\begin{aligned} 0 = \langle R_i(\Delta t, \Delta m), V_j(t, m) \rangle \\ + \langle V_i(t, m), R_j(\Delta t, \Delta m) \rangle \\ + \Delta t \langle R_i(\Delta t, \Delta m), R_j(\Delta t, \Delta m) \rangle. \end{aligned} \tag{119}$$

Now let

$$\Delta t \to 0, \qquad \Delta m \to 0, \qquad \frac{\Delta m}{\Delta t} \to \dot{m}.$$

For the first component of R_i,

$$\frac{V_i(t + \Delta t, m + \Delta m) - V_i(t, m + \Delta m)}{\Delta t} \longrightarrow \partial_t V_i(t, m). \tag{120}$$

For the second component, differentiability with respect to m gives

$$V_i(t, m + \Delta m) - V_i(t, m) = D_m V_i(t, m)[\Delta m] + o(\|\Delta m\|). \tag{121}$$

Therefore,

$$\begin{aligned} \frac{V_i(t, m + \Delta m) - V_i(t, m)}{\Delta t} &= D_m V_i(t, m) \left[\frac{\Delta m}{\Delta t} \right] + \frac{o(\|\Delta m\|)}{\Delta t} \\ &\longrightarrow D_m V_i(t, m)[\dot{m}], \end{aligned} \tag{122}$$

Therefore,

$$R_i(\Delta t, \Delta m) \longrightarrow \partial_t V_i(t, m) + D_m V_i(t, m)[\dot{m}]. \tag{123}$$

Equivalently, by the chain rule, with m understood to vary along the considered trajectory with velocity $\dot{m}$,

$$\frac{d}{dt}V_i(t,m) = \partial_t V_i(t,m) + D_m V_i(t,m)[\dot{m}]. \tag{124}$$

Consequently,

$$\left\langle \frac{d}{dt}V_i(t,m), V_j(t,m) \right\rangle + \left\langle V_i(t,m), \frac{d}{dt}V_j(t,m) \right\rangle = 0. \tag{125}$$

In particular, when $i = j$,

$$\left\langle \frac{d}{dt}V_i(t,m), V_i(t,m) \right\rangle = 0. \tag{126}$$

SGD. Suppose that

$$m = \alpha \sum_i M_i(t,m), \qquad \alpha > 0, \tag{127}$$

where

$$\langle M_i(t,m), M_j(t,m) \rangle_F = \delta_{ij}. \tag{128}$$

For the coefficient-wise potential,

$$\mathcal{K}_t(m) = \sum_i |a_i(t,m)|. \tag{129}$$

the time-variation term is

$$\partial_t \mathcal{K}_t(m) = \sum_i \operatorname{sign}(a_i(t,m)) \left\langle \frac{d}{dt}M_i(t,m), m \right\rangle_F. \tag{130}$$

Since

$$a_i(t,m) = \alpha > 0, \tag{131}$$

we have

$$\operatorname{sign}(a_i(t,m)) = 1. \tag{132}$$

Therefore,

$$\begin{aligned} \partial_t \mathcal{K}_t(m) &= \sum_i \left\langle \frac{d}{dt}M_i(t,m), m \right\rangle_F \\ &= \alpha \sum_i \sum_j \left\langle \frac{d}{dt}M_i(t,m), M_j(t,m) \right\rangle_F. \end{aligned} \tag{133}$$

By applying (125) to the selected orthonormal rank-one system $\{M_i(t,m)\}_i$, we have

$$\left\langle \frac{d}{dt}M_i(t,m), M_j(t,m) \right\rangle_F + \left\langle M_i(t,m), \frac{d}{dt}M_j(t,m) \right\rangle_F = 0. \tag{134}$$

Hence,

$$\begin{aligned} &\sum_i \sum_j \left\langle \frac{d}{dt}M_i(t,m), M_j(t,m) \right\rangle_F \\ &= -\sum_i \sum_j \left\langle M_i(t,m), \frac{d}{dt}M_j(t,m) \right\rangle_F \\ &= -\sum_j \sum_i \left\langle \frac{d}{dt}M_j(t,m), M_i(t,m) \right\rangle_F \\ &= -\sum_i \sum_j \left\langle \frac{d}{dt}M_i(t,m), M_j(t,m) \right\rangle_F. \end{aligned} \tag{135}$$

Therefore,

$$\sum_i \sum_j \left\langle \frac{d}{dt}M_i(t,m), M_j(t,m) \right\rangle_F = 0, \tag{136}$$

and hence

$$\partial_t \mathcal{K}_t(m) = 0. \tag{137}$$

Lion. For Lion, we consider the coordinate-wise orthonormal basis

$$\{M_i\}_i, \tag{138}$$

which is independent of both time and the momentum argument. An arbitrary argument m is written as

$$m = \sum_i a_i(t, m) M_i. \tag{139}$$

The coefficient-wise potential is

$$\mathcal{K}_t(m) = \sum_i |a_i(t, m)|. \tag{140}$$

Since the selected basis is fixed,

$$\frac{d}{dt} M_i = 0 \tag{141}$$

for every i. Therefore,

$$\begin{aligned} \partial_t \mathcal{K}_t(m) &= \sum_i \operatorname{sign}(a_i(t, m)) \left\langle \frac{d}{dt} M_i, m \right\rangle_F \\ &= 0. \end{aligned} \tag{142}$$

Hence,

$$\partial_t \mathcal{K}_t(m) = 0. \tag{143}$$

Muon. Suppose that

$$m = \sum_i \sigma_i(t, m) M_i(t, m), \qquad \sigma_i(t, m) > 0, \tag{144}$$

where

$$M_i(t, m) = u_i(t, m) v_i(t, m)^\top. \tag{145}$$

Assume

$$\langle u_i(t, m), u_j(t, m) \rangle = \delta_{ij}, \qquad \langle v_i(t, m), v_j(t, m) \rangle = \delta_{ij}. \tag{146}$$

Along the considered variation,

$$\begin{aligned} \frac{d}{dt} M_i(t, m) = \frac{d}{dt} u_i(t, m)\, v_i(t, m)^\top \\ + u_i(t, m) \left(\frac{d}{dt} v_i(t, m) \right)^\top. \end{aligned} \tag{147}$$

The time-variation term is

$$\partial_t \mathcal{K}_t(m) = \sum_i \operatorname{sign}(\sigma_i(t, m)) \left\langle \frac{d}{dt} M_i(t, m), m \right\rangle_F. \tag{148}$$

Since

$$\sigma_i(t, m) > 0, \tag{149}$$

we have

$$\operatorname{sign}(\sigma_i(t, m)) = 1. \tag{150}$$

Thus,

$$\partial_t \mathcal{K}_t(m) = \sum_i \left\langle \frac{d}{dt} M_i(t, m), m \right\rangle_F. \tag{151}$$

Substituting

$$m = \sum_j \sigma_j(t, m) M_j(t, m), \tag{152}$$

we obtain

$$\partial_t \mathcal{K}_t(m) = \sum_i \sum_j \sigma_j(t, m) \left\langle \frac{d}{dt} M_i(t, m), M_j(t, m) \right\rangle_F. \tag{153}$$

Since

$$M_j(t,m) = u_j(t,m)v_j(t,m)^\top, \tag{154}$$

we have

$$\begin{aligned}&\left\langle \frac{d}{dt}M_i(t,m), M_j(t,m)\right\rangle_F \\ &= \left\langle \frac{d}{dt}u_i(t,m), u_j(t,m)\right\rangle \langle v_i(t,m), v_j(t,m)\rangle \\ &\quad + \langle u_i(t,m), u_j(t,m)\rangle \left\langle \frac{d}{dt}v_i(t,m), v_j(t,m)\right\rangle.\end{aligned} \tag{155}$$

Therefore,

$$\partial_t \mathcal{K}_t(m) = \sum_i \sigma_i(t,m)\left[\left\langle \frac{d}{dt}u_i(t,m), u_i(t,m)\right\rangle + \left\langle \frac{d}{dt}v_i(t,m), v_i(t,m)\right\rangle\right]. \tag{156}$$

Applying (126) to the orthonormal families $\{u_i(t,m)\}_i$ and $\{v_i(t,m)\}_i$, respectively, gives

$$\left\langle \frac{d}{dt}u_i(t,m), u_i(t,m)\right\rangle = 0, \qquad \left\langle \frac{d}{dt}v_i(t,m), v_i(t,m)\right\rangle = 0. \tag{157}$$

Hence,

$$\partial_t \mathcal{K}_t(m) = 0. \tag{158}$$

A.4 Vanishing Time-Variation Terms of $\mathcal{K}_t$ and $\mathcal{K}_t^*$

The previous subsection shows the differential cancellation for the standard representations considered above wherever the selected representation is continuously differentiable:

$$\partial_t \mathcal{K}_t(m) = 0. \tag{159}$$

This time-independence is also seen directly from the closed-form potentials $\sqrt{n}\|m\|_F$, $\|m\|_1$, and $\|m\|_*$ for balanced SGD, Lion, and Muon, respectively.

Thus, for each representation separately, the induced potential is invariant with respect to t at every admissible m. Therefore, it can be identified with a fixed convex function $\mathcal{K}$:

$$\mathcal{K}_t(m) = \mathcal{K}(m) \qquad \text{for all admissible } m. \tag{160}$$

Consequently, its Fenchel conjugate is also independent of t:

$$\begin{aligned}\mathcal{K}_t^*(y) &= \sup_m \{\langle y, m\rangle_F - \mathcal{K}_t(m)\} \\ &= \sup_m \{\langle y, m\rangle_F - \mathcal{K}(m)\} \\ &= \mathcal{K}^*(y).\end{aligned} \tag{161}$$

Hence, on the effective domain of $\mathcal{K}^*$,

$$\partial_t \mathcal{K}_t^*(y) = 0. \tag{162}$$

Thus, for each of the standard representations considered above,

$$\partial_t \mathcal{K}_t(m) = 0, \qquad \partial_t \mathcal{K}_t^*(y) = 0, \qquad y \in \mathrm{dom}(\mathcal{K}^*). \tag{163}$$

A.5 Induced Constrained Problems for SGD, Lion, and Muon

The preceding discussion shows that, for the SGD, Lion, and Muon representations, the induced potential is time-independent. Although the selected orthonormal rank-one system may change as its argument varies for SGD and Muon, and hence may vary along an optimization trajectory, the corresponding time-variation terms of $\mathcal{K}_t$ and $\mathcal{K}_t^*$ vanish for all three representations.

Accordingly, each representation can be associated with a fixed convex potential $\mathcal{K}$. The choice of $\mathcal{K}$ determines the corresponding Fenchel conjugate $\mathcal{K}^*$ and, under decoupled weight decay, the associated constrained problem.

Here, for a set $\mathcal{C}$, we denote by $\chi_{\mathcal{C}}$ its indicator function, defined as

$$\chi_{\mathcal{C}}(y) = \begin{cases} 0, & y \in \mathcal{C}, \\ +\infty, & y \notin \mathcal{C}. \end{cases} \tag{164}$$

For the balanced SGD representation, the induced potential is proportional to the Frobenius norm:

$$\mathcal{K}_{\mathrm{SGD}}(m) = \sqrt{n}\|m\|_F. \tag{165}$$

Its Fenchel conjugate is the indicator function of a Frobenius-norm ball:

$$\mathcal{K}^*_{\mathrm{SGD}}(y) = \chi_{\{\|y\|_F \le \sqrt{n}\}}(y). \tag{166}$$

Thus, under decoupled weight decay, the corresponding constrained problem is

$$\min_X F(X) \qquad \text{s.t.} \qquad \|\lambda X\|_F \le \sqrt{n}. \tag{167}$$

Equivalently,

$$\|X\|_F \le \frac{\sqrt{n}}{\lambda}. \tag{168}$$

For Lion, the coordinate-wise sign structure corresponds to

$$\mathcal{K}_{\mathrm{Lion}}(m) = \|m\|_1, \qquad \mathrm{sign}(m) \in \partial \mathcal{K}_{\mathrm{Lion}}(m), \tag{169}$$

where the zero entries are selected according to $\mathrm{sign}(0) = 0$. Its Fenchel conjugate is the indicator function of an ℓ_∞-norm ball:

$$\mathcal{K}^*_{\mathrm{Lion}}(y) = \chi_{\{\|y\|_\infty \le 1\}}(y). \tag{170}$$

Therefore, the corresponding constrained problem is

$$\min_X F(X) \qquad \text{s.t.} \qquad \|\lambda X\|_\infty \le 1, \tag{171}$$

or equivalently,

$$\|X\|_\infty \le \frac{1}{\lambda}. \tag{172}$$

For Muon, the rank-one singular-vector representation corresponds to the nuclear norm:

$$\mathcal{K}_{\mathrm{Muon}}(m) = \|m\|_*. \tag{173}$$

Its Fenchel conjugate is the indicator function of the spectral-norm ball:

$$\mathcal{K}^*_{\mathrm{Muon}}(y) = \chi_{\{\|y\|_{\mathrm{op}} \le 1\}}(y). \tag{174}$$

Thus, the corresponding constrained problem is

$$\min_X F(X) \qquad \text{s.t.} \qquad \|\lambda X\|_{\mathrm{op}} \le 1, \tag{175}$$

or equivalently,

$$\|X\|_{\mathrm{op}} \le \frac{1}{\lambda}. \tag{176}$$

Consequently, SGD, Lion, and Muon admit the same potential–conjugate formulation, while the induced feasible geometry, and hence the corresponding KKT characterization, differ according to the choice of $\mathcal{K}$: Frobenius-norm constrained optimization for the balanced SGD representation, ℓ_∞-constrained optimization for Lion, and spectral-norm constrained optimization for Muon.

B Convergence Analysis for Non-Momentum Dynamics with Time-Dependent Potentials

We consider the non-momentum dynamics

$$W_{t+1} = W_t - \eta_t \xi_t, \qquad g_t = \nabla \mathcal{L}(W_t), \tag{177}$$

where ξ_t is generated from a possibly time-dependent potential $\mathcal{K}_t$. For each time t and argument g, let

$$\mathcal{B}_t(g) = \{M_i(t,g)\}_{i=1}^{r(t,g)}$$

denote the orthonormal rank-one system selected according to the corresponding selection rule. We assume that

$$g = \sum_{i=1}^{r(t,g)} \langle g, M_i(t,g)\rangle_F M_i(t,g), \tag{178}$$

with

$$\langle M_i(t,g), M_j(t,g)\rangle_F = \delta_{ij}, \qquad \operatorname{rank}(M_i(t,g)) = 1. \tag{179}$$

Since $\mathbb{R}^{m\times n}$ has dimension $N := mn$, every such orthonormal rank-one system satisfies

$$r(t,g) \le N. \tag{180}$$

The corresponding coefficient-wise potential is

$$\mathcal{K}_t(g) = \sum_{i=1}^{r(t,g)} |\langle g, M_i(t,g)\rangle_F|. \tag{181}$$

As in Section 3, we consider selection rules for which this construction defines a convex potential $\mathcal{K}_t$ and for which the signed combination below is a valid subgradient. At the current gradient g_t, define

$$I_t = \{i \in \{1,\dots,r(t,g_t)\} : \langle g_t, M_i(t,g_t)\rangle_F \neq 0\}, \qquad N_t = |I_t|. \tag{182}$$

Using the admissible subgradient selection specified in Section 3, we choose

$$\xi_t \in \partial \mathcal{K}_t(g_t) \tag{183}$$

as

$$\xi_t = \sum_{i\in I_t} \operatorname{sign}\left(\langle g_t, M_i(t,g_t)\rangle_F\right) M_i(t,g_t). \tag{184}$$

Since the coefficients outside I_t vanish, the selected representation of g_t can equivalently be written as

$$g_t = \sum_{i\in I_t} \langle g_t, M_i(t,g_t)\rangle_F M_i(t,g_t). \tag{185}$$

Hence,

$$\{M_i(t,g_t)\}_{i\in I_t}$$

is the active orthonormal rank-one system representing g_t. By construction,

$$\langle g_t, \xi_t\rangle_F = \left\langle g_t, \sum_{i\in I_t} \operatorname{sign}\left(\langle g_t, M_i(t,g_t)\rangle_F\right) M_i(t,g_t)\right\rangle_F \tag{186}$$

$$= \sum_{i\in I_t} |\langle g_t, M_i(t,g_t)\rangle_F|. \tag{187}$$

Also, since the elements $\{M_i(t,g_t)\}_{i\in I_t}$ are orthonormal,

$$\|\xi_t\|_F^2 = \left\| \sum_{i\in I_t} \operatorname{sign}\left(\langle g_t, M_i(t,g_t)\rangle_F\right) M_i(t,g_t)\right\|_F^2 \tag{188}$$

$$= \sum_{i\in I_t} 1 = N_t. \tag{189}$$

Define

$$\mu_t = \frac{1}{N_t} \sum_{i \in I_t} |\langle g_t, M_i(t, g_t) \rangle_F| \,, \tag{190}$$

whenever $N_t > 0$. Then

$$\langle g_t, \xi_t \rangle_F = N_t \mu_t, \qquad \|\xi_t\|_F^2 = N_t. \tag{191}$$

If $N_t = 0$, then the orthonormal representation above implies $g_t = 0$. In this case, $\xi_t = 0$, so $W_{t+1} = W_t$, and the sequence has already reached a first-order stationary point. Hence, it suffices to consider the case $N_t > 0$.

Assume that $\mathcal{L}$ is L-smooth. Then

$$\mathcal{L}(W_{t+1}) \leq \mathcal{L}(W_t) + \langle \nabla \mathcal{L}(W_t), W_{t+1} - W_t \rangle_F + \frac{L}{2} \|W_{t+1} - W_t\|_F^2 \tag{192}$$

$$= \mathcal{L}(W_t) - \eta_t \langle g_t, \xi_t \rangle_F + \frac{L \eta_t^2}{2} \|\xi_t\|_F^2 \tag{193}$$

$$= \mathcal{L}(W_t) - \eta_t N_t \mu_t + \frac{L \eta_t^2}{2} N_t \tag{194}$$

$$= \mathcal{L}(W_t) - \eta_t N_t \left(\mu_t - \frac{L \eta_t}{2} \right). \tag{195}$$

Therefore, if

$$0 < \eta_t < \frac{2 \mu_t}{L} \qquad \text{for every } t \text{ such that } N_t > 0, \tag{196}$$

then the loss decreases at step t. In particular, choosing

$$\eta_t = \theta_t \frac{\mu_t}{L}, \qquad 0 < \theta_{\min} \leq \theta_t \leq \theta_{\max} < 2, \tag{197}$$

gives

$$\mathcal{L}(W_{t+1}) \leq \mathcal{L}(W_t) - \frac{N_t \mu_t^2}{2L} (2 - \theta_t) \theta_t. \tag{198}$$

Define

$$D_t = \frac{N_t \mu_t^2}{2L} (2 - \theta_t) \theta_t. \tag{199}$$

Then

$$\mathcal{L}(W_{t+1}) \leq \mathcal{L}(W_t) - D_t. \tag{200}$$

Summing from $t = 0$ to $T - 1$, we obtain

$$\sum_{t=0}^{T-1} D_t \leq \mathcal{L}(W_0) - \mathcal{L}(W_T). \tag{201}$$

If $\mathcal{L}$ is bounded below by $\mathcal{L}(W^*)$, then

$$\sum_{t=0}^{T-1} D_t \leq \mathcal{L}(W_0) - \mathcal{L}(W^*) < \infty. \tag{202}$$

Hence,

$$\sum_{t=0}^{\infty} D_t < \infty, \qquad D_t \to 0. \tag{203}$$

It remains to relate D_t to the gradient norm. Since

$$\mu_t = \frac{1}{N_t} \sum_{i \in I_t} |\langle g_t, M_i(t, g_t) \rangle_F| \,,$$

we have

$$D_t = \frac{(2 - \theta_t) \theta_t}{2 L N_t} \left(\sum_{i \in I_t} |\langle g_t, M_i(t, g_t) \rangle_F| \right)^2. \tag{204}$$

By the inequality $(\sum_i |a_i|)^2 \geq \sum_i |a_i|^2$, applied to the coefficients

$$a_i = \langle g_t, M_i(t, g_t) \rangle_F,$$

we have

$$D_t \geq \frac{(2-\theta_t)\,\theta_t}{2LN_t} \sum_{i \in I_t} |\langle g_t, M_i(t, g_t) \rangle_F|^2 . \tag{205}$$

Because g_t is represented by the orthonormal system $\{M_i(t, g_t)\}_{i \in I_t}$, orthonormality gives

$$\sum_{i \in I_t} |\langle g_t, M_i(t, g_t) \rangle_F|^2 = \|g_t\|_F^2. \tag{206}$$

Therefore,

$$D_t \geq (2-\theta_t)\,\theta_t \frac{\|g_t\|_F^2}{2LN_t}. \tag{207}$$

Since an orthonormal family in $\mathbb{R}^{m \times n}$ contains at most $N = mn$ elements, we have $N_t \leq N$.

$$D_t \geq (2-\theta_t)\,\theta_t \frac{\|g_t\|_F^2}{2LN}. \tag{208}$$

Since

$$0 < \theta_{\min} \leq \theta_t \leq \theta_{\max} < 2,$$

there exists a constant

$$c_\theta := \min_{\theta \in [\theta_{\min}, \theta_{\max}]} \theta(2-\theta) > 0.$$

Hence

$$D_t \geq \frac{c_\theta}{2LN} \|g_t\|_F^2. \tag{209}$$

Because $D_t \to 0$, it follows that

$$\|g_t\|_F^2 \to 0. \tag{210}$$

Thus, for the non-momentum dynamics with the possibly time-dependent potential $\mathcal{K}_t$, the sequence is asymptotically first-order stationary:

$$\|\nabla \mathcal{L}(W_t)\|_F \to 0. \tag{211}$$

C Common-Domain Regime

Let

$$\mathcal{X} = \mathbb{R}^{d_1 \times d_2}, \qquad N := d_1 d_2,$$

equipped with the Frobenius inner product and norm. As in Section 3, for each time t and argument m, let

$$\mathcal{B}_t(m) = \{M_i(t, m)\}_{i=1}^{r(t,m)}$$

denote the selected orthonormal rank-one system representing m. Thus,

$$m = \sum_{i=1}^{r(t,m)} \langle m, M_i(t, m) \rangle_F M_i(t, m), \tag{212}$$

with

$$\langle M_i(t, m), M_j(t, m) \rangle_F = \delta_{ij}, \qquad \operatorname{rank}(M_i(t, m)) = 1. \tag{213}$$

Since the selected system is orthonormal in the N-dimensional space $\mathcal{X}$,

$$r(t, m) \leq N. \tag{214}$$

The corresponding coefficient-wise potential is

$$\mathcal{K}_t(m) := \sum_{i=1}^{r(t,m)} |\langle m, M_i(t, m) \rangle_F| . \tag{215}$$

For the analysis in this section, we consider selection rules for which, for each fixed t, the resulting potential $\mathcal{K}_t$ is proper, closed, convex, and positively homogeneous on $\mathcal{X}$. Define

$$\mathcal{C}_t := \{y \in \mathcal{X} : \langle y, m\rangle_F \le \mathcal{K}_t(m) \quad \text{for all } m \in \mathcal{X}\}. \tag{216}$$

We denote by $\chi_{\mathcal{C}_t}$ the indicator function of $\mathcal{C}_t$, defined by

$$\chi_{\mathcal{C}_t}(y) := \begin{cases} 0, & y \in \mathcal{C}_t, \\ +\infty, & y \notin \mathcal{C}_t. \end{cases} \tag{217}$$

We first characterize the Fenchel conjugate of $\mathcal{K}_t$. By definition,

$$\mathcal{K}_t^*(y) = \sup_{m\in\mathcal{X}} \{\langle y, m\rangle_F - \mathcal{K}_t(m)\}. \tag{218}$$

If $y \in \mathcal{C}_t$, then by the definition of $\mathcal{C}_t$,

$$\langle y, m\rangle_F \le \mathcal{K}_t(m), \qquad \forall m \in \mathcal{X}. \tag{219}$$

Therefore,

$$\langle y, m\rangle_F - \mathcal{K}_t(m) \le 0, \qquad \forall m \in \mathcal{X}. \tag{220}$$

Since equality is attained at $m = 0$,

$$\mathcal{K}_t^*(y) = 0. \tag{221}$$

Conversely, suppose $y \notin \mathcal{C}_t$. By the definition of $\mathcal{C}_t$, there exists $m_0 \in \mathcal{X}$ such that

$$\langle y, m_0\rangle_F > \mathcal{K}_t(m_0). \tag{222}$$

For any $s > 0$, since $m_0 \in \mathcal{X}$ and $\mathcal{X}$ is a vector space, we have $sm_0 \in \mathcal{X}$. Hence, using positive homogeneity of $\mathcal{K}_t$,

$$\langle y, sm_0\rangle_F - \mathcal{K}_t(sm_0) = s\left(\langle y, m_0\rangle_F - \mathcal{K}_t(m_0)\right) \tag{223}$$

$$\to +\infty \qquad \text{as } s \to \infty. \tag{224}$$

Since each sm_0 is admissible in the supremum defining $\mathcal{K}_t^*(y)$, it follows that

$$\mathcal{K}_t^*(y) = +\infty. \tag{225}$$

Combining the two cases,

$$\mathcal{K}_t^*(y) = \begin{cases} 0, & y \in \mathcal{C}_t, \\ +\infty, & y \notin \mathcal{C}_t. \end{cases} \tag{226}$$

Equivalently,

$$\mathcal{K}_t^* = \chi_{\mathcal{C}_t}. \tag{227}$$

We next establish time-uniform bounds on the dual feasible sets $\mathcal{C}_t$. For any $m \in \mathcal{X}$, define

$$a_i(t, m) := \langle m, M_i(t, m)\rangle_F. \tag{228}$$

Since the selected orthonormal rank-one system represents m, orthonormality gives

$$\|m\|_F^2 = \sum_{i=1}^{r(t,m)} |a_i(t, m)|^2. \tag{229}$$

Therefore,

$$\|m\|_F \le \sum_{i=1}^{r(t,m)} |a_i(t, m)| = \mathcal{K}_t(m). \tag{230}$$

Moreover, since each $M_i(t, m)$ has rank one and unit Frobenius norm,

$$\|M_i(t, m)\|_* = \|M_i(t, m)\|_F = 1. \tag{231}$$

Hence, using the representation of m and the triangle inequality for the nuclear norm,

$$\|m\|_* = \left\| \sum_{i=1}^{r(t,m)} a_i(t,m) M_i(t,m) \right\|_* \tag{232}$$

$$\leq \sum_{i=1}^{r(t,m)} |a_i(t,m)| \, \|M_i(t,m)\|_* \tag{233}$$

$$= \mathcal{K}_t(m). \tag{234}$$

On the other hand, by the Cauchy–Schwarz inequality,

$$\mathcal{K}_t(m) = \sum_{i=1}^{r(t,m)} |a_i(t,m)| \tag{235}$$

$$\leq \sqrt{r(t,m)} \left(\sum_{i=1}^{r(t,m)} |a_i(t,m)|^2 \right)^{1/2} \tag{236}$$

$$\leq \sqrt{N} \, \|m\|_F. \tag{237}$$

Hence,

$$\|m\|_F \leq \|m\|_* \leq \mathcal{K}_t(m) \leq \sqrt{N} \, \|m\|_F, \qquad \forall t, \ \forall m \in \mathcal{X}. \tag{238}$$

We now use (238) to obtain uniform bounds on the sets $\mathcal{C}_t$. Suppose first that $\|y\|_{\mathrm{op}} \leq 1$. For any $m \in \mathcal{X}$, let

$$m = \sum_{j=1}^{\mathrm{rank}(m)} \sigma_j u_j v_j^\top$$

be a singular value decomposition, where $\|u_j\|_2 = \|v_j\|_2 = 1$. For each j,

$$\langle y, u_j v_j^\top \rangle_F = \mathrm{tr}\big(y^\top u_j v_j^\top\big) \tag{239}$$

$$= v_j^\top y^\top u_j \tag{240}$$

$$= u_j^\top y v_j. \tag{241}$$

Therefore,

$$\left|\langle y, m\rangle_F\right| = \left|\left\langle y, \sum_{j=1}^{\operatorname{rank}(m)} \sigma_j u_j v_j^\top \right\rangle_F\right| \tag{242}$$

$$= \left|\sum_{j=1}^{\operatorname{rank}(m)} \sigma_j \left\langle y, u_j v_j^\top \right\rangle_F\right| \tag{243}$$

$$= \left|\sum_{j=1}^{\operatorname{rank}(m)} \sigma_j u_j^\top y v_j\right| \tag{244}$$

$$\le \sum_{j=1}^{\operatorname{rank}(m)} \sigma_j \left|u_j^\top y v_j\right| \tag{245}$$

$$\le \sum_{j=1}^{\operatorname{rank}(m)} \sigma_j \|u_j\|_2 \, \|y v_j\|_2 \tag{246}$$

$$\le \sum_{j=1}^{\operatorname{rank}(m)} \sigma_j \|u_j\|_2 \, \|y\|_{\mathrm{op}} \|v_j\|_2 \tag{247}$$

$$= \|y\|_{\mathrm{op}} \sum_{j=1}^{\operatorname{rank}(m)} \sigma_j \tag{248}$$

$$= \|y\|_{\mathrm{op}} \|m\|_*. \tag{249}$$

Therefore,

$$\langle y, m\rangle_F \le \left|\langle y, m\rangle_F\right| \tag{250}$$

$$\le \|y\|_{\mathrm{op}} \|m\|_* \tag{251}$$

$$\le \|m\|_* \tag{252}$$

$$\le \mathcal{K}_t(m). \tag{253}$$

Hence, by the definition of $\mathcal{C}_t$,

$$\|y\|_{\mathrm{op}} \le 1 \quad \Longrightarrow \quad y \in \mathcal{C}_t, \qquad \forall t \ge 0. \tag{254}$$

Conversely, suppose $y \in \mathcal{C}_t$. Then

$$\langle y, m\rangle_F \le \mathcal{K}_t(m) \le \sqrt{N}\, \|m\|_F, \qquad \forall m \in \mathcal{X}. \tag{255}$$

Taking

$$m = \begin{cases} y/\|y\|_F, & y \ne 0, \\ 0, & y = 0, \end{cases}$$

gives

$$\|y\|_F \le \sqrt{N}. \tag{256}$$

Therefore, for every $t \ge 0$,

$$\{y \in \mathcal{X} : \|y\|_{\mathrm{op}} \le 1\} \subseteq \mathcal{C}_t \subseteq \left\{y \in \mathcal{X} : \|y\|_F \le \sqrt{N}\right\}. \tag{257}$$

In particular, the spectral-norm unit ball is contained in every time-dependent dual feasible set $\mathcal{C}_t$. Therefore, it forms a common feasible region shared by all $\mathcal{C}_t$:

$$\{y \in \mathcal{X} : \|y\|_{\mathrm{op}} \le 1\} \subseteq \bigcap_{t \ge 0} \mathcal{C}_t. \tag{258}$$

Moreover, this framework-wide universal common region coincides with the dual feasible set induced by Muon. Indeed, for Muon,

$$\mathcal{K}_{\mathrm{Muon}}(m) = \|m\|_*, \tag{259}$$

and therefore

$$\mathcal{C}_{\mathrm{Muon}} = \{y \in \mathcal{X} : \|y\|_{\mathrm{op}} \le 1\}. \tag{260}$$

Thus, the spectral-norm unit ball identified above as the universal common feasible region of the admissible framework is exactly Muon's dual feasible set.

Fix $\lambda > 0$, and let $\{X_t\}_{t\ge 0}$ be the parameter trajectory generated by the momentum dynamics in Appendix D with this fixed value of λ. We impose the following sufficient condition for the common-domain regime:

$$\sup_{t\ge 0} \lambda \|X_t\|_{\mathrm{op}} \le 1. \tag{261}$$

This is a condition on the trajectory generated with the fixed parameter λ. Since the trajectory itself may depend on λ, boundedness alone does not justify choosing a smaller value of λ after obtaining a trajectory bound.

By (261), we have

$$\|\lambda X_t\|_{\mathrm{op}} \le 1, \qquad \forall t \ge 0.$$

Hence, (258) yields

$$\lambda X_t \in \bigcap_{s\ge 0} \mathcal{C}_s, \qquad \forall t \ge 0. \tag{262}$$

The norm condition (261) is sufficient for this common-domain membership; it is not asserted to be necessary.

Using (227), we obtain

$$\mathcal{K}_s^*(\lambda X_t) = 0, \qquad \forall s, t \ge 0. \tag{263}$$

In particular,

$$\mathcal{K}_t^*(\lambda X_{t+1}) = \mathcal{K}_{t+1}^*(\lambda X_{t+1}) = 0, \qquad \forall t \ge 0. \tag{264}$$

Therefore,

$$\mathcal{K}_{t+1}^*(\lambda X_{t+1}) - \mathcal{K}_t^*(\lambda X_{t+1}) = 0, \qquad \forall t \ge 0. \tag{265}$$

Thus, under the sufficient condition (261), the dual-potential variation term appearing in the subsequent momentum convergence analysis vanishes.

C.1 Static Potentials and Native Dual Feasible Sets

The common-domain construction above is useful for genuinely time-dependent potentials. When the induced potential is static, $\mathcal{K}_t \equiv \mathcal{K}$, the dual feasible set is also static, $\mathcal{C}_t \equiv \mathcal{C}$, and the common-domain condition reduces to

$$\lambda X_t \in \mathcal{C}, \qquad \forall t \ge 0. \tag{266}$$

For the static geometries considered in this work, $\mathcal{C}$ is exactly the native dual feasible set of the corresponding potential.

For balanced SGD, with the integer n fixed as in Section 3.2, the induced potential is

$$\mathcal{K}_{\mathrm{SGD}}(M) = \sqrt{n}\,\|M\|_F, \tag{267}$$

and hence

$$\mathcal{C}_{\mathrm{SGD}} = \left\{Y \in \mathcal{X} : \|Y\|_F \le \sqrt{n}\right\}. \tag{268}$$

For Lion,

$$\mathcal{K}_{\mathrm{Lion}}(M) = \|M\|_1, \tag{269}$$

so that

$$\mathcal{C}_{\mathrm{Lion}} = \{Y \in \mathcal{X} : \|Y\|_\infty \le 1\}. \tag{270}$$

For Muon,

$$\mathcal{K}_{\mathrm{Muon}}(M) = \|M\|_*, \tag{271}$$

and therefore

$$\mathcal{C}_{\mathrm{Muon}} = \{Y \in \mathcal{X} : \|Y\|_{\mathrm{op}} \le 1\}. \tag{272}$$

For SPU, the architectural partition is fixed, so the induced geometry is also static. Under a row split, writing

$$M = \begin{bmatrix} M^{(1)} \\ M^{(2)} \end{bmatrix},$$

the induced potential is

$$\mathcal{K}^{\mathrm{row}}_{\mathrm{SPU}}(M) = \|M^{(1)}\|_* + \|M^{(2)}\|_*. \tag{273}$$

Accordingly, writing

$$Y = \begin{bmatrix} Y^{(1)} \\ Y^{(2)} \end{bmatrix}$$

according to the same partition used by SPU, the corresponding native dual feasible set is

$$\mathcal{C}^{\mathrm{row}}_{\mathrm{SPU}} = \left\{ Y \in \mathcal{X} : \|Y^{(1)}\|_{\mathrm{op}} \le 1, \quad \|Y^{(2)}\|_{\mathrm{op}} \le 1 \right\}. \tag{274}$$

Under a column split, writing

$$M = \begin{bmatrix} M^{(1)} & M^{(2)} \end{bmatrix},$$

the induced potential is

$$\mathcal{K}^{\mathrm{col}}_{\mathrm{SPU}}(M) = \|M^{(1)}\|_* + \|M^{(2)}\|_*. \tag{275}$$

Writing

$$Y = \begin{bmatrix} Y^{(1)} & Y^{(2)} \end{bmatrix},$$

the corresponding native dual feasible set is

$$\mathcal{C}^{\mathrm{col}}_{\mathrm{SPU}} = \left\{ Y \in \mathcal{X} : \|Y^{(1)}\|_{\mathrm{op}} \le 1, \quad \|Y^{(2)}\|_{\mathrm{op}} \le 1 \right\}. \tag{276}$$

Because the coefficient-wise potential decomposes across the fixed blocks, nuclear-norm/spectral-norm duality applies independently to each block. Thus, the SPU dual geometry is given by the product of the spectral-norm unit balls associated with its fixed blocks.

We next show that, for a static potential, its native dual feasible set is forward invariant under the parameter dynamics. Let

$$U_t := \nabla\mathcal{K}(\widetilde{M}_{t+1}) \in \partial\mathcal{K}(\widetilde{M}_{t+1}) \tag{277}$$

denote the selected update subgradient.

Recall that, for a proper, closed, convex, positively homogeneous degree-one potential,

$$\mathcal{K}^* = \chi_{\mathcal{C}}, \tag{278}$$

where $\mathcal{C}$ is its dual feasible set. Equivalently,

$$\mathcal{K}^*(Y) = \begin{cases} 0, & Y \in \mathcal{C}, \\ +\infty, & Y \notin \mathcal{C}. \end{cases} \tag{279}$$

Thus,

$$\operatorname{dom}\mathcal{K}^* = \mathcal{C}. \tag{280}$$

We first verify that the selected update subgradient U_t belongs to $\mathcal{C}$. Since

$$U_t \in \partial\mathcal{K}(\widetilde{M}_{t+1}), \tag{281}$$

the Fenchel–Young equality gives

$$\mathcal{K}(\widetilde{M}_{t+1}) + \mathcal{K}^*(U_t) = \left\langle U_t, \widetilde{M}_{t+1} \right\rangle_F. \tag{282}$$

The right-hand side of (282) is finite, and $\mathcal{K}(\widetilde{M}_{t+1})$ is finite for the static potentials considered here. Therefore,

$$\mathcal{K}^*(U_t) < \infty. \tag{283}$$

Since $\operatorname{dom} \mathcal{K}^* = \mathcal{C}$, it follows that

$$U_t \in \mathcal{C}. \tag{284}$$

Thus, for a static potential, the selected update subgradient always belongs to the corresponding native dual feasible set.

We now show that this property makes $\mathcal{C}$ invariant under the parameter dynamics. For a static potential, the parameter update in Appendix D becomes

$$X_{t+1} = X_t + \eta_t \left(U_t - \lambda X_{t+1}\right). \tag{285}$$

Moving the implicit decay term to the left-hand side gives

$$\left(1 + \lambda\eta_t\right) X_{t+1} = X_t + \eta_t U_t. \tag{286}$$

Since $\lambda > 0$ and $\eta_t > 0$, we may divide by $1 + \lambda\eta_t$ and multiply by λ, yielding

$$\lambda X_{t+1} = \frac{1}{1 + \lambda\eta_t} \lambda X_t + \frac{\lambda\eta_t}{1 + \lambda\eta_t} U_t. \tag{287}$$

Define

$$\alpha_t := \frac{1}{1 + \lambda\eta_t}. \tag{288}$$

Then

$$1 - \alpha_t = \frac{\lambda\eta_t}{1 + \lambda\eta_t}. \tag{289}$$

Because $\lambda > 0$ and $\eta_t > 0$,

$$0 < \alpha_t < 1, \qquad 0 < 1 - \alpha_t < 1, \tag{290}$$

and the two coefficients satisfy

$$\alpha_t + (1 - \alpha_t) = 1. \tag{291}$$

Hence, (287) can be written as

$$\lambda X_{t+1} = \alpha_t \lambda X_t + (1 - \alpha_t) U_t. \tag{292}$$

Therefore, λX_{t+1} is a convex combination of λX_t and U_t.

By its definition as the intersection of the closed half-spaces

$$\left\{Y \in \mathcal{X} : \langle Y, M\rangle_F \le \mathcal{K}(M)\right\}, \qquad M \in \mathcal{X},$$

the set $\mathcal{C}$ is closed and convex. Suppose that

$$\lambda X_t \in \mathcal{C}. \tag{293}$$

By (284), we also have

$$U_t \in \mathcal{C}. \tag{294}$$

Since both points belong to $\mathcal{C}$, convexity of $\mathcal{C}$ and the convex-combination representation above imply

$$\lambda X_{t+1} = \alpha_t \lambda X_t + (1 - \alpha_t) U_t \in \mathcal{C}. \tag{295}$$

We have therefore established the one-step implication

$$\lambda X_t \in \mathcal{C} \quad \Longrightarrow \quad \lambda X_{t+1} \in \mathcal{C}. \tag{296}$$

Applying this implication inductively gives

$$\lambda X_0 \in \mathcal{C} \quad \Longrightarrow \quad \lambda X_t \in \mathcal{C}, \qquad \forall t \ge 0. \tag{297}$$

Therefore, for balanced SGD, Lion, Muon, and SPU with its fixed architectural partition, the common-domain condition used in Appendix D does not require a separate trajectory-wide assumption once the corresponding static native dual feasible set is used. In fact, under the stated parameter dynamics, this common-domain condition is equivalent to

$$\lambda X_0 \in \mathcal{C}, \tag{298}$$

because (297) guarantees

$$\lambda X_t \in \mathcal{C}, \qquad \forall t \ge 0, \tag{299}$$

while the reverse implication follows immediately by taking $t = 0$.

Consequently, since $\mathcal{K}_t = \mathcal{K}_{t+1} = \mathcal{K}$ in the static case, we also have

$$\mathcal{K}_t^* = \mathcal{K}_{t+1}^* = \mathcal{K}^* = \chi_{\mathcal{C}}. \tag{300}$$

Together with $\lambda X_{t+1} \in \mathcal{C}$, this gives

$$\mathcal{K}_t^*(\lambda X_{t+1}) = \mathcal{K}_{t+1}^*(\lambda X_{t+1}) = 0, \qquad \forall t \geq 0. \tag{301}$$

Hence, the dual-potential variation term appearing in the momentum convergence analysis vanishes identically:

$$\mathcal{K}_{t+1}^*(\lambda X_{t+1}) - \mathcal{K}_t^*(\lambda X_{t+1}) = 0, \qquad \forall t \geq 0. \tag{302}$$

This static argument is distinct from the genuinely time-dependent case. When $\mathcal{K}_t$ varies with t, the selected update subgradient at time t belongs to $\mathcal{C}_t$, but this alone does not imply membership in $\mathcal{C}_s$ for other times s. The framework-wide common-domain condition developed above is therefore retained for the general time-dependent setting.

D Convergence Analysis for Momentum Dynamics with Time-Dependent Potentials

Our analysis builds on the discrete-time Lyapunov construction for Lion-$\mathcal{K}$ of Chen et al. (2024) and its matrix-valued formulation in Chen et al. (2026). The Lyapunov function in (329) adapts this construction to the time-dependent potential $\mathcal{K}_t$. Under the assumptions stated below, we additionally control the discrepancy between the frozen and actual Lyapunov values in (371) to account for changes in the potential.

We consider the deterministic discrete-time momentum dynamics with time-dependent potential $\mathcal{K}_t$:

$$M_{t+1} = \beta_2 M_t - (1-\beta_2)\nabla\mathcal{F}(X_t), \tag{303}$$

$$\widetilde{M}_{t+1} = \beta_1 M_t - (1-\beta_1)\nabla\mathcal{F}(X_t), \tag{304}$$

$$X_{t+1} = X_t + \eta_t \left(\nabla\mathcal{K}_t(\widetilde{M}_{t+1}) - \lambda X_{t+1}\right). \tag{305}$$

We work on $\mathcal{X} = \mathbb{R}^{d_1 \times d_2}$, equipped with the Frobenius inner product and norm, and write $N := d_1 d_2$. All inner products below are Frobenius inner products; the subscript F is omitted when no confusion can arise. The initial states $X_0, M_0 \in \mathcal{X}$ are finite. We assume that $\mathcal{F} : \mathcal{X} \to \mathbb{R}$ is differentiable and L-smooth, and that

$$\mathcal{F}^\star := \inf_{X \in \mathcal{X}} \mathcal{F}(X) > -\infty. \tag{306}$$

Convexity of $\mathcal{F}$ is not required.

The parameters are fixed and satisfy

$$0 \leq \beta_1 < \beta_2 < 1, \qquad \lambda > 0, \qquad \eta_t > 0, \qquad \forall t \geq 0. \tag{307}$$

The learning-rate schedule satisfies

$$\eta_{t+1} \leq \eta_t, \qquad \forall t \geq 0, \tag{308}$$

and

$$\sum_{t=0}^{\infty} \eta_t = \infty, \qquad \sum_{t=0}^{\infty} \eta_t^2 < \infty. \tag{309}$$

In particular, square summability implies $\eta_t \to 0$.

We consider the admissible selection rules specified in Section 3. As in Appendix C, we restrict attention to rules for which, at each fixed t, the resulting potential $\mathcal{K}_t$ is proper, closed, convex, and positively homogeneous of degree one. The update direction $\nabla\mathcal{K}_t(\widetilde{M}_{t+1})$ denotes the selected subgradient specified by these rules:

$$\nabla\mathcal{K}_t(\widetilde{M}_{t+1}) \in \partial\mathcal{K}_t(\widetilde{M}_{t+1}).$$

We use the uniform potential bounds and conjugate representation established in Appendix C, in particular (238), (227), and (257).

For the trajectory generated with the fixed value of λ, we impose the common-domain condition

$$\lambda X_t \in \mathcal{C}_s, \qquad \forall s, t \geq 0. \tag{310}$$

For static potentials $\mathcal{K}_t \equiv \mathcal{K}$, condition (310) is equivalent, under the stated parameter dynamics, to $\lambda X_0 \in \mathcal{C}$, where $\mathcal{C} = \operatorname{dom} \mathcal{K}^*$; see Appendix C.1. As shown in Appendix C, the condition (261) is sufficient for (310). By the conjugate representation (227), this implies

$$\mathcal{K}_s^*(\lambda X_t) = 0, \qquad \forall s, t \geq 0. \tag{311}$$

We also impose two trajectory-level conditions on the Fenchel–Young gaps. Define

$$G_t^{(t)} := \mathcal{K}_t(\widetilde{M}_{t+1}) + \mathcal{K}_t^*(\lambda X_{t+1}) - \langle \lambda X_{t+1}, \widetilde{M}_{t+1} \rangle_F, \qquad t \geq 0, \tag{312}$$

and let

$$Z_t := (X_t, M_t, \widetilde{M}_t), \qquad t \geq 1.$$

First, we assume the following no-spike condition: for every $\varepsilon > 0$, there exist $\rho > 0$ and an integer $T_\varepsilon \geq 0$ such that, for all integers $s \geq t \geq T_\varepsilon$ satisfying

$$\sum_{r=t}^{s-1} \eta_r \leq \rho, \tag{313}$$

one has

$$|G_s^{(s)} - G_t^{(t)}| \leq \varepsilon. \tag{314}$$

Second, we assume that there exists a single proper, lower-semicontinuous, convex, positively homogeneous potential $\mathcal{K}_\infty$, common to all convergent full-state subsequences, with the following property. Define

$$G_\infty(X, \widetilde{M}) := \mathcal{K}_\infty(\widetilde{M}) + \mathcal{K}_\infty^*(\lambda X) - \langle \lambda X, \widetilde{M} \rangle_F. \tag{315}$$

For every sequence of indices $t_j \to \infty$ such that

$$Z_{t_j+1} \to (X_\infty, M_\infty, \widetilde{M}_\infty),$$

we assume that

$$G_{t_j}^{(t_j)} \to G_\infty(X_\infty, \widetilde{M}_\infty). \tag{316}$$

We write this symbolically as

$$\mathcal{K}_t \to \mathcal{K}_\infty. \tag{317}$$

Here this notation refers specifically to the Fenchel–Young gap convergence in (316). It does not assert pointwise or uniform convergence of the potentials or their conjugates. The no-spike and limiting-gap conditions are explicit trajectory-level assumptions.

Under the dynamics (303)–(305) and the preceding assumptions on the initial states, objective, parameters, learning-rate schedule, potential, and selected subgradient, when $\mathcal{K}_t \equiv \mathcal{K}$ and $\lambda X_0 \in \mathcal{C} := \operatorname{dom} \mathcal{K}^*$, Appendix C.1 guarantees the common-domain condition (310). Both gap conditions then hold with $\mathcal{K}_\infty = \mathcal{K}$. The limiting-gap condition follows from the sufficient condition given below, while the no-spike condition is verified in Appendix D.2.

Consequently, under the static conditions just stated, the convergence conclusion established below, (462), applies: $d(X_t, \mathrm{KKT}) \to 0$, where KKT is defined in (454) with $\mathcal{K}_\infty = \mathcal{K}$. This static specialization applies to balanced SGD, Lion, Muon, and SPU with a fixed architectural partition, using their respective native dual feasible sets in Appendix C.1.

A sufficient condition for (316), under the common-domain condition, is that $\mathcal{K}_t$ converges uniformly on bounded sets to a finite-valued potential $\mathcal{K}_\infty$. Indeed, consider any convergent full-state subsequence

$$Z_{t_j+1} \to (X_\infty, M_\infty, \widetilde{M}_\infty).$$

Since $\mathcal{K}_\infty$ is finite-valued and convex, it is continuous. Uniform convergence on bounded sets therefore gives

$$\mathcal{K}_{t_j}(\widetilde{M}_{t_j+1}) \to \mathcal{K}_\infty(\widetilde{M}_\infty).$$

Moreover, the common-domain condition implies $\mathcal{K}^*_{t_j}(\lambda X_{t_j+1}) = 0$. By the Fenchel–Young inequality,

$$\langle \lambda X_{t_j+1}, M\rangle_F \le \mathcal{K}_{t_j}(M), \qquad \forall M \in \mathcal{X}.$$

Passing to the limit for each fixed M, we obtain

$$\langle \lambda X_\infty, M\rangle_F \le \mathcal{K}_\infty(M), \qquad \forall M \in \mathcal{X}.$$

Since $\mathcal{K}_\infty$ is finite-valued and positively homogeneous of degree one, $\mathcal{K}_\infty(0) = 0$. The definition of the Fenchel conjugate therefore yields

$$\mathcal{K}^*_\infty(\lambda X_\infty) = \sup_{M\in\mathcal{X}} \{\langle \lambda X_\infty, M\rangle_F - \mathcal{K}_\infty(M)\} = 0.$$

Together with continuity of the inner product, these observations establish (316). This is a sufficient condition for the limiting-gap assumption; it is not asserted to hold for every admissible selection rule.

Before introducing the Lyapunov function, we establish uniform bounds on the generated states.

By the common-domain condition (310), we have $\lambda X_t \in \mathcal{C}_t$ for every $t \ge 0$. The uniform outer bound in (257) therefore gives

$$\lambda\|X_t\|_F \le \sqrt{N}, \qquad \forall t \ge 0. \tag{318}$$

Define $B_X := \sqrt{N}/\lambda < \infty$. Then

$$\|X_t\|_F \le B_X, \qquad \forall t \ge 0. \tag{319}$$

Since $\mathcal{F}$ is L-smooth, its gradient is L-Lipschitz continuous; that is,

$$\|\nabla\mathcal{F}(X) - \nabla\mathcal{F}(Y)\|_F \le L\|X - Y\|_F, \qquad \text{for all } X, Y \in \mathcal{X}.$$

Taking $X = X_t$ and $Y = 0$, we obtain

$$\|\nabla\mathcal{F}(X_t) - \nabla\mathcal{F}(0)\|_F \le L\|X_t\|_F.$$

Thus, by the triangle inequality and (319),

$$\begin{aligned}\|\nabla\mathcal{F}(X_t)\|_F &= \big\|\nabla\mathcal{F}(0) + \big(\nabla\mathcal{F}(X_t) - \nabla\mathcal{F}(0)\big)\big\|_F \\ &\le \|\nabla\mathcal{F}(0)\|_F + \|\nabla\mathcal{F}(X_t) - \nabla\mathcal{F}(0)\|_F \\ &\le \|\nabla\mathcal{F}(0)\|_F + L\|X_t\|_F \\ &\le \|\nabla\mathcal{F}(0)\|_F + LB_X. \end{aligned} \tag{320}$$

Define

$$B_\nabla := \|\nabla\mathcal{F}(0)\|_F + LB_X.$$

Since $\mathcal{F}$ is differentiable at 0, $\|\nabla\mathcal{F}(0)\|_F$ is finite. Together with $B_X < \infty$, this gives $B_\nabla < \infty$. Therefore,

$$\|\nabla\mathcal{F}(X_t)\|_F \le B_\nabla, \qquad \forall t \ge 0. \tag{321}$$

We next establish a uniform bound on the momentum. For $t \ge 1$, repeatedly substituting the momentum recursion

$$M_{t+1} = \beta_2 M_t - (1-\beta_2)\nabla\mathcal{F}(X_t)$$

gives

$$M_t = \beta_2^t M_0 - (1-\beta_2)\sum_{k=0}^{t-1} \beta_2^{t-1-k}\nabla\mathcal{F}(X_k).$$

Taking Frobenius norms and applying the triangle inequality, we obtain

$$\begin{aligned}\|M_t\|_F &\le \beta_2^t\|M_0\|_F + (1-\beta_2)\sum_{k=0}^{t-1}\beta_2^{t-1-k}\|\nabla\mathcal{F}(X_k)\|_F \\ &\le \beta_2^t\|M_0\|_F + (1-\beta_2)B_\nabla\sum_{k=0}^{t-1}\beta_2^{t-1-k}. \end{aligned} \tag{322}$$

Here the second inequality follows from (321), applied to each $k = 0, \dots, t-1$.

Since $0 \le \beta_2 < 1$, the geometric-series identity gives

$$\sum_{k=0}^{t-1} \beta_2^{t-1-k} = \sum_{j=0}^{t-1} \beta_2^j = \frac{1-\beta_2^t}{1-\beta_2}.$$

Consequently,

$$(1-\beta_2) \sum_{k=0}^{t-1} \beta_2^{t-1-k} = 1 - \beta_2^t.$$

Substituting this identity into the preceding momentum bound, we obtain

$$\|M_t\|_F \le \beta_2^t \|M_0\|_F + (1-\beta_2^t) B_\nabla.$$

Because $0 \le \beta_2^t \le 1$, both coefficients on the right-hand side lie in $[0, 1]$. Hence,

$$\beta_2^t \|M_0\|_F + (1-\beta_2^t) B_\nabla \le \|M_0\|_F + B_\nabla.$$

Define

$$B_M := \|M_0\|_F + B_\nabla < \infty.$$

It follows that

$$\|M_t\|_F \le B_M, \qquad \forall t \ge 1.$$

For $t = 0$, the same bound holds because

$$\|M_0\|_F \le \|M_0\|_F + B_\nabla = B_M.$$

Thus,

$$\|M_t\|_F \le B_M, \qquad \forall t \ge 0. \tag{323}$$

Hence, $\{M_t\}_{t \ge 0}$ is bounded.

We now bound the filtered momentum. By (304),

$$\widetilde{M}_{t+1} = \beta_1 M_t - (1-\beta_1) \nabla \mathcal{F}(X_t).$$

Since $0 \le \beta_1 < 1$, taking Frobenius norms and applying the triangle inequality yields

$$\|\widetilde{M}_{t+1}\|_F \le \beta_1 \|M_t\|_F + (1-\beta_1) \|\nabla \mathcal{F}(X_t)\|_F.$$

Using (323) and (321), we obtain

$$\|\widetilde{M}_{t+1}\|_F \le \beta_1 B_M + (1-\beta_1) B_\nabla.$$

By the definition of B_M,

$$B_\nabla \le \|M_0\|_F + B_\nabla = B_M.$$

Therefore,

$$\begin{aligned} \|\widetilde{M}_{t+1}\|_F &\le \beta_1 B_M + (1-\beta_1) B_\nabla \\ &\le \beta_1 B_M + (1-\beta_1) B_M \\ &= B_M. \end{aligned} \tag{324}$$

Thus,

$$\|\widetilde{M}_{t+1}\|_F \le B_M, \qquad \forall t \ge 0. \tag{325}$$

Equivalently,

$$\|\widetilde{M}_t\|_F \le B_M, \qquad \forall t \ge 1.$$

Hence, $\{\widetilde{M}_t\}_{t \ge 1}$ is also bounded.

Combining these estimates, we have

$$\|X_t\|_F + \|M_t\|_F + \|\widetilde{M}_t\|_F \le B_X + B_M + B_M = B_X + 2B_M, \qquad \forall t \ge 1.$$

Since B_X and B_M are finite and independent of t, taking the supremum over $t \geq 1$ gives

$$\sup_{t\geq 1}\Big(\|X_t\|_F + \|M_t\|_F + \|\widetilde{M}_t\|_F\Big) \leq B_X + 2B_M < \infty. \tag{326}$$

Therefore, the full state trajectory $\{Z_t\}_{t\geq 1}$ is bounded.

For the Lyapunov analysis, define

$$c_t := \frac{\eta_t\lambda\beta_1}{\eta_t\lambda(1-\beta_1)+(1-\beta_2)}, \qquad b_t := \frac{\beta_1(1-\beta_2)}{(\beta_2-\beta_1)\big(\eta_t\lambda(1-\beta_1)+(1-\beta_2)\big)}, \qquad a_t := c_t + 1.$$

Since

$$\frac{d}{d\eta}\left(\frac{\eta\lambda\beta_1}{\eta\lambda(1-\beta_1)+(1-\beta_2)}\right) = \frac{\lambda\beta_1(1-\beta_2)}{(\eta\lambda(1-\beta_1)+(1-\beta_2))^2} \geq 0, \tag{327}$$

it follows that

$$c_{t+1} \leq c_t, \qquad \forall t \geq 0. \tag{328}$$

The time-dependent Lyapunov function is

$$\begin{aligned}\mathcal{H}_t &:= \mathcal{F}(X_t) - \mathcal{F}^\star + \frac{1}{\lambda}\mathcal{K}_t^*(\lambda X_t) + \frac{c_t}{\lambda}\left(\mathcal{K}_t^*(\lambda X_t) + \mathcal{K}_t(M_t) - \langle \lambda X_t, M_t\rangle\right)\\ &= \mathcal{F}(X_t) - \mathcal{F}^\star + \frac{a_t}{\lambda}\mathcal{K}_t^*(\lambda X_t) + \frac{c_t}{\lambda}\mathcal{K}_t(M_t) - c_t\langle X_t, M_t\rangle.\end{aligned} \tag{329}$$

Since $\mathcal{F}$ is real-valued, (306) ensures that $\mathcal{F}^\star$ is finite. By L-smoothness, the Cauchy–Schwarz inequality, and (319),

$$0 \leq \mathcal{F}(X_t) - \mathcal{F}^\star \leq \mathcal{F}(0) - \mathcal{F}^\star + \|\nabla\mathcal{F}(0)\|_F B_X + \frac{L}{2}B_X^2 =: B_F < \infty.$$

The common-domain condition (310) gives $\mathcal{K}_t^*(\lambda X_t) = 0$, and the Fenchel–Young inequality ensures that the gap in (329) is nonnegative. Using the uniform potential bound (238), the state bounds (319) and (323), and $0 \leq c_t \leq c_0$ from the definition of c_t and (328), we obtain

$$\begin{aligned}0 \leq \mathcal{H}_t &\leq B_F + \frac{c_t}{\lambda}\sqrt{N}\,B_M + c_t B_X B_M\\ &\leq B_F + \frac{c_0}{\lambda}(\sqrt{N} + \lambda B_X)B_M =: B_H < \infty, \qquad \forall t \geq 0.\end{aligned} \tag{330}$$

Because the potential may change between two consecutive iterations, we separate the within-step change evaluated with $\mathcal{K}_t$ from the explicit potential variation from $\mathcal{K}_t$ to $\mathcal{K}_{t+1}$. For this purpose, we introduce the frozen next Lyapunov value

$$\begin{aligned}\widehat{\mathcal{H}}_{t+1}^{(t)} &:= \mathcal{F}(X_{t+1}) - \mathcal{F}^\star + \frac{1}{\lambda}\mathcal{K}_t^*(\lambda X_{t+1})\\ &\quad + \frac{c_{t+1}}{\lambda}\left(\mathcal{K}_t^*(\lambda X_{t+1}) + \mathcal{K}_t(M_{t+1}) - \langle \lambda X_{t+1}, M_{t+1}\rangle\right).\end{aligned} \tag{331}$$

This is not the actual value $\mathcal{H}_{t+1}$. It uses the next variables (X_{t+1}, M_{t+1}), but keeps the potential frozen at $\mathcal{K}_t$. In contrast, the actual next Lyapunov value is

$$\begin{aligned}\mathcal{H}_{t+1} &:= \mathcal{F}(X_{t+1}) - \mathcal{F}^\star + \frac{1}{\lambda}\mathcal{K}_{t+1}^*(\lambda X_{t+1})\\ &\quad + \frac{c_{t+1}}{\lambda}\left(\mathcal{K}_{t+1}^*(\lambda X_{t+1}) + \mathcal{K}_{t+1}(M_{t+1}) - \langle \lambda X_{t+1}, M_{t+1}\rangle\right).\end{aligned} \tag{332}$$

Since the potentials may be nondifferentiable, the notation $\nabla\mathcal{K}_t(M)$ denotes a selected subgradient:

$$\nabla\mathcal{K}_t(M) \in \partial\mathcal{K}_t(M). \tag{333}$$

By (227), $\mathcal{K}_t^* = \chi_{\mathcal{C}_t}$. The common-domain condition gives $\lambda X_{t+1} \in \mathcal{C}_t$, and hence

$$0 \in \partial\mathcal{K}_t^*(\lambda X_{t+1}). \tag{334}$$

Throughout the Lyapunov analysis, we choose

$$\nabla\mathcal{K}_t^*(\lambda X_{t+1}) := 0, \qquad \forall t \geq 0. \tag{335}$$

Define the residual terms associated with the potential evaluated at time t by

$$\Gamma_t^{(t)} := \left\langle \nabla\mathcal{K}_t(\widetilde{M}_{t+1}) - \lambda X_{t+1}, \widetilde{M}_{t+1} - \nabla\mathcal{K}_t^*(\lambda X_{t+1}) \right\rangle, \tag{336}$$

$$\Delta_t^{(t)} := \left\langle \nabla\mathcal{K}_t(\widetilde{M}_{t+1}) - \nabla\mathcal{K}_t(M_{t+1}), \widetilde{M}_{t+1} - M_{t+1} \right\rangle. \tag{337}$$

We next relate $\Gamma_t^{(t)}$ to the frozen Fenchel–Young gap $G_t^{(t)}$ defined in (312). By the Fenchel–Young inequality,

$$G_t^{(t)} \geq 0. \tag{338}$$

Let

$$z_{t+1}^{(t)} := \nabla\mathcal{K}_t^*(\lambda X_{t+1}) \in \partial\mathcal{K}_t^*(\lambda X_{t+1}). \tag{339}$$

By Fenchel conjugacy,

$$\lambda X_{t+1} \in \partial\mathcal{K}_t(z_{t+1}^{(t)}). \tag{340}$$

Equivalently,

$$\mathcal{K}_t(z_{t+1}^{(t)}) + \mathcal{K}_t^*(\lambda X_{t+1}) - \left\langle \lambda X_{t+1}, z_{t+1}^{(t)} \right\rangle = 0. \tag{341}$$

We recall the monotonicity property of the subdifferential of a convex function and derive it from the subgradient inequalities. Let f be a proper convex function, and suppose that

$$u \in \partial f(x), \qquad v \in \partial f(y). \tag{342}$$

By the definition of a subgradient,

$$f(y) \geq f(x) + \langle u, y - x\rangle_F, \tag{343}$$

and

$$f(x) \geq f(y) + \langle v, x - y\rangle_F. \tag{344}$$

Adding these two inequalities cancels the function values and gives

$$\begin{aligned} 0 &\geq \langle u, y - x\rangle_F + \langle v, x - y\rangle_F \\ &= -\langle u - v, x - y\rangle_F. \end{aligned} \tag{345}$$

Consequently,

$$\langle u - v, x - y\rangle_F \geq 0. \tag{346}$$

This is the monotonicity property of ∂f.

We apply this property to the fixed-time convex potential $\mathcal{K}_t$. The selected update subgradient satisfies

$$\nabla\mathcal{K}_t(\widetilde{M}_{t+1}) \in \partial\mathcal{K}_t(\widetilde{M}_{t+1}), \tag{347}$$

while (340) gives

$$\lambda X_{t+1} \in \partial\mathcal{K}_t(z_{t+1}^{(t)}). \tag{348}$$

Thus, taking

$$\begin{aligned} x &= \widetilde{M}_{t+1}, \quad u = \nabla\mathcal{K}_t(\widetilde{M}_{t+1}), \\ y &= z_{t+1}^{(t)}, \qquad v = \lambda X_{t+1}, \end{aligned} \tag{349}$$

the monotonicity inequality yields

$$\Gamma_t^{(t)} = \left\langle \nabla\mathcal{K}_t(\widetilde{M}_{t+1}) - \lambda X_{t+1}, \widetilde{M}_{t+1} - z_{t+1}^{(t)} \right\rangle_F \geq 0. \tag{350}$$

Moreover, using (341) in the definition of $G_t^{(t)}$, we obtain

$$\begin{aligned} G_t^{(t)} &= \mathcal{K}_t(\widetilde{M}_{t+1}) + \mathcal{K}_t^*(\lambda X_{t+1}) - \left\langle \lambda X_{t+1}, \widetilde{M}_{t+1} \right\rangle \\ &= \mathcal{K}_t(\widetilde{M}_{t+1}) - \mathcal{K}_t(z_{t+1}^{(t)}) - \left\langle \lambda X_{t+1}, \widetilde{M}_{t+1} - z_{t+1}^{(t)} \right\rangle \\ &\leq \left\langle \nabla\mathcal{K}_t(\widetilde{M}_{t+1}), \widetilde{M}_{t+1} - z_{t+1}^{(t)} \right\rangle - \left\langle \lambda X_{t+1}, \widetilde{M}_{t+1} - z_{t+1}^{(t)} \right\rangle \\ &= \Gamma_t^{(t)}. \end{aligned} \tag{351}$$

Similarly, the selected subgradients satisfy

$$\nabla\mathcal{K}_t(\widetilde{M}_{t+1}) \in \partial\mathcal{K}_t(\widetilde{M}_{t+1}), \qquad \nabla\mathcal{K}_t(M_{t+1}) \in \partial\mathcal{K}_t(M_{t+1}). \tag{352}$$

Applying the same monotonicity property with

$$\begin{aligned} x &= \widetilde{M}_{t+1}, \quad u = \nabla\mathcal{K}_t(\widetilde{M}_{t+1}), \\ y &= M_{t+1}, \quad v = \nabla\mathcal{K}_t(M_{t+1}), \end{aligned} \tag{353}$$

we obtain

$$\Delta_t^{(t)} = \left\langle \nabla\mathcal{K}_t(\widetilde{M}_{t+1}) - \nabla\mathcal{K}_t(M_{t+1}), \widetilde{M}_{t+1} - M_{t+1} \right\rangle_F \geq 0. \tag{354}$$

Therefore,

$$0 \leq G_t^{(t)} \leq \Gamma_t^{(t)}. \tag{355}$$

For convenience, define

$$\delta_t := \nabla\mathcal{K}_t(\widetilde{M}_{t+1}) - \lambda X_{t+1}. \tag{356}$$

By the parameter update in (305), we have

$$X_{t+1} - X_t = \eta_t \delta_t. \tag{357}$$

The frozen next value $\widehat{\mathcal{H}}_{t+1}^{(t)}$ uses the coefficient c_{t+1}. To derive its one-step bound, define the auxiliary coefficient-frozen value

$$\begin{aligned} \overline{\mathcal{H}}_{t+1}^{(t)} := \mathcal{F}(X_{t+1}) - \mathcal{F}^\star + \frac{1}{\lambda}\mathcal{K}_t^*(\lambda X_{t+1}) \\ + \frac{c_t}{\lambda}\left(\mathcal{K}_t^*(\lambda X_{t+1}) + \mathcal{K}_t(M_{t+1}) - \langle \lambda X_{t+1}, M_{t+1}\rangle\right). \end{aligned} \tag{358}$$

By the Fenchel–Young inequality,

$$\mathcal{K}_t^*(\lambda X_{t+1}) + \mathcal{K}_t(M_{t+1}) - \langle \lambda X_{t+1}, M_{t+1}\rangle \geq 0. \tag{359}$$

Since $c_{t+1} \leq c_t$ by (328), we obtain

$$\widehat{\mathcal{H}}_{t+1}^{(t)} \leq \overline{\mathcal{H}}_{t+1}^{(t)}. \tag{360}$$

By the L-smoothness of $\mathcal{F}$,

$$\mathcal{F}(X_{t+1}) - \mathcal{F}(X_t) \leq \langle \nabla\mathcal{F}(X_t), X_{t+1} - X_t\rangle + \frac{L}{2}\|X_{t+1} - X_t\|_F^2. \tag{361}$$

By the subgradient inequalities for the convex functions $\mathcal{K}_t^*$ and $\mathcal{K}_t$,

$$\mathcal{K}_t^*(\lambda X_{t+1}) - \mathcal{K}_t^*(\lambda X_t) \leq \langle \lambda\nabla\mathcal{K}_t^*(\lambda X_{t+1}), X_{t+1} - X_t\rangle, \tag{362}$$

$$\mathcal{K}_t(M_{t+1}) - \mathcal{K}_t(M_t) \leq \langle \nabla\mathcal{K}_t(M_{t+1}), M_{t+1} - M_t\rangle. \tag{363}$$

Moreover,

$$\langle X_{t+1}, M_{t+1}\rangle - \langle X_t, M_t\rangle = \langle M_t, X_{t+1} - X_t\rangle + \langle X_{t+1}, M_{t+1} - M_t\rangle. \tag{364}$$

Applying (361), (363), and (364) to $\overline{\mathcal{H}}_{t+1}^{(t)} - \mathcal{H}_t$ yields

$$\begin{aligned} \overline{\mathcal{H}}_{t+1}^{(t)} - \mathcal{H}_t \leq \langle \nabla\mathcal{F}(X_t) + a_t\nabla\mathcal{K}_t^*(\lambda X_{t+1}) - c_t M_t, X_{t+1} - X_t\rangle \\ + \left\langle \frac{c_t}{\lambda}\nabla\mathcal{K}_t(M_{t+1}) - c_t X_{t+1}, M_{t+1} - M_t \right\rangle \\ + \frac{L}{2}\|X_{t+1} - X_t\|_F^2. \end{aligned} \tag{365}$$

From the momentum definitions in (303) and (304), we have

$$M_{t+1} - M_t = \frac{1-\beta_2}{\beta_2 - \beta_1}\left(\widetilde{M}_{t+1} - M_{t+1}\right). \tag{366}$$

The definitions of a_t, b_t, and c_t imply

$$c_t = (a_t + b_t)\beta_1 - b_t\beta_2, \qquad c_t = a_t - 1, \qquad \frac{c_t(1-\beta_2)}{\lambda(\beta_2-\beta_1)} = \eta_t b_t. \tag{367}$$

Consequently,

$$\nabla\mathcal{F}(X_t)+a_t\nabla\mathcal{K}_t^*(\lambda X_{t+1})-c_t M_t = -a_t\left(\widetilde{M}_{t+1} - \nabla\mathcal{K}_t^*(\lambda X_{t+1})\right) - b_t\left(\widetilde{M}_{t+1} - M_{t+1}\right). \tag{368}$$

Substituting (357), (366), (367), and (368) into (365), we obtain

$$\begin{aligned}
\widehat{\mathcal{H}}_{t+1}^{(t)} - \mathcal{H}_t \le \overline{\mathcal{H}}_{t+1}^{(t)} - \mathcal{H}_t &\le -\eta_t a_t\left\langle \delta_t, \widetilde{M}_{t+1} - \nabla\mathcal{K}_t^*(\lambda X_{t+1})\right\rangle \\
&\quad - \eta_t b_t\left\langle \delta_t, \widetilde{M}_{t+1} - M_{t+1}\right\rangle \\
&\quad + \eta_t b_t\left\langle \nabla\mathcal{K}_t(M_{t+1}) - \lambda X_{t+1}, \widetilde{M}_{t+1} - M_{t+1}\right\rangle \\
&\quad + \frac{L}{2}\eta_t^2\left\|\delta_t\right\|_F^2 \\
&= -\eta_t\left(a_t\Gamma_t^{(t)} + b_t\Delta_t^{(t)}\right) + \frac{L}{2}\eta_t^2\left\|\delta_t\right\|_F^2.
\end{aligned} \tag{369}$$

Finally, by (360) and (369),

$$\eta_t\left(a_t\Gamma_t^{(t)} + b_t\Delta_t^{(t)}\right) \le \mathcal{H}_t - \widehat{\mathcal{H}}_{t+1}^{(t)} + \frac{L}{2}\eta_t^2\left\|\nabla\mathcal{K}_t(\widetilde{M}_{t+1}) - \lambda X_{t+1}\right\|_F^2. \tag{370}$$

The remaining contribution caused by the time dependence of the potential is the discrepancy between $\widehat{\mathcal{H}}_{t+1}^{(t)}$ and $\mathcal{H}_{t+1}$. It remains to relate the frozen next value $\widehat{\mathcal{H}}_{t+1}^{(t)}$ to the actual next value $\mathcal{H}_{t+1}$. By direct subtraction of (331) from (332), we have

$$\begin{aligned}
\mathcal{H}_{t+1} - \widehat{\mathcal{H}}_{t+1}^{(t)} &= \frac{1}{\lambda}\left[\mathcal{K}_{t+1}^*(\lambda X_{t+1}) - \mathcal{K}_t^*(\lambda X_{t+1})\right] \\
&\quad + \frac{c_{t+1}}{\lambda}\left[\mathcal{K}_{t+1}^*(\lambda X_{t+1}) - \mathcal{K}_t^*(\lambda X_{t+1})\right] \\
&\quad + \frac{c_{t+1}}{\lambda}\left[\mathcal{K}_{t+1}(M_{t+1}) - \mathcal{K}_t(M_{t+1})\right] \\
&= \frac{a_{t+1}}{\lambda}\left[\mathcal{K}_{t+1}^*(\lambda X_{t+1}) - \mathcal{K}_t^*(\lambda X_{t+1})\right] \\
&\quad + \frac{c_{t+1}}{\lambda}\left[\mathcal{K}_{t+1}(M_{t+1}) - \mathcal{K}_t(M_{t+1})\right].
\end{aligned} \tag{371}$$

Therefore,

$$\mathcal{H}_t - \widehat{\mathcal{H}}_{t+1}^{(t)} = \mathcal{H}_t - \mathcal{H}_{t+1} + \mathcal{H}_{t+1} - \widehat{\mathcal{H}}_{t+1}^{(t)}.$$

Substituting (371) into (370) yields the time-dependent one-step inequality

$$\begin{aligned}
\eta_t\left(a_t\Gamma_t^{(t)} + b_t\Delta_t^{(t)}\right) &\le \mathcal{H}_t - \mathcal{H}_{t+1} + \frac{L}{2}\eta_t^2\left\|\nabla\mathcal{K}_t(\widetilde{M}_{t+1}) - \lambda X_{t+1}\right\|_F^2 \\
&\quad + \frac{a_{t+1}}{\lambda}\left[\mathcal{K}_{t+1}^*(\lambda X_{t+1}) - \mathcal{K}_t^*(\lambda X_{t+1})\right] \\
&\quad + \frac{c_{t+1}}{\lambda}\left[\mathcal{K}_{t+1}(M_{t+1}) - \mathcal{K}_t(M_{t+1})\right].
\end{aligned} \tag{372}$$

We now apply the common-domain condition (310). By Appendix C, for every pair of indices $s, t \ge 0$,

$$\mathcal{K}_s^*(\lambda X_t) = 0. \tag{373}$$

In particular, taking $s = t$ and $s = t+1$ at the point X_{t+1}, we have

$$\mathcal{K}_t^*(\lambda X_{t+1}) = 0, \qquad \mathcal{K}_{t+1}^*(\lambda X_{t+1}) = 0, \qquad \forall t \ge 0. \tag{374}$$

Therefore, the dual variation term in (372) vanishes:

$$\frac{a_{t+1}}{\lambda}\left[\mathcal{K}_{t+1}^*(\lambda X_{t+1}) - \mathcal{K}_t^*(\lambda X_{t+1})\right] = 0, \qquad \forall t \ge 0. \tag{375}$$

We now use the uniform upper bound on weighted signed primal variation proved in Appendix D.1. Under the conditions stated there, there exists a finite constant C_{var}, independent of T, such that

$$\sum_{t=0}^{T-1} \frac{c_{t+1}}{\lambda} \left[\mathcal{K}_{t+1}(M_{t+1}) - \mathcal{K}_t(M_{t+1})\right] \le C_{\mathrm{var}} < \infty, \qquad \forall T \ge 1. \tag{376}$$

This estimate follows from the proof in Appendix D.1, with the explicit constant in (464). Thus, the explicit primal variation term in the summed version of (372) is uniformly bounded above along the generated trajectory. It remains to bound the discrete smoothness remainder. To avoid confusion between the momentum variable M and the notation used for the selected orthonormal rank-one system, we denote the system selected at time t and argument M by

$$\mathcal{V}_t(M) = \{V_i(t,M)\}_{i=1}^{r(t,M)}.$$

Thus, by the admissible subgradient selection specified in Section 3, for any $M \in \mathcal{X}$, we use

$$\nabla\mathcal{K}_t(M) = \sum_{i=1}^{r(t,M)} \operatorname{sign}\left(\langle M, V_i(t,M)\rangle_F\right) V_i(t,M). \tag{377}$$

Since $\mathcal{V}_t(M)$ is an orthonormal rank-one system,

$$\|\nabla\mathcal{K}_t(M)\|_F^2 = \sum_{i=1}^{r(t,M)} \left|\operatorname{sign}\left(\langle M, V_i(t,M)\rangle_F\right)\right|^2 \tag{378}$$

$$\le r(t,M) \le N. \tag{379}$$

Moreover, under the common-domain regime,

$$\mathcal{K}_t^*(\lambda X_{t+1}) = 0. \tag{380}$$

By (227), this implies

$$\lambda X_{t+1} \in \mathcal{C}_t. \tag{381}$$

Hence, by the uniform outer bound (257),

$$\|\lambda X_{t+1}\|_F^2 \le N. \tag{382}$$

For any $A, B \in \mathcal{X}$, the triangle inequality gives

$$\|A - B\|_F \le \|A\|_F + \|B\|_F. \tag{383}$$

Since both sides are nonnegative, squaring yields

$$\|A - B\|_F^2 \le \|A\|_F^2 + 2\|A\|_F\|B\|_F + \|B\|_F^2. \tag{384}$$

Moreover,

$$0 \le (\|A\|_F - \|B\|_F)^2 = \|A\|_F^2 - 2\|A\|_F\|B\|_F + \|B\|_F^2, \tag{385}$$

so

$$2\|A\|_F\|B\|_F \le \|A\|_F^2 + \|B\|_F^2. \tag{386}$$

Combining these inequalities, we obtain

$$\|A - B\|_F^2 \le 2\|A\|_F^2 + 2\|B\|_F^2. \tag{387}$$

Applying this inequality with

$$A = \nabla\mathcal{K}_t(\widetilde{M}_{t+1}), \qquad B = \lambda X_{t+1}, \tag{388}$$

and using the bounds

$$\|\nabla\mathcal{K}_t(\widetilde{M}_{t+1})\|_F^2 \le N, \qquad \|\lambda X_{t+1}\|_F^2 \le N \tag{389}$$

established above, we conclude that

$$\begin{aligned}\left\|\nabla\mathcal{K}_t(\widetilde{M}_{t+1}) - \lambda X_{t+1}\right\|_F^2 &\le 2\|\nabla\mathcal{K}_t(\widetilde{M}_{t+1})\|_F^2 + 2\|\lambda X_{t+1}\|_F^2 \\ &\le 2N + 2N = 4N.\end{aligned} \tag{390}$$

Summing (372) from $t = 0$ to $T - 1$, and using the vanishing of the dual variation term, we obtain

$$\sum_{t=0}^{T-1} \eta_t \left(a_t \Gamma_t^{(t)} + b_t \Delta_t^{(t)} \right) \le \mathcal{H}_0 - \mathcal{H}_T + \frac{L}{2} \sum_{t=0}^{T-1} \eta_t^2 \left\| \nabla \mathcal{K}_t(\widetilde{M}_{t+1}) - \lambda X_{t+1} \right\|_F^2 + \sum_{t=0}^{T-1} \frac{c_{t+1}}{\lambda} \left[\mathcal{K}_{t+1}(M_{t+1}) - \mathcal{K}_t(M_{t+1}) \right]. \tag{391}$$

By the weighted signed primal-variation bound (376), established in Appendix D.1, we have

$$\sum_{t=0}^{T-1} \frac{c_{t+1}}{\lambda} \left[\mathcal{K}_{t+1}(M_{t+1}) - \mathcal{K}_t(M_{t+1}) \right] \le C_{\mathrm{var}}.$$

Therefore, (391) gives

$$\begin{aligned} \sum_{t=0}^{T-1} \eta_t \left(a_t \Gamma_t^{(t)} + b_t \Delta_t^{(t)} \right) &\le \mathcal{H}_0 - \mathcal{H}_T + \frac{L}{2} \sum_{t=0}^{T-1} \eta_t^2 \left\| \nabla \mathcal{K}_t(\widetilde{M}_{t+1}) - \lambda X_{t+1} \right\|_F^2 + C_{\mathrm{var}} \\ &\le \mathcal{H}_0 - \mathcal{H}_T + 2LN \sum_{t=0}^{T-1} \eta_t^2 + C_{\mathrm{var}} \end{aligned} \tag{392}$$

where the last inequality follows from (390). By (330), we have

$$\mathcal{H}_0 - \mathcal{H}_T \le \mathcal{H}_0.$$

Together with (309) and (376), the preceding estimate gives

$$\sum_{t=0}^{T-1} \eta_t \left(a_t \Gamma_t^{(t)} + b_t \Delta_t^{(t)} \right) \le \mathcal{H}_0 + 2LN \sum_{t=0}^{\infty} \eta_t^2 + C_{\mathrm{var}} < \infty. \tag{393}$$

The bound is independent of T. Since the summands on the left are nonnegative, letting $T \to \infty$ yields

$$\sum_{t=0}^{\infty} \eta_t \left(a_t \Gamma_t^{(t)} + b_t \Delta_t^{(t)} \right) < \infty. \tag{394}$$

Since

$$a_t \ge 1, \qquad b_t \ge 0, \qquad \Gamma_t^{(t)} \ge 0, \qquad \Delta_t^{(t)} \ge 0, \tag{395}$$

we also have

$$\sum_{t=0}^{\infty} \eta_t \Gamma_t^{(t)} \le \sum_{t=0}^{\infty} \eta_t a_t \Gamma_t^{(t)} \le \sum_{t=0}^{\infty} \eta_t \left(a_t \Gamma_t^{(t)} + b_t \Delta_t^{(t)} \right) < \infty. \tag{396}$$

Combining (396) and (355), we obtain

$$\sum_{t=0}^{\infty} \eta_t G_t^{(t)} < \infty. \tag{397}$$

We next prove that the weighted summability of the Fenchel–Young gaps, together with the no-spike condition, implies $G_t^{(t)} \to 0$. Recall that we have already shown

$$\sum_{t=0}^{\infty} \eta_t G_t^{(t)} < \infty, \qquad G_t^{(t)} \ge 0. \tag{398}$$

By (309), we have

$$\sum_{t=0}^{\infty} \eta_t = \infty, \qquad \eta_t \to 0, \tag{399}$$

where the second conclusion follows from square summability.

Recall the no-spike condition (313)–(314) imposed at the beginning of this appendix. This no-spike condition is the discrete-time analogue of asymptotic uniform continuity of the Fenchel–Young gap with respect to algorithmic time

$$\tau_t := \sum_{r=0}^{t-1} \eta_r. \tag{400}$$

Indeed,

$$\sum_{r=t}^{s-1} \eta_r = \tau_s - \tau_t, \tag{401}$$

so (314) requires the gap to vary uniformly little over sufficiently short algorithmic-time intervals. It therefore rules out arbitrarily narrow gap spikes. This is a trajectory-level regularity assumption and does not assume that $G_t^{(t)}$ converges or vanishes.

We claim that

$$G_t^{(t)} \to 0. \tag{402}$$

We prove this by contradiction. Suppose that (402) is false. Then there exist a constant $\varepsilon_0 > 0$ and a subsequence $\{t_j\}_{j\geq 1}$ with $t_j \to \infty$ such that

$$G_{t_j}^{(t_j)} \geq \varepsilon_0, \qquad \forall j. \tag{403}$$

Apply the local uniform continuity condition (314) with $\varepsilon = \varepsilon_0/2$. Let the corresponding constants be $\rho > 0$ and $T_{\varepsilon_0} \geq 0$. Hence, whenever $s \geq t \geq T_{\varepsilon_0}$ and

$$\sum_{q=t}^{s-1} \eta_q \leq \rho,$$

we have

$$\left| G_s^{(s)} - G_t^{(t)} \right| \leq \frac{\varepsilon_0}{2}. \tag{404}$$

For all sufficiently large j, we have $t_j \geq T_{\varepsilon_0}$. Since

$$\sum_{t=0}^{\infty} \eta_t = \infty, \qquad \eta_t \to 0, \qquad \eta_{t+1} \leq \eta_t,$$

we may choose $m_j \geq t_j$ such that

$$\frac{\rho}{2} \leq \sum_{r=t_j}^{m_j} \eta_r \leq \rho. \tag{405}$$

Indeed, define m_j as the first index $m \geq t_j$ for which

$$\sum_{r=t_j}^{m} \eta_r \geq \frac{\rho}{2}. \tag{406}$$

Such an index exists because the tail sum of the learning rates diverges:

$$\sum_{r=t_j}^{\infty} \eta_r = \infty.$$

Thus, starting from t_j, the partial sums must eventually reach the level $\rho/2$.

By the definition of m_j, the index m_j is the first time at which the partial sum reaches or exceeds $\rho/2$. Therefore, the partial sum up to the previous index $m_j - 1$ cannot have already reached $\rho/2$. Otherwise, $m_j - 1$ would also satisfy the defining condition (406), contradicting the minimality of m_j. Hence,

$$\sum_{r=t_j}^{m_j-1} \eta_r < \frac{\rho}{2}. \tag{407}$$

Adding the last term η_{m_j} gives

$$\sum_{r=t_j}^{m_j} \eta_r = \sum_{r=t_j}^{m_j-1} \eta_r + \eta_{m_j} < \frac{\rho}{2} + \eta_{m_j}. \tag{408}$$

Since $m_j \geq t_j$ and the learning-rate sequence is nonincreasing, we have

$$\eta_{m_j} \leq \eta_{t_j}.$$

Moreover, $t_j \to \infty$ and $\eta_{t_j} \to 0$. Hence, for all sufficiently large j,

$$\eta_{t_j} \leq \frac{\rho}{2}, \qquad \text{and therefore} \qquad \eta_{m_j} \leq \frac{\rho}{2}.$$

Substituting this into (408), we obtain

$$\sum_{r=t_j}^{m_j} \eta_r < \frac{\rho}{2} + \frac{\rho}{2} = \rho.$$

Together with the defining inequality (406), this proves

$$\frac{\rho}{2} \leq \sum_{r=t_j}^{m_j} \eta_r \leq \rho,$$

which is (405).

Now fix any $r \in \{t_j, \dots, m_j\}$. By (405),

$$\sum_{q=t_j}^{r-1} \eta_q \leq \sum_{q=t_j}^{m_j} \eta_q \leq \rho.$$

Therefore, applying (404) with $t = t_j$ and $s = r$, we get

$$\left| G_r^{(r)} - G_{t_j}^{(t_j)} \right| \leq \frac{\varepsilon_0}{2}. \tag{409}$$

From this absolute-value bound,

$$G_r^{(r)} - G_{t_j}^{(t_j)} \geq -\left| G_r^{(r)} - G_{t_j}^{(t_j)} \right|.$$

Hence,

$$\begin{aligned} G_r^{(r)} &\geq G_{t_j}^{(t_j)} - \left| G_r^{(r)} - G_{t_j}^{(t_j)} \right| \\ &\geq G_{t_j}^{(t_j)} - \frac{\varepsilon_0}{2} \\ &\geq \varepsilon_0 - \frac{\varepsilon_0}{2} = \frac{\varepsilon_0}{2}, \end{aligned} \tag{410}$$

where the last inequality uses (403). Thus, on the whole block $\{t_j, \dots, m_j\}$, the gap is bounded from below by $\varepsilon_0/2$.

Consequently,

$$\begin{aligned} \sum_{r=t_j}^{m_j} \eta_r G_r^{(r)} &\geq \sum_{r=t_j}^{m_j} \eta_r \frac{\varepsilon_0}{2} \\ &= \frac{\varepsilon_0}{2} \sum_{r=t_j}^{m_j} \eta_r \\ &\geq \frac{\varepsilon_0}{2} \cdot \frac{\rho}{2} = \frac{\varepsilon_0 \rho}{4}. \end{aligned} \tag{411}$$

Since the subsequence $\{t_j\}$ is infinite and $t_j \to \infty$, we may pass to a further subsequence, if necessary, so that the corresponding blocks

$$\{t_j, \dots, m_j\}$$

are pairwise disjoint. Summing (411) over these infinitely many disjoint blocks gives

$$\sum_{t=0}^{\infty} \eta_t G_t^{(t)} = \infty.$$

This contradicts the weighted summability (398). Therefore, the contradiction assumption was false, and we conclude that

$$G_t^{(t)} \to 0.$$

Now let

$$Z_{t_j+1} = (X_{t_j+1}, M_{t_j+1}, \widetilde{M}_{t_j+1}) \to Z_\infty := (X_\infty, M_\infty, \widetilde{M}_\infty) \tag{412}$$

be any convergent subsequence. Since (402) holds along the whole sequence, we have

$$\lim_{j\to\infty} G_{t_j}^{(t_j)} = 0. \tag{413}$$

The limiting-gap assumption (316) applies to the convergent full-state subsequence chosen above, with the single potential $\mathcal{K}_\infty$ fixed at the beginning of this appendix.

Combining (413) and (316), we obtain

$$G_\infty(X_\infty, \widetilde{M}_\infty) = 0. \tag{414}$$

By the definition of G_∞, this is precisely the limiting Fenchel–Young equality

$$\mathcal{K}_\infty(\widetilde{M}_\infty) + \mathcal{K}_\infty^*(\lambda X_\infty) - \left\langle \lambda X_\infty, \widetilde{M}_\infty \right\rangle = 0. \tag{415}$$

Hence, by Fenchel conjugacy,

$$\lambda X_\infty \in \partial \mathcal{K}_\infty(\widetilde{M}_\infty). \tag{416}$$

We next show that the filtered momentum becomes asymptotically equal to the negative gradient, in the sense that $\widetilde{M}_t + \nabla \mathcal{F}(X_t) \to 0$ along the generated trajectory. This step converts the limiting Fenchel consistency obtained above into the KKT-type stationarity condition.

Recall that the update direction is uniformly bounded. Indeed, by (390),

$$\left\| \nabla \mathcal{K}_t(\widetilde{M}_{t+1}) - \lambda X_{t+1} \right\|_F^2 \le 4N. \tag{417}$$

Therefore,

$$\left\| \nabla \mathcal{K}_t(\widetilde{M}_{t+1}) - \lambda X_{t+1} \right\|_F \le 2\sqrt{N}. \tag{418}$$

Since the parameter update is

$$X_{t+1} - X_t = \eta_t \left(\nabla \mathcal{K}_t(\widetilde{M}_{t+1}) - \lambda X_{t+1} \right), \tag{419}$$

we obtain

$$\|X_{t+1} - X_t\|_F \le 2\sqrt{N}\, \eta_t. \tag{420}$$

Because $\mathcal{F}$ is L-smooth, its gradient is L-Lipschitz. Hence,

$$\|\nabla \mathcal{F}(X_{t+1}) - \nabla \mathcal{F}(X_t)\|_F \le L \|X_{t+1} - X_t\|_F. \tag{421}$$

Combining this with (420) gives

$$\|\nabla \mathcal{F}(X_{t+1}) - \nabla \mathcal{F}(X_t)\|_F \le 2L\sqrt{N}\, \eta_t. \tag{422}$$

We now control the Polyak momentum tracking error. Let

$$A_t := \|\nabla \mathcal{F}(X_t) + M_t\|_F. \tag{423}$$

Using the deterministic momentum recursion

$$M_t = \beta_2 M_{t-1} - (1-\beta_2)\nabla\mathcal{F}(X_{t-1}), \tag{424}$$

we have

$$\begin{aligned}\nabla\mathcal{F}(X_t) + M_t &= \nabla\mathcal{F}(X_t) + \beta_2 M_{t-1} - (1-\beta_2)\nabla\mathcal{F}(X_{t-1}) \\ &= \nabla\mathcal{F}(X_t) - \nabla\mathcal{F}(X_{t-1}) + \beta_2\left(\nabla\mathcal{F}(X_{t-1}) + M_{t-1}\right).\end{aligned} \tag{425}$$

Taking Frobenius norms and using (422), we obtain

$$A_t \le 2L\sqrt{N}\,\eta_{t-1} + \beta_2 A_{t-1}. \tag{426}$$

Iterating (426) yields

$$A_t \le 2L\sqrt{N}\sum_{k=1}^{t}\beta_2^{t-k}\eta_{k-1} + \beta_2^t A_0. \tag{427}$$

Since $0 \le \beta_2 < 1$, the second term satisfies

$$\beta_2^t A_0 \to 0. \tag{428}$$

It remains to show that the geometric convolution term vanishes. Since $\eta_t \to 0$, for every $\varepsilon > 0$ there exists $K \ge 1$ such that

$$\eta_{k-1} \le \varepsilon, \qquad \forall k \ge K. \tag{429}$$

For $t \ge K$, split the convolution as

$$\sum_{k=1}^{t}\beta_2^{t-k}\eta_{k-1} = \sum_{k=1}^{K-1}\beta_2^{t-k}\eta_{k-1} + \sum_{k=K}^{t}\beta_2^{t-k}\eta_{k-1}. \tag{430}$$

The first term on the right-hand side of (430) converges to zero as $t \to \infty$, since it is a finite sum and each factor β_2^{t-k} converges to zero. For the second term, we have

$$\sum_{k=K}^{t}\beta_2^{t-k}\eta_{k-1} \le \varepsilon\sum_{k=K}^{t}\beta_2^{t-k} \le \frac{\varepsilon\left(1-\beta_2^{t-K+1}\right)}{1-\beta_2} \le \frac{\varepsilon}{1-\beta_2}. \tag{431}$$

Because $\varepsilon > 0$ is arbitrary, it follows that

$$\sum_{k=1}^{t}\beta_2^{t-k}\eta_{k-1} \to 0. \tag{432}$$

Therefore, from (427),

$$\|\nabla\mathcal{F}(X_t) + M_t\|_F \to 0. \tag{433}$$

We now pass from M_t to the filtered momentum $\widetilde{M}_t$. By the definition of the filtered momentum,

$$\widetilde{M}_t = \beta_1 M_{t-1} - (1-\beta_1)\nabla\mathcal{F}(X_{t-1}). \tag{434}$$

Hence,

$$\nabla\mathcal{F}(X_t) + \widetilde{M}_t = \nabla\mathcal{F}(X_t) - \nabla\mathcal{F}(X_{t-1}) + \beta_1\left(\nabla\mathcal{F}(X_{t-1}) + M_{t-1}\right). \tag{435}$$

Taking norms and using (422), we obtain

$$\|\nabla\mathcal{F}(X_t) + \widetilde{M}_t\|_F \le 2L\sqrt{N}\,\eta_{t-1} + \beta_1\|\nabla\mathcal{F}(X_{t-1}) + M_{t-1}\|_F. \tag{436}$$

Since $\eta_{t-1} \to 0$, and (433) gives

$$\|\nabla\mathcal{F}(X_{t-1}) + M_{t-1}\|_F \to 0, \tag{437}$$

we conclude from (436) that

$$\|\nabla\mathcal{F}(X_t) + \widetilde{M}_t\|_F \to 0. \tag{438}$$

The preceding estimates show that the position increments and the momentum tracking errors vanish asymptotically. From (420) and $\eta_t \to 0$, we have

$$\|X_{t+1} - X_t\|_F \to 0. \tag{439}$$

Moreover, (433) gives

$$\|\nabla \mathcal{F}(X_t) + M_t\|_F \to 0. \tag{440}$$

This corresponds to the vanishing of the momentum residual in the continuous-time equation

$$\dot{M}_t = -\nabla \mathcal{F}(X_t) - M_t. \tag{441}$$

Finally, (438) gives

$$\|\nabla \mathcal{F}(X_t) + \widetilde{M}_t\|_F \to 0. \tag{442}$$

We now combine these momentum tracking limits with the limiting Fenchel consistency established in (416) to obtain the KKT stationarity condition.

Returning to the arbitrary convergent full-state subsequence

$$Z_{t_j+1} \to (X_\infty, M_\infty, \widetilde{M}_\infty)$$

considered above, (438) and the continuity of $\nabla \mathcal{F}$ imply

$$\widetilde{M}_\infty = -\nabla \mathcal{F}(X_\infty). \tag{443}$$

Combining (443) with (416), we obtain

$$\lambda X_\infty \in \partial \mathcal{K}_\infty \left(-\nabla \mathcal{F}(X_\infty)\right). \tag{444}$$

By Fenchel conjugacy, for any proper closed convex function $\mathcal{K}$, the following conditions are equivalent:

$$Y \in \partial \mathcal{K}(Z) \iff \mathcal{K}(Z) + \mathcal{K}^*(Y) = \langle Y, Z \rangle \iff Z \in \partial \mathcal{K}^*(Y).$$

Applying this equivalence with

$$\mathcal{K} = \mathcal{K}_\infty, \qquad Z = -\nabla \mathcal{F}(X_\infty), \qquad Y = \lambda X_\infty,$$

and using (444), we obtain

$$-\nabla \mathcal{F}(X_\infty) \in \partial \mathcal{K}_\infty^*(\lambda X_\infty). \tag{445}$$

Equivalently,

$$0 \in \nabla \mathcal{F}(X_\infty) + \partial \mathcal{K}_\infty^*(\lambda X_\infty). \tag{446}$$

We now interpret the preceding limiting relations as the KKT conditions of the limiting constrained problem induced by $\mathcal{K}_\infty$.

Let

$$C_\infty := \operatorname{dom} \mathcal{K}_\infty^* = \{Y : \mathcal{K}_\infty^*(Y) < \infty\}.$$

Here C_∞ is the limiting dual feasible set, that is, the set of dual variables for which the limiting conjugate potential is finite. As in Appendix C, we recall that χ_C denotes the convex-analytic indicator function of a set C:

$$\chi_C(Y) := \begin{cases} 0, & Y \in C, \\ +\infty, & Y \notin C. \end{cases}$$

Since $\mathcal{K}_\infty$ is proper, lower-semicontinuous, convex, and positively homogeneous, the same argument as in Appendix C shows that its Fenchel conjugate is the indicator function of its effective domain C_∞, namely

$$\mathcal{K}_\infty^* = \chi_{C_\infty}.$$

Equivalently, this means that

$$\mathcal{K}_\infty^*(Y) = \begin{cases} 0, & Y \in C_\infty, \\ +\infty, & Y \notin C_\infty. \end{cases}$$

Consider the limiting constrained problem

$$\min_X \mathcal{F}(X) \qquad \text{subject to} \qquad \lambda X \in C_\infty. \tag{447}$$

Equivalently, this can be written as the extended-valued optimization problem

$$\min_X \left\{ \mathcal{F}(X) + \frac{1}{\lambda} \mathcal{K}_\infty^*(\lambda X) \right\}. \tag{448}$$

Indeed, since $\lambda > 0$ and $\mathcal{K}_\infty^* = \chi_{C_\infty}$, the term

$$\frac{1}{\lambda}\mathcal{K}_\infty^*(\lambda X)$$

is equal to 0 when $\lambda X \in C_\infty$, and is $+\infty$ otherwise. Hence the extended-valued formulation (448) is exactly the constrained problem (447).

First, the limiting Fenchel–Young equality implies primal feasibility. Indeed,

$$\mathcal{K}_\infty(\widetilde{M}_\infty) + \mathcal{K}_\infty^*(\lambda X_\infty) - \left\langle \lambda X_\infty, \widetilde{M}_\infty \right\rangle = 0 \tag{449}$$

is finite. Therefore,

$$\mathcal{K}_\infty^*(\lambda X_\infty) < \infty. \tag{450}$$

By the definition of

$$C_\infty = \{Y : \mathcal{K}_\infty^*(Y) < \infty\},$$

we obtain

$$\lambda X_\infty \in C_\infty. \tag{451}$$

Second, the stationarity condition has already been established in (446).

We interpret (446) using normal cones. Since $\mathcal{K}_\infty^* = \chi_{C_\infty}$ is proper, closed, and convex, the set C_∞ is nonempty, closed, and convex. For a nonempty closed convex set $C \subseteq \mathcal{X}$, define

$$N_C(Y) := \begin{cases} \{V \in \mathcal{X} : \langle V, Z - Y\rangle_F \le 0 \text{ for all } Z \in C\}, & Y \in C, \\ \emptyset, & Y \notin C. \end{cases}$$

Since C is nonempty, closed, and convex, its indicator function χ_C is proper, closed, and convex. We now verify the standard identity

$$\partial\chi_C(Y) = N_C(Y).$$

Suppose first that $Y \in C$. By the definition of the subdifferential of a convex function,

$$V \in \partial\chi_C(Y)$$

if and only if

$$\chi_C(Z) \ge \chi_C(Y) + \langle V, Z - Y\rangle_F, \qquad \forall Z \in \mathcal{X}.$$

Since $Y \in C$, we have $\chi_C(Y) = 0$. If $Z \in C$, then $\chi_C(Z) = 0$, and hence the subgradient inequality is equivalent to

$$\langle V, Z - Y\rangle_F \le 0, \qquad \forall Z \in C.$$

If $Z \notin C$, then $\chi_C(Z) = +\infty$, so the subgradient inequality is automatically satisfied. Therefore,

$$V \in \partial\chi_C(Y) \Longleftrightarrow V \in N_C(Y).$$

If $Y \notin C$, then $Y \notin \operatorname{dom}\chi_C$, so

$$\partial\chi_C(Y) = \emptyset = N_C(Y).$$

Thus,

$$\partial\chi_C(Y) = N_C(Y), \qquad \forall Y \in \mathcal{X}.$$

Since $\mathcal{K}_\infty^* = \chi_{C_\infty}$, (445) and (451) give

$$\lambda X_\infty \in C_\infty, \qquad -\nabla\mathcal{F}(X_\infty) \in N_{C_\infty}(\lambda X_\infty).$$

Define the feasible set in the parameter space by

$$Q_\infty := \{X \in \mathcal{X} : \lambda X \in C_\infty\}. \tag{452}$$

For $X \in Q_\infty$, the definition of the normal cone and $\lambda > 0$ imply

$$\begin{aligned} V \in N_{Q_\infty}(X) &\Longleftrightarrow \langle V, Z - X\rangle_F \le 0 \quad \text{for all } Z \in Q_\infty \\ &\Longleftrightarrow \langle V, \lambda Z - \lambda X\rangle_F \le 0 \quad \text{for all } Z \in Q_\infty \\ &\Longleftrightarrow \langle V, Y - \lambda X\rangle_F \le 0 \quad \text{for all } Y \in C_\infty \\ &\Longleftrightarrow V \in N_{C_\infty}(\lambda X). \end{aligned}$$

Thus,

$$N_{Q_\infty}(X) = N_{C_\infty}(\lambda X), \qquad X \in Q_\infty.$$

Consequently,

$$X_\infty \in Q_\infty, \qquad 0 \in \nabla\mathcal{F}(X_\infty) + N_{Q_\infty}(X_\infty). \tag{453}$$

This is the first-order normal-cone stationarity condition for the constrained problem (447). It applies without assuming convexity of $\mathcal{F}$. The argument above establishes this condition for the parameter component of every convergent full-state subsequence.

We now extend the conclusion to every cluster point of the parameter sequence. Let $\bar{X}$ be any cluster point of $\{X_t\}_{t\geq 0}$, and choose indices $n_j \to \infty$ such that

$$X_{n_j} \to \bar{X}.$$

By (433), (438), and the continuity of $\nabla\mathcal{F}$,

$$M_{n_j} \to -\nabla\mathcal{F}(\bar{X}), \qquad \widetilde{M}_{n_j} \to -\nabla\mathcal{F}(\bar{X}).$$

Hence,

$$Z_{n_j} \to \left(\bar{X}, -\nabla\mathcal{F}(\bar{X}), -\nabla\mathcal{F}(\bar{X})\right).$$

After discarding finitely many indices, define $s_j := n_j - 1 \geq 0$. Since $s_j \to \infty$ and

$$Z_{s_j+1} = Z_{n_j} \to \left(\bar{X}, -\nabla\mathcal{F}(\bar{X}), -\nabla\mathcal{F}(\bar{X})\right),$$

the sequence $\{s_j\}$ satisfies the hypothesis of the limiting-gap assumption (316). Applying the preceding full-state limit argument therefore yields

$$\bar{X} \in Q_\infty, \qquad 0 \in \nabla\mathcal{F}(\bar{X}) + N_{Q_\infty}(\bar{X}).$$

Equivalently,

$$\lambda\bar{X} \in C_\infty, \qquad 0 \in \nabla\mathcal{F}(\bar{X}) + \partial\mathcal{K}_\infty^*(\lambda\bar{X}).$$

Thus every cluster point of the parameter sequence satisfies the KKT conditions of the same limiting constrained problem.

Let

$$\mathrm{KKT} := \{X : \lambda X \in C_\infty \text{ and } 0 \in \nabla\mathcal{F}(X) + \partial\mathcal{K}_\infty^*(\lambda X)\} \tag{454}$$

denote the KKT set of the limiting constrained problem (447). The preceding argument shows that every limit point of $\{X_t\}_{t\geq 0}$ belongs to KKT.

We now show convergence to this KKT set. Define

$$d(X, \mathrm{KKT}) := \inf_{Y\in\mathrm{KKT}} \|X - Y\|_F. \tag{455}$$

Suppose, for contradiction, that

$$d(X_t, \mathrm{KKT}) \not\to 0. \tag{456}$$

Then there exist $\varepsilon > 0$ and a subsequence $\{X_{t_j}\}_{j\geq 1}$ such that

$$d(X_{t_j}, \mathrm{KKT}) \geq \varepsilon, \qquad \forall j. \tag{457}$$

Since the generated trajectory is bounded, there exists a further subsequence, still denoted by $\{X_{t_j}\}_{j\geq 1}$, and a point X_∞ such that

$$X_{t_j} \to X_\infty. \tag{458}$$

By the limit-point argument above,

$$X_\infty \in \mathrm{KKT}\,. \tag{459}$$

Therefore, by the definition of distance to a set,

$$d(X_{t_j}, \mathrm{KKT}) \leq \|X_{t_j} - X_\infty\|_F \to 0. \tag{460}$$

This contradicts

$$d(X_{t_j}, \mathrm{KKT}) \geq \varepsilon. \tag{461}$$

Hence,

$$d(X_t, \mathrm{KKT}) \to 0. \tag{462}$$

Consequently, the generated trajectory converges to the KKT set of the limiting constrained problem induced by $\mathcal{K}_\infty$.

D.1 A Uniform Upper Bound on Weighted Signed Primal Variation

We establish the cumulative bound used in (376):

$$V_T := \sum_{t=0}^{T-1} \frac{c_{t+1}}{\lambda} \left[\mathcal{K}_{t+1}(M_{t+1}) - \mathcal{K}_t(M_{t+1})\right] \le C_{\mathrm{var}} \qquad \forall T \ge 1.$$

Here C_{var} is a finite constant independent of T.

Inside the brackets, the same M_{t+1} is supplied to both functions:

$$\mathcal{K}_{t+1}(M_{t+1}) - \mathcal{K}_t(M_{t+1})$$

Thus, this difference measures the change in the value caused by the change in the potential itself, with the argument held fixed.

We derive this cumulative upper bound from the potential, dynamics, and learning-rate conditions recalled below. Consequently, (376) need not be imposed as a separate assumption.

We retain the sign of each difference. Positive and negative contributions can therefore cancel.

We use the following dynamics:

$$\begin{aligned} g_t &:= \nabla \mathcal{F}(X_t), \\ M_{t+1} &= \beta_2 M_t - (1-\beta_2) g_t, \\ \widetilde{M}_{t+1} &= \beta_1 M_t - (1-\beta_1) g_t, \\ u_t &\in \partial \mathcal{K}_t(\widetilde{M}_{t+1}), \\ X_{t+1} &= X_t + \eta_t (u_t - \lambda X_{t+1}). \end{aligned}$$

Here u_t is the selected subgradient denoted by $\nabla \mathcal{K}_t(\widetilde{M}_{t+1})$ in the main analysis. As in Appendix C and the opening of Appendix D, each $\mathcal{K}_t$ is convex and positively homogeneous. By (238), $0 \le \mathcal{K}_t(H) \le U\|H\|_F$, where $U := \sqrt{N}$. The dynamics above are the exact-gradient dynamics (303)–(305), with fixed $0 \le \beta_1 < \beta_2 < 1$ and $\lambda > 0$ as in (307). We use L-smoothness in the form $\|\nabla \mathcal{F}(X) - \nabla \mathcal{F}(Y)\|_F \le L\|X - Y\|_F$.

The learning rates satisfy $\eta_t > 0$ and $\eta_{t+1} \le \eta_t$, as specified in (307) and (308). We also write $S_2 := \sum_{t=0}^{\infty} \eta_t^2 < \infty$, using the square-summability condition in (309).

Under these conditions and with finite initial states X_0, M_0, we show that V_T is bounded above uniformly in T, with the explicit constant given in (464). The argument does not require the common-domain assumption. All norms and inner products below are Frobenius norms and inner products. Auxiliary symbols introduced below are local to this subsection; in particular, A_T below is distinct from the tracking error A_t in the main analysis, and D_t denotes a scalar potential increment.

First, define

$$w_t := \frac{c_{t+1}}{\lambda}, \qquad q_t := \mathcal{K}_t(M_t)$$

The quantity q_t is the value of the time-t potential evaluated at the time-t momentum.

Also define

$$D_t := \mathcal{K}_t(M_{t+1}) - \mathcal{K}_t(M_t).$$

In D_t, the potential $\mathcal{K}_t$ is held fixed and only its argument changes. We then have

$$\begin{aligned}(q_{t+1} - q_t) - D_t &= \mathcal{K}_{t+1}(M_{t+1}) - \mathcal{K}_t(M_t) \\ &\quad - [\mathcal{K}_t(M_{t+1}) - \mathcal{K}_t(M_t)] \\ &= \mathcal{K}_{t+1}(M_{t+1}) - \mathcal{K}_t(M_{t+1}).\end{aligned}$$

Therefore, the exact decomposition is

$$\boxed{V_T = A_T - E_T}$$

where

$$A_T := \sum_{t=0}^{T-1} w_t(q_{t+1} - q_t), \qquad E_T := \sum_{t=0}^{T-1} w_t D_t.$$

To control A_T, we use boundedness of q_t and cancellation in the weighted differences. For E_T, we use the Lipschitz property of $\mathcal{K}_t$ and the momentum increments. We first establish the Lipschitz and boundedness estimates needed for these arguments.

Convexity of $\mathcal{K}_t$ means that, for arbitrary A, B and $\theta \in [0, 1]$, the following inequality holds:

$$\mathcal{K}_t(\theta A + (1-\theta)B) \le \theta \mathcal{K}_t(A) + (1-\theta)\mathcal{K}_t(B)$$

In particular, taking $\theta = \frac{1}{2}$ gives

$$\mathcal{K}_t\left(\frac{A+B}{2}\right) \le \frac{1}{2}\mathcal{K}_t(A) + \frac{1}{2}\mathcal{K}_t(B).$$

Here positive homogeneity means positive homogeneity of degree one: for every positive α,

$$\mathcal{K}_t(\alpha H) = \alpha \mathcal{K}_t(H)$$

Thus, multiplying the argument by a positive scalar multiplies the function value by the same scalar. Substituting $\alpha = 2$ and $H = (A + B)/2$ into this definition gives

$$\mathcal{K}_t(A+B) = 2\mathcal{K}_t\left(\frac{A+B}{2}\right).$$

Combining these two properties, we obtain

$$\begin{aligned}\mathcal{K}_t(A+B) &= 2\mathcal{K}_t\left(\frac{A+B}{2}\right) && \text{(positive homogeneity)} \\ &\le 2\left(\frac{1}{2}\mathcal{K}_t(A) + \frac{1}{2}\mathcal{K}_t(B)\right) && \text{(convexity)} \\ &= \mathcal{K}_t(A) + \mathcal{K}_t(B).\end{aligned}$$

Hence,

$$\boxed{\mathcal{K}_t(A+B) \le \mathcal{K}_t(A) + \mathcal{K}_t(B).}$$

Applying this property to $A = B + (A - B)$ yields

$$\mathcal{K}_t(A) \le \mathcal{K}_t(B) + \mathcal{K}_t(A - B).$$

Subtracting $\mathcal{K}_t(B)$ from both sides and using the existing bound $\mathcal{K}_t(H) \le U\|H\|_F$, we obtain

$$\mathcal{K}_t(A) - \mathcal{K}_t(B) \le \mathcal{K}_t(A - B) \le U\|A - B\|_F.$$

Interchanging A and B gives

$$\mathcal{K}_t(B) - \mathcal{K}_t(A) \le U\|B - A\|_F = U\|A - B\|_F.$$

Together, these inequalities give

$$-U\|A - B\|_F \le \mathcal{K}_t(A) - \mathcal{K}_t(B) \le U\|A - B\|_F.$$

Therefore,

$$\boxed{|\mathcal{K}_t(A) - \mathcal{K}_t(B)| \le U\|A - B\|_F.}$$

This Lipschitz bound is uniform in time.

By definition, u is a subgradient of the convex function $\mathcal{K}_t$ at H if

$$\boxed{\mathcal{K}_t(Y) \ge \mathcal{K}_t(H) + \langle u, Y - H\rangle_F \qquad \forall Y}$$

For a differentiable convex function, $u = \nabla\mathcal{K}_t(H)$. At a nondifferentiable point, we use a vector satisfying this inequality.

In the dynamics under consideration, we have selected

$$u_t \in \partial\mathcal{K}_t(\widetilde{M}_{t+1})$$

We may therefore substitute

$$H = \widetilde{M}_{t+1}, \qquad Y = H + Z$$

into the definition above. This gives

$$\mathcal{K}_t(H + Z) \ge \mathcal{K}_t(H) + \langle u_t, Z\rangle_F.$$

Rearranging yields

$$\langle u_t, Z\rangle_F \le \mathcal{K}_t(H + Z) - \mathcal{K}_t(H).$$

This is the source of the first inequality below. The second follows from the Lipschitz property just proved:

$$\begin{aligned}\mathcal{K}_t(H + Z) - \mathcal{K}_t(H) &\le |\mathcal{K}_t(H + Z) - \mathcal{K}_t(H)| \\ &\le U\|(H + Z) - H\|_F \\ &= U\|Z\|_F.\end{aligned}$$

Consequently,

$$\boxed{\langle u_t, Z\rangle_F \le \mathcal{K}_t(H + Z) - \mathcal{K}_t(H) \le U\|Z\|_F.}$$

This inequality holds for every Z. If $u_t \neq 0$, choose

$$Z = \frac{u_t}{\|u_t\|_F}$$

Then

$$\left\langle u_t, \frac{u_t}{\|u_t\|_F} \right\rangle_F = \frac{\|u_t\|_F^2}{\|u_t\|_F} = \|u_t\|_F, \qquad \left\| \frac{u_t}{\|u_t\|_F} \right\|_F = 1.$$

It follows that

$$\boxed{\|u_t\|_F \le U.}$$

The same conclusion holds when $u_t = 0$.

We next establish bounds on the parameter, gradient, and momentum sequences.

Rearranging the original update gives

$$(1 + \lambda\eta_t)X_{t+1} = X_t + \eta_t u_t,$$

and hence

$$X_{t+1} = \frac{X_t + \eta_t u_t}{1 + \lambda\eta_t}.$$

Define

$$\alpha_t := \frac{1}{1 + \lambda\eta_t}$$

Then $0 < \alpha_t < 1$ and

$$1 - \alpha_t = \frac{\lambda\eta_t}{1 + \lambda\eta_t}.$$

Therefore,

$$X_{t+1} = \alpha_t X_t + (1 - \alpha_t)\frac{u_t}{\lambda}.$$

To obtain a norm bound from this convex combination, we explicitly apply the triangle inequality:

$$\begin{aligned}
\|X_{t+1}\|_F &= \left\| \alpha_t X_t + (1 - \alpha_t)\frac{u_t}{\lambda} \right\|_F \\
&\le \|\alpha_t X_t\|_F + \left\| (1 - \alpha_t)\frac{u_t}{\lambda} \right\|_F \quad \text{(triangle inequality)} \\
&= \alpha_t \|X_t\|_F + (1 - \alpha_t)\frac{\|u_t\|_F}{\lambda} \\
&\le \alpha_t \|X_t\|_F + (1 - \alpha_t)\frac{U}{\lambda}.
\end{aligned}$$

Now set

$$r_t := \max\left\{ \|X_t\|_F, \frac{U}{\lambda} \right\}$$

Since $\|X_t\|_F \le r_t$ and $U/\lambda \le r_t$, we have

$$\begin{aligned}\alpha_t\|X_t\|_F + (1-\alpha_t)\frac{U}{\lambda} &\le \alpha_t r_t + (1-\alpha_t)r_t \\ &= [\alpha_t + (1-\alpha_t)]r_t \\ &= r_t.\end{aligned}$$

Thus,

$$\boxed{\|X_{t+1}\|_F \le \max\left\{\|X_t\|_F, \frac{U}{\lambda}\right\}.}$$

Define

$$\boxed{R := \max\left\{\|X_0\|_F, \frac{U}{\lambda}\right\}}$$

We use this same constant R throughout the argument. By definition, $\|X_0\|_F \le R$ and $U/\lambda \le R$. If $\|X_t\|_F \le R$, then

$$\|X_{t+1}\|_F \le \max\left\{\|X_t\|_F, \frac{U}{\lambda}\right\} \le R.$$

Mathematical induction therefore gives

$$\boxed{\|X_t\|_F \le R \qquad \forall t \ge 0.}$$

The constant R is an explicit bound derived here, not an additional assumed bound.

The original update is

$$X_{t+1} = X_t + \eta_t(u_t - \lambda X_{t+1})$$

Subtracting X_t from both sides immediately yields

$$\boxed{X_{t+1} - X_t = \eta_t(u_t - \lambda X_{t+1})}$$

Consequently, the corresponding norm identity is

$$\boxed{\|X_{t+1} - X_t\|_F = \eta_t\|u_t - \lambda X_{t+1}\|_F}$$

We can also compute the increment from the expression containing the denominator:

$$\begin{aligned}X_{t+1} - X_t &= \frac{X_t + \eta_t u_t}{1+\lambda\eta_t} - X_t \\ &= \frac{X_t + \eta_t u_t - (1+\lambda\eta_t)X_t}{1+\lambda\eta_t} \\ &= \frac{\eta_t u_t - \lambda\eta_t X_t}{1+\lambda\eta_t} \\ &= \frac{\eta_t}{1+\lambda\eta_t}(u_t - \lambda X_t).\end{aligned}$$

Thus the two exact expressions are

$$\boxed{X_{t+1} - X_t = \eta_t(u_t - \lambda X_{t+1}) = \frac{\eta_t}{1 + \lambda\eta_t}(u_t - \lambda X_t).}$$

The expression without the denominator contains X_{t+1}, whereas the expression with the denominator contains X_t. The time index and the denominator must be kept together as displayed; combining the denominator from one expression with the time index from the other does not give the same identity.

We now bound the increment:

$$\begin{aligned}
\|X_{t+1} - X_t\|_F &= \eta_t\|u_t - \lambda X_{t+1}\|_F \\
&\le \eta_t\left(\|u_t\|_F + \lambda\|X_{t+1}\|_F\right) && \text{(triangle inequality)} \\
&\le \eta_t(U + \lambda R) && \text{(the two bounds above).}
\end{aligned}$$

We abbreviate the recurring constant $U + \lambda R$ by

$$\boxed{B := U + \lambda R}$$

It follows that

$$\boxed{\|X_{t+1} - X_t\|_F \le B\eta_t.}$$

This definition introduces no additional assumption on B. By L-smoothness,

$$\begin{aligned}
\|g_{t+1} - g_t\|_F &= \|\nabla\mathcal{F}(X_{t+1}) - \nabla\mathcal{F}(X_t)\|_F \\
&\le L\|X_{t+1} - X_t\|_F \\
&\le LB\eta_t.
\end{aligned}$$

That is,

$$\boxed{\|g_{t+1} - g_t\|_F \le LB\eta_t.}$$

In addition to the gradient increments, we need a bound on g_t itself. Add and subtract the same reference gradient $\nabla\mathcal{F}(0)$:

$$g_t = \nabla\mathcal{F}(0) + \big[\nabla\mathcal{F}(X_t) - \nabla\mathcal{F}(0)\big].$$

Then

$$\begin{aligned}
\|g_t\|_F &= \|\nabla\mathcal{F}(0) + [\nabla\mathcal{F}(X_t) - \nabla\mathcal{F}(0)]\|_F \\
&\le \|\nabla\mathcal{F}(0)\|_F + \|\nabla\mathcal{F}(X_t) - \nabla\mathcal{F}(0)\|_F && \text{(triangle inequality)} \\
&\le \|\nabla\mathcal{F}(0)\|_F + L\|X_t - 0\|_F && (L\text{-smoothness}) \\
&= \|\nabla\mathcal{F}(0)\|_F + L\|X_t\|_F \\
&\le \|\nabla\mathcal{F}(0)\|_F + LR.
\end{aligned}$$

Give the final constant a name:

$$\boxed{G_R := \|\nabla\mathcal{F}(0)\|_F + LR, \qquad \|g_t\|_F \le G_R.}$$

Here 0 is simply a convenient fixed reference point. We do not assume $\nabla\mathcal{F}(0) = 0$.

For the momentum, the intermediate estimates are

$$\begin{aligned}\|M_{t+1}\|_F &= \|\beta_2 M_t - (1-\beta_2) g_t\|_F \\ &\le \beta_2 \|M_t\|_F + (1-\beta_2)\|g_t\|_F \\ &\le \beta_2 \|M_t\|_F + (1-\beta_2) G_R.\end{aligned}$$

Set

$$M_{\max} := \max\{\|M_0\|_F, G_R\}$$

If $\|M_t\|_F \le M_{\max}$, then

$$\begin{aligned}\|M_{t+1}\|_F &\le \beta_2 M_{\max} + (1-\beta_2) M_{\max} \\ &= M_{\max}.\end{aligned}$$

The bound also holds at initialization. Induction therefore yields

$$\|M_t\|_F \le M_{\max} \qquad \forall t.$$

Consequently,

$$0 \le q_t = \mathcal{K}_t(M_t) \le U\|M_t\|_F \le U M_{\max}.$$

$$\boxed{C_K := U M_{\max}, \qquad 0 \le q_t \le C_K.}$$

We now show that the learning-rate-weighted momentum increments have a finite sum.

Define

$$\beta := \beta_2, \qquad e_t := M_t + g_t$$

The error e_t measures the deviation of M_t from the current negative gradient $-g_t$. The momentum recursion gives

$$\begin{aligned}e_{t+1} &= M_{t+1} + g_{t+1} \\ &= \beta M_t - (1-\beta) g_t + g_{t+1} \\ &= \beta(M_t + g_t) + (g_{t+1} - g_t) \\ &= \beta e_t + (g_{t+1} - g_t).\end{aligned}$$

Also,

$$\begin{aligned}M_{t+1} - M_t &= \beta M_t - (1-\beta) g_t - M_t \\ &= -(1-\beta)(M_t + g_t) \\ &= -(1-\beta) e_t.\end{aligned}$$

Hence,

$$\|M_{t+1} - M_t\|_F = (1-\beta)\|e_t\|_F.$$

Repeated substitution in the error recursion gives

$$e_t = \beta^t e_0 + \sum_{s=0}^{t-1} \beta^{t-1-s}(g_{s+1} - g_s).$$

Using the triangle inequality and the preceding gradient-increment bound, we obtain

$$\begin{aligned}\|e_t\|_F &\le \beta^t\|e_0\|_F + \sum_{s=0}^{t-1}\beta^{t-1-s}\|g_{s+1}-g_s\|_F \\ &\le \beta^t\|e_0\|_F + LB\sum_{s=0}^{t-1}\beta^{t-1-s}\eta_s.\end{aligned}$$

For an integer $n \ge 1$, let

$$S_{M,n} := \sum_{t=0}^{n}\eta_{t+1}\|M_{t+1}-M_t\|_F$$

Substituting the preceding inequality gives

$$\begin{aligned}S_{M,n} \le{}& (1-\beta)\|e_0\|_F\sum_{t=0}^{n}\eta_{t+1}\beta^t \\ &+ (1-\beta)LB\sum_{t=0}^{n}\sum_{s=0}^{t-1}\eta_{t+1}\beta^{t-1-s}\eta_s.\end{aligned}$$

We handle the two terms separately.

Since the learning rates are nonincreasing, $\eta_{t+1} \le \eta_0$ for every $t \ge 0$. Therefore,

$$\begin{aligned}&(1-\beta)\|e_0\|_F\sum_{t=0}^{n}\eta_{t+1}\beta^t \\ &\le (1-\beta)\|e_0\|_F\sum_{t=0}^{n}\eta_0\beta^t \\ &= (1-\beta)\|e_0\|_F\eta_0\sum_{t=0}^{n}\beta^t \\ &= (1-\beta)\|e_0\|_F\eta_0\frac{1-\beta^{n+1}}{1-\beta} \\ &= \eta_0\|e_0\|_F(1-\beta^{n+1}) \\ &\le \boxed{\eta_0\|e_0\|_F}.\end{aligned}$$

For the second term, write

$$a_{t,s} := \eta_{t+1}\beta^{t-1-s}\eta_s \qquad (0 \le s < t)$$

We prove by mathematical induction that, for every integer $n \ge 1$,

$$\boxed{\sum_{t=0}^{n}\sum_{s=0}^{t-1}a_{t,s} = \sum_{s=0}^{n-1}\sum_{t=s+1}^{n}a_{t,s}} \tag{463}$$

At $t = 0$, the inner sum is empty and is defined to be zero.

For the base case $n = 1$, the left-hand side is

$$\sum_{t=0}^{1}\sum_{s=0}^{t-1} a_{t,s} = 0 + a_{1,0} = a_{1,0},$$

and the right-hand side is

$$\sum_{s=0}^{0}\sum_{t=s+1}^{1} a_{t,s} = \sum_{t=1}^{1} a_{t,0} = a_{1,0}.$$

Thus (463) holds for $n = 1$.

For the induction hypothesis, suppose that, for some integer $n \geq 1$,

$$\sum_{t=0}^{n}\sum_{s=0}^{t-1} a_{t,s} = \sum_{s=0}^{n-1}\sum_{t=s+1}^{n} a_{t,s}$$

For the induction step, separate the last term, corresponding to $t = n + 1$, on the left-hand side:

$$\sum_{t=0}^{n+1}\sum_{s=0}^{t-1} a_{t,s} = \sum_{t=0}^{n}\sum_{s=0}^{t-1} a_{t,s} + \sum_{s=0}^{n} a_{n+1,s}.$$

Applying the induction hypothesis to the first sum gives

$$= \sum_{s=0}^{n-1}\sum_{t=s+1}^{n} a_{t,s} + \sum_{s=0}^{n} a_{n+1,s}.$$

Separate the term $s = n$ in the second sum:

$$= \sum_{s=0}^{n-1}\sum_{t=s+1}^{n} a_{t,s} + \sum_{s=0}^{n-1} a_{n+1,s} + a_{n+1,n}.$$

Both of the first two sums run over $s = 0, \ldots, n-1$, so we can group terms with the same s:

$$= \sum_{s=0}^{n-1}\left(\sum_{t=s+1}^{n} a_{t,s} + a_{n+1,s}\right) + a_{n+1,n}.$$

The term $a_{n+1,s}$ in parentheses adds the $t = n + 1$ term to the inner sum. Thus the expression becomes

$$= \sum_{s=0}^{n-1}\sum_{t=s+1}^{n+1} a_{t,s} + a_{n+1,n}.$$

The final term can be written as

$$a_{n+1,n} = \sum_{t=n+1}^{n+1} a_{t,n}$$

which is exactly the $s = n$ term of the outer sum. Therefore,

$$= \sum_{s=0}^{n} \sum_{t=s+1}^{n+1} a_{t,s}.$$

In other words,

$$\boxed{\sum_{t=0}^{n+1} \sum_{s=0}^{t-1} a_{t,s} = \sum_{s=0}^{n} \sum_{t=s+1}^{n+1} a_{t,s}}$$

so the identity holds for $n+1$ as well. By mathematical induction, (463) holds for every integer $n \geq 1$.

This rearrangement is an identity for finite sums. It requires neither nonnegativity of the terms nor convergence of an infinite series. Nonnegativity will be used later, when we pass from the resulting bound to the infinite sum.

Substituting the original summands back into the identity gives

$$\begin{aligned}
&\sum_{t=0}^{n} \sum_{s=0}^{t-1} \eta_{t+1} \beta^{t-1-s} \eta_s \\
&= \sum_{s=0}^{n-1} \sum_{t=s+1}^{n} \eta_{t+1} \beta^{t-1-s} \eta_s \\
&= \sum_{s=0}^{n-1} \eta_s \sum_{t=s+1}^{n} \eta_{t+1} \beta^{t-1-s}.
\end{aligned}$$

In the last line, s is fixed within the inner sum, so η_s can be taken outside it. In that inner sum, $t \geq s+1$. Since the learning rates are nonincreasing,

$$\eta_{t+1} \leq \eta_s.$$

Therefore,

$$\begin{aligned}
&\sum_{s=0}^{n-1} \eta_s \sum_{t=s+1}^{n} \eta_{t+1} \beta^{t-1-s} \\
&\leq \sum_{s=0}^{n-1} \eta_s \sum_{t=s+1}^{n} \eta_s \beta^{t-1-s} \\
&= \sum_{s=0}^{n-1} \eta_s^2 \sum_{t=s+1}^{n} \beta^{t-1-s}.
\end{aligned}$$

For the inner geometric sum, substitute $k = t-1-s$. When $t = s+1$, we have $k = 0$; when $t = n$, we have $k = n-1-s$. Hence,

$$\sum_{t=s+1}^{n} \beta^{t-1-s} = \sum_{k=0}^{n-1-s} \beta^k = \frac{1-\beta^{n-s}}{1-\beta} \leq \frac{1}{1-\beta}.$$

The second term is consequently bounded by

$$
\begin{aligned}
&(1-\beta)LB\sum_{t=0}^{n}\sum_{s=0}^{t-1}\eta_{t+1}\beta^{t-1-s}\eta_s \\
&\le (1-\beta)LB\sum_{s=0}^{n-1}\eta_s^2\frac{1}{1-\beta} \\
&= LB\sum_{s=0}^{n-1}\eta_s^2 \\
&\le LBS_2.
\end{aligned}
$$

Combining this with the bound on the first term, we obtain, for every $n \ge 1$,

$$\boxed{S_{M,n} \le \eta_0\|e_0\|_F + LBS_2.}$$

We now pass to the infinite sum. Each summand

$$\eta_{t+1}\|M_{t+1} - M_t\|_F$$

is nonnegative, so $S_{M,n}$ is nondecreasing in n. It is also bounded above by the finite constant just obtained. It therefore has a finite limit, and

$$\boxed{S_M := \sum_{t=0}^{\infty}\eta_{t+1}\|M_{t+1} - M_t\|_F \le \eta_0\|e_0\|_F + LBS_2 < \infty.}$$

We have not presumed an interchange of infinite sums. We rearranged finite sums exactly, obtained a uniform bound, and then took the limit of the nonnegative partial sums.

We next relate the weights in the variation sum to the learning rates.

Recall the coefficient definition

$$c_t = \frac{\lambda\beta_1\eta_t}{(1-\beta_2) + \lambda(1-\beta_1)\eta_t}$$

Thus,

$$w_t = \frac{c_{t+1}}{\lambda} = \frac{\beta_1\eta_{t+1}}{(1-\beta_2) + \lambda(1-\beta_1)\eta_{t+1}}.$$

Since the denominator is at least $1-\beta_2 > 0$,

$$0 \le w_t \le \frac{\beta_1\eta_{t+1}}{1-\beta_2}.$$

$$\boxed{\kappa := \frac{\beta_1}{1-\beta_2}, \qquad 0 \le w_t \le \kappa\eta_{t+1}.}$$

For the function

$$f(\eta) := \frac{\beta_1\eta}{(1-\beta_2) + \lambda(1-\beta_1)\eta}$$

we have

$$f'(\eta) = \frac{\beta_1(1-\beta_2)}{\left[(1-\beta_2)+\lambda(1-\beta_1)\eta\right]^2} \geq 0.$$

Thus f is nondecreasing. Since $\eta_{t+2} \leq \eta_{t+1}$,

$$w_{t+1} = f(\eta_{t+2}) \leq f(\eta_{t+1}) = w_t.$$

That is,

$$\boxed{w_{t+1} \leq w_t.}$$

Furthermore, $\sum_t \eta_t^2 < \infty$ implies $\eta_t^2 \to 0$, and hence $\eta_t \to 0$. Consequently,

$$0 \leq w_t \leq \kappa\eta_{t+1} \to 0$$

which also gives $w_t \to 0$.

We return to A_T and apply summation by parts.

Recall that

$$A_T := \sum_{t=0}^{T-1} w_t(q_{t+1} - q_t)$$

Splitting the sum into two parts gives

$$A_T = \sum_{t=0}^{T-1} w_t q_{t+1} - \sum_{t=0}^{T-1} w_t q_t.$$

In the first sum, change the index by setting $j = t+1$:

$$\sum_{t=0}^{T-1} w_t q_{t+1} = \sum_{j=1}^{T} w_{j-1} q_j = w_{T-1}q_T + \sum_{j=1}^{T-1} w_{j-1} q_j.$$

In the second sum, separate the first term:

$$\sum_{t=0}^{T-1} w_t q_t = w_0 q_0 + \sum_{j=1}^{T-1} w_j q_j.$$

Subtracting these two expressions yields

$$\boxed{A_T = w_{T-1}q_T - w_0 q_0 + \sum_{j=1}^{T-1}(w_{j-1} - w_j)q_j.}$$

This is the Abel summation identity. We have already established that

$$0 \leq q_j \leq C_K, \qquad w_{j-1} - w_j \geq 0$$

Thus the individual terms satisfy

$$w_{T-1}q_T \le w_{T-1}C_K,$$

$$-w_0 q_0 \le 0,$$

$$(w_{j-1} - w_j)q_j \le (w_{j-1} - w_j)C_K.$$

Substituting these estimates gives

$$\begin{aligned} A_T &\le w_{T-1}C_K + \sum_{j=1}^{T-1}(w_{j-1} - w_j)C_K \\ &= C_K \left[w_{T-1} + \sum_{j=1}^{T-1}(w_{j-1} - w_j) \right]. \end{aligned}$$

Here,

$$\begin{aligned} \sum_{j=1}^{T-1}(w_{j-1} - w_j) &= (w_0 - w_1) + (w_1 - w_2) + \cdots + (w_{T-2} - w_{T-1}) \\ &= w_0 - w_{T-1}. \end{aligned}$$

Therefore,

$$\begin{aligned} A_T &\le C_K[w_{T-1} + w_0 - w_{T-1}] \\ &= \boxed{w_0 C_K}. \end{aligned}$$

When $T = 1$, the intermediate sum is empty and the same identity and bound remain valid.

It remains to control E_T and combine the two estimates.

Recall that

$$E_T := \sum_{t=0}^{T-1} w_t D_t, \qquad D_t := \mathcal{K}_t(M_{t+1}) - \mathcal{K}_t(M_t)$$

The quantity D_t is real-valued, so its magnitude is expressed by the absolute value $|D_t|$. By the Lipschitz property of the same fixed potential $\mathcal{K}_t$,

$$\boxed{|D_t| = |\mathcal{K}_t(M_{t+1}) - \mathcal{K}_t(M_t)| \le U\|M_{t+1} - M_t\|_F.}$$

For every real number E_T, we have

$$-E_T \le |E_T|$$

and therefore

$$A_T - E_T \le A_T + |E_T|$$

The next step separately uses the triangle inequality:

$$
\begin{aligned}
|E_T| &= \left| \sum_{t=0}^{T-1} w_t D_t \right| \\
&\le \sum_{t=0}^{T-1} |w_t D_t| \quad \text{(triangle inequality)} \\
&= \sum_{t=0}^{T-1} w_t |D_t| \quad (w_t \ge 0)
\end{aligned}
$$

Writing both steps explicitly gives

$$
\boxed{A_T - E_T \le A_T + |E_T| \le A_T + \sum_{t=0}^{T-1} w_t |D_t|.}
$$

The inequality

$$
E_T \le \sum_{t=0}^{T-1} w_t |D_t|
$$

is also valid. However, the term appearing in V_T is $-E_T$. An upper bound on E_T alone does not give an upper bound on $-E_T$. We must therefore either pass through the absolute value, as above, or bound $-E_T$ directly.

For every real number D_t,

$$
-D_t \le |D_t|
$$

and, since $w_t \ge 0$,

$$
-w_t D_t \le w_t |D_t|.
$$

Consequently,

$$
\begin{aligned}
V_T &= A_T - E_T \\
&= A_T + \sum_{t=0}^{T-1} w_t (-D_t) \\
&\le A_T + \sum_{t=0}^{T-1} w_t |D_t| \\
&\le w_0 C_K + U \sum_{t=0}^{T-1} w_t \|M_{t+1} - M_t\|_F \\
&\le w_0 C_K + \kappa U \sum_{t=0}^{T-1} \eta_{t+1} \|M_{t+1} - M_t\|_F \\
&\le w_0 C_K + \kappa U \sum_{t=0}^{\infty} \eta_{t+1} \|M_{t+1} - M_t\|_F \\
&\le w_0 C_K + \kappa U \left(\eta_0 \|e_0\|_F + LBS_2 \right).
\end{aligned}
$$

Here the first inequality uses $-D_t \le |D_t|$ and $w_t \ge 0$. The next uses $A_T \le w_0 C_K$ and $|D_t| \le U\|M_{t+1} - M_t\|_F$, followed by $w_t \le \kappa \eta_{t+1}$. The finite sum is then bounded by the corresponding

infinite sum of nonnegative terms, and the last inequality uses the weighted momentum-increment estimate established above. Thus, defining

$$\boxed{C_{\mathrm{var}} := w_0 C_K + \kappa U\left(\eta_0 \|e_0\|_F + LBS_2\right)} \tag{464}$$

we obtain

$$\boxed{V_T \le C_{\mathrm{var}} < \infty \qquad \forall T \ge 1.} \tag{465}$$

All constants have been defined above and are independent of T. This proves the weighted signed primal-variation bound (376).

The common-domain, no-spike, and limiting-gap conditions are not used in this argument. They retain their separate roles in the main convergence analysis.

D.2 The No-Spike Condition for a Static Potential

We verify that the no-spike condition (314) holds automatically when $\mathcal{K}_t \equiv \mathcal{K}$.

Under the common-domain condition, which in the static case follows from $\lambda X_0 \in \mathcal{C}$ by Appendix C.1,

$$\mathcal{K}^*(\lambda X_t) = 0, \qquad \forall t \ge 0, \tag{466}$$

so

$$G_t^{(t)} = \mathcal{K}(\widetilde{M}_{t+1}) - \left\langle \lambda X_{t+1}, \widetilde{M}_{t+1} \right\rangle_F. \tag{467}$$

Moreover, convexity, positive homogeneity, and (238) imply that $\mathcal{K}$ is $\sqrt{N}$-Lipschitz. Indeed, convexity and positive homogeneity imply subadditivity, and hence

$$\mathcal{K}(A) \le \mathcal{K}(B) + \mathcal{K}(A-B) \le \mathcal{K}(B) + \sqrt{N}\|A-B\|_F. \tag{468}$$

Interchanging A and B gives

$$\mathcal{K}(B) \le \mathcal{K}(A) + \sqrt{N}\|A-B\|_F. \tag{469}$$

Therefore,

$$|\mathcal{K}(A) - \mathcal{K}(B)| \le \sqrt{N}\|A-B\|_F. \tag{470}$$

Let

$$E_t := \widetilde{M}_t + \nabla\mathcal{F}(X_t), \qquad t \ge 1. \tag{471}$$

By (438),

$$\|E_t\|_F \to 0. \tag{472}$$

The derivation of (438) does not use the no-spike condition.

Let $\rho > 0$ be arbitrary for the moment, and let $s \ge t \ge 0$ be arbitrary integers satisfying

$$\sum_{r=t}^{s-1} \eta_r \le \rho. \tag{473}$$

By telescoping, the triangle inequality, (420), and the monotonicity of the learning rates,

$$\begin{aligned} \|X_{s+1} - X_{t+1}\|_F &\le 2\sqrt{N} \sum_{r=t+1}^{s} \eta_r \\ &\le 2\sqrt{N} \sum_{r=t}^{s-1} \eta_r \\ &\le 2\sqrt{N}\,\rho. \end{aligned} \tag{474}$$

By L-smoothness, it follows that

$$\|\nabla\mathcal{F}(X_{s+1}) - \nabla\mathcal{F}(X_{t+1})\|_F \le 2L\sqrt{N}\,\rho. \tag{475}$$

Therefore,

$$\|\widetilde{M}_{s+1} - \widetilde{M}_{t+1}\|_F \le \|E_{s+1}\|_F + \|E_{t+1}\|_F + 2L\sqrt{N}\,\rho. \tag{476}$$

Using the uniform bounds (319) and (325), we obtain

$$\begin{aligned} |G_s^{(s)} - G_t^{(t)}| \le (\sqrt{N} + \lambda B_X)\|\widetilde{M}_{s+1} - \widetilde{M}_{t+1}\|_F \\ + \lambda B_M \|X_{s+1} - X_{t+1}\|_F. \end{aligned} \tag{477}$$

Consequently,

$$\begin{aligned} |G_s^{(s)} - G_t^{(t)}| \le (\sqrt{N} + \lambda B_X)\left(\|E_{s+1}\|_F + \|E_{t+1}\|_F + 2L\sqrt{N}\,\rho\right) \\ + 2\lambda B_M \sqrt{N}\,\rho. \end{aligned} \tag{478}$$

Given $\varepsilon > 0$, recall that $E_t \to 0$ means that, for every $\delta > 0$, there exists an integer $T \ge 1$ such that

$$r \ge T \quad \Longrightarrow \quad \|E_r\|_F \le \delta. \tag{479}$$

To make the contribution of the two E-terms at most $\varepsilon/2$, choose

$$\delta := \frac{\varepsilon}{4(\sqrt{N} + \lambda B_X)}. \tag{480}$$

Accordingly, choose an integer $T_\varepsilon \ge 0$ sufficiently large so that

$$\|E_r\|_F \le \frac{\varepsilon}{4(\sqrt{N} + \lambda B_X)}, \qquad \forall r \ge T_\varepsilon + 1. \tag{481}$$

Define

$$C_{\mathrm{ns}} := L(\sqrt{N} + \lambda B_X) + \lambda B_M. \tag{482}$$

The remaining ρ-dependent contribution is

$$2\sqrt{N}\,\rho C_{\mathrm{ns}}. \tag{483}$$

Since the preceding estimate holds for every $\rho > 0$, when $C_{\mathrm{ns}} > 0$, choose

$$\rho := \frac{\varepsilon}{4\sqrt{N}\,C_{\mathrm{ns}}}. \tag{484}$$

Then the ρ-dependent contribution satisfies

$$2\sqrt{N}\,\rho C_{\mathrm{ns}} = \frac{\varepsilon}{2}. \tag{485}$$

If $C_{\mathrm{ns}} = 0$, the ρ-dependent contribution vanishes identically, so any $\rho > 0$ may be chosen.

Then, whenever $s \ge t \ge T_\varepsilon$ and

$$\sum_{r=t}^{s-1} \eta_r \le \rho, \tag{486}$$

we have

$$\begin{aligned} |G_s^{(s)} - G_t^{(t)}| &\le (\sqrt{N} + \lambda B_X)\left(\|E_{s+1}\|_F + \|E_{t+1}\|_F\right) \\ &\quad + 2\sqrt{N}\,\rho C_{\mathrm{ns}} \\ &\le \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned} \tag{487}$$

Thus the no-spike condition holds for a static potential.

D.3 Relation to the Practical Single-β Nesterov Form

The implemented algorithms use a single momentum coefficient β. To distinguish the practical momentum variable from the notation used in the convergence analysis above, let P_t denote the practical momentum and let G_t denote the corresponding gradient. The implemented Nesterov form is

$$P_t = \beta P_{t-1} + (1-\beta)G_t, \tag{488}$$
$$N_t = G_t + \beta(P_t - G_t). \tag{489}$$

Expanding the second expression and substituting the momentum recursion gives

$$N_t = \beta P_t + (1-\beta)G_t \tag{490}$$
$$= \beta\left(\beta P_{t-1} + (1-\beta)G_t\right) + (1-\beta)G_t \tag{491}$$
$$= \beta^2 P_{t-1} + (1-\beta^2)G_t. \tag{492}$$

Now set

$$G_t = \nabla F(X_t), \qquad M_t = -P_{t-1}.$$

Then, with

$$\beta_2 = \beta, \qquad \beta_1 = \beta^2,$$

the two momentum variables used in the convergence analysis above satisfy

$$M_{t+1} = \beta_2 M_t - (1-\beta_2)\nabla F(X_t) = -P_t, \tag{493}$$
$$\widetilde{M}_{t+1} = \beta_1 M_t - (1-\beta_1)\nabla F(X_t) = -N_t. \tag{494}$$

Hence, up to the sign convention used for the update direction and the one-step indexing of the momentum variable, the practical single-β Nesterov form is exactly the specialization

$$\boxed{\beta_2 = \beta, \qquad \beta_1 = \beta^2}$$

of the two-coefficient momentum parameterization analyzed above.

Moreover, for $0 < \beta < 1$,

$$0 \le \beta_1 = \beta^2 < \beta = \beta_2 < 1,$$

so the standing condition $0 \le \beta_1 < \beta_2 < 1$ is automatically satisfied. The relation between the sign convention for the theoretical update direction and that used in the implemented parameter update is made explicit in Appendix D.4.

D.4 Relation to the Implemented Decoupled-Weight-Decay Update

The parameter dynamics analyzed in Appendix D use the implicit weight-decay parameterization

$$X_{t+1} = X_t + \eta_t\left(D_t - \lambda X_{t+1}\right), \qquad D_t := \nabla \mathcal{K}_t(\widetilde{M}_{t+1}). \tag{495}$$

This parameterization is exactly equivalent, under a reparameterization of the learning rate, to the explicit decoupled-weight-decay parameterization for the same selected-subgradient update-direction sequence. To distinguish the learning rate in the implicit parameterization from that in the corresponding explicit decoupled-weight-decay parameterization, we denote the latter by h_t throughout this subsection.

Indeed,

$$(1+\eta_t\lambda)X_{t+1} = X_t + \eta_t D_t, \tag{496}$$

and hence

$$X_{t+1} = \frac{1}{1+\eta_t\lambda}X_t + \frac{\eta_t}{1+\eta_t\lambda}D_t. \tag{497}$$

Define

$$h_t := \frac{\eta_t}{1+\eta_t\lambda}. \tag{498}$$

Then

$$1 - h_t\lambda = \frac{1}{1+\eta_t\lambda}, \tag{499}$$

so the update becomes

$$X_{t+1} = (1 - h_t\lambda)X_t + h_t D_t. \tag{500}$$

To express the same update-direction sequence using the sign convention of the implemented algorithms, define

$$U_t := -D_t. \tag{501}$$

For the concrete update rules whose practical implementations are related here, namely balanced SGD, Lion, Muon, SPU, and idealized ROBN, the selected update map is odd under sign reversal at each fixed time:

$$\nabla\mathcal{K}_t(-M) = -\nabla\mathcal{K}_t(M). \tag{502}$$

This follows from $\operatorname{sign}(-M) = -\operatorname{sign}(M)$ for Lion, from the normalized direction $M/\|M\|_F$ for balanced SGD, and from $\operatorname{Polar}(-M) = -\operatorname{Polar}(M)$ for Muon and, blockwise, for SPU. For the idealized ROBN update, with the same sampled orthogonal matrix A_t held fixed, sign reversal leaves the row- or column-normalization denominators unchanged and negates the resulting normalized update. Since the preceding subsection gives

$$\widetilde{M}_{t+1} = -N_t, \tag{503}$$

we therefore have

$$\begin{aligned} U_t &= -D_t \\ &= -\nabla\mathcal{K}_t(\widetilde{M}_{t+1}) \\ &= -\nabla\mathcal{K}_t(-N_t) \\ &= \nabla\mathcal{K}_t(N_t). \end{aligned} \tag{504}$$

The recursion then becomes

$$X_{t+1} = (1 - h_t\lambda)X_t - h_t U_t, \tag{505}$$

which is exactly the corresponding explicit decoupled-weight-decay parameterization for the same update-direction sequence. Thus, under the learning-rate reparameterization above and the identification $U_t = -D_t$, the implicit parameterization used in the convergence analysis and the corresponding explicit decoupled-weight-decay parameterization generate exactly the same parameter iterates from the same initialization for the same update-direction sequence.

Conversely, whenever $h_t\lambda < 1$,

$$\eta_t = \frac{h_t}{1 - h_t\lambda}. \tag{506}$$

The condition $h_t\lambda < 1$ ensures that the inverse transformation is well-defined with a positive denominator. Therefore, the two parameterizations are in one-to-one correspondence.

Moreover, if

$$0 \le h_t\lambda \le \rho < 1 \qquad \text{for all } t, \tag{507}$$

then

$$h_t \le \eta_t = \frac{h_t}{1 - h_t\lambda} \le \frac{1}{1 - \rho} h_t. \tag{508}$$

Thus, the explicit and implicit learning rates are uniformly comparable up to a fixed multiplicative constant. Let

$$C_\rho := \frac{1}{1 - \rho} < \infty. \tag{509}$$

Then

$$0 \le h_t \le \eta_t \le C_\rho h_t. \tag{510}$$

First, if $h_t \to 0$, then

$$0 \le \eta_t \le C_\rho h_t \to 0, \tag{511}$$

so $\eta_t \to 0$. Conversely, if $\eta_t \to 0$, then

$$0 \le h_t \le \eta_t \to 0, \tag{512}$$

so $h_t \to 0$. Hence,

$$h_t \to 0 \iff \eta_t \to 0. \tag{513}$$

Next, since $h_t \le \eta_t$,

$$\sum_{t=0}^{\infty} h_t = \infty \quad \Longrightarrow \quad \sum_{t=0}^{\infty} \eta_t = \infty. \tag{514}$$

Conversely, if

$$\sum_{t=0}^{\infty} h_t < \infty, \tag{515}$$

then

$$\sum_{t=0}^{\infty} \eta_t \le C_\rho \sum_{t=0}^{\infty} h_t < \infty. \tag{516}$$

Taking the contrapositive of the preceding implication gives

$$\sum_{t=0}^{\infty} \eta_t = \infty \quad \Longrightarrow \quad \sum_{t=0}^{\infty} h_t = \infty. \tag{517}$$

Therefore,

$$\sum_{t=0}^{\infty} h_t = \infty \iff \sum_{t=0}^{\infty} \eta_t = \infty. \tag{518}$$

For the square-summability condition, squaring the uniform comparison gives

$$h_t^2 \le \eta_t^2 \le C_\rho^2 h_t^2. \tag{519}$$

Hence,

$$\sum_{t=0}^{\infty} h_t^2 < \infty \iff \sum_{t=0}^{\infty} \eta_t^2 < \infty. \tag{520}$$

Finally, define

$$f(h) := \frac{h}{1 - \lambda h}. \tag{521}$$

For $0 \le h < 1/\lambda$,

$$f'(h) = \frac{1}{(1 - \lambda h)^2} > 0, \tag{522}$$

so f is strictly increasing. Therefore, if

$$h_{t+1} \le h_t, \tag{523}$$

then

$$\eta_{t+1} = f(h_{t+1}) \le f(h_t) = \eta_t. \tag{524}$$

Hence, a nonincreasing explicit learning-rate schedule $\{h_t\}$ induces a nonincreasing implicit learning-rate sequence $\{\eta_t\}$.

E Applicability of the Momentum Convergence Analysis to ROBN

We explain how ROBN fits the convergence analysis in Appendix D and identify the conditions under which its convergence conclusion applies.

For the idealized ROBN update, with an exactly orthogonal A_t and exact row- or column-normalization ($\varepsilon = 0$), at each fixed t, the sampled orthogonal matrix is held fixed, and the resulting coefficient-wise potential $\mathcal{K}_t$ is a norm. Hence, $\mathcal{K}_t$ is proper, closed, convex, and positively homogeneous. The associated normalized update direction is the selected subgradient of $\mathcal{K}_t$ specified by the selection rule in Section 3. Equivalently, at each fixed t, the selected ROBN direction is a maximizer of the linear objective induced by the momentum over the dual unit ball associated with $\mathcal{K}_t$, or, up to sign, a corresponding linear minimization oracle (LMO) direction. ROBN does not use a single

norm fixed across iterations: $\mathcal{K}_t$ is determined by the selection rule through the sampled A_t and generally changes across iterations. Thus, ROBN is not obtained by fixing a single norm and repeatedly applying its associated LMO. Appendix D is formulated to account for such time variation under its stated assumptions. The finite-step Newton–Schulz orthogonalization used by ROBN-Haar and the ε-stabilized normalization used in the practical implementation are implementation approximations and are not covered by this exact selected-subgradient convergence result. Thus, these idealized ROBN potentials satisfy the structural assumptions of Appendix C, as well as the uniform bounds in (238).

Consider an idealized ROBN trajectory satisfying the assumptions required for the weighted signed primal-variation bound in Appendix D.1, including the exact-gradient dynamics, parameter conditions, smoothness condition, and learning-rate conditions. The bound proved there gives

$$\sum_{t=0}^{T-1} \frac{c_{t+1}}{\lambda} \left[\mathcal{K}_{t+1}(M_{t+1}) - \mathcal{K}_t(M_{t+1})\right] \le C_{\mathrm{var}} < \infty, \qquad \forall T \ge 1, \tag{525}$$

where C_{var} is independent of T and may be chosen as in (464). Thus the weighted signed primal-variation bound (376) is a derived result for these idealized ROBN trajectories. No additional upper bound on the unweighted cumulative primal variation is required.

To apply the full convergence result of Appendix D, we retain its remaining standing assumptions. Unlike the static setting of Appendix C.1, ROBN generally induces a time-dependent potential because its orthogonal matrix A_t is resampled across iterations. Therefore, the static forward-invariance argument does not imply the common-domain condition from the initialization alone. In particular, the trajectory generated with the fixed value of λ must satisfy the common-domain condition (310). The no-spike condition (313)–(314) and the limiting-gap condition (316), with a single common limiting potential $\mathcal{K}_\infty$, are explicit assumptions on the gap behavior along the ROBN trajectory.

The weighted signed primal-variation estimate does not by itself imply the no-spike and limiting-gap conditions. Accordingly, we consider ROBN trajectories for which these two remaining gap conditions of Appendix D hold. These conditions concern only the gap behavior along the trajectory and do not require convergence of the full sequence of potentials.

In particular, the per-step random resampling used by ROBN is compatible with this formulation: the convergence conclusion applies conditionally to those ROBN trajectories that satisfy the common-domain condition and whose induced gap sequence satisfies the no-spike and limiting-gap assumptions stated above.

Under all the standing assumptions of Appendix D, the preceding applicability argument allows us to use its conclusion (462) for the corresponding ROBN trajectory:

$$d(X_t, \mathrm{KKT}) \to 0. \tag{526}$$

where KKT is the set defined in (454) for the constrained problem induced by $\mathcal{K}_\infty$.

F ALGORITHMIC DETAILS OF RANDOM ORTHOGONAL BASIS NORMALIZATION (ROBN)

Algorithm 1: ROBN-Haar for Matrix Parameters

given: learning rates $\{\eta_t\}_{t\in\mathbb{N}} \subset \mathbb{R}_{>0}$, momentum coefficient $\beta \in [0, 1)$, weight decay $\lambda \geq 0$, matrix parameters $\{X_1^{(\ell)}\}_{\ell=1}^{L}$ with $X_1^{(\ell)} \in \mathbb{R}^{m_\ell \times n_\ell}$, initial momenta $M_0^{(\ell)} = 0$ for all ℓ, number of Newton–Schulz steps K, numerical stability constant $\varepsilon > 0$

initialize time step $t \leftarrow 1$;

while *not converged* **do**

  compute stochastic gradients $G_t^{(\ell)} \leftarrow \text{STOCHASTICGRADIENT}(X_t^{(\ell)})$ for all $\ell = 1, \ldots, L$;

  for $\ell = 1, \ldots, L$ **do**

    $M_t^{(\ell)} \leftarrow \beta M_{t-1}^{(\ell)} + (1-\beta) G_t^{(\ell)}$;

    $N_t^{(\ell)} \leftarrow G_t^{(\ell)} + \beta \left(M_t^{(\ell)} - G_t^{(\ell)} \right)$; `// Nesterov momentum direction`

    $\Delta_t^{(\ell)} \leftarrow N_t^{(\ell)}$; `// raw update direction`

    if $m_\ell \geq n_\ell$ **then**

      sample $Z_t^{(\ell)} \in \mathbb{R}^{m_\ell \times m_\ell}$ with i.i.d. standard Gaussian entries;

      $A_t^{(\ell)} \leftarrow \text{NEWTONSCHULZORTHOGONALIZE}(Z_t^{(\ell)}, K)$;

      $\widetilde{\Delta}_t^{(\ell)} \leftarrow (A_t^{(\ell)})^\top \Delta_t^{(\ell)}$;

      $\widehat{\Delta}_{t,i:}^{(\ell)} \leftarrow \dfrac{\widetilde{\Delta}_{t,i:}^{(\ell)}}{\|\widetilde{\Delta}_{t,i:}^{(\ell)}\|_2 + \varepsilon}$ for $i = 1, \ldots, m_\ell$;

      $U_t^{(\ell)} \leftarrow A_t^{(\ell)} \widehat{\Delta}_t^{(\ell)}$;

    else

      sample $Z_t^{(\ell)} \in \mathbb{R}^{n_\ell \times n_\ell}$ with i.i.d. standard Gaussian entries;

      $A_t^{(\ell)} \leftarrow \text{NEWTONSCHULZORTHOGONALIZE}(Z_t^{(\ell)}, K)$;

      $\widetilde{\Delta}_t^{(\ell)} \leftarrow \Delta_t^{(\ell)} (A_t^{(\ell)})^\top$;

      $\widehat{\Delta}_{t,:j}^{(\ell)} \leftarrow \dfrac{\widetilde{\Delta}_{t,:j}^{(\ell)}}{\|\widetilde{\Delta}_{t,:j}^{(\ell)}\|_2 + \varepsilon}$ for $j = 1, \ldots, n_\ell$;

      $U_t^{(\ell)} \leftarrow \widehat{\Delta}_t^{(\ell)} A_t^{(\ell)}$;

    end

    $X_{t+1}^{(\ell)} \leftarrow (1 - \eta_t \lambda) X_t^{(\ell)} - \eta_t U_t^{(\ell)}$;

  end

  update all remaining non-matrix parameters using an auxiliary optimizer, if any;

  $t \leftarrow t + 1$;

end

return *optimized parameters* $\{X_t^{(\ell)}\}_{\ell=1}^{L}$;

ROBN-Haar. ROBN-Haar samples Z_t with i.i.d. standard Gaussian entries and approximately orthogonalizes it using the Newton–Schulz iteration:

$$A_t = \mathrm{NS}(Z_t).$$

This produces a Haar-like random orthogonal matrix intended to approximate uniform sampling over the orthogonal group, with a new A_t sampled across iterations.

Algorithm 2: ROBN-τ for Matrix Parameters

given: learning rates $\{\eta_t\}_{t\in\mathbb{N}} \subset \mathbb{R}_{>0}$, momentum coefficient $\beta \in [0,1)$, weight decay $\lambda \geq 0$, matrix parameters $\{X_1^{(\ell)}\}_{\ell=1}^{L}$ with $X_1^{(\ell)} \in \mathbb{R}^{m_\ell \times n_\ell}$, initial momenta $M_0^{(\ell)} = 0$, near-identity scale $\tau \geq 0$, numerical stability constant $\varepsilon > 0$

initialize time step $t \leftarrow 1$;

while *not converged* **do**

  compute stochastic gradients $G_t^{(\ell)} \leftarrow \text{STOCHASTICGRADIENT}(X_t^{(\ell)})$ for all $\ell = 1, \ldots, L$;

  for $\ell = 1, \ldots, L$ **do**

    $M_t^{(\ell)} \leftarrow \beta M_{t-1}^{(\ell)} + (1-\beta) G_t^{(\ell)}$;

    $N_t^{(\ell)} \leftarrow G_t^{(\ell)} + \beta \left(M_t^{(\ell)} - G_t^{(\ell)} \right)$; `// Nesterov momentum direction`

    $\Delta_t^{(\ell)} \leftarrow N_t^{(\ell)}$; `// raw update direction`

    if $m_\ell \geq n_\ell$ **then**

      sample $Z_t^{(\ell)} \in \mathbb{R}^{m_\ell \times m_\ell}$ with i.i.d. standard Gaussian entries;

      $S_t^{(\ell)} \leftarrow \dfrac{Z_t^{(\ell)} - (Z_t^{(\ell)})^\top}{\sqrt{2m_\ell}}$;

      $A_t^{(\ell)} \leftarrow \exp\left(\tau S_t^{(\ell)}\right)$;

      $\widetilde{\Delta}_t^{(\ell)} \leftarrow (A_t^{(\ell)})^\top \Delta_t^{(\ell)}$;

      $\widehat{\Delta}_{t,i:}^{(\ell)} \leftarrow \dfrac{\widetilde{\Delta}_{t,i:}^{(\ell)}}{\|\widetilde{\Delta}_{t,i:}^{(\ell)}\|_2 + \varepsilon}$ for $i = 1, \ldots, m_\ell$;

      $U_t^{(\ell)} \leftarrow A_t^{(\ell)} \widehat{\Delta}_t^{(\ell)}$;

    else

      sample $Z_t^{(\ell)} \in \mathbb{R}^{n_\ell \times n_\ell}$ with i.i.d. standard Gaussian entries;

      $S_t^{(\ell)} \leftarrow \dfrac{Z_t^{(\ell)} - (Z_t^{(\ell)})^\top}{\sqrt{2n_\ell}}$;

      $A_t^{(\ell)} \leftarrow \exp\left(\tau S_t^{(\ell)}\right)$;

      $\widetilde{\Delta}_t^{(\ell)} \leftarrow \Delta_t^{(\ell)} (A_t^{(\ell)})^\top$;

      $\widehat{\Delta}_{t,:j}^{(\ell)} \leftarrow \dfrac{\widetilde{\Delta}_{t,:j}^{(\ell)}}{\|\widetilde{\Delta}_{t,:j}^{(\ell)}\|_2 + \varepsilon}$ for $j = 1, \ldots, n_\ell$;

      $U_t^{(\ell)} \leftarrow \widehat{\Delta}_t^{(\ell)} A_t^{(\ell)}$;

    end

    $X_{t+1}^{(\ell)} \leftarrow (1 - \eta_t \lambda) X_t^{(\ell)} - \eta_t U_t^{(\ell)}$;

  end

  update all remaining non-matrix parameters using an auxiliary optimizer, if any;

  $t \leftarrow t + 1$;

end

return *optimized parameters* $\{X_t^{(\ell)}\}_{\ell=1}^{L}$;

ROBN-τ. ROBN-τ instead samples orthogonal matrices concentrated near the identity. Let

$$Z_t \in \mathbb{R}^{d \times d}$$

have i.i.d. standard Gaussian entries and define the skew-symmetric matrix

$$S_t = \frac{Z_t - Z_t^\top}{\sqrt{2d}}.$$

We then set

$$A_t = \exp(\tau S_t).$$

Since $S_t^\top = -S_t$, A_t is orthogonal. The parameter τ controls the scale of the deviation from the identity: $\tau = 0$ gives $A_t = I$, while larger values of τ produce orthogonal matrices with stronger deviations from the identity.

F.1 Right-Side One-Sided Orthogonality for $r < c$

When $r < c$, let $A_t \in \mathbb{R}^{c\times c}$ be a randomly sampled orthogonal matrix and define

$$\widetilde{m}_t = m_t A_t^\top, \qquad D_t = \operatorname{diag}\left(\|\widetilde{m}_{t,:1}\|_2, \ldots, \|\widetilde{m}_{t,:c}\|_2\right).$$

Then

$$\Delta_t^{\mathrm{ROBN}} = m_t A_t^\top D_t^\dagger A_t. \tag{527}$$

Define the active column index set

$$\mathcal{I}_t^{\mathrm{col}} = \{j \in \{1, \ldots, c\} : \|\widetilde{m}_{t,:j}\|_2 > 0\}.$$

For $j \in \mathcal{I}_t^{\mathrm{col}}$, let

$$p_j(t) = \frac{\widetilde{m}_{t,:j}}{\|\widetilde{m}_{t,:j}\|_2}, \qquad q_j(t) = A_t^\top e_j.$$

Then

$$\Delta_t^{\mathrm{ROBN}} = \sum_{j\in\mathcal{I}_t^{\mathrm{col}}} p_j(t) q_j(t)^\top.$$

Hence, for all $j, k \in \mathcal{I}_t^{\mathrm{col}}$,

$$\|p_j(t)\|_2 = 1, \qquad \langle q_j(t), q_k(t)\rangle = \delta_{jk}.$$

Thus, the active ROBN orthonormal rank-one system is right-side one-sided orthogonal.

G ALGORITHMIC DETAILS OF SPLIT POLAR UPDATE (SPU)

Algorithm 3: Split Polar Update (SPU) for Matrix Parameters

given: learning rates $\{\eta_t\}_{t\in\mathbb{N}} \subset \mathbb{R}_{>0}$, momentum coefficient $\beta \in [0,1)$, weight decay $\lambda \geq 0$, matrix parameters $\{X_1^{(\ell)}\}_{\ell=1}^{L}$ with $X_1^{(\ell)} \in \mathbb{R}^{m_\ell \times n_\ell}$, initial momenta $M_0^{(\ell)} = 0$ for all ℓ, number of Newton–Schulz steps K

initialize time step $t \leftarrow 1$;

while *not converged* **do**

  compute stochastic gradients $G_t^{(\ell)} \leftarrow \textsc{StochasticGradient}(X_t^{(\ell)})$ for all $\ell = 1, \ldots, L$;

  for $\ell = 1, \ldots, L$ **do**

    $M_t^{(\ell)} \leftarrow \beta M_{t-1}^{(\ell)} + (1-\beta) G_t^{(\ell)}$;

    $N_t^{(\ell)} \leftarrow G_t^{(\ell)} + \beta \left(M_t^{(\ell)} - G_t^{(\ell)} \right)$; `// Nesterov momentum direction`

    if $m_\ell \geq n_\ell$ *and* m_ℓ *is even* **then**

      split $N_t^{(\ell)}$ along the row dimension;

      $N_t^{(\ell)} = \begin{bmatrix} N_{t,\mathrm{top}}^{(\ell)} \\ N_{t,\mathrm{bottom}}^{(\ell)} \end{bmatrix}$;

      $\widehat{N}_{t,\mathrm{top}}^{(\ell)} \leftarrow \textsc{NewtonSchulzNormalize}(N_{t,\mathrm{top}}^{(\ell)}, K)$;

      $\widehat{N}_{t,\mathrm{bottom}}^{(\ell)} \leftarrow \textsc{NewtonSchulzNormalize}(N_{t,\mathrm{bottom}}^{(\ell)}, K)$;

      $\widetilde{U}_t^{(\ell)} \leftarrow \begin{bmatrix} \widehat{N}_{t,\mathrm{top}}^{(\ell)} \\ \widehat{N}_{t,\mathrm{bottom}}^{(\ell)} \end{bmatrix}$;

    else if $m_\ell < n_\ell$ *and* n_ℓ *is even* **then**

      split $N_t^{(\ell)}$ along the column dimension;

      $N_t^{(\ell)} = \begin{bmatrix} N_{t,\mathrm{left}}^{(\ell)} & N_{t,\mathrm{right}}^{(\ell)} \end{bmatrix}$;

      $\widehat{N}_{t,\mathrm{left}}^{(\ell)} \leftarrow \textsc{NewtonSchulzNormalize}(N_{t,\mathrm{left}}^{(\ell)}, K)$;

      $\widehat{N}_{t,\mathrm{right}}^{(\ell)} \leftarrow \textsc{NewtonSchulzNormalize}(N_{t,\mathrm{right}}^{(\ell)}, K)$;

      $\widetilde{U}_t^{(\ell)} \leftarrow \begin{bmatrix} \widehat{N}_{t,\mathrm{left}}^{(\ell)} & \widehat{N}_{t,\mathrm{right}}^{(\ell)} \end{bmatrix}$;

    else

      $\widetilde{U}_t^{(\ell)} \leftarrow \textsc{NewtonSchulzNormalize}(N_t^{(\ell)}, K)$; `// Muon fallback`

    end

    $s_\ell \leftarrow \sqrt{\max\left\{1, \frac{m_\ell}{n_\ell}\right\}}$;

    $U_t^{(\ell)} \leftarrow s_\ell \widetilde{U}_t^{(\ell)}$;

    $X_{t+1}^{(\ell)} \leftarrow (1 - \eta_t \lambda) X_t^{(\ell)} - \eta_t U_t^{(\ell)}$;

  end

  update all remaining non-matrix parameters using an auxiliary optimizer, if any;

  $t \leftarrow t + 1$;

end

return *optimized parameters* $\{X_t^{(\ell)}\}_{\ell=1}^{L}$;

Why SPU is One-Sided Orthogonal. To characterize the underlying geometry of SPU, we consider exact blockwise polar normalization. For a matrix $M = U\Sigma V^\top$, this normalization returns its polar factor $UV^\top$. The finite-step Newton–Schulz iteration used in Algorithm 3 approximates this polar factor in practice.

Let r and c denote the numbers of rows and columns, respectively. We first consider the case $r \geq c$ with even r. After splitting along the row dimension and applying this normalization independently

to the two blocks, write

$$\widehat{m}_t^{(1)} = \sum_{i=1}^{k_1} u_i^{(1)} (v_i^{(1)})^\top, \qquad \widehat{m}_t^{(2)} = \sum_{j=1}^{k_2} u_j^{(2)} (v_j^{(2)})^\top, \tag{528}$$

where the singular-vector systems are orthonormal within each block.

Embed the left factors into the full row space as

$$p_i^{(1)} = \begin{bmatrix} u_i^{(1)} \\ 0 \end{bmatrix}, \qquad p_j^{(2)} = \begin{bmatrix} 0 \\ u_j^{(2)} \end{bmatrix}, \tag{529}$$

and let

$$q_i^{(1)} = v_i^{(1)}, \qquad q_j^{(2)} = v_j^{(2)}. \tag{530}$$

Then the concatenated SPU update admits the rank-one representation

$$\Delta_t^{\mathrm{SPU}} = \sum_{i=1}^{k_1} p_i^{(1)} (q_i^{(1)})^\top + \sum_{j=1}^{k_2} p_j^{(2)} (q_j^{(2)})^\top. \tag{531}$$

Within each block, the left factors are orthonormal. Moreover, the zero-padding gives

$$\left\langle p_i^{(1)}, p_j^{(2)} \right\rangle = 0 \qquad \text{for all } i, j. \tag{532}$$

Hence, collecting all left factors into P_t,

$$P_t^\top P_t = I. \tag{533}$$

The right factors from different blocks, however, need not satisfy

$$\left\langle q_i^{(1)}, q_j^{(2)} \right\rangle = 0. \tag{534}$$

Thus SPU realizes left-sided one-sided orthogonality when the matrix is split along the row dimension.

When $r < c$ with even c, the argument is symmetric. After splitting along the column dimension, embed the right factors into the full column space as

$$q_i^{(1)} = \begin{bmatrix} v_i^{(1)} \\ 0 \end{bmatrix}, \qquad q_j^{(2)} = \begin{bmatrix} 0 \\ v_j^{(2)} \end{bmatrix}. \tag{535}$$

The disjoint supports imply

$$\left\langle q_i^{(1)}, q_j^{(2)} \right\rangle = 0, \tag{536}$$

and therefore, after collecting the right factors into Q_t,

$$Q_t^\top Q_t = I. \tag{537}$$

The left factors from different blocks need not be mutually orthogonal, so SPU realizes right-sided one-sided orthogonality in this case.

When the larger dimension cannot be evenly split, SPU falls back to the standard Muon update.

H Experimental Details, Hyperparameter Selection, and Additional Results

This appendix provides the complete experimental configurations and hyperparameter-selection procedures used for the vision and language-modeling experiments in Section 5. Hyperparameter sweeps are performed using a single random seed, after which the selected settings are fixed for the final three-seed comparisons.

H.1 ImageNet-1K Experiments

Model architecture. We use a ViT-B/16-style architecture for ImageNet-1K classification (Russakovsky et al., 2015; Dosovitskiy et al., 2021). Input images have resolution 224×224 and are divided into non-overlapping 16×16 patches. The model uses embedding dimension 768, 12 Transformer blocks, 12 attention heads, and an MLP hidden dimension of 3072.

After the final Transformer block, layer normalization is applied and the patch representations are averaged over the token dimension. The resulting representation is passed to a linear classification head over the 1000 ImageNet classes. Standard dropout is set to zero.

Training protocol. All models are trained for 150 epochs with a global batch size of 1024. We use a step-wise cosine learning-rate schedule with 7 warmup epochs, with the learning rate decaying to zero by the end of training. Training is performed using bfloat16 automatic mixed precision.

The final comparison is performed over three random seeds,

$$\{0, 1, 2\}.$$

For a fixed seed, all compared optimizers start from the same model initialization. All methods use the same architecture, training duration, learning-rate schedule, data pipeline, and evaluation procedure.

Data augmentation and regularization. For training, we use random resized cropping to 224×224, random horizontal flipping, and normalization using the standard ImageNet mean and standard deviation. For validation, each image is resized to 256, center-cropped to 224×224, and normalized using the same ImageNet statistics.

We use Mixup with

$$\alpha_{\mathrm{mixup}} = 0.8,$$

CutMix with

$$\alpha_{\mathrm{cutmix}} = 1.0,$$

label smoothing with coefficient 0.1, and stochastic depth with a drop-path rate of 0.1. Standard dropout is set to 0 (Touvron et al., 2021).

Optimizer parameter groups. For Muon, ROBN-Haar, ROBN-τ, and SPU, the corresponding matrix update is applied to all two-dimensional weight matrices inside the Transformer blocks. All remaining parameters, including the patch projection, positional embedding, classification head, normalization parameters, and biases, are updated using an auxiliary AdamW optimizer.

Muon, ROBN-Haar, ROBN-τ, and SPU use exactly the same parameter grouping and auxiliary AdamW configuration. Their comparison therefore changes the matrix update geometry while keeping the remaining optimizer configuration fixed.

For the AdamW baseline, all trainable model parameters are optimized directly using AdamW.

Final optimizer settings. The final ImageNet-1K comparison uses the hyperparameters summarized in Table 2.

For AdamW and the auxiliary AdamW components, we use

$$\beta_1 = 0.9, \qquad \beta_2 = 0.999, \qquad \epsilon_{\mathrm{Adam}} = 10^{-9}.$$

Table 2: Optimizer hyperparameters used in the ImageNet-1K comparison.

Method	Primary LR	Weight decay	Momentum	Aux. AdamW LR
AdamW	7×10^{-4}	0.25	–	–
Muon	3×10^{-3}	0.10	0.95	10^{-3}
ROBN-Haar	3×10^{-3}	0.10	0.95	10^{-3}
ROBN-τ	3×10^{-3}	0.10	0.95	10^{-3}
SPU	3×10^{-3}	0.10	0.95	10^{-3}

The auxiliary AdamW group used by Muon, ROBN-Haar, ROBN-τ, and SPU has zero weight decay. We use momentum coefficient

$$\beta = 0.95$$

for Muon, ROBN-Haar, ROBN-τ, and SPU. For ROBN-τ, we use

$$\tau = 0.5.$$

Newton–Schulz operations use $K = 5$ iterations.

H.1.1 Hyperparameter Selection

All hyperparameter sweeps are conducted using a single random seed (0) before the final three-seed comparison. The sweep runs are used only for hyperparameter selection and are not reused in the final three-seed comparison. After selecting the hyperparameters, the final comparison is run from scratch with random seeds $\{0, 1, 2\}$. AdamW and Muon are tuned independently. ROBN-Haar, ROBN-τ, and SPU are not separately tuned; instead, they use the hyperparameters selected for Muon.

Learning-rate sweep. For AdamW, we sweep the learning rate over

$$\left\{10^{-4}, 3 \times 10^{-4}, 5 \times 10^{-4}, 7 \times 10^{-4}\right\}$$

with zero weight decay. We additionally test 10^{-3}, which diverges during training and is therefore excluded from the completed sweep results. Among the stable runs, 7×10^{-4} achieves the best validation performance and is selected for the subsequent weight-decay sweep.

For Muon, we sweep the primary learning rate over

$$\left\{7 \times 10^{-4}, 10^{-3}, 3 \times 10^{-3}, 5 \times 10^{-3}, 7 \times 10^{-3}\right\}$$

with zero weight decay. We select

$$\eta_{\mathrm{Muon}} = 3 \times 10^{-3},$$

which achieves both the highest best validation accuracy and the lowest best validation loss.

The complete learning-rate sweep results are reported in Table 3.

Weight-decay sweep. With the AdamW learning rate fixed at

$$\eta_{\mathrm{AdamW}} = 7 \times 10^{-4},$$

we sweep weight decay over

$$\{0.05,\ 0.07,\ 0.10,\ 0.12,\ 0.15,\ 0.17,\ 0.20,\ 0.25,\ 0.30\}.$$

Weight decay 0.25 achieves both the highest best validation accuracy and the lowest best validation loss, and is therefore selected:

$$\lambda_{\mathrm{AdamW}} = 0.25.$$

With the Muon learning rate fixed at

$$\eta_{\mathrm{Muon}} = 3 \times 10^{-3},$$

we sweep weight decay over

$$\{0.03,\ 0.05,\ 0.07,\ 0.10,\ 0.12,\ 0.15\}.$$

Table 3: Single-seed learning-rate sweeps on ImageNet-1K. Weight decay is fixed to zero. Bold values indicate the learning rates selected for the subsequent weight-decay sweeps.

Method	Learning rate	Final val. loss	Final val. acc.	Best val. loss	Best val. acc.
AdamW	1×10^{-4}	1.1318	0.7589	1.1227	0.7594
AdamW	3×10^{-4}	1.1313	0.7647	1.1057	0.7653
AdamW	5×10^{-4}	1.1341	0.7644	1.1115	0.7647
AdamW	$\mathbf{7 \times 10^{-4}}$	1.1259	**0.7651**	**1.1024**	**0.7658**
AdamW	1×10^{-3}	Diverged during training			
Muon	7×10^{-4}	1.1086	0.7701	1.0824	0.7704
Muon	1×10^{-3}	1.1064	0.7718	1.0679	0.7731
Muon	$\mathbf{3 \times 10^{-3}}$	1.1266	**0.7753**	**1.0608**	**0.7760**
Muon	5×10^{-3}	1.1450	0.7731	1.0613	0.7733
Muon	7×10^{-3}	1.1506	0.7729	1.0636	0.7733

Table 4: Single-seed weight-decay sweeps on ImageNet-1K using the learning rates selected in Table 3.

Method	Weight decay	Final val. loss	Final val. acc.	Best val. loss	Best val. acc.
AdamW	0.05	1.1041	0.7737	1.0604	0.7746
AdamW	0.07	1.0826	0.7774	1.0378	0.7784
AdamW	0.10	1.0378	0.7818	1.0079	0.7824
AdamW	0.12	1.0171	0.7845	0.9994	0.7855
AdamW	0.15	0.9883	0.7885	0.9756	0.7887
AdamW	0.17	0.9744	0.7901	0.9691	0.7904
AdamW	0.20	0.9604	0.7909	0.9567	0.7909
AdamW	**0.25**	**0.9464**	**0.7914**	**0.9400**	**0.7919**
AdamW	0.30	0.9537	0.7890	0.9533	0.7890
Muon	0.03	1.0285	0.7932	0.9605	0.7934
Muon	0.05	0.9742	0.8003	0.9279	0.8003
Muon	0.07	0.9363	0.8028	0.9133	0.8028
Muon	**0.10**	0.9174	**0.8041**	**0.8924**	**0.8043**
Muon	0.12	0.9140	0.8028	0.9062	0.8030
Muon	0.15	**0.9040**	0.8031	0.8982	0.8031

Weight decay 0.10 achieves both the highest best validation accuracy and the lowest best validation loss, and is therefore selected:

$$\lambda_{\mathrm{Muon}} = 0.10.$$

The complete weight-decay sweep results are reported in Table 4.

The resulting primary optimizer settings used in the final comparison are therefore

$$\eta_{\mathrm{AdamW}} = 7 \times 10^{-4}, \qquad \lambda_{\mathrm{AdamW}} = 0.25,$$

and

$$\eta_{\mathrm{Muon}} = 3 \times 10^{-3}, \qquad \lambda_{\mathrm{Muon}} = 0.10.$$

ROBN-Haar, ROBN-τ, and SPU use the selected Muon settings

$$\eta = 3 \times 10^{-3}, \qquad \lambda = 0.10, \qquad \beta = 0.95,$$

without additional method-specific tuning.

H.2 FineWeb-Edu Language-Modeling Experiments

Model architecture and dataset. We train a GPT-style autoregressive Transformer (Radford et al., 2019) on FineWeb-Edu (Penedo et al., 2024). The model contains 24 Transformer blocks,

hidden dimension 768, 12 attention heads, and context length 2048. The MLP hidden dimension is four times the model dimension. The vocabulary contains $50{,}257$ tokens, and the token embedding is tied to the output language-modeling head.

Training sequences are sampled from the FineWeb-Edu training split by randomly selecting starting positions with replacement and extracting contiguous sequences of length 2048. Each model is trained from scratch until approximately 10^{10} training tokens have been processed. Training therefore does not correspond to a sequential epoch over the training corpus; sequences are sampled from randomly selected starting positions throughout training.

Training protocol. The effective global batch contains 384 sequences per optimizer step, with sequence length 2048. We use gradient accumulation and clip the global gradient norm at 1.0. Training is performed using bfloat16 automatic mixed precision.

The learning rate is linearly warmed up for the first 1000 optimizer steps and subsequently follows a cosine schedule to zero. Validation is performed every 500 optimizer steps using 100 validation batches.

The final comparison is performed over

$$\{0, 1, 2\}$$

random seeds. For every fixed seed, all optimizers are initialized using the same random seed and receive exactly the same deterministic sequence of training token batches. In addition, every optimizer and every seed is evaluated on the same fixed validation batches using a common evaluation seed.

Optimizer parameter groups. For Muon, ROBN-Haar, ROBN-τ, and SPU, all two-dimensional weight matrices inside the Transformer blocks are updated using the corresponding matrix optimizer.

Token embeddings, positional embeddings, and the tied language-modeling head are also two-dimensional parameters, but are excluded from the Transformer-block matrix optimizer and updated using auxiliary AdamW. They receive the same weight decay as the two-dimensional Transformer-block matrices. Layer-normalization parameters, biases, and all remaining parameters of dimension smaller than two are also updated using auxiliary AdamW, but receive zero weight decay.

For the AdamW baseline, all two-dimensional model parameters, including the Transformer-block matrices, token and positional embeddings, and the tied language-modeling head, are updated using AdamW with weight decay. Parameters of dimension smaller than two form a separate zero-weight-decay AdamW group. The same AdamW learning rate is used for both parameter groups.

Final optimizer settings. After hyperparameter selection, we fix the optimizer settings for the final three-seed language-modeling comparison as summarized in Table 5.

Table 5: Final optimizer hyperparameters used for the three-seed FineWeb-Edu comparison.

Method	Transformer-block 2D LR	2D WD	$<$2D WD	Momentum	Aux. AdamW LR
AdamW	5×10^{-3}	0.10	0	–	–
Muon	10^{-2}	0.10	0	0.95	5×10^{-3}
ROBN-Haar	10^{-2}	0.10	0	0.95	5×10^{-3}
ROBN-τ	10^{-2}	0.10	0	0.95	5×10^{-3}
SPU	10^{-2}	0.10	0	0.95	5×10^{-3}

For the AdamW baseline, the learning rate 5×10^{-3} is used for all parameter groups. For Muon, ROBN-Haar, ROBN-τ, and SPU, the Transformer-block 2D learning rate in Table 5 is used by the corresponding matrix optimizer, while all remaining parameters are updated using auxiliary AdamW with learning rate 5×10^{-3}.

For all methods, two-dimensional parameters receive weight decay 0.10, whereas parameters of dimension smaller than two receive zero weight decay. Both the AdamW baseline and all auxiliary AdamW parameter groups use

$$\beta_1 = 0.9, \qquad \beta_2 = 0.95, \qquad \epsilon_{\mathrm{Adam}} = 10^{-8}.$$

For Muon, ROBN-Haar, ROBN-τ, and SPU, the momentum coefficient of the corresponding matrix optimizer is

$$\beta = 0.95.$$

For ROBN-τ, we use

$$\tau = 0.5.$$

Newton–Schulz operations use $K = 5$ iterations.

H.2.1 Hyperparameter Selection

As in the vision experiments, all learning-rate and weight-decay sweeps are conducted using a single random seed (0) before the final three-seed comparison. AdamW and Muon are tuned independently. ROBN-Haar, ROBN-τ, and SPU inherit the hyperparameters selected for Muon and are not separately tuned.

Within each sweep, all candidate hyperparameters are evaluated on the same sampled validation batches at each evaluation checkpoint. These sweep-specific validation batches differ from the common fixed validation batches used for the final three-seed comparison. Consequently, validation losses reported in the sweep stage need not coincide numerically with the final-comparison losses, even at the same hyperparameter values.

Learning-rate sweep. For AdamW, we sweep the learning rate over

$$\left\{3 \times 10^{-4}, 5 \times 10^{-4}, 7 \times 10^{-4}, 10^{-3}, 3 \times 10^{-3}, 5 \times 10^{-3}, 7 \times 10^{-3}, 10^{-2}, 3 \times 10^{-2}\right\}$$

with two-dimensional weight decay fixed at 0.10. We select

$$\eta_{\mathrm{AdamW}} = 5 \times 10^{-3},$$

which achieves the lowest best validation loss.

For Muon, we sweep the learning rate applied to the two-dimensional Transformer-block matrices over

$$\left\{5 \times 10^{-3}, 10^{-2}, 2 \times 10^{-2}, 5 \times 10^{-2}\right\}$$

with two-dimensional weight decay fixed at 0.10. We select

$$\eta_{\mathrm{Muon}} = 10^{-2},$$

which achieves the lowest best validation loss.

The complete learning-rate sweep results are reported in Table 6.

Table 6: Single-seed learning-rate sweeps for the FineWeb-Edu language-modeling experiments. Two-dimensional weight decay is fixed at 0.10. Bold values indicate the learning rates selected for the subsequent weight-decay sweeps.

Method	Learning rate	Final val. loss	Best val. loss	Best val. ppl.
AdamW	3×10^{-4}	2.9743	2.9743	19.58
AdamW	5×10^{-4}	2.9079	2.9079	18.32
AdamW	7×10^{-4}	2.8844	2.8844	17.89
AdamW	10^{-3}	2.8756	2.8756	17.74
AdamW	3×10^{-3}	2.8237	2.8237	16.84
AdamW	$\mathbf{5 \times 10^{-3}}$	**2.8158**	**2.8158**	**16.71**
AdamW	7×10^{-3}	2.9996	2.9996	20.08
AdamW	10^{-2}	2.8335	2.8335	17.00
AdamW	3×10^{-2}	6.9205	6.9193	1011.59
Muon	5×10^{-3}	2.8060	2.8060	16.54
Muon	$\mathbf{10^{-2}}$	**2.7904**	**2.7904**	**16.29**
Muon	2×10^{-2}	2.7916	2.7916	16.31
Muon	5×10^{-2}	2.8156	2.8156	16.70

Weight-decay sweep. With the AdamW learning rate fixed at

$$\eta_{\mathrm{AdamW}} = 5 \times 10^{-3},$$

we sweep the weight decay applied to two-dimensional parameters over

$$\{0.05,\ 0.10,\ 0.15,\ 0.20,\ 0.25\}.$$

Weight decay 0.10 achieves the lowest best validation loss and is therefore selected:

$$\lambda_{\mathrm{AdamW}} = 0.10.$$

With the Muon learning rate for two-dimensional Transformer-block matrices fixed at

$$\eta_{\mathrm{Muon}} = 10^{-2},$$

we sweep the weight decay applied to two-dimensional parameters over

$$\{0.05,\ 0.10,\ 0.15,\ 0.20\}.$$

Weight decay 0.10 likewise achieves the lowest best validation loss and is selected:

$$\lambda_{\mathrm{Muon}} = 0.10.$$

The complete weight-decay sweep results are reported in Table 7.

Table 7: Single-seed weight-decay sweeps for the FineWeb-Edu language-modeling experiments using the learning rates selected in Table 6. Weight decay is applied to two-dimensional parameters, while parameters of dimension smaller than two receive zero weight decay.

Method	Weight decay	Final val. loss	Best val. loss	Best val. ppl.
AdamW	0.05	2.8258	2.8258	16.87
AdamW	**0.10**	**2.8143**	**2.8143**	**16.68**
AdamW	0.15	2.8162	2.8162	16.71
AdamW	0.20	2.8145	2.8145	16.69
AdamW	0.25	2.8215	2.8215	16.80
Muon	0.05	2.7950	2.7950	16.36
Muon	**0.10**	**2.7902**	**2.7902**	**16.28**
Muon	0.15	2.7957	2.7957	16.37
Muon	0.20	2.8040	2.8040	16.51

The hyperparameters selected for the final comparison are therefore

$$\eta_{\mathrm{AdamW}} = 5 \times 10^{-3}, \qquad \lambda_{\mathrm{AdamW}} = 0.10,$$

for AdamW, and

$$\eta_{\mathrm{Muon}} = 10^{-2}, \qquad \lambda_{\mathrm{Muon}} = 0.10$$

for the Muon matrix update.

ROBN-Haar, ROBN-τ, and SPU use the same Transformer-block 2D learning rate, two-dimensional weight decay, and momentum coefficient selected for Muon,

$$\eta = 10^{-2}, \qquad \lambda = 0.10, \qquad \beta = 0.95,$$

without additional method-specific tuning. For Muon, ROBN-Haar, ROBN-τ, and SPU, the auxiliary AdamW learning rate is fixed at 5×10^{-3}.

H.3 OpenWebText GPT-2 Small Language-Modeling Experiments

Model architecture and dataset. We train GPT-2 Small (Radford et al., 2019) on OpenWebText (Gokaslan & Cohen, 2019). The model contains 12 Transformer blocks, hidden dimension 768, 12 attention heads, and context length 1024. The MLP hidden dimension is four times the model dimension. The vocabulary contains 50,257 tokens, and the token embedding is tied to the output language-modeling head. The resulting model contains approximately 124M parameters.

Training sequences are sampled from the OpenWebText training corpus by randomly selecting starting positions with replacement and extracting contiguous sequences of length 1024. The training corpus is not truncated to an 8B-token subset. Instead, training sequences are sampled from the full training corpus until each model has processed approximately 8×10^9 tokens in total. Evaluation is performed on a fixed held-out OpenWebText validation split that is not used for training.

Training protocol. The effective global batch contains 768 sequences per optimizer step, with sequence length 1024, corresponding to $786{,}432$ processed tokens per optimizer step. Training for the 8B-token target therefore takes $10{,}173$ optimizer steps and processes $8{,}000{,}372{,}736$ tokens in total. We use gradient accumulation and clip the global gradient norm at 1.0. Training is performed using bfloat16 automatic mixed precision.

The learning rate is linearly warmed up for the first 800 optimizer steps and subsequently follows a cosine schedule to zero. Validation is performed every 500 optimizer steps using 100 validation batches.

The final comparison is performed over

$$\{0, 1, 2\}$$

random seeds. For every fixed seed, all optimizers are initialized using the same random seed and receive exactly the same deterministic sequence of training token batches. Every optimizer and every seed is evaluated on the same fixed validation batches using a common evaluation seed.

Optimizer parameter groups. For Muon, ROBN-Haar, ROBN-τ, and SPU, all two-dimensional weight matrices inside the Transformer blocks are updated using the corresponding matrix optimizer. Token embeddings, positional embeddings, and the tied language-modeling head are updated using auxiliary AdamW. These two-dimensional auxiliary parameters receive the same weight decay as the Transformer-block matrices. Layer-normalization parameters, biases, and all remaining parameters of dimension smaller than two are updated using auxiliary AdamW with zero weight decay.

For the AdamW baseline, all two-dimensional parameters, including the Transformer-block matrices, token and positional embeddings, and the tied language-modeling head, are updated using AdamW with weight decay. Parameters of dimension smaller than two form a separate zero-weight-decay AdamW group. The same AdamW learning rate is used for both parameter groups.

Final optimizer settings. The final three-seed OpenWebText comparison uses the hyperparameters summarized in Table 8.

Table 8: Final optimizer hyperparameters used for the three-seed OpenWebText GPT-2 Small comparison.

Method	Transformer-block 2D LR	2D WD	<2D WD	Momentum	Aux. AdamW LR
AdamW	5×10^{-3}	0.20	0	–	–
Muon	1.5×10^{-2}	0.10	0	0.95	5×10^{-3}
ROBN-Haar	1.5×10^{-2}	0.10	0	0.95	5×10^{-3}
ROBN-τ	1.5×10^{-2}	0.10	0	0.95	5×10^{-3}
SPU	1.5×10^{-2}	0.10	0	0.95	5×10^{-3}

For both the AdamW baseline and all auxiliary AdamW parameter groups, we use

$$\beta_1 = 0.9, \qquad \beta_2 = 0.95, \qquad \epsilon_{\mathrm{Adam}} = 10^{-8}.$$

For Muon, ROBN-Haar, ROBN-τ, and SPU, the momentum coefficient of the corresponding matrix optimizer is

$$\beta = 0.95.$$

For ROBN-τ, we use

$$\tau = 0.5.$$

Newton–Schulz operations use $K = 5$ iterations.

H.3.1 Hyperparameter Selection

All learning-rate and weight-decay sweeps are conducted using a single random seed (0) before the final three-seed comparison. AdamW and Muon are tuned independently. ROBN-Haar, ROBN-τ, and SPU inherit the hyperparameters selected for Muon and are not separately tuned.

Within each sweep, all candidate hyperparameters are evaluated on the same sampled validation batches at each evaluation checkpoint. These sweep-specific validation batches differ from the common fixed validation batches used for the final three-seed comparison. Consequently, validation losses reported in the sweep stage need not coincide numerically with the final-comparison losses, even at the same hyperparameter values.

Learning-rate sweep. For AdamW, we sweep the learning rate over

$$\left\{3 \times 10^{-4}, 5 \times 10^{-4}, 7 \times 10^{-4}, 10^{-3}, 3 \times 10^{-3}, 5 \times 10^{-3}, 7 \times 10^{-3}, 10^{-2}\right\}$$

with two-dimensional weight decay fixed at 0.10. We select

$$\eta_{\mathrm{AdamW}} = 5 \times 10^{-3},$$

which achieves the lowest best validation loss.

For Muon, we sweep the learning rate applied to the two-dimensional Transformer-block matrices over

$$\left\{5 \times 10^{-3}, 7 \times 10^{-3}, 10^{-2}, 1.5 \times 10^{-2}, 3 \times 10^{-2}, 5 \times 10^{-2}\right\}$$

with two-dimensional weight decay fixed at 0.10. We select

$$\eta_{\mathrm{Muon}} = 1.5 \times 10^{-2},$$

which achieves the lowest best validation loss.

The complete learning-rate sweep results are reported in Table 9.

Table 9: Single-seed learning-rate sweeps for the OpenWebText GPT-2 Small experiments. Two-dimensional weight decay is fixed at 0.10. Bold values indicate the learning rates selected for the subsequent weight-decay sweeps.

Method	Learning rate	Final val. loss	Best val. loss	Best val. ppl.
AdamW	3×10^{-4}	3.2346	3.2187	25.00
AdamW	5×10^{-4}	3.1580	3.1419	23.15
AdamW	7×10^{-4}	3.1203	3.1046	22.30
AdamW	10^{-3}	3.0994	3.0828	21.82
AdamW	3×10^{-3}	3.0403	3.0231	20.55
AdamW	$\mathbf{5 \times 10^{-3}}$	**3.0323**	**3.0155**	**20.40**
AdamW	7×10^{-3}	3.0348	3.0181	20.45
AdamW	10^{-2}	3.0396	3.0225	20.54
Muon	5×10^{-3}	3.0259	3.0092	20.27
Muon	7×10^{-3}	3.0159	2.9989	20.06
Muon	10^{-2}	3.0111	2.9939	19.96
Muon	$\mathbf{1.5 \times 10^{-2}}$	**2.9993**	**2.9827**	**19.74**
Muon	3×10^{-2}	3.0019	2.9849	19.78
Muon	5×10^{-2}	3.0165	3.0002	20.09

Weight-decay sweep. With the AdamW learning rate fixed at

$$\eta_{\mathrm{AdamW}} = 5 \times 10^{-3},$$

we sweep the weight decay applied to two-dimensional parameters over

$$\{0.05,\ 0.10,\ 0.15,\ 0.20,\ 0.25,\ 0.30\}.$$

Weight decay 0.20 achieves the lowest best validation loss and is therefore selected:

$$\lambda_{\mathrm{AdamW}} = 0.20.$$

With the Muon learning rate fixed at

$$\eta_{\mathrm{Muon}} = 1.5 \times 10^{-2},$$

we sweep the weight decay applied to two-dimensional parameters over

$$\{0.05,\ 0.10,\ 0.15,\ 0.20,\ 0.25,\ 0.30\}.$$

Weight decay 0.10 achieves the lowest best validation loss and is therefore selected:

$$\lambda_{\mathrm{Muon}} = 0.10.$$

The complete weight-decay sweep results are reported in Table 10.

Table 10: Single-seed weight-decay sweeps for the OpenWebText GPT-2 Small experiments using the learning rates selected in Table 9. Weight decay is applied to two-dimensional parameters, while parameters of dimension smaller than two receive zero weight decay.

Method	Weight decay	Final val. loss	Best val. loss	Best val. ppl.
AdamW	0.05	3.0451	3.0286	20.67
AdamW	0.10	3.0370	3.0211	20.51
AdamW	0.15	3.0308	3.0141	20.37
AdamW	**0.20**	**3.0295**	**3.0128**	**20.34**
AdamW	0.25	3.0322	3.0158	20.41
AdamW	0.30	3.0312	3.0146	20.38
Muon	0.05	2.9988	2.9946	19.98
Muon	**0.10**	**2.9953**	**2.9932**	**19.95**
Muon	0.15	2.9989	2.9977	20.04
Muon	0.20	3.0134	3.0119	20.33
Muon	0.25	3.0244	3.0235	20.56
Muon	0.30	3.0361	3.0347	20.79

The hyperparameters selected for the final OpenWebText comparison are therefore

$$\eta_{\mathrm{AdamW}} = 5 \times 10^{-3}, \qquad \lambda_{\mathrm{AdamW}} = 0.20,$$

for AdamW, and

$$\eta_{\mathrm{Muon}} = 1.5 \times 10^{-2}, \qquad \lambda_{\mathrm{Muon}} = 0.10$$

for the Muon matrix update.

ROBN-Haar, ROBN-τ, and SPU use the same Transformer-block 2D learning rate, two-dimensional weight decay, and momentum coefficient selected for Muon,

$$\eta = 1.5 \times 10^{-2}, \qquad \lambda = 0.10, \qquad \beta = 0.95,$$

without additional method-specific tuning. For Muon, ROBN-Haar, ROBN-τ, and SPU, the auxiliary AdamW learning rate is fixed at 5×10^{-3}.

Reporting. For ImageNet-1K, we report validation loss and validation accuracy over the three random seeds. For both FineWeb-Edu and OpenWebText, we report validation loss and perplexity using the common fixed validation batches defined for each setting. Aggregate curves report the mean and standard deviation across the three seeds.

H.4 Additional Validation Results

The main results in Section 5.1 report validation performance at the final training checkpoint, providing a synchronized comparison under the same training budget. For completeness, we additionally report the best validation performance attained during ImageNet-1K training and late-stage validation trajectories for ImageNet-1K, FineWeb-Edu, and OpenWebText.

ImageNet-1K. For ImageNet-1K, the best validation accuracy and the lowest validation loss are selected independently for each random seed and therefore need not occur at the same epoch. Table 11 reports the resulting three-seed averages.

For best validation accuracy, ROBN-Haar improves over AdamW, ROBN-τ further improves performance, Muon achieves higher accuracy, and SPU attains the highest value. For best validation loss, SPU attains the lowest value, followed closely by ROBN-τ, which also improves over Muon.

Table 11: Best ImageNet-1K validation performance attained during training over three random seeds. Best validation accuracy and lowest validation loss are selected independently for each seed and may occur at different epochs. Values are mean ± standard deviation. Higher validation accuracy and lower validation loss are better.

Method	Best Val. Acc. (%) ↑	Best Val. Loss ↓
AdamW	79.274 ± 0.111	0.9412 ± 0.0016
ROBN-Haar	79.668 ± 0.181	0.9207 ± 0.0088
ROBN-τ	80.212 ± 0.135	0.8943 ± 0.0064
Muon	80.439 ± 0.052	0.9029 ± 0.0054
SPU	$\mathbf{80.590 \pm 0.065}$	$\mathbf{0.8918 \pm 0.0004}$

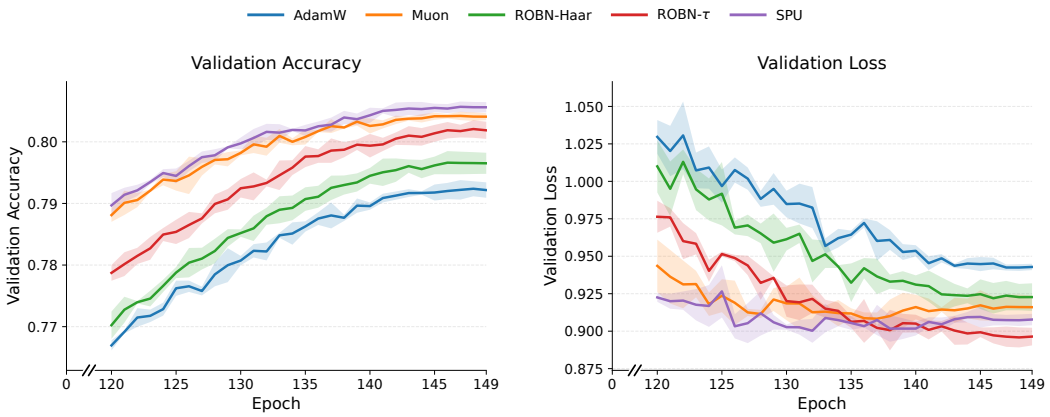


Figure 2: Late-stage ImageNet-1K validation accuracy and validation loss over three random seeds. Curves show the mean across seeds, and shaded regions indicate standard deviation.

ImageNet-1K validation trajectories. Figure 2 shows the late-stage validation-accuracy and validation-loss trajectories over the three random seeds. SPU attains the highest late-stage validation accuracy, while ROBN-τ achieves the lowest final validation loss. SPU attains the lowest best-validation loss over the full training run, as reported in Table 11.

Language modeling. For both FineWeb-Edu and OpenWebText, the lowest logged validation loss for every method and every random seed occurs at the final evaluation checkpoint. Consequently, the best and final language-modeling results coincide with those reported in Table 1, and we therefore do not report separate best-checkpoint tables.

Figure 3 shows the late-stage validation-loss trajectories for both language-modeling settings. Toward the end of training, SPU attains the lowest mean validation loss in both FineWeb-Edu and OpenWebText, followed by Muon, consistent with the final-checkpoint results in Table 1.

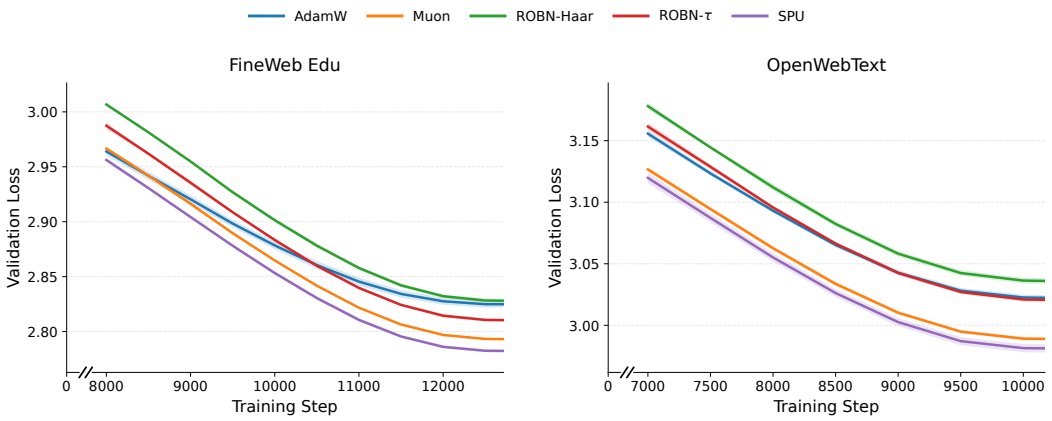


Figure 3: Late-stage validation loss on FineWeb-Edu and OpenWebText over three random seeds. Curves show the mean across seeds, and shaded regions indicate standard deviation.