Rank-One Potential Geometry for Normalized Optimizers

Insung Yun DLSTJD0209@HANYANG.AC.KR
Department of Artificial Intelligence, Hanyang University, Seoul, Korea

Sungyoon Lee SUNGYOONLEE@HANYANG.AC.KR
Department of Computer Science, Hanyang University, Seoul, Korea

Abstract

Recent normalized optimizers such as Lion and Muon highlight the importance of geometry in modern optimizer design. We propose a unified framework that extends the Lion-$\mathcal{K}$ perspective to a broader class of normalized update rules by representing the momentum variable in an orthonormal rank-one system and defining an ℓ_1-type coefficient potential, thereby covering SGD, Lion, and Muon within a single geometric view. We further study a regime with explicit time-dependent potentials, which is not covered by the static formulation, and show empirically that the resulting optimizer remains stable at ImageNet scale. On ViT-Base trained on ImageNet-1k, the proposed variant converges reliably and outperforms AdamW in our comparison, suggesting a route toward more systematic normalized-optimizer design.

1. Introduction

Modern optimizers increasingly use normalized or geometry-aware update directions. Methods such as momentum SGD [8, 9], Lion [3], and Muon [5] modify gradients or momentum in different ways, but all impose geometry on the update space. The Lion-$\mathcal{K}$ viewpoint formalizes this connection through nonsmooth potentials and a Lyapunov-based constrained-optimization interpretation [1].

In this work, we extend the Lion-$\mathcal{K}$ perspective through a rank-one orthonormal representation of the momentum variable and a coefficient-wise ℓ_1-type potential. With appropriate choices of the rank-one basis, SGD, Lion, and Muon can be expressed within a single potential-based framework. This places these seemingly different normalized update rules under the same geometric structure and allows their behavior to be interpreted through the Lion-$\mathcal{K}$ constrained-optimization and Lyapunov viewpoint [1], while also covering the spectral-norm-constrained interpretation of Muon [2].

We then go beyond the standard regime in which the explicit time-dependence of the potential does not contribute to the dynamics. Although the underlying basis or geometry may evolve over time, SGD, Lion, and Muon correspond to cases where this explicit contribution vanishes in the Lion-$\mathcal{K}$ Lyapunov formulation. Our framework naturally suggests a broader setting in which this term is nonzero. We first establish a convergence result for the non-momentum case with explicit time-dependent potentials in Appendix B. For the momentum setting, Appendices D isolate and control the additional time-variation terms introduced by replacing the static potential K with K_t; once these terms are controlled, the remaining stationarity argument follows the standard Lion-$\mathcal{K}$/Muon analysis.

Empirically, the corresponding time-dependent momentum-based variant remains stable at ImageNet scale. On ViT-Base trained on ImageNet-1k, it converges reliably and outperforms a learning-

© I. Yun & S. Lee.

rate- and weight-decay-tuned AdamW baseline [6, 7]. These results suggest that Lion-$\mathcal{K}$-style normalized optimizer frameworks can extend beyond the standard static setting.

2. A Unified Rank-One Potential Framework

We first introduce a unified notation for normalized update rules. Let m_t denote the momentum or update variable at time t. We represent m_t using **a rank-one orthonormal system**

$$m_t = \sum_i a_i(t) M_i(t), \qquad a_i(t) = \langle m_t, M_i(t) \rangle, \tag{1}$$

where

$$\langle M_i(t), M_j(t) \rangle = \delta_{ij}, \qquad \operatorname{rank}(M_i(t)) = 1. \tag{2}$$

Here, $\langle \cdot, \cdot \rangle$ denotes the Euclidean inner product for vectors and the Frobenius inner product for matrices. Given this representation, we define the coefficient-wise potential

$$K_t(m) = \sum_i |\langle m, M_i(t) \rangle|. \tag{3}$$

When the rank-one system is chosen adaptively, $K_t(m)$ is understood through the corresponding admissible representation rule, rather than by freezing an arbitrary basis at the current trajectory point. In particular, along the trajectory,

$$K_t(m_t) = \sum_i |a_i(t)|. \tag{4}$$

The corresponding normalized update direction is determined by a subgradient of this potential with respect to m_t.

It is important to distinguish two notions of time dependence. The basis used to represent the current momentum variable may depend on m_t, and hence may change along the trajectory. This does not necessarily mean that the potential itself has explicit time dependence. For example, the nuclear norm can be written using the singular-vector basis of its argument, but the nuclear norm is a fixed function. In contrast, an explicitly time-dependent potential K_t changes as a function of time even when its argument is held fixed. Along a trajectory, in the selected-subgradient sense,

$$\frac{d}{dt} K_t(m_t) = \langle \partial_m K_t(m_t), \dot{m}_t \rangle + \partial_t K_t(m_t). \tag{5}$$

Here we use the zero subgradient selection at zero coefficients. Specifically,

$$\operatorname{sign}(x) := \begin{cases} 1, & x > 0, \\ 0, & x = 0, \\ -1, & x < 0. \end{cases} \tag{6}$$

With this convention, we denote the selected momentum subgradient by

$$\partial_m K_t(m_t) := \sum_i \operatorname{sign}(\langle m_t, M_i(t) \rangle)\, M_i(t) \in \partial K_t(m_t). \tag{7}$$

The explicit time derivative along the chosen rank-one representation is

$$\partial_t K_t(m_t) := \left\langle m_t, \sum_i \operatorname{sign}(\langle m_t, M_i(t)\rangle)\, \dot{M}_i(t) \right\rangle. \tag{8}$$

Thus coefficients with $\langle m_t, M_i(t)\rangle = 0$ are inactive and contribute neither to the selected subgradient nor to the explicit time derivative. The standard examples below, namely SGD, Lion, and Muon, correspond to the regime in which this explicit time derivative vanishes. In contrast, the time-dependent extension proposed in this work allows this term to be nonzero.

2.1. Recovering SGD, Lion, and Muon

We now show how SGD, Lion, and Muon arise from different choices of the rank-one system.

SGD. For SGD-type updates, choose a rank-one orthonormal system adapted to the direction of m_t so that the coefficients are balanced:

$$a_i(t) = \alpha_t, \qquad \alpha_t = \frac{\|m_t\|}{\sqrt{n}}, \qquad i = 1, \dots, n. \tag{9}$$

Equivalently, any coefficient not included in the balanced representation is zero and therefore inactive under the zero-selection convention above. For $m_t \neq 0$, the selected subgradient becomes

$$\partial_m K_t(m_t) = \sum_{i=1}^{n} M_i(t) = \frac{\sqrt{n}}{\|m_t\|} m_t. \tag{10}$$

Thus the induced normalized direction is aligned with m_t, recovering the SGD-type update direction. The existence of such a balanced rank-one representation is shown in Appendix A.1.

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

$$M_{ij} = e_i e_j^\top, \qquad a_{ij}(t) = \langle m_t, M_{ij}\rangle = (m_t)_{ij}, \tag{11}$$

where e_i and e_j are standard basis vectors. Then

$$K(m_t) = \sum_{i,j} |a_{ij}(t)| = \sum_{i,j} |(m_t)_{ij}| = \|m_t\|_1. \tag{12}$$

Consequently, using the zero-selection convention specified above at zero entries, the selected subgradient is

$$\partial_m K(m_t) = \operatorname{sign}(m_t). \tag{13}$$

This recovers the sign-type normalized update used by Lion.

Muon. Muon corresponds to a rank-one system aligned with the singular directions of m_t. Let

$$M_i(t) = p_i(t) q_i(t)^\top, \qquad I_t := \{i : a_i(t) \neq 0\}. \tag{14}$$

Assume that, on the active set I_t,

$$\langle p_i(t), p_j(t)\rangle = \delta_{ij}, \qquad \langle q_i(t), q_j(t)\rangle = \delta_{ij}, \qquad i, j \in I_t. \tag{15}$$

Then the active rank-one system is orthonormal under the Frobenius inner product and corresponds to a singular-vector representation of m_t. Hence

$$K(m_t) = \sum_{i \in I_t} |a_i(t)| = \|m_t\|_*. \tag{16}$$

Thus the coefficient-wise potential recovers the nuclear-norm geometry underlying Muon. The verification is given in Appendix A.2.

Although the representing basis may be fixed, as in Lion, or depend on the current momentum variable, as in SGD and Muon, the induced potentials in these three cases are time-independent functions. Hence

$$\partial_t K_t(m_t) = 0 \qquad \text{for SGD, Lion, and Muon.} \tag{17}$$

This zero-explicit-time-contribution property is verified in Appendix A.3.

3. One-Sided Rank-One Instantiations

We next describe two concrete one-sided rank-one instantiations suggested by the proposed framework: Random Orthogonal Basis Normalization (ROBN) and the Split Newton–Schulz Update (SNSU). Both are motivated by the observation that useful normalized update rules may preserve only one side of the rank-one orthogonality structure, rather than the full two-sided singular-vector structure of Muon.

Random Orthogonal Basis Normalization. ROBN instantiates an explicitly time-dependent coefficient-wise potential; its algorithmic details are provided in Appendix F. For a rank-one system and active index set

$$M_i(t) = p_i(t) q_i(t)^\top, \qquad I_t := \{i : a_i(t) \neq 0\}, \tag{18}$$

ROBN preserves a one-sided orthogonality condition of the form

$$\langle p_i(t), p_j(t) \rangle = \delta_{ij} \quad \text{or} \quad \langle q_i(t), q_j(t) \rangle = \delta_{ij}, \qquad i, j \in I_t. \tag{19}$$

Unlike SGD, Lion, and Muon, the random orthogonal basis is resampled over time. Thus the induced coefficient-wise potential generally satisfies

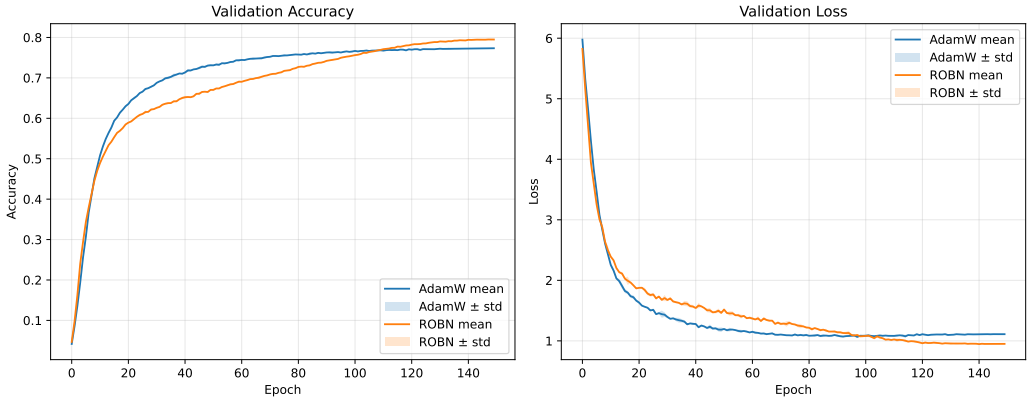
$$\partial_t K_t(m_t) \neq 0. \tag{20}$$

This makes ROBN a direct test case for the explicitly time-dependent potential regime discussed above.

Split Newton–Schulz Update. SNSU splits an update matrix into two blocks, applies Newton–Schulz normalization to each block, and concatenates the results. This construction admits a one-sided rank-one interpretation while using a fixed split pattern, so the associated potential is time-independent. Details and an auxiliary ImageNet-scale comparison with Muon are given in Appendix H.

Table 1: ImageNet-1K results with ViT-B/16. ROBN improves validation accuracy over AdamW and shows lower late-stage validation loss.

Method	Best Acc.	Final Acc.
AdamW	0.7735 ± 0.0006	0.7733 ± 0.0004
ROBN	$\mathbf{0.7950 \pm 0.0013}$	$\mathbf{0.7949 \pm 0.0011}$



Solid lines indicate mean validation performance; shaded regions indicate ±1 standard deviation across seeds.

4. Experiments

We empirically evaluate ROBN, the explicitly time-dependent one-sided rank-one instantiation introduced in Section 3. The goal is not to claim a new state-of-the-art training recipe, but to test whether the time-dependent update-rule view can produce stable and competitive optimization behavior at ImageNet scale.

We compare ROBN against AdamW on ImageNet-1K using a ViT-B/16 backbone [4]. The training settings are provided in Appendix G.1.

Baseline and proposed method. The AdamW baseline uses learning rate 3×10^{-4} and weight decay 0.1, selected by the hyperparameter sweeps in Appendix G.2. ROBN uses learning rate 1.5×10^{-3}. Unless otherwise specified, all non-optimizer training settings, including the data pipeline, batch size, schedule, augmentation, regularization, and training horizon, are kept identical between AdamW and ROBN.

4.1. Results

Table 1 summarizes the ImageNet-1K results with a ViT-B/16 backbone. ROBN achieves higher best and final validation accuracy than AdamW, while also showing lower late-stage validation loss.

The learning curves show a clear two-phase behavior. ROBN improves more slowly than AdamW in the early stage, but continues to make progress after AdamW has largely saturated. This suggests that the explicitly time-dependent potential setting, $\partial_t K_t(m_t) \neq 0$, can still lead to stable optimization at ImageNet scale.

5. Discussion

The main empirical finding is that stable optimization can be obtained even with an explicitly time-dependent coefficient-wise potential, $\partial_t K_t(m_t) \neq 0$. ROBN tests this case at ImageNet scale and shows that such time dependence does not necessarily destabilize training.

The results also provide a useful perspective on why Muon performs well. Muon corresponds to a two-sided rank-one structure, where both the left and right rank-one factors are orthogonal. In contrast, ROBN and SNSU only preserve a one-sided orthogonality structure. Nevertheless, their final validation performance remains comparable to Muon, although the ROBN run uses a longer training horizon. This suggests that an important part of Muon's practical behavior may come from the orthogonality of the rank-one factors, rather than from requiring the full two-sided singular-vector geometry.

More concretely, for the active set $I_t = \{i : a_i(t) \neq 0\}$, ROBN and SNSU satisfy a one-sided condition of the form $\langle p_i(t), p_j(t)\rangle = \delta_{ij}$ or $\langle q_i(t), q_j(t)\rangle = \delta_{ij}$ for $i, j \in I_t$. The SNSU experiment is particularly informative: even after splitting the update matrix and weakening the full Muon-style two-sided structure, the optimizer preserves nearly the same ImageNet-scale validation behavior. This suggests that one side of the rank-one orthogonality may already be sufficient to retain much of the benefit of normalized update rules.

This interpretation is suggestive rather than conclusive, since learning rate, weight decay, epoch budget, and late-stage dynamics may also contribute. Nevertheless, the results support the view that the proposed update-rule framework can capture a broader family of structured normalized updates beyond the standard time-independent nuclear-norm case.

References

[1] Lizhang Chen, Bo Liu, Kaizhao Liang, and qiang liu. Lion secretly solves a constrained optimization: As lyapunov predicts. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=e4xS9ZarDr.

[2] 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.

[3] Xiangning Chen, Chen Liang, Da Huang, Esteban Real, Kaiyuan Wang, Hieu Pham, Xuanyi Dong, Thang Luong, Cho-Jui Hsieh, Yifeng Lu, and Quoc V Le. Symbolic discovery of optimization algorithms. In *Thirty-seventh Conference on Neural Information Processing Systems*, 2023. URL https://openreview.net/forum?id=ne6zeqLFCZ.

[4] 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.

[5] 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/.

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

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

[8] 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.

[9] 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, editors, *Proceedings of the 30th International Conference on Machine Learning*, volume 28 of *Proceedings of Machine Learning Research*, pages 1139–1147, Atlanta, Georgia, USA, 17–19 Jun 2013. PMLR. URL https://proceedings.mlr.press/v28/sutskever13.html.

Appendix A. Static Examples and Balanced Rank-One Representations

A.1. Balanced Rank-One Representation of SGD

Let

$$m \in \mathbb{R}^{p\times q}. \tag{21}$$

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

Let

$$B = m^\top m \quad \text{or} \quad B = mm^\top. \tag{22}$$

Then B is symmetric positive semidefinite. Let

$$\lambda_1, \dots, \lambda_N \tag{23}$$

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

$$\operatorname{rank}(m) = \operatorname{rank}(B) \le n \le N. \tag{24}$$

Define

$$d = \left(\frac{\operatorname{tr}(B)}{n}, \dots, \frac{\operatorname{tr}(B)}{n}, 0, \dots, 0\right) \in \mathbb{R}^N, \tag{25}$$

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

Since

$$\sum_{i=1}^{N} d_i = \operatorname{tr}(B) = \sum_{i=1}^{N} \lambda_i, \tag{26}$$

and d is majorized by the eigenvalue vector λ, the Schur–Horn theorem implies that there exists an orthogonal matrix Q such that

$$\operatorname{diag}(Q^\top BQ) = d. \tag{27}$$

Let

$$Q = \begin{bmatrix} u_1 & u_2 & \cdots & u_N \end{bmatrix}. \tag{28}$$

Then

$$u_i^\top B u_i = d_i. \tag{29}$$

Moreover,

$$\frac{\operatorname{tr}(B)}{n} = \frac{\|m\|_F^2}{n}. \tag{30}$$

Define

$$\alpha = \frac{\|m\|_F}{\sqrt{n}}. \tag{31}$$

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

$$\sqrt{d_i} = \alpha. \tag{32}$$

Let

$$I = \{i : d_i \neq 0\}. \tag{33}$$

Then $|I| = n$.

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

$$u_i^\top m^\top m u_i = d_i. \tag{34}$$

Hence

$$\|m u_i\|_2 = \sqrt{d_i} = \alpha. \tag{35}$$

Define

$$v_i = \frac{m u_i}{\alpha}. \tag{36}$$

Then

$$\|v_i\|_2 = 1. \tag{37}$$

Moreover,

$$m u_i = \alpha v_i. \tag{38}$$

Define

$$M_i = v_i u_i^\top, \qquad i \in I. \tag{39}$$

Then each M_i is rank one. Also,

$$\langle M_i, M_j \rangle_F = \mathrm{tr}(M_i^\top M_j) = \mathrm{tr}(u_i v_i^\top v_j u_j^\top) = (v_i^\top v_j)(u_j^\top u_i). \tag{40}$$

Since the vectors u_i are orthonormal, we obtain

$$\langle M_i, M_j \rangle_F = \delta_{ij}. \tag{41}$$

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

Furthermore,

$$\alpha \sum_{i \in I} M_i = \alpha \sum_{i \in I} v_i u_i^\top = \sum_{i \in I} m u_i u_i^\top. \tag{42}$$

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

$$\|m u_i\|_2^2 = u_i^\top m^\top m u_i = u_i^\top B u_i = d_i = 0. \tag{43}$$

Thus

$$m u_i = 0 \qquad \text{for } i \notin I. \tag{44}$$

Therefore,

$$\sum_{i \in I} m u_i u_i^\top = \sum_{i=1}^{N} m u_i u_i^\top. \tag{45}$$

Since Q is orthogonal,

$$\sum_{i=1}^{N} u_i u_i^\top = Q Q^\top = I. \tag{46}$$

Hence

$$\alpha \sum_{i \in I} M_i = m \sum_{i=1}^{N} u_i u_i^\top = m Q Q^\top = m. \tag{47}$$

Thus,

$$m = \alpha \sum_{i \in I} M_i. \tag{48}$$

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

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

Hence

$$\|m^\top u_i\|_2 = \sqrt{d_i} = \alpha. \tag{50}$$

Define

$$v_i = \frac{m^\top u_i}{\alpha}. \tag{51}$$

Then

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

Moreover,

$$m^\top u_i = \alpha v_i. \tag{53}$$

Define

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

Then each M_i is rank one. Also,

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

Since the vectors u_i are orthonormal, we obtain

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

Furthermore,

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

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

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

Thus

$$m^\top u_i = 0 \qquad \text{for } i \notin I. \tag{59}$$

Therefore,

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

Since Q is orthogonal,

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

Hence

$$\alpha \sum_{i \in I} M_i = \left(\sum_{i=1}^{N} u_i u_i^\top \right) m = QQ^\top m = m. \tag{62}$$

Thus,

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

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

$$m = \alpha \sum_{i \in I} M_i, \qquad \alpha = \frac{\|m\|_F}{\sqrt{n}}, \tag{64}$$

where

$$\langle M_i, M_j \rangle_F = \delta_{ij}, \qquad \operatorname{rank}(M_i) = 1. \tag{65}$$

A.2. Rank-One Matrix Systems and the Nuclear Norm

Suppose

$$m = \sum_i a_i p_i q_i^\top. \tag{66}$$

Let

$$I := \{i : \, a_i \neq 0\} \tag{67}$$

be the active index set, and define

$$M_i = p_i q_i^\top, \qquad i \in I. \tag{68}$$

Assume that, for all $i, j \in I$,

$$\langle p_i, p_j \rangle = \delta_{ij}, \qquad \langle q_i, q_j \rangle = \delta_{ij}. \tag{69}$$

First, each M_i is rank-one. Moreover, for $i, j \in I$,

$$\begin{aligned}
\langle M_i, M_j \rangle_F &= \left\langle p_i q_i^\top, \, p_j q_j^\top \right\rangle_F && (70)\\
&= \operatorname{tr}\Big((p_i q_i^\top)^\top p_j q_j^\top \Big) && (71)\\
&= \operatorname{tr}\Big(q_i p_i^\top p_j q_j^\top \Big) && (72)\\
&= \langle p_i, p_j \rangle \langle q_i, q_j \rangle && (73)\\
&= \delta_{ij}. && (74)
\end{aligned}$$

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

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

$$\widetilde{p}_i = \operatorname{sign}(a_i) p_i. \tag{75}$$

Then $\{\widetilde{p}_i\}_{i \in I}$ remains orthonormal, and

$$m = \sum_{i \in I} |a_i| \, \widetilde{p}_i q_i^\top. \tag{76}$$

We now verify that this representation is a thin singular-value decomposition. For each $j \in I$,

$$mq_j = \sum_{i\in I} |a_i|\, \widetilde{p}_i q_i^\top q_j \tag{77}$$

$$= \sum_{i\in I} |a_i|\, \widetilde{p}_i \langle q_i, q_j\rangle \tag{78}$$

$$= |a_j|\, \widetilde{p}_j. \tag{79}$$

Similarly,

$$m^\top \widetilde{p}_j = \sum_{i\in I} |a_i|\, q_i \widetilde{p}_i^\top \widetilde{p}_j \tag{80}$$

$$= \sum_{i\in I} |a_i|\, q_i \langle \widetilde{p}_i, \widetilde{p}_j\rangle \tag{81}$$

$$= |a_j|\, q_j. \tag{82}$$

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

$$m^\top m q_j = |a_j|^2 q_j. \tag{83}$$

Thus the nonzero singular values of m are precisely $\{|a_i| : i \in I\}$, possibly together with additional zero singular values after extending the orthonormal systems to full orthonormal bases. Hence

$$\|m\|_* = \sum_{i\in I} |a_i| = \sum_i |a_i|. \tag{84}$$

Thus, for this rank-one orthonormal system, the coefficient-wise potential

$$K(m) = \sum_i |a_i| \tag{85}$$

coincides with the nuclear norm:

$$K(m) = \|m\|_*. \tag{86}$$

A.3. Zero Explicit Time Contribution of Standard Examples

We first record a simple consequence of differentiating an orthonormal basis. Let

$$\langle V_i(t), V_j(t)\rangle = \delta_{ij}. \tag{87}$$

Then, for $h > 0$,

$$\begin{aligned}\langle V_i(t+h), V_j(t+h)\rangle &= \left\langle V_i(t) + h\frac{V_i(t+h)-V_i(t)}{h}, V_j(t) + h\frac{V_j(t+h)-V_j(t)}{h}\right\rangle \\ &= \langle V_i(t), V_j(t)\rangle + h\left\langle V_i(t), \frac{V_j(t+h)-V_j(t)}{h}\right\rangle \\ &\quad + h\left\langle V_j(t), \frac{V_i(t+h)-V_i(t)}{h}\right\rangle \\ &\quad + h^2\left\langle \frac{V_i(t+h)-V_i(t)}{h}, \frac{V_j(t+h)-V_j(t)}{h}\right\rangle.\end{aligned} \tag{88}$$

Since

$$\langle V_i(t+h), V_j(t+h)\rangle = \langle V_i(t), V_j(t)\rangle = \delta_{ij}, \tag{89}$$

we obtain

$$\begin{aligned} 0 = h\left[\left\langle V_i(t), \frac{V_j(t+h)-V_j(t)}{h}\right\rangle + \left\langle V_j(t), \frac{V_i(t+h)-V_i(t)}{h}\right\rangle\right] \\ + h^2\left\langle \frac{V_i(t+h)-V_i(t)}{h}, \frac{V_j(t+h)-V_j(t)}{h}\right\rangle. \end{aligned} \tag{90}$$

Dividing by h and taking $h \to 0$ gives

$$\langle V_i(t), \dot{V}_j(t)\rangle + \langle V_j(t), \dot{V}_i(t)\rangle = 0. \tag{91}$$

Equivalently,

$$\langle \dot{V}_i(t), V_j(t)\rangle + \langle V_i(t), \dot{V}_j(t)\rangle = 0. \tag{92}$$

In particular, when $i = j$, we have

$$\langle \dot{V}_i(t), V_i(t)\rangle = 0. \tag{93}$$

SGD. Suppose that

$$m = \alpha \sum_i M_i, \qquad \alpha > 0, \tag{94}$$

where

$$\langle M_i, M_j\rangle_F = \delta_{ij}. \tag{95}$$

For the coefficient-wise potential

$$K_t(m) = \sum_i |a_i|, \tag{96}$$

the explicit time-variation term is

$$\partial_t K_t(m) = \sum_i \operatorname{sign}(a_i)\langle \dot{M}_i, m\rangle_F. \tag{97}$$

Since $a_i = \alpha > 0$, we have

$$\operatorname{sign}(a_i) = 1. \tag{98}$$

Therefore,

$$\begin{aligned} \partial_t K_t(m) &= \sum_i \langle \dot{M}_i, m\rangle_F \\ &= \sum_i \left\langle \dot{M}_i, \alpha \sum_j M_j \right\rangle_F \\ &= \alpha \sum_i \sum_j \langle \dot{M}_i, M_j\rangle_F. \end{aligned} \tag{99}$$

By applying (92) to the orthonormal family $\{M_i(t)\}_i$, we have

$$\langle \dot{M}_i, M_j\rangle_F + \langle M_i, \dot{M}_j\rangle_F = 0. \tag{100}$$

Hence,

$$\begin{aligned}\sum_i \sum_j \langle \dot{M}_i, M_j \rangle_F &= -\sum_i \sum_j \langle M_i, \dot{M}_j \rangle_F \\ &= -\sum_j \sum_i \langle \dot{M}_j, M_i \rangle_F \\ &= -\sum_i \sum_j \langle \dot{M}_i, M_j \rangle_F.\end{aligned} \tag{101}$$

Therefore,

$$\sum_i \sum_j \langle \dot{M}_i, M_j \rangle_F = 0. \tag{102}$$

Hence

$$\partial_t K_t(m) = 0. \tag{103}$$

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

$$\{M_i\}_i, \tag{104}$$

which is fixed in time. The momentum variable is written as

$$m = \sum_i a_i M_i. \tag{105}$$

The coefficient-wise potential is defined by

$$K_t(m) = \sum_i |a_i|. \tag{106}$$

Since the basis is fixed, we have

$$\dot{M}_i = 0 \tag{107}$$

for every i. Therefore, the explicit time-variation term satisfies

$$\begin{aligned}\partial_t K_t(m) &= \sum_i \operatorname{sign}(a_i) \langle \dot{M}_i, m \rangle_F \\ &= 0.\end{aligned} \tag{108}$$

Hence,

$$\partial_t K_t(m) = 0. \tag{109}$$

Muon. Suppose that

$$m = \sum_i \sigma_i M_i, \qquad \sigma_i > 0, \tag{110}$$

where

$$M_i = u_i v_i^\top. \tag{111}$$

Assume

$$\langle u_i, u_j \rangle = \delta_{ij}, \qquad \langle v_i, v_j \rangle = \delta_{ij}. \tag{112}$$

Then

$$\dot{M}_i = \dot{u}_i v_i^\top + u_i \dot{v}_i^\top. \tag{113}$$

The explicit time-variation term is

$$\partial_t K_t(m) = \sum_i \operatorname{sign}(\sigma_i) \langle \dot{M}_i, m \rangle_F. \tag{114}$$

Since

$$\sigma_i > 0, \tag{115}$$

we have

$$\operatorname{sign}(\sigma_i) = 1. \tag{116}$$

Thus,

$$\partial_t K_t(m) = \sum_i \langle \dot{M}_i, m \rangle_F. \tag{117}$$

Substituting

$$m = \sum_j \sigma_j M_j, \tag{118}$$

we get

$$\begin{aligned} \partial_t K_t(m) &= \sum_i \left\langle \dot{M}_i, \sum_j \sigma_j M_j \right\rangle_F \\ &= \sum_i \sum_j \sigma_j \langle \dot{M}_i, M_j \rangle_F. \end{aligned} \tag{119}$$

Since

$$M_j = u_j v_j^\top, \tag{120}$$

we have

$$\begin{aligned} \langle \dot{M}_i, M_j \rangle_F &= \left\langle \dot{u}_i v_i^\top + u_i \dot{v}_i^\top, u_j v_j^\top \right\rangle_F \\ &= \langle \dot{u}_i, u_j \rangle \langle v_i, v_j \rangle + \langle u_i, u_j \rangle \langle \dot{v}_i, v_j \rangle. \end{aligned} \tag{121}$$

Therefore,

$$\begin{aligned} \partial_t K_t(m) &= \sum_i \sum_j \sigma_j \left[\langle \dot{u}_i, u_j \rangle \langle v_i, v_j \rangle + \langle u_i, u_j \rangle \langle \dot{v}_i, v_j \rangle \right] \\ &= \sum_i \sigma_i \left[\langle \dot{u}_i, u_i \rangle + \langle \dot{v}_i, v_i \rangle \right]. \end{aligned} \tag{122}$$

By applying (93) to the orthonormal families $\{u_i(t)\}_i$ and $\{v_i(t)\}_i$, respectively, we have

$$\langle \dot{u}_i, u_i \rangle = 0, \qquad \langle \dot{v}_i, v_i \rangle = 0. \tag{123}$$

Hence

$$\partial_t K_t(m) = 0. \tag{124}$$

A.4. Vanishing Time-Variation Terms of K_t and K_t^*

Consequently, for each of the representations considered above, including SGD, Lion, and Muon, the explicit time-variation term vanishes for every momentum variable m admitting the corresponding representation:

$$\partial_t K_t(m) = 0. \tag{125}$$

Equivalently, the time-dependent basis only changes the representation of m, but not the value of the induced potential.

Therefore, the induced potential can be identified with a fixed convex function K, independent of the chosen time-dependent basis. That is,

$$K_t(m) = K(m) \qquad \text{for all } m. \tag{126}$$

Consequently, its Fenchel conjugate is also time-independent:

$$\begin{aligned} K_t^*(y) &= \sup_m \{\langle y, m\rangle_F - K_t(m)\} \\ &= \sup_m \{\langle y, m\rangle_F - K(m)\} \\ &= K^*(y). \end{aligned} \tag{127}$$

Hence,

$$\partial_t K_t^*(y) = 0 \qquad \text{for all } y. \tag{128}$$

Thus, for the representations considered above, both explicit time-variation terms vanish:

$$\partial_t K_t(m) = 0, \qquad \partial_t K_t^*(y) = 0. \tag{129}$$

Therefore, no additional explicit time-variation terms arise in the derivative of the Lion-$\mathcal{K}$ energy function. Accordingly, under the remaining assumptions of the Lion-$\mathcal{K}$ framework, the same Lyapunov monotonicity argument applies to the SGD, Lion, and Muon representations considered here.

A.5. Implications for SGD, Lion, and Muon

The preceding discussion shows that the SGD, Lion, and Muon representations all belong to the static-potential regime of the Lion-$\mathcal{K}$ framework. In particular, although the corresponding rank-one basis may be written in a time-dependent form, the induced potential itself is time-independent. Hence no additional explicit time-variation terms appear in the derivative of the Lion-$\mathcal{K}$ energy function.

Therefore, under the remaining assumptions of the Lion-$\mathcal{K}$ framework, the same Lyapunov monotonicity and convergence arguments apply to the SGD, Lion, and Muon representations considered here. The only difference is the choice of the convex potential K, which determines the corresponding Fenchel conjugate K^* and hence the implicit constrained problem.

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

$$\chi_{\mathcal{C}}(y) = \begin{cases} 0, & y \in \mathcal{C}, \\ +\infty, & y \notin \mathcal{C}. \end{cases} \tag{130}$$

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

$$K_{\mathrm{SGD}}(m) = \sqrt{n}\|m\|_F. \tag{131}$$

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

$$K^*_{\mathrm{SGD}}(y) = \chi_{\{\|y\|_F \le \sqrt{n}\}}(y). \tag{132}$$

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

$$\min_X F(X) \qquad \text{s.t.} \qquad \|\lambda X\|_F \le \sqrt{n}. \tag{133}$$

Equivalently,

$$\|X\|_F \le \frac{\sqrt{n}}{\lambda}. \tag{134}$$

For Lion, the coordinate-wise sign structure corresponds to

$$K_{\mathrm{Lion}}(m) = \|m\|_1, \qquad \nabla K_{\mathrm{Lion}}(m) = \mathrm{sign}(m). \tag{135}$$

Its Fenchel conjugate is the indicator function of an ℓ_∞-norm ball:

$$K^*_{\mathrm{Lion}}(y) = \chi_{\{\|y\|_\infty \le 1\}}(y). \tag{136}$$

Therefore, the corresponding constrained problem is

$$\min_X F(X) \qquad \text{s.t.} \qquad \|\lambda X\|_\infty \le 1, \tag{137}$$

or equivalently,

$$\|X\|_\infty \le \frac{1}{\lambda}. \tag{138}$$

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

$$K_{\mathrm{Muon}}(m) = \|m\|_{\mathrm{tr}}. \tag{139}$$

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

$$K^*_{\mathrm{Muon}}(y) = \chi_{\{\|y\|_{\mathrm{op}} \le 1\}}(y). \tag{140}$$

Thus, the corresponding constrained problem is

$$\min_X F(X) \qquad \text{s.t.} \qquad \|\lambda X\|_{\mathrm{op}} \le 1, \tag{141}$$

or equivalently,

$$\|X\|_{\mathrm{op}} \le \frac{1}{\lambda}. \tag{142}$$

Consequently, SGD, Lion, and Muon can all be treated within the same Lion-$\mathcal{K}$ convergence framework. The Lyapunov monotonicity argument is shared across the three cases, while the induced constrained problem and the corresponding KKT characterization differ according to the choice of K: Frobenius-norm constrained optimization for the balanced SGD representation, ℓ_∞-constrained optimization for Lion, and spectral-norm constrained optimization for Muon.

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

We consider the non-momentum dynamics

$$W_{t+1} = W_t - \eta_t \xi_t, \qquad g_t = \nabla \mathcal{L}(W_t), \tag{143}$$

where ξ_t is generated from a time-dependent potential K_t. At each time t, let

$$\beta_t = \{M_i(t)\}_{i=1}^N, \qquad N = mn,$$

be a Frobenius-orthonormal basis of the whole matrix space $\mathbb{R}^{m\times n}$. That is,

$$\langle M_i(t), M_j(t)\rangle_F = \delta_{ij}, \quad rank(M_i) = 1 \qquad \text{for all } i, j \in \{1, \dots, N\}. \tag{144}$$

The basis β_t may depend on time. Since β_t is a basis of the whole space $\mathbb{R}^{m\times n}$, every matrix $g_t \in \mathbb{R}^{m\times n}$ admits the expansion

$$g_t = \sum_{i=1}^N \langle g_t, M_i(t)\rangle_F M_i(t). \tag{145}$$

Define the time-dependent potential

$$K_t(g) = \sum_{i=1}^N |\langle g, M_i(t)\rangle_F| \,. \tag{146}$$

Since K_t is an ℓ_1-type potential, it may be nonsmooth when some coefficients vanish. We therefore choose the subgradient

$$\xi_t \in \partial K_t(g_t) \tag{147}$$

as

$$\xi_t = \sum_{i\in I_t} \operatorname{sign}\left(\langle g_t, M_i(t)\rangle_F\right) M_i(t), \tag{148}$$

where

$$I_t = \{i : \langle g_t, M_i(t)\rangle_F \neq 0\}\,, \qquad N_t = |I_t|. \tag{149}$$

By construction,

$$\langle g_t, \xi_t\rangle_F = \left\langle g_t, \sum_{i\in I_t} \operatorname{sign}\left(\langle g_t, M_i(t)\rangle_F\right) M_i(t) \right\rangle_F \tag{150}$$

$$= \sum_{i\in I_t} |\langle g_t, M_i(t)\rangle_F| \,. \tag{151}$$

Also, since the $M_i(t)$'s are orthonormal,

$$\|\xi_t\|_F^2 = \left\| \sum_{i\in I_t} \operatorname{sign}\left(\langle g_t, M_i(t)\rangle_F\right) M_i(t) \right\|_F^2 \tag{152}$$

$$= \sum_{i\in I_t} 1 = N_t. \tag{153}$$

Define

$$\mu_t = \frac{1}{N_t} \sum_{i \in I_t} |\langle g_t, M_i(t) \rangle_F| \,, \tag{154}$$

whenever $N_t > 0$. Then

$$\langle g_t, \xi_t \rangle_F = N_t \mu_t, \qquad \|\xi_t\|_F^2 = N_t. \tag{155}$$

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

$$\mathcal{L}(W_{t+1}) \leq \mathcal{L}(W_t) + \langle \nabla \mathcal{L}(W_t), W_{t+1} - W_t \rangle_F + \frac{L}{2} \|W_{t+1} - W_t\|_F^2 \tag{156}$$

$$= \mathcal{L}(W_t) - \eta_t \langle g_t, \xi_t \rangle_F + \frac{L\eta_t^2}{2} \|\xi_t\|_F^2 \tag{157}$$

$$= \mathcal{L}(W_t) - \eta_t N_t \mu_t + \frac{L\eta_t^2}{2} N_t \tag{158}$$

$$= \mathcal{L}(W_t) - \eta_t N_t \left(\mu_t - \frac{L\eta_t}{2} \right). \tag{159}$$

Therefore, if

$$0 < \eta_t < \frac{2\mu_t}{L} \qquad \text{for all } t \tag{160}$$

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

$$\eta_t = \theta_t \frac{\mu_t}{L}, \qquad 0 < \theta_{min} \leq \theta_t \leq \theta_{max} < 2, \tag{161}$$

gives

$$\mathcal{L}(W_{t+1}) \leq \mathcal{L}(W_t) - \frac{N_t \mu_t^2}{2L} (2 - \theta_t) \, \theta_t. \tag{162}$$

Define

$$D_t = \frac{N_t \mu_t^2}{2L} (2 - \theta_t) \, \theta_t. \tag{163}$$

Then

$$\mathcal{L}(W_{t+1}) \leq \mathcal{L}(W_t) - D_t. \tag{164}$$

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

$$\sum_{t=0}^{T-1} D_t \leq \mathcal{L}(W_0) - \mathcal{L}(W_T). \tag{165}$$

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

$$\sum_{t=0}^{T-1} D_t \leq \mathcal{L}(W_0) - \mathcal{L}(W^*) < \infty. \tag{166}$$

Hence,

$$\sum_{t=0}^{\infty} D_t < \infty, \qquad D_t \to 0. \tag{167}$$

It remains to relate D_t to the gradient norm. Since

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

we have

$$D_t = \frac{(2-\theta_t)\,\theta_t}{2LN_t} \left(\sum_{i \in I_t} |\langle g_t, M_i(t) \rangle_F| \right)^2 . \tag{168}$$

By the inequality $(\sum_i |a_i|)^2 \geq \sum_i |a_i|^2$, applied to the coefficients $a_i = \langle g_t, M_i(t) \rangle_F$, we have

$$D_t \geq \frac{(2-\theta_t)\,\theta_t}{2LN_t} \sum_{i \in I_t} |\langle g_t, M_i(t) \rangle_F|^2 . \tag{169}$$

Since the coefficients outside I_t are zero and $\beta_t = \{M_i(t)\}_{i=1}^N$ is a Frobenius-orthonormal basis of the whole space $\mathbb{R}^{m \times n}$, Parseval's identity gives

$$\sum_{i \in I_t} |\langle g_t, M_i(t) \rangle_F|^2 = \|g_t\|_F^2. \tag{170}$$

Therefore,

$$D_t \geq (2-\theta_t)\,\theta_t \frac{\|g_t\|_F^2}{2LN_t}. \tag{171}$$

Since $N_t \leq N$, we further obtain

$$D_t \geq (2-\theta_t)\,\theta_t \frac{\|g_t\|_F^2}{2LN}. \tag{172}$$

Because $D_t \to 0$, it follows that

$$\|g_t\|_F^2 \to 0. \tag{173}$$

Equivalently,

$$\nabla \mathcal{L}(W_t) \to 0. \tag{174}$$

Thus, for the non-momentum dynamics with the time-dependent potential K_t, the iterates converge to first-order stationary points in the sense that

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

Appendix C. Common-Domain Regime

Let $\mathcal{X}$ be a finite-dimensional Hilbert space equipped with inner product $\langle \cdot, \cdot \rangle$. For each time t, let

$$\beta_t = \{M_i(t)\}_{i=1}^N$$

be an orthonormal basis of $\mathcal{X}$. We define the time-dependent coefficient ℓ_1-type potential by

$$K_t(m) := \sum_{i=1}^N |\langle m, M_i(t) \rangle| . \tag{175}$$

For fixed t, K_t is a norm on $\mathcal{X}$. We first characterize its Fenchel conjugate. By definition,

$$K_t^*(y) = \sup_{m \in \mathcal{X}} \{\langle y, m\rangle - K_t(m)\}. \tag{176}$$

Since β_t is an orthonormal basis, every $m \in \mathcal{X}$ can be written uniquely as

$$m = \sum_{i=1}^{N} a_i M_i(t), \qquad a_i = \langle m, M_i(t)\rangle. \tag{177}$$

Hence, by (175),

$$K_t(m) = \sum_{i=1}^{N} |a_i|. \tag{178}$$

Define

$$\mathcal{C}_t := \left\{ y \in \mathcal{X} : \max_{1 \le i \le N} |\langle y, M_i(t)\rangle| \le 1 \right\}. \tag{179}$$

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

$$\chi_{\mathcal{C}_t}(y) := \begin{cases} 0, & y \in \mathcal{C}_t, \\ +\infty, & y \notin \mathcal{C}_t. \end{cases} \tag{180}$$

We now show that $K_t^* = \chi_{\mathcal{C}_t}$. First, suppose $y \in \mathcal{C}_t$. Then

$$|\langle y, M_i(t)\rangle| \le 1, \qquad i = 1, \dots, N. \tag{181}$$

For any $m = \sum_{i=1}^{N} a_i M_i(t)$, we have

$$\begin{aligned} \langle y, m\rangle &= \sum_{i=1}^{N} a_i \langle y, M_i(t)\rangle && (182) \\ &\le \sum_{i=1}^{N} |a_i|\, |\langle y, M_i(t)\rangle| && (183) \\ &\le \sum_{i=1}^{N} |a_i| && (184) \\ &= K_t(m). && (185) \end{aligned}$$

Therefore,

$$\langle y, m\rangle - K_t(m) \le 0, \qquad \forall m \in \mathcal{X}. \tag{186}$$

Since equality is attained at $m = 0$, it follows from (176) that

$$K_t^*(y) = 0. \tag{187}$$

Conversely, suppose $y \notin \mathcal{C}_t$. Then there exists $j \in \{1, \dots, N\}$ such that

$$|\langle y, M_j(t)\rangle| > 1. \tag{188}$$

For $s > 0$, choose

$$m_s = s \ \mathrm{sign}\left(\langle y, M_j(t)\rangle\right) M_j(t). \tag{189}$$

Then

$$K_t(m_s) = s, \tag{190}$$

and

$$\langle y, m_s\rangle = s\,|\langle y, M_j(t)\rangle|\,. \tag{191}$$

Hence

$$\langle y, m_s\rangle - K_t(m_s) = s\left(|\langle y, M_j(t)\rangle| - 1\right) \tag{192}$$

$$\to +\infty \qquad \text{as } s \to \infty. \tag{193}$$

Thus,

$$K_t^*(y) = +\infty. \tag{194}$$

Combining the two cases, we obtain

$$K_t^*(y) = \begin{cases} 0, & \max\limits_{1\le i\le N} |\langle y, M_i(t)\rangle| \le 1, \\ +\infty, & \text{otherwise.} \end{cases} \tag{195}$$

Equivalently,

$$K_t^*(y) = \chi_{\mathcal{C}_t}(y). \tag{196}$$

We now show that the dual potential vanishes along the trajectory when λ is sufficiently small. Since β_t is orthonormal, for any $y \in \mathcal{X}$ and any i,

$$|\langle y, M_i(t)\rangle| \le \|y\|\,\|M_i(t)\| = \|y\|. \tag{197}$$

Hence,

$$\|y\| \le 1 \quad \Longrightarrow \quad y \in \mathcal{C}_t, \qquad \forall t \ge 0. \tag{198}$$

By (196), this implies

$$\|y\| \le 1 \quad \Longrightarrow \quad K_t^*(y) = 0, \qquad \forall t \ge 0. \tag{199}$$

Assume that the trajectory is bounded:

$$\sup_{t\ge 0} \|x_t\| \le R \tag{200}$$

for some $R > 0$. If

$$\lambda \le \frac{1}{R}, \tag{201}$$

then

$$\|\lambda x_t\| \le \lambda R \le 1, \qquad \forall t \ge 0. \tag{202}$$

Therefore,

$$K_t^*(\lambda x_t) = 0, \qquad \forall t \ge 0. \tag{203}$$

Moreover, for each fixed $t \ge 0$, we have

$$\|\lambda x_t\| \le 1. \tag{204}$$

Hence, by (199),

$$K_s^*(\lambda x_t) = 0, \qquad \forall s \geq 0. \tag{205}$$

Therefore, the explicit time-variation of the dual potential vanishes:

$$\partial_s K_s^*(\lambda x_t)\big|_{s=t} = 0, \qquad \forall t \geq 0. \tag{206}$$

Equivalently, we write

$$\partial_t K_t^*(\lambda x_t) = 0, \qquad \forall t \geq 0. \tag{207}$$

Thus, in the subsequent Lyapunov analysis, the only remaining explicit time-variation term comes from the primal momentum potential $K_t(m_t)$.

Appendix D. Convergence Analysis for Momentum Dynamics with Time-Dependent Potentials

We consider the deterministic discrete-time momentum dynamics with time-dependent potential K_t:

$$M_{t+1} = \beta_2 M_t - (1-\beta_2)\nabla\mathcal{F}(X_t), \tag{208}$$

$$\widetilde{M}_{t+1} = \beta_1 M_t - (1-\beta_1)\nabla\mathcal{F}(X_t), \tag{209}$$

$$X_{t+1} = X_t + \eta_t\left(\nabla K_t(\widetilde{M}_{t+1}) - \lambda X_{t+1}\right). \tag{210}$$

We assume throughout this analysis that

$$0 \leq \beta_1 < \beta_2 < 1, \qquad \lambda > 0, \qquad \eta_t > 0, \qquad \forall t \geq 0, \tag{211}$$

and that $\mathcal{F}$ is L-smooth. We analyze these dynamics through the following time-dependent Lyapunov function.

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

We assume that the learning-rate schedule is nonincreasing:

$$\eta_{t+1} \leq \eta_t, \qquad \forall t \geq 0. \tag{212}$$

Since

$$\frac{d}{d\eta}\left(\frac{\eta\lambda\beta_1}{\eta\lambda(1-\beta_1)+(1-\beta_2)}\right) = \frac{\lambda\beta_1(1-\beta_2)}{\left(\eta\lambda(1-\beta_1)+(1-\beta_2)\right)^2} \geq 0, \tag{213}$$

it follows that

$$c_{t+1} \leq c_t, \qquad \forall t \geq 0. \tag{214}$$

The time-dependent Lyapunov function is

$$\begin{aligned}\mathcal{H}_t &:= \mathcal{F}(X_t) - \mathcal{F}^\star + \frac{1}{\lambda}K_t^*(\lambda X_t) + \frac{c_t}{\lambda}\left(K_t^*(\lambda X_t) + K_t(M_t) - \langle \lambda X_t, M_t\rangle\right) \\ &= \mathcal{F}(X_t) - \mathcal{F}^\star + \frac{a_t}{\lambda}K_t^*(\lambda X_t) + \frac{c_t}{\lambda}K_t(M_t) - c_t\langle X_t, M_t\rangle.\end{aligned} \tag{215}$$

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

$$\widehat{\mathcal{H}}_{t+1}^{(t)} := \mathcal{F}(X_{t+1}) - \mathcal{F}^\star + \frac{1}{\lambda} K_t^*(\lambda X_{t+1}) + \frac{c_{t+1}}{\lambda} \left(K_t^*(\lambda X_{t+1}) + K_t(M_{t+1}) - \langle \lambda X_{t+1}, M_{t+1} \rangle \right). \tag{216}$$

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 K_t. In contrast, the actual next Lyapunov value is

$$\mathcal{H}_{t+1} := \mathcal{F}(X_{t+1}) - \mathcal{F}^\star + \frac{1}{\lambda} K_{t+1}^*(\lambda X_{t+1}) + \frac{c_{t+1}}{\lambda} \left(K_{t+1}^*(\lambda X_{t+1}) + K_{t+1}(M_{t+1}) - \langle \lambda X_{t+1}, M_{t+1} \rangle \right). \tag{217}$$

Since the coefficient-wise ℓ_1-type potentials considered below may be nondifferentiable, we use the notation $\nabla K_t(\cdot)$ and $\nabla K_t^*(\cdot)$ to denote selected subgradients at the corresponding evaluation points whenever ordinary gradients do not exist. Thus, throughout this appendix,

$$\nabla K_t(M) \in \partial K_t(M), \qquad \nabla K_t^*(Y) \in \partial K_t^*(Y), \tag{218}$$

whenever the corresponding subdifferential is nonempty. In particular, K_t^* is an extended-valued conjugate of the coefficient-wise ℓ_1-type potential. Hence $\partial K_t^*(Y)$ is used only at points where this subdifferential is well-defined and nonempty. In the present proof, all occurrences of $\nabla K_t^*(\lambda X_{t+1})$ are used under the common-domain regime imposed below, which ensures that

$$K_t^*(\lambda X_{t+1}) = 0 < \infty.$$

For these indicator-type conjugates, this guarantees that

$$\partial K_t^*(\lambda X_{t+1}) \neq \emptyset.$$

We then choose an arbitrary element of this subdifferential and denote it by $\nabla K_t^*(\lambda X_{t+1})$.

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

$$\Gamma_t^{(t)} := \left\langle \nabla K_t(\widetilde{M}_{t+1}) - \lambda X_{t+1}, \widetilde{M}_{t+1} - \nabla K_t^*(\lambda X_{t+1}) \right\rangle, \tag{219}$$

$$\Delta_t^{(t)} := \left\langle \nabla K_t(\widetilde{M}_{t+1}) - \nabla K_t(M_{t+1}), \widetilde{M}_{t+1} - M_{t+1} \right\rangle. \tag{220}$$

For convenience, define

$$\delta_t := \nabla K_t(\widetilde{M}_{t+1}) - \lambda X_{t+1}. \tag{221}$$

By the parameter update in (210), we have

$$X_{t+1} - X_t = \eta_t \delta_t. \tag{222}$$

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

$$\overline{\mathcal{H}}_{t+1}^{(t)} := \mathcal{F}(X_{t+1}) - \mathcal{F}^\star + \frac{1}{\lambda} K_t^*(\lambda X_{t+1}) + \frac{c_t}{\lambda} \left(K_t^*(\lambda X_{t+1}) + K_t(M_{t+1}) - \langle \lambda X_{t+1}, M_{t+1} \rangle \right). \tag{223}$$

By the Fenchel–Young inequality,

$$K_t^*(\lambda X_{t+1}) + K_t(M_{t+1}) - \langle \lambda X_{t+1}, M_{t+1} \rangle \geq 0. \tag{224}$$

Since $c_{t+1} \leq c_t$ by (214), we obtain

$$\widehat{\mathcal{H}}_{t+1}^{(t)} \leq \overline{\mathcal{H}}_{t+1}^{(t)}. \tag{225}$$

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

$$\mathcal{F}(X_{t+1}) - \mathcal{F}(X_t) \leq \langle \nabla \mathcal{F}(X_t), X_{t+1} - X_t \rangle + \frac{L}{2} \|X_{t+1} - X_t\|_F^2 . \tag{226}$$

By the subgradient inequalities for the convex functions K_t^* and K_t,

$$K_t^*(\lambda X_{t+1}) - K_t^*(\lambda X_t) \leq \langle \lambda \nabla K_t^*(\lambda X_{t+1}), X_{t+1} - X_t \rangle , \tag{227}$$

$$K_t(M_{t+1}) - K_t(M_t) \leq \langle \nabla K_t(M_{t+1}), M_{t+1} - M_t \rangle . \tag{228}$$

Moreover,

$$\langle X_{t+1}, M_{t+1} \rangle - \langle X_t, M_t \rangle = \langle M_t, X_{t+1} - X_t \rangle + \langle X_{t+1}, M_{t+1} - M_t \rangle . \tag{229}$$

Applying (226), (228), and (229) to $\overline{\mathcal{H}}_{t+1}^{(t)} - \mathcal{H}_t$ yields

$$\begin{aligned} \overline{\mathcal{H}}_{t+1}^{(t)} - \mathcal{H}_t \leq{}& \langle \nabla \mathcal{F}(X_t) + a_t \nabla K_t^*(\lambda X_{t+1}) - c_t M_t, X_{t+1} - X_t \rangle \\ &+ \left\langle \frac{c_t}{\lambda} \nabla K_t(M_{t+1}) - c_t X_{t+1}, M_{t+1} - M_t \right\rangle \\ &+ \frac{L}{2} \|X_{t+1} - X_t\|_F^2 . \end{aligned} \tag{230}$$

From the momentum definitions in (210), we have

$$M_{t+1} - M_t = \frac{1 - \beta_2}{\beta_2 - \beta_1} \left(\widetilde{M}_{t+1} - M_{t+1} \right). \tag{231}$$

The definitions of a_t, b_t, and c_t imply

$$c_t = (a_t + b_t)\beta_1 - b_t \beta_2, \qquad c_t = a_t - 1, \qquad \frac{c_t(1 - \beta_2)}{\lambda(\beta_2 - \beta_1)} = \eta_t b_t. \tag{232}$$

Consequently,

$$\nabla \mathcal{F}(X_t) + a_t \nabla K_t^*(\lambda X_{t+1}) - c_t M_t = -a_t \left(\widetilde{M}_{t+1} - \nabla K_t^*(\lambda X_{t+1}) \right) - b_t \left(\widetilde{M}_{t+1} - M_{t+1} \right). \tag{233}$$

Substituting (222), (231), (232), and (233) into (230), we obtain

$$\begin{aligned} \widehat{\mathcal{H}}_{t+1}^{(t)} - \mathcal{H}_t \leq \overline{\mathcal{H}}_{t+1}^{(t)} - \mathcal{H}_t \leq{}& -\eta_t a_t \left\langle \delta_t, \widetilde{M}_{t+1} - \nabla K_t^*(\lambda X_{t+1}) \right\rangle \\ &- \eta_t b_t \left\langle \delta_t, \widetilde{M}_{t+1} - M_{t+1} \right\rangle \\ &+ \eta_t b_t \left\langle \nabla K_t(M_{t+1}) - \lambda X_{t+1}, \widetilde{M}_{t+1} - M_{t+1} \right\rangle \\ &+ \frac{L}{2} \eta_t^2 \|\delta_t\|_F^2 \\ ={}& -\eta_t \left(a_t \Gamma_t^{(t)} + b_t \Delta_t^{(t)} \right) + \frac{L}{2} \eta_t^2 \|\delta_t\|_F^2 . \end{aligned} \tag{234}$$

Finally, by (225) and (234),

$$\eta_t \left(a_t \Gamma_t^{(t)} + b_t \Delta_t^{(t)} \right) \leq \mathcal{H}_t - \widehat{\mathcal{H}}_{t+1}^{(t)} + \frac{L}{2} \eta_t^2 \left\| \nabla K_t(\widetilde{M}_{t+1}) - \lambda X_{t+1} \right\|_F^2 . \tag{235}$$

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 (216) from (217), we have

$$\begin{aligned} \mathcal{H}_{t+1} - \widehat{\mathcal{H}}_{t+1}^{(t)} &= \frac{1}{\lambda} \left[K_{t+1}^*(\lambda X_{t+1}) - K_t^*(\lambda X_{t+1}) \right] \\ &\quad + \frac{c_{t+1}}{\lambda} \left[K_{t+1}^*(\lambda X_{t+1}) - K_t^*(\lambda X_{t+1}) \right] \\ &\quad + \frac{c_{t+1}}{\lambda} \left[K_{t+1}(M_{t+1}) - K_t(M_{t+1}) \right] \\ &= \frac{a_{t+1}}{\lambda} \left[K_{t+1}^*(\lambda X_{t+1}) - K_t^*(\lambda X_{t+1}) \right] \\ &\quad + \frac{c_{t+1}}{\lambda} \left[K_{t+1}(M_{t+1}) - K_t(M_{t+1}) \right] . \end{aligned} \tag{236}$$

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 (236) into (235) yields the time-dependent one-step inequality

$$\begin{aligned} \eta_t \left(a_t \Gamma_t^{(t)} + b_t \Delta_t^{(t)} \right) &\leq \mathcal{H}_t - \mathcal{H}_{t+1} + \frac{L}{2} \eta_t^2 \left\| \nabla K_t(\widetilde{M}_{t+1}) - \lambda X_{t+1} \right\|_F^2 \\ &\quad + \frac{a_{t+1}}{\lambda} \left[K_{t+1}^*(\lambda X_{t+1}) - K_t^*(\lambda X_{t+1}) \right] \\ &\quad + \frac{c_{t+1}}{\lambda} \left[K_{t+1}(M_{t+1}) - K_t(M_{t+1}) \right] . \end{aligned} \tag{237}$$

We now impose the following assumptions.

First, by the common-domain argument established in Appendix C, suppose that the generated parameter trajectory is bounded:

$$\sup_{t \geq 0} \|X_t\|_F \leq R \tag{238}$$

for some $R > 0$, and choose

$$\lambda \leq \frac{1}{R}. \tag{239}$$

Then, for every pair of indices $s, t \geq 0$,

$$K_s^*(\lambda X_t) = 0. \tag{240}$$

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

$$K_t^*(\lambda X_{t+1}) = 0, \qquad K_{t+1}^*(\lambda X_{t+1}) = 0, \qquad \forall t \geq 0. \tag{241}$$

Therefore, the dual variation term in (237) vanishes:

$$\frac{a_{t+1}}{\lambda} \left[K_{t+1}^*(\lambda X_{t+1}) - K_t^*(\lambda X_{t+1}) \right] = 0, \qquad \forall t \geq 0. \tag{242}$$

Second, we assume that the cumulative weighted primal variation is uniformly bounded above. Specifically,

$$\exists\, C_{\text{var}} < \infty \quad \text{such that} \quad \sum_{t=0}^{T-1} \frac{c_{t+1}}{\lambda} \left[K_{t+1}(M_{t+1}) - K_t(M_{t+1})\right] \le C_{\text{var}}, \qquad \forall T \ge 1. \tag{243}$$

Thus, the explicit primal variation term in the summed version of (237) is uniformly bounded above along the generated trajectory.

It remains to bound the discrete smoothness remainder. Assume that K_t is a coefficient-wise ℓ_1-type potential induced by a Frobenius-orthonormal basis $\{V_i(t)\}_{i=1}^N$:

$$K_t(M) = \sum_{i=1}^{N} |\langle M, V_i(t)\rangle_F|.$$

At differentiable points,

$$\nabla K_t(M) = \sum_{i=1}^{N} \operatorname{sign}(\langle M, V_i(t)\rangle_F) V_i(t).$$

At nondifferentiable points, $\nabla K_t(M)$ denotes the selected subgradient specified above. In either case,

$$\|\nabla K_t(M)\|_F^2 \le N.$$

Moreover, from $K_t^*(\lambda X_{t+1}) = 0$, we have

$$|\langle \lambda X_{t+1}, V_i(t)\rangle_F| \le 1, \qquad i = 1, \dots, N.$$

Since $\{V_i(t)\}_{i=1}^N$ is Frobenius-orthonormal, this implies

$$\|\lambda X_{t+1}\|_F^2 = \sum_{i=1}^{N} |\langle \lambda X_{t+1}, V_i(t)\rangle_F|^2 \le N.$$

Therefore,

$$\begin{aligned} \left\|\nabla K_t(\widetilde{M}_{t+1}) - \lambda X_{t+1}\right\|_F^2 &\le 2\|\nabla K_t(\widetilde{M}_{t+1})\|_F^2 + 2\|\lambda X_{t+1}\|_F^2 \\ &\le 2N + 2N = 4N. \end{aligned} \tag{244}$$

Summing (237) from $t = 0$ to $T - 1$, and using the vanishing of the dual variation term, we obtain

$$\begin{aligned} \sum_{t=0}^{T-1} \eta_t \left(a_t \Gamma_t^{(t)} + b_t \Delta_t^{(t)}\right) \le \mathcal{H}_0 - \mathcal{H}_T + \frac{L}{2} \sum_{t=0}^{T-1} \eta_t^2 \left\|\nabla K_t(\widetilde{M}_{t+1}) - \lambda X_{t+1}\right\|_F^2 \\ + \sum_{t=0}^{T-1} \frac{c_{t+1}}{\lambda} \left[K_{t+1}(M_{t+1}) - K_t(M_{t+1})\right]. \end{aligned} \tag{245}$$

By the cumulative weighted variation assumption (243), we have

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

Therefore, (245) gives

$$\begin{aligned}\sum_{t=0}^{T-1} \eta_t \left(a_t \Gamma_t^{(t)} + b_t \Delta_t^{(t)}\right) &\le \mathcal{H}_0 - \mathcal{H}_T + \frac{L}{2} \sum_{t=0}^{T-1} \eta_t^2 \left\| \nabla K_t(\widetilde{M}_{t+1}) - \lambda X_{t+1} \right\|_F^2 + C_{\mathrm{var}} \\ &\le \mathcal{H}_0 - \mathcal{H}_T + 2LN \sum_{t=0}^{T-1} \eta_t^2 + C_{\mathrm{var}} \end{aligned} \tag{246}$$

where the last inequality follows from (244). Assume further that

$$\sum_{t=0}^{\infty} \eta_t^2 < \infty \tag{247}$$

and that $\mathcal{H}_t$ is bounded from below along the generated trajectory:

$$\inf_{t \ge 0} \mathcal{H}_t > -\infty. \tag{248}$$

Then, letting $T \to \infty$, we obtain

$$\sum_{t=0}^{\infty} \eta_t \left(a_t \Gamma_t^{(t)} + b_t \Delta_t^{(t)}\right) < \infty. \tag{249}$$

Indeed,

$$\mathcal{H}_0 - \mathcal{H}_T \le \mathcal{H}_0 - \inf_{s \ge 0} \mathcal{H}_s < \infty, \tag{250}$$

$$2LN \sum_{t=0}^{\infty} \eta_t^2 < \infty, \tag{251}$$

and, by assumption,

$$C_{\mathrm{var}} < \infty. \tag{252}$$

Therefore,

$$\mathcal{H}_0 - \mathcal{H}_T + 2LN \sum_{t=0}^{T-1} \eta_t^2 + C_{\mathrm{var}} \le \mathcal{H}_0 - \inf_{s \ge 0} \mathcal{H}_s + 2LN \sum_{t=0}^{\infty} \eta_t^2 + C_{\mathrm{var}} < \infty. \tag{253}$$

Hence, the right-hand side remains uniformly bounded as $T \to \infty$.

Since

$$a_t \ge 1, \qquad b_t \ge 0, \qquad \Gamma_t^{(t)} \ge 0, \qquad \Delta_t^{(t)} \ge 0, \tag{254}$$

we also have

$$\sum_{t=0}^{\infty} \eta_t \Gamma_t^{(t)} \le \sum_{t=0}^{\infty} \eta_t a_t \Gamma_t^{(t)} \le \sum_{t=0}^{\infty} \eta_t \left(a_t \Gamma_t^{(t)} + b_t \Delta_t^{(t)}\right) < \infty. \tag{255}$$

We next relate $\Gamma_t^{(t)}$ to the frozen Fenchel–Young gap. Define

$$G_t^{(t)} := K_t(\widetilde{M}_{t+1}) + K_t^*(\lambda X_{t+1}) - \left\langle \lambda X_{t+1}, \widetilde{M}_{t+1} \right\rangle. \tag{256}$$

By the Fenchel–Young inequality,

$$G_t^{(t)} \geq 0. \tag{257}$$

Let

$$z_{t+1}^{(t)} := \nabla K_t^*(\lambda X_{t+1}) \in \partial K_t^*(\lambda X_{t+1}). \tag{258}$$

By Fenchel conjugacy,

$$\lambda X_{t+1} \in \partial K_t(z_{t+1}^{(t)}). \tag{259}$$

Equivalently,

$$K_t(z_{t+1}^{(t)}) + K_t^*(\lambda X_{t+1}) - \left\langle \lambda X_{t+1}, z_{t+1}^{(t)} \right\rangle = 0. \tag{260}$$

Since

$$\nabla K_t(\widetilde{M}_{t+1}) \in \partial K_t(\widetilde{M}_{t+1}), \tag{261}$$

the monotonicity of ∂K_t gives

$$\Gamma_t^{(t)} = \left\langle \nabla K_t(\widetilde{M}_{t+1}) - \lambda X_{t+1}, \widetilde{M}_{t+1} - z_{t+1}^{(t)} \right\rangle \geq 0. \tag{262}$$

Moreover, using (260) in the definition of $G_t^{(t)}$, we obtain

$$\begin{aligned} G_t^{(t)} &= K_t(\widetilde{M}_{t+1}) + K_t^*(\lambda X_{t+1}) - \left\langle \lambda X_{t+1}, \widetilde{M}_{t+1} \right\rangle \\ &= K_t(\widetilde{M}_{t+1}) - K_t(z_{t+1}^{(t)}) - \left\langle \lambda X_{t+1}, \widetilde{M}_{t+1} - z_{t+1}^{(t)} \right\rangle \\ &\leq \left\langle \nabla K_t(\widetilde{M}_{t+1}), \widetilde{M}_{t+1} - z_{t+1}^{(t)} \right\rangle - \left\langle \lambda X_{t+1}, \widetilde{M}_{t+1} - z_{t+1}^{(t)} \right\rangle \\ &= \Gamma_t^{(t)}. \end{aligned} \tag{263}$$

Therefore,

$$0 \leq G_t^{(t)} \leq \Gamma_t^{(t)}. \tag{264}$$

Combining (255) and (264), we obtain

$$\sum_{t=0}^{\infty} \eta_t G_t^{(t)} < \infty. \tag{265}$$

Moreover, since

$$a_t \geq 1, \qquad b_t \geq 0, \qquad \Gamma_t^{(t)} \geq 0, \qquad \Delta_t^{(t)} \geq 0, \tag{266}$$

the bound (264) implies

$$G_t^{(t)} \leq \Gamma_t^{(t)} \leq a_t \Gamma_t^{(t)} + b_t \Delta_t^{(t)}. \tag{267}$$

Therefore, the one-step inequality (237), together with the vanishing dual variation term and the remainder bound (244), yields

$$\mathcal{H}_{t+1} - \mathcal{H}_t \leq -\eta_t G_t^{(t)} + 2LN\eta_t^2 + R_t, \tag{268}$$

where

$$R_t := \frac{c_{t+1}}{\lambda}\left[K_{t+1}(M_{t+1}) - K_t(M_{t+1})\right]. \tag{269}$$

We now strengthen the limit-set argument by using a discrete Barbalat-type no-spike condition. Recall that we have already shown

$$\sum_{t=0}^{\infty} \eta_t G_t^{(t)} < \infty, \qquad G_t^{(t)} \geq 0. \tag{270}$$

Assume also that

$$\sum_{t=0}^{\infty} \eta_t = \infty, \qquad \eta_t \to 0. \tag{271}$$

Let

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

Assume that the generated trajectory is bounded and asymptotically regular, in the sense that there exists a constant $C_Z > 0$ such that

$$\|Z_{t+1} - Z_t\| \leq \eta_t C_Z \qquad \text{for all sufficiently large } t. \tag{272}$$

Moreover, assume that the time-dependent Fenchel–Young gaps are locally uniformly continuous along the generated trajectory in the following sense: for every $\varepsilon > 0$, there exist $\rho > 0$ and $T_\varepsilon \geq 0$ such that, whenever $s \geq t \geq T_\varepsilon$ and

$$\sum_{r=t}^{s-1} \eta_r \leq \rho, \tag{273}$$

one has

$$\left| G_s^{(s)} - G_t^{(t)} \right| \leq \varepsilon. \tag{274}$$

This condition rules out arbitrarily narrow spikes of the time-dependent Fenchel–Young gap along the trajectory. It is satisfied, for example, when the trajectory is bounded, (272) holds, and the potentials K_t stabilize locally uniformly along the trajectory.

We claim that

$$G_t^{(t)} \to 0. \tag{275}$$

We prove this by contradiction. Suppose that (275) 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

$$G_{t_j}^{(t_j)} \geq \varepsilon_0, \qquad \forall j. \tag{276}$$

Apply the local uniform continuity condition (274) 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

$$\left| G_s^{(s)} - G_t^{(t)} \right| \leq \frac{\varepsilon_0}{2}. \tag{277}$$

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

$$\frac{\rho}{2} \leq \sum_{r=t_j}^{m_j} \eta_r \leq \rho. \tag{278}$$

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

$$\sum_{r=t_j}^{m} \eta_r \geq \frac{\rho}{2}. \tag{279}$$

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 (279), contradicting the minimality of m_j. Hence,

$$\sum_{r=t_j}^{m_j-1} \eta_r < \frac{\rho}{2}. \tag{280}$$

Adding the last term η_{m_j} gives

$$\begin{aligned} \sum_{r=t_j}^{m_j} \eta_r &= \sum_{r=t_j}^{m_j-1} \eta_r + \eta_{m_j} \\ &< \frac{\rho}{2} + \eta_{m_j}. \end{aligned} \tag{281}$$

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 (281), we obtain

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

Together with the defining inequality (279), this proves

$$\frac{\rho}{2} \le \sum_{r=t_j}^{m_j} \eta_r \le \rho,$$

which is (278).

Now fix any $r \in \{t_j, \dots, m_j\}$. By (278),

$$\sum_{q=t_j}^{r-1} \eta_q \le \sum_{q=t_j}^{m_j} \eta_q \le \rho.$$

Therefore, applying (277) with $t = t_j$ and $s = r$, we get

$$\left| G_r^{(r)} - G_{t_j}^{(t_j)} \right| \le \frac{\varepsilon_0}{2}. \tag{282}$$

From this absolute-value bound,

$$G_r^{(r)} - G_{t_j}^{(t_j)} \ge - \left| G_r^{(r)} - G_{t_j}^{(t_j)} \right|.$$

Hence,

$$\begin{aligned} G_r^{(r)} &\ge G_{t_j}^{(t_j)} - \left| G_r^{(r)} - G_{t_j}^{(t_j)} \right| \\ &\ge G_{t_j}^{(t_j)} - \frac{\varepsilon_0}{2} \\ &\ge \varepsilon_0 - \frac{\varepsilon_0}{2} = \frac{\varepsilon_0}{2}, \end{aligned} \tag{283}$$

where the last inequality uses (276). Thus, on the whole block $\{t_j, \dots, m_j\}$, the gap is bounded from below by $\varepsilon_0/2$.

Consequently,

$$\begin{aligned} \sum_{r=t_j}^{m_j} \eta_r G_r^{(r)} &\ge \sum_{r=t_j}^{m_j} \eta_r \frac{\varepsilon_0}{2} \\ &= \frac{\varepsilon_0}{2} \sum_{r=t_j}^{m_j} \eta_r \\ &\ge \frac{\varepsilon_0}{2} \cdot \frac{\rho}{2} = \frac{\varepsilon_0 \rho}{4}. \end{aligned} \tag{284}$$

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 (284) over these infinitely many disjoint blocks gives

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

This contradicts the weighted summability

$$\sum_{t=0}^{\infty} \eta_t G_t^{(t)} < \infty. \tag{285}$$

Therefore, the contradiction assumption was false, and we conclude that

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

Now let

$$Z_{t_j+1} = (X_{t_j+1}, M_{t_j+1}, \widetilde{M}_{t_j+1}) \to Z_\infty := (X_\infty, M_\infty, \widetilde{M}_\infty) \tag{286}$$

be any convergent subsequence. Since (275) holds along the whole sequence, we have

$$\lim_{j\to\infty} G_{t_j}^{(t_j)} = 0. \tag{287}$$

Finally, assume that the time-dependent potentials stabilize along the trajectory to a limiting potential K_∞, where K_∞ is proper, lower-semicontinuous, and convex. We write this symbolically as

$$K_t \to K_\infty, \tag{288}$$

in the following Fenchel–Young gap sense: for every convergent subsequence as above,

$$G_{t_j}^{(t_j)} \to G_\infty(X_\infty, \widetilde{M}_\infty), \tag{289}$$

where

$$G_\infty(X, \widetilde{M}) := K_\infty(\widetilde{M}) + K_\infty^*(\lambda X) - \left\langle \lambda X, \widetilde{M} \right\rangle. \tag{290}$$

This is the sense in which the notation $K_t \to K_\infty$ is used here: the Fenchel–Young gaps induced by K_t converge along trajectory subsequences to the Fenchel–Young gap induced by K_∞.

Combining (287) and (289), we obtain

$$G_\infty(X_\infty, \widetilde{M}_\infty) = 0. \tag{291}$$

By the definition of G_∞, this is precisely the limiting Fenchel–Young equality

$$K_\infty(\widetilde{M}_\infty) + K_\infty^*(\lambda X_\infty) - \left\langle \lambda X_\infty, \widetilde{M}_\infty \right\rangle = 0. \tag{292}$$

Hence, by Fenchel conjugacy,

$$\lambda X_\infty \in \partial K_\infty(\widetilde{M}_\infty). \tag{293}$$

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 (244),

$$\left\| \nabla K_t(\widetilde{M}_{t+1}) - \lambda X_{t+1} \right\|_F^2 \le 4N. \tag{294}$$

Therefore,

$$\left\|\nabla K_t(\widetilde{M}_{t+1}) - \lambda X_{t+1}\right\|_F \le 2\sqrt{N}. \tag{295}$$

Since the parameter update is

$$X_{t+1} - X_t = \eta_t \left(\nabla K_t(\widetilde{M}_{t+1}) - \lambda X_{t+1}\right), \tag{296}$$

we obtain

$$\|X_{t+1} - X_t\|_F \le 2\sqrt{N}\,\eta_t. \tag{297}$$

Because $\mathcal{F}$ is L-smooth, its gradient is L-Lipschitz. Hence,

$$\|\nabla\mathcal{F}(X_{t+1}) - \nabla\mathcal{F}(X_t)\|_F \le L\|X_{t+1} - X_t\|_F. \tag{298}$$

Combining this with (297) gives

$$\|\nabla\mathcal{F}(X_{t+1}) - \nabla\mathcal{F}(X_t)\|_F \le 2L\sqrt{N}\,\eta_t. \tag{299}$$

We now control the Polyak momentum tracking error. Let

$$A_t := \|\nabla\mathcal{F}(X_t) + M_t\|_F. \tag{300}$$

Using the deterministic momentum recursion

$$M_t = \beta_2 M_{t-1} - (1-\beta_2)\nabla\mathcal{F}(X_{t-1}), \tag{301}$$

we have

$$\begin{aligned}\nabla\mathcal{F}(X_t) + M_t &= \nabla\mathcal{F}(X_t) + \beta_2 M_{t-1} - (1-\beta_2)\nabla\mathcal{F}(X_{t-1}) \\ &= \nabla\mathcal{F}(X_t) - \nabla\mathcal{F}(X_{t-1}) + \beta_2\left(\nabla\mathcal{F}(X_{t-1}) + M_{t-1}\right).\end{aligned} \tag{302}$$

Taking Frobenius norms and using (299), we obtain

$$A_t \le 2L\sqrt{N}\,\eta_{t-1} + \beta_2 A_{t-1}. \tag{303}$$

Iterating (303) yields

$$A_t \le 2L\sqrt{N}\sum_{k=1}^{t}\beta_2^{t-k}\eta_{k-1} + \beta_2^t A_0. \tag{304}$$

Since $0 \le \beta_2 < 1$, the second term satisfies

$$\beta_2^t A_0 \to 0. \tag{305}$$

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

$$\eta_{k-1} \le \varepsilon, \qquad \forall k \ge K. \tag{306}$$

For $t \ge K$, split the convolution as

$$\sum_{k=1}^{t}\beta_2^{t-k}\eta_{k-1} = \sum_{k=1}^{K-1}\beta_2^{t-k}\eta_{k-1} + \sum_{k=K}^{t}\beta_2^{t-k}\eta_{k-1}. \tag{307}$$

The first term on the right-hand side of (307) 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

$$\sum_{k=K}^{t} \beta_2^{t-k} \eta_{k-1} \leq \varepsilon \sum_{k=K}^{t} \beta_2^{t-k} \leq \frac{\varepsilon \left(1 - \beta_2^{t-K+1}\right)}{1 - \beta_2} \leq \frac{\varepsilon}{1 - \beta_2}. \tag{308}$$

Because $\varepsilon > 0$ is arbitrary, it follows that

$$\sum_{k=1}^{t} \beta_2^{t-k} \eta_{k-1} \to 0. \tag{309}$$

Therefore, from (304),

$$\|\nabla \mathcal{F}(X_t) + M_t\|_F \to 0. \tag{310}$$

We now pass from M_t to the filtered momentum $\widetilde{M}_t$. By the definition of the filtered momentum,

$$\widetilde{M}_t = \beta_1 M_{t-1} - (1 - \beta_1) \nabla \mathcal{F}(X_{t-1}). \tag{311}$$

Hence,

$$\nabla \mathcal{F}(X_t) + \widetilde{M}_t = \nabla \mathcal{F}(X_t) - \nabla \mathcal{F}(X_{t-1}) + \beta_1 \left(\nabla \mathcal{F}(X_{t-1}) + M_{t-1}\right). \tag{312}$$

Taking norms and using (299), we obtain

$$\|\nabla \mathcal{F}(X_t) + \widetilde{M}_t\|_F \leq 2L\sqrt{N}\, \eta_{t-1} + \beta_1 \|\nabla \mathcal{F}(X_{t-1}) + M_{t-1}\|_F. \tag{313}$$

Since $\eta_{t-1} \to 0$, and (310) gives

$$\|\nabla \mathcal{F}(X_{t-1}) + M_{t-1}\|_F \to 0, \tag{314}$$

we conclude from (313) that

$$\|\nabla \mathcal{F}(X_t) + \widetilde{M}_t\|_F \to 0. \tag{315}$$

The preceding estimates also show that the discrete dynamics are asymptotically stationary. From (297) and $\eta_t \to 0$, we have

$$\|X_{t+1} - X_t\|_F \to 0. \tag{316}$$

This is the discrete-time analogue of the vanishing position velocity $\dot{X}_t \to 0$. Moreover, (310) gives

$$\|\nabla \mathcal{F}(X_t) + M_t\|_F \to 0. \tag{317}$$

This corresponds to the vanishing of the momentum residual in the continuous-time equation

$$\dot{M}_t = -\nabla \mathcal{F}(X_t) - M_t. \tag{318}$$

Finally, (315) gives

$$\|\nabla \mathcal{F}(X_t) + \widetilde{M}_t\|_F \to 0. \tag{319}$$

Thus both the position update and the momentum dynamics become asymptotically stationary, and the remaining limiting condition is the Fenchel consistency between $\widetilde{M}_t$ and λX_t.

Finally, let

$$(X_{t_j}, M_{t_j}, \widetilde{M}_{t_j}) \to (X_\infty, M_\infty, \widetilde{M}_\infty) \tag{320}$$

be any convergent subsequence. By (315) and the continuity of $\nabla\mathcal{F}$, we have

$$\widetilde{M}_\infty = -\nabla\mathcal{F}(X_\infty). \tag{321}$$

Combining (321) with the limiting Fenchel consistency condition

$$\lambda X_\infty \in \partial K_\infty(\widetilde{M}_\infty), \tag{322}$$

we obtain

$$\lambda X_\infty \in \partial K_\infty\left(-\nabla\mathcal{F}(X_\infty)\right). \tag{323}$$

We now interpret this condition as the KKT condition of the limiting constrained problem induced by K_∞.

Let

$$C_\infty := \operatorname{dom} K_\infty^* = \{Y : K_\infty^*(Y) < \infty\}.$$

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:

$$\chi_C(Y) := \begin{cases} 0, & Y \in C, \\ +\infty, & Y \notin C. \end{cases}$$

We assume that the limiting potential remains a coefficient-wise ℓ_1-type norm. Hence its Fenchel conjugate is the indicator function of C_∞, namely

$$K_\infty^* = \chi_{C_\infty}.$$

Equivalently, this means that

$$K_\infty^*(Y) = \begin{cases} 0, & Y \in C_\infty, \\ +\infty, & Y \notin C_\infty. \end{cases}$$

Consider the limiting constrained problem

$$\min_X \mathcal{F}(X) \qquad \text{subject to} \qquad \lambda X \in C_\infty. \tag{324}$$

Equivalently, this can be written as the extended-valued optimization problem

$$\min_X \left\{ \mathcal{F}(X) + \frac{1}{\lambda} K_\infty^*(\lambda X) \right\}. \tag{325}$$

Indeed, since $\lambda > 0$ and $K_\infty^* = \chi_{C_\infty}$, the term

$$\frac{1}{\lambda} K_\infty^*(\lambda X)$$

is equal to 0 when $\lambda X \in C_\infty$, and is $+\infty$ otherwise. Hence the extended-valued formulation (325) is exactly the constrained problem (324).

First, the limiting Fenchel–Young equality implies primal feasibility. Indeed,

$$K_\infty(\widetilde{M}_\infty) + K_\infty^*(\lambda X_\infty) - \left\langle \lambda X_\infty, \widetilde{M}_\infty \right\rangle = 0 \tag{326}$$

is finite. Therefore,

$$K_\infty^*(\lambda X_\infty) < \infty. \tag{327}$$

By the definition of

$$C_\infty = \{Y : K_\infty^*(Y) < \infty\},$$

we obtain

$$\lambda X_\infty \in C_\infty. \tag{328}$$

Second, we derive the stationarity condition. Recall from (293) that

$$\lambda X_\infty \in \partial K_\infty(\widetilde{M}_\infty). \tag{329}$$

By Fenchel conjugacy, for any proper closed convex function K, the following conditions are equivalent:

$$Y \in \partial K(Z) \iff K(Z) + K^*(Y) = \langle Y, Z\rangle \iff Z \in \partial K^*(Y).$$

Applying this equivalence with $K = K_\infty$, $Z = \widetilde{M}_\infty$, and $Y = \lambda X_\infty$, we obtain

$$\widetilde{M}_\infty \in \partial K_\infty^*(\lambda X_\infty). \tag{330}$$

On the other hand, by (321), the filtered momentum satisfies

$$\widetilde{M}_\infty = -\nabla \mathcal{F}(X_\infty). \tag{331}$$

Substituting this identity into (330), we obtain

$$-\nabla \mathcal{F}(X_\infty) \in \partial K_\infty^*(\lambda X_\infty). \tag{332}$$

Equivalently,

$$0 \in \nabla \mathcal{F}(X_\infty) + \partial K_\infty^*(\lambda X_\infty). \tag{333}$$

This inclusion is precisely the first-order stationarity condition for the extended-valued problem (325). Indeed, since $\mathcal{F}$ is differentiable and the map $X \mapsto \lambda X$ is linear, the subdifferential chain rule gives

$$\partial_X \left[\mathcal{F}(X) + \frac{1}{\lambda} K_\infty^*(\lambda X) \right] = \nabla \mathcal{F}(X) + \partial K_\infty^*(\lambda X).$$

Therefore, (333) is equivalent to

$$0 \in \partial_X \left[\mathcal{F}(X_\infty) + \frac{1}{\lambda} K_\infty^*(\lambda X_\infty) \right].$$

Since $K_\infty^* = \chi_{C_\infty}$, $\partial K_\infty^*(\lambda X_\infty)$ is the normal cone to C_∞ at λX_∞. Consequently, (333) is the normal-cone form of the KKT stationarity condition for the constrained problem (324). Together with the primal feasibility condition (328), this shows that X_∞ satisfies the KKT conditions of

the limiting constrained problem. Therefore, every limit point X_∞ is a KKT point of the limiting constrained problem.

Let

$$\mathrm{KKT} := \{X : \lambda X \in C_\infty \text{ and } 0 \in \nabla\mathcal{F}(X) + \partial K_\infty^*(\lambda X)\} \tag{334}$$

denote the KKT set of the limiting constrained problem (324). 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

$$d(X, \mathrm{KKT}) := \inf_{Y\in\mathrm{KKT}} \|X - Y\|_F. \tag{335}$$

Suppose, for contradiction, that

$$d(X_t, \mathrm{KKT}) \not\to 0. \tag{336}$$

Then there exist $\varepsilon > 0$ and a subsequence $\{X_{t_j}\}_{j\geq 1}$ such that

$$d(X_{t_j}, \mathrm{KKT}) \geq \varepsilon, \qquad \forall j. \tag{337}$$

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

$$X_{t_j} \to X_\infty. \tag{338}$$

By the limit-point argument above,

$$X_\infty \in \mathrm{KKT}\,. \tag{339}$$

Therefore, by the definition of distance to a set,

$$d(X_{t_j}, \mathrm{KKT}) \leq \|X_{t_j} - X_\infty\|_F \to 0. \tag{340}$$

This contradicts

$$d(X_{t_j}, \mathrm{KKT}) \geq \varepsilon. \tag{341}$$

Hence,

$$d(X_t, \mathrm{KKT}) \to 0. \tag{342}$$

Consequently, the generated trajectory converges to the KKT set of the limiting constrained problem induced by K_∞.

Appendix E. Verification of the ROBN Variation Assumption

We verify that the variation assumption required in Appendix D follows from a natural unweighted variation condition for ROBN.

Recall that the weighted primal variation assumption in Appendix D is

$$\sup_{T\geq 1} \sum_{t=0}^{T-1} \frac{c_{t+1}}{\lambda} \left[K_{t+1}(M_{t+1}) - K_t(M_{t+1})\right] < \infty. \tag{343}$$

For ROBN, motivated by the diagnostic evidence in Appendix I, we assume that the corresponding unweighted cumulative primal variation is uniformly bounded above:

$$\exists\, C_{\mathrm{unw}} < \infty \quad \text{such that} \quad \sum_{t=0}^{T-1} \left[K_{t+1}(M_{t+1}) - K_t(M_{t+1})\right] \leq C_{\mathrm{unw}}, \qquad \forall T \geq 1. \tag{344}$$

We now show that this implies (343).

Define

$$d_t := K_{t+1}(M_{t+1}) - K_t(M_{t+1}), \qquad w_t := \frac{c_{t+1}}{\lambda}. \tag{345}$$

Also define the partial sums

$$S_T := \sum_{t=0}^{T-1} d_t, \qquad T \geq 1, \qquad S_0 := 0. \tag{346}$$

Then (344) states that

$$S_T \leq C_{\text{unw}}, \qquad \forall T \geq 1. \tag{347}$$

By the definition of c_t in Appendix D, we have $c_t \geq 0$. Moreover, since the learning-rate schedule is nonincreasing, Appendix D shows that $c_{t+1} \leq c_t$. Therefore,

$$w_t \geq 0, \qquad w_{t+1} \leq w_t, \qquad \forall t \geq 0. \tag{348}$$

Using Abel summation, for every $T \geq 1$,

$$\sum_{t=0}^{T-1} w_t d_t = w_{T-1} S_T + \sum_{t=0}^{T-2} (w_t - w_{t+1}) S_{t+1}. \tag{349}$$

Since $w_t - w_{t+1} \geq 0$ and $S_T \leq C_{\text{unw}}$, we obtain

$$\begin{aligned} \sum_{t=0}^{T-1} w_t d_t &\leq C_{\text{unw}} w_{T-1} + C_{\text{unw}} \sum_{t=0}^{T-2} (w_t - w_{t+1}) \\ &= C_{\text{unw}} \left[w_{T-1} + \sum_{t=0}^{T-2} (w_t - w_{t+1}) \right] \\ &= C_{\text{unw}} w_0 < \infty. \end{aligned} \tag{350}$$

Substituting the definitions of d_t and w_t, this gives

$$\sum_{t=0}^{T-1} \frac{c_{t+1}}{\lambda} \left[K_{t+1}(M_{t+1}) - K_t(M_{t+1}) \right] \leq C_{\text{unw}} \frac{c_1}{\lambda}, \qquad \forall T \geq 1. \tag{351}$$

Therefore,

$$\sup_{T \geq 1} \sum_{t=0}^{T-1} \frac{c_{t+1}}{\lambda} \left[K_{t+1}(M_{t+1}) - K_t(M_{t+1}) \right] < \infty. \tag{352}$$

Thus ROBN satisfies the weighted variation assumption (243) in Appendix D, with

$$C_{\text{var}} := C_{\text{unw}} \frac{c_1}{\lambda} < \infty. \tag{353}$$

The remaining assumptions, namely the no-spike/local uniform continuity condition for the Fenchel–Young gap and the stabilization condition $K_t \to K_\infty$ in the Fenchel–Young gap sense, are imposed as trajectory-level regularity assumptions for ROBN. In practice, ROBN is implemented as

a finite-step, discrete-time random resampling algorithm that generates time-dependent orthogonal matrices. When the generated ROBN trajectory is idealized through an asymptotic convergence theorem, it is therefore natural to assume that the effective potentials stabilize along the trajectory in the Fenchel–Young gap sense and that the Fenchel–Young gap does not develop arbitrarily narrow spikes.

These assumptions require neither global convergence of K_t as functions nor convergence of the entire sequence of random orthogonal matrices. Rather, they only require the stabilization needed along the ROBN trajectory considered in the convergence argument. Under these trajectory-level regularity assumptions, together with the unweighted variation bound verified above, the weighted variation assumption in Appendix D holds with finite C_{var}, and the convergence-to-KKT-set conclusion of Appendix D applies to the corresponding ROBN trajectory.

Appendix F. Algorithmic Details of Random Orthogonal Basis Normalization (ROBN)

Algorithm 1: Random Orthogonal-Basis Normalization for Matrix Parameters

given: learning rates $\{\eta_t\}_{t\in\mathbb{N}} \subset \mathbb{R}_{>0}$, momentum coefficient $\beta \in [0,1)$, 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 $\ell = 1, \dots, L$;

  for $\ell = 1, \dots, L$ **do**

    $M_t^{(\ell)} \leftarrow \beta M_{t-1}^{(\ell)} + G_t^{(\ell)}$;

    $N_t^{(\ell)} \leftarrow G_t^{(\ell)} + \beta M_t^{(\ell)}$; `// Nesterov momentum direction`

    $\Delta_t^{(\ell)} \leftarrow N_t^{(\ell)}$; `// raw update direction`

    if $m_\ell \geq n_\ell$ **then**

      sample $Z_t^{(\ell)} \in \mathbb{R}^{m_\ell \times m_\ell}$;

      $A_t^{(\ell)} \leftarrow \textsc{NewtonSchulzOrthogonalize}(Z_t^{(\ell)}, K)$;

      $\widetilde{\Delta}_t^{(\ell)} \leftarrow (A_t^{(\ell)})^\top \Delta_t^{(\ell)}$;

      $\widehat{\Delta}_{t,i:}^{(\ell)} \leftarrow \dfrac{\widetilde{\Delta}_{t,i:}^{(\ell)}}{\|\widetilde{\Delta}_{t,i:}^{(\ell)}\|_2 + \varepsilon}$ for $i = 1, \dots, m_\ell$;

      $D_t^{(\ell)} \leftarrow A_t^{(\ell)} \widehat{\Delta}_t^{(\ell)}$;

    else

      sample $Z_t^{(\ell)} \in \mathbb{R}^{n_\ell \times n_\ell}$;

      $A_t^{(\ell)} \leftarrow \textsc{NewtonSchulzOrthogonalize}(Z_t^{(\ell)}, K)$;

      $\widetilde{\Delta}_t^{(\ell)} \leftarrow \Delta_t^{(\ell)} (A_t^{(\ell)})^\top$;

      $\widehat{\Delta}_{t,:j}^{(\ell)} \leftarrow \dfrac{\widetilde{\Delta}_{t,:j}^{(\ell)}}{\|\widetilde{\Delta}_{t,:j}^{(\ell)}\|_2 + \varepsilon}$ for $j = 1, \dots, n_\ell$;

      $D_t^{(\ell)} \leftarrow \widehat{\Delta}_t^{(\ell)} A_t^{(\ell)}$;

    end

    $X_{t+1}^{(\ell)} \leftarrow X_t^{(\ell)} - \eta_t D_t^{(\ell)}$;

  end

  update all remaining non-matrix parameters using an auxiliary optimizer, if any;

  $t \leftarrow t + 1$;

end

return *optimized parameters* $\{X_t^{(\ell)}\}_{\ell=1}^{L}$;

For the case $m \geq n$, let

$$\widetilde{\Delta}_t = A_t^\top \Delta_t, \tag{354}$$

and let $p_i(t)$ denote the i-th column of A_t. ROBN normalizes each row of $\widetilde{\Delta}_t$ and produces

$$D_t = A_t \widehat{\Delta}_t = \sum_i p_i(t) \widehat{\Delta}_{t,i:}, \qquad \widehat{\Delta}_{t,i:} = \frac{\widetilde{\Delta}_{t,i:}}{\|\widetilde{\Delta}_{t,i:}\|_2 + \varepsilon}. \tag{355}$$

Conditioned on A_t, this update is associated with the time-indexed potential

$$K_t(\Delta) = \sum_i \left\| (A_t^\top \Delta)_{i:} \right\|_2. \tag{356}$$

For nonzero rows and $\varepsilon = 0$, the corresponding row-normalized direction is an element of the subdifferential:

$$A_t \widehat{\Delta}_t \in \partial_\Delta K_t(\Delta_t), \qquad \widehat{\Delta}_{t,i:} = \frac{\widetilde{\Delta}_{t,i:}}{\|\widetilde{\Delta}_{t,i:}\|_2}. \tag{357}$$

The small constant $\varepsilon > 0$ is used only for numerical stability, preventing division by zero when $\|\widetilde{\Delta}_{t,i:}\|_2 = 0$. For nonzero rows, the update approaches the exact subgradient direction as $\varepsilon \to 0$.

Since A_t is sampled across iterations, the induced potential changes with t. In particular, for a fixed argument Δ, one generally has

$$K_{t+1}(\Delta) \neq K_t(\Delta). \tag{358}$$

Thus, in the discrete-time ROBN algorithm, the explicit time dependence appears through the changing random orthogonal basis A_t.

The case $m < n$ is analogous. Conditioning on A_t, the induced potential is

$$K_t(\Delta) = \sum_{j=1}^{n} \left\| (\Delta A_t^\top)_{:j} \right\|_2, \tag{359}$$

and the ROBN direction satisfies

$$\widehat{\Delta}_t A_t \in \partial_\Delta K_t(\Delta_t) \tag{360}$$

for nonzero columns when $\varepsilon = 0$.

Appendix G. Experimental Details and Hyperparameter Selection

G.1. Training Details

Training protocol. All models are trained for 150 epochs with a global batch size of 1024. We use a cosine learning-rate schedule with 7 warmup epochs. All results are reported over three random seeds, $\{0, 1, 2\}$.

Optimizer hyperparameters. For AdamW, we use the standard Adam hyperparameters

$$\beta_1 = 0.9, \qquad \beta_2 = 0.999, \qquad \epsilon_{\text{Adam}} = 10^{-9}.$$

For ROBN, we use momentum coefficient

$$\beta = 0.95,$$

numerical stability constant

$$\epsilon_{\text{ROBN}} = 10^{-8},$$

and

$$K = 5$$

Newton–Schulz steps. Unless otherwise specified, all non-optimizer training settings are kept identical across methods.

Data augmentation and regularization. For training, we use random resized cropping, random horizontal flipping, and ImageNet normalization. For validation, images are resized to 256, center-cropped to 224×224, and normalized using the same ImageNet statistics. We additionally use Mixup with $\alpha = 0.8$, CutMix with $\alpha = 1.0$, label smoothing with coefficient 0.1, and stochastic depth with drop-path rate 0.1. Dropout is set to 0.0.

G.2. Hyperparameter Selection

Before the final comparison, we perform single-seed learning-rate and weight-decay sweeps for the AdamW baseline. We first sweep the learning rate over $\{3 \times 10^{-4}, 5 \times 10^{-4}, 7 \times 10^{-4}\}$ without weight decay and select 3×10^{-4} according to validation accuracy. Using this learning rate, we then perform single-seed weight-decay sweeps and select 0.1 as the final baseline value. The selected hyperparameters are then used in the three-seed comparison reported in Table 1.

Table 2: Learning-rate sweep for the AdamW baseline. The best learning rate is selected according to validation accuracy.

Learning rate	Final val. loss	Final val. acc.	Best val. loss	Best val. acc.
3.0×10^{-4}	**1.0806**	**0.7648**	**1.0792**	**0.7650**
5.0×10^{-4}	1.0971	0.7628	1.0956	0.7630
7.0×10^{-4}	1.0988	0.7591	1.0963	0.7593

Table 3: First weight-decay sweep for AdamW with learning rate 3×10^{-4}.

Weight decay	Final val. loss	Final val. acc.	Best val. loss	Best val. acc.
1.0×10^{-2}	1.0760	0.7694	1.0722	0.7697
3.0×10^{-2}	1.0855	0.7656	1.0806	0.7667
5.0×10^{-2}	1.0742	0.7698	1.0615	0.7702
7.0×10^{-2}	1.0642	0.7702	**1.0489**	0.7704
1.0×10^{-1}	**1.0548**	**0.7724**	1.0493	**0.7729**

Table 4: Second weight-decay sweep for AdamW with learning rate 3×10^{-4}.

Weight decay	Final val. loss	Final val. acc.	Best val. loss	Best val. acc.
1.0×10^{-1}	1.0455	**0.7748**	1.0407	**0.7748**
1.2×10^{-1}	1.0492	0.7733	1.0432	0.7733
1.5×10^{-1}	**1.0282**	0.7730	**1.0256**	0.7733

Across the two weight-decay sweeps, weight decay 0.1 achieves the highest validation accuracy and is therefore used for the AdamW baseline in the final comparison.

Appendix H. Auxiliary Experiment: Split Newton–Schulz Update

This appendix provides an auxiliary ImageNet-1K experiment using a ViT-B/16 backbone. The goal is to test a simple split variant of the update rule, which we call the *Split Newton–Schulz Update* (SNSU).

Given an update matrix G, if $G \in \mathbb{R}^{2n\times m}$, we split it into upper and lower blocks,

$$G = \begin{bmatrix} G_{\text{top}} \\ G_{\text{bottom}} \end{bmatrix}, \qquad G_{\text{top}}, G_{\text{bottom}} \in \mathbb{R}^{n\times m}. \tag{361}$$

We apply the Newton–Schulz procedure to each block separately and concatenate the resulting matrices:

$$\Delta = \begin{bmatrix} \mathrm{NS}(G_{\text{top}}) \\ \mathrm{NS}(G_{\text{bottom}}) \end{bmatrix}. \tag{362}$$

If the update matrix does not have the form $2n \times m$, SNSU uses the same update as Muon.

Rank-one interpretation. SNSU also admits a rank-one interpretation within the one-sided orthogonality framework. Suppose $G \in \mathbb{R}^{2n\times m}$ and write

$$G = \begin{bmatrix} G_{\text{top}} \\ G_{\text{bottom}} \end{bmatrix}, \qquad G_{\text{top}}, G_{\text{bottom}} \in \mathbb{R}^{n\times m}. \tag{363}$$

After applying Newton–Schulz normalization separately to each block, let

$$\mathrm{NS}(G_{\text{top}}) = \sum_i \sigma_i^{\text{top}} u_i^{\text{top}} v_i^{\text{top}\top}, \qquad \mathrm{NS}(G_{\text{bottom}}) = \sum_j \sigma_j^{\text{bottom}} u_j^{\text{bottom}} v_j^{\text{bottom}\top} \tag{364}$$

be rank-one singular expansions of the two normalized blocks, where $\{u_i^{\text{top}}\}_i$, $\{u_j^{\text{bottom}}\}_j$, $\{v_i^{\text{top}}\}_i$, and $\{v_j^{\text{bottom}}\}_j$ are orthonormal systems in their corresponding spaces.

We embed the two normalized blocks into $\mathbb{R}^{2n\times m}$ by zero-padding. Define

$$p_i^{\text{top}} = \begin{bmatrix} u_i^{\text{top}} \\ 0 \end{bmatrix}, \qquad q_i^{\text{top}} = v_i^{\text{top}}, \tag{365}$$

and

$$p_j^{\text{bottom}} = \begin{bmatrix} 0 \\ u_j^{\text{bottom}} \end{bmatrix}, \qquad q_j^{\text{bottom}} = v_j^{\text{bottom}}. \tag{366}$$

Then the SNSU update can be written as

$$\Delta = \sum_i \sigma_i^{\text{top}} p_i^{\text{top}} q_i^{\text{top}\top} + \sum_j \sigma_j^{\text{bottom}} p_j^{\text{bottom}} q_j^{\text{bottom}\top}. \tag{367}$$

Equivalently, setting

$$M_i^{\text{top}} = p_i^{\text{top}} q_i^{\text{top}\top}, \qquad M_j^{\text{bottom}} = p_j^{\text{bottom}} q_j^{\text{bottom}\top}, \tag{368}$$

we obtain

$$\Delta = \sum_i a_i^{\text{top}} M_i^{\text{top}} + \sum_j a_j^{\text{bottom}} M_j^{\text{bottom}}, \qquad a_i^{\text{top}} = \sigma_i^{\text{top}}, \quad a_j^{\text{bottom}} = \sigma_j^{\text{bottom}}. \tag{369}$$

Now combine the top and bottom rank-one systems into a single global indexing. That is, we identify

$$\{p_k(t)\}_k = \{p_i^{\text{top}}\}_i \cup \{p_j^{\text{bottom}}\}_j, \qquad \{q_k(t)\}_k = \{q_i^{\text{top}}\}_i \cup \{q_j^{\text{bottom}}\}_j, \tag{370}$$

and

$$\{a_k(t)\}_k = \{a_i^{\text{top}}\}_i \cup \{a_j^{\text{bottom}}\}_j = \{\sigma_i^{\text{top}}\}_i \cup \{\sigma_j^{\text{bottom}}\}_j. \tag{371}$$

With this notation,

$$\Delta = \sum_k a_k(t) p_k(t) q_k(t)^\top. \tag{372}$$

The zero-padding makes the top and bottom left factors have disjoint row supports:

$$\left\langle p_i^{\text{top}}, p_j^{\text{bottom}} \right\rangle = 0 \qquad \text{for all } i, j. \tag{373}$$

Together with the orthonormality of the left singular vectors within each block, this implies that the combined left factors are orthonormal. Therefore, for the global active index set

$$I_t := \{k : a_k(t) \neq 0\}, \tag{374}$$

we have

$$\langle p_k(t), p_\ell(t) \rangle = \delta_{k\ell}, \qquad k, \ell \in I_t. \tag{375}$$

Thus SNSU is an instance of the one-sided rank-one orthogonality framework, with the left-side orthogonality condition holding throughout. Since the split pattern is fixed rather than time-resampled, the corresponding potential is time-independent:

$$\partial_t K_t(m_t) = 0. \tag{376}$$

We compare SNSU with Muon under the same ViT-B/16 ImageNet-1K setting as Appendix G.1. Both Muon and SNSU use learning rate 3×10^{-3}, weight decay 0.07, and are trained for 150 epochs. All other hyperparameters and training settings are kept identical to Appendix G.1. This auxiliary experiment is conducted with a single seed, so the results should be interpreted as qualitative evidence rather than as a statistically robust comparison.

As shown in Table 5 and Figure 1, SNSU exhibits validation behavior that is nearly identical to Muon, even though the update matrix is split into two blocks and Newton–Schulz normalization is applied separately to each block. The single-seed results show slightly higher best and final validation accuracy for SNSU, but the main purpose of this auxiliary experiment is not to claim a statistically robust improvement. Rather, the result suggests that such a split Newton–Schulz construction can preserve the essential ImageNet-scale training behavior of Muon within the time-independent potential setting $\partial_t K_t(m_t) = 0$.

Method	Best Val. Acc.	Final Val. Acc.
Muon	0.7975	0.7973
SNSU	**0.8016**	**0.8014**

Table 5: Auxiliary single-seed ImageNet-1K results with a ViT-B/16 backbone.

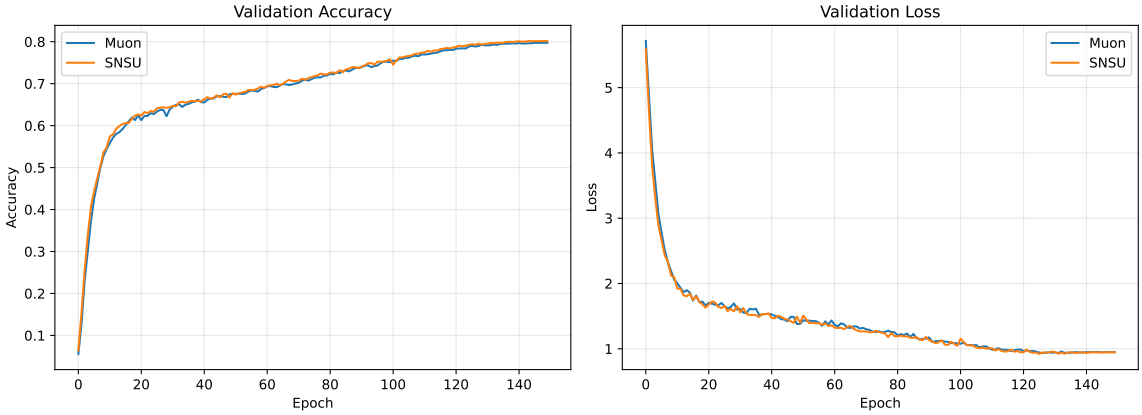


Figure 1: Auxiliary single-seed validation curves for Muon and SNSU on ImageNet-1K using a ViT-B/16 backbone.

Appendix I. Additional Diagnostic for ROBN Primal-Variation Boundedness

We provide an additional diagnostic check for the unweighted primal-variation condition used to verify the ROBN variation assumption. The purpose of this experiment is not to prove the condition in full generality, but to check whether the cumulative basis-induced primal variation remains small in a controlled high-momentum training regime.

We consider a synthetic multi-class classification task generated by `make_classification`. The dataset contains 600,000 samples with input dimension 512 and 15 classes. The data are split into training and validation sets with a validation ratio of 0.1. Input features are scaled using a robust scaler, and labels are one-hot encoded.

The model is a fully connected MLP with 10 hidden layers, hidden width 1024, LeakyReLU activations, no bias terms, and an output softmax layer. We train using the ROBN update rule with momentum coefficient $\beta = 0.95$, batch size 4096, learning rate 10^{-3}, weight decay 0.07, and no learning rate scheduler. The experiment is run with distributed data parallel training over four GPUs.

During training, we monitor the cumulative primal-variation term

$$\sum_{t=0}^{T-1} V_t, \qquad V_t := K_{t+1}(M_{t+1}) - K_t(M_{t+1}). \tag{377}$$

This is the unweighted quantity whose uniform upper boundedness is sufficient for the weighted variation assumption in Appendix E.

At the end of training, the ROBN model reaches training loss 0.06453 and training accuracy 0.98188, indicating that the monitored trajectory corresponds to a normal optimization run rather than a degenerate one. The measured cumulative primal-variation statistic is

$$\sum_{t=0}^{T-1} V_t = 0.006165, \qquad T = 19954. \tag{378}$$

Equivalently,

$$\sum_{t=0}^{T-1} [K_{t+1}(M_{t+1}) - K_t(M_{t+1})] = 0.006165. \tag{379}$$

Thus, in this controlled high-momentum ROBN run, the cumulative basis-induced primal variation remains finite and numerically very small over the monitored training horizon. This diagnostic does not prove the idealized uniform bound

$$\sup_{T\geq 1} \sum_{t=0}^{T-1} [K_{t+1}(M_{t+1}) - K_t(M_{t+1})] < \infty, \tag{380}$$

but it provides empirical support for the plausibility of this assumption in the tested ROBN regime.

LLM Usage

Large language models were used only as writing aids to improve clarity, grammar, and organization. The theoretical ideas, proof strategy, experimental design, and interpretation of results were developed and directed by the authors. All technical content and final manuscript decisions were reviewed and approved by the authors.