Rank-One Potential Geometry for Normalized Optimizers

author names withheld

Under Review for the Workshop on High-dimensional Learning Dynamics, 2026

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-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 C and 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-

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

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

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,

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

where the second term denotes the explicit time-variation of the potential. The standard cases discussed below correspond to the regime in which this explicit contribution vanishes, while our proposed extension also allows it 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:

$$a_i(t) = \alpha_t, \qquad \alpha_t = \frac{\|m_t\|}{\sqrt{n}}, \qquad i = 1, \dots, n. \tag{6}$$

For $m_t \neq 0$, the corresponding subgradient contains

$$\sum_{i=1}^{n} M_i(t) = \frac{\sqrt{n}}{\|m_t\|} m_t \in \partial_m K_t(m_t). \tag{7}$$

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

$$M_{ij} = e_i e_j^\top, \qquad a_{ij}(t) = \langle m_t, M_{ij} \rangle = (m_t)_{ij}, \tag{8}$$

where e_i and e_j are standard basis vectors. Then

$$K(m_t) = \sum_{i,j} |a_{ij}(t)| = \sum_{i,j} |(m_t)_{ij}| = \|m_t\|_1. \tag{9}$$

Consequently, the element-wise sign direction is one valid subgradient:

$$\operatorname{sign}(m_t) \in \partial_m K(m_t), \tag{10}$$

with the usual set-valued interpretation at zero entries. 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

$$M_i(t) = p_i(t) q_i(t)^\top, \qquad I_t := \{i : a_i(t) \neq 0\}. \tag{11}$$

Assume that, on the active set I_t,

$$\langle p_i(t), p_j(t) \rangle = \delta_{ij}, \qquad \langle q_i(t), q_j(t) \rangle = \delta_{ij}, \qquad i, j \in I_t. \tag{12}$$

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

$$K(m_t) = \sum_{i \in I_t} |a_i(t)| = \|m_t\|_*. \tag{13}$$

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

$$\partial_t K_t(m) = 0 \qquad \text{for SGD, Lion, and Muon.} \tag{14}$$

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 E. For a rank-one system and active index set

$$M_i(t) = p_i(t)q_i(t)^\top, \qquad I_t := \{i : a_i(t) \neq 0\}, \tag{15}$$

ROBN preserves a one-sided orthogonality condition of the form

$$\langle p_i(t), p_j(t)\rangle = \delta_{ij} \quad \text{or} \quad \langle q_i(t), q_j(t)\rangle = \delta_{ij}, \qquad i, j \in I_t. \tag{16}$$

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

$$\partial_t K_t(m_t) \neq 0. \tag{17}$$

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 G.

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 F.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 F.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.

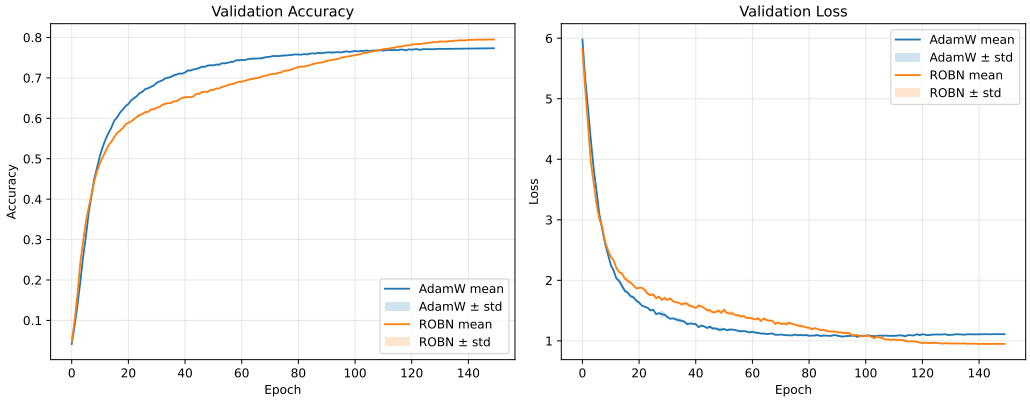
4.1. Results

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.

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.

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

$$m \in \mathbb{R}^{p\times q}. \tag{18}$$

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

$$B = m^\top m \quad \text{or} \quad B = mm^\top. \tag{19}$$

Then B is symmetric positive semidefinite. Let

$$\lambda_1, \dots, \lambda_N \tag{20}$$

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

$$\operatorname{rank}(m) = \operatorname{rank}(B) \le n. \tag{21}$$

Define

$$d = \left(\frac{\operatorname{tr}(B)}{n}, \dots, \frac{\operatorname{tr}(B)}{n}, 0, \dots, 0\right) \in \mathbb{R}^N, \tag{22}$$

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

Since

$$\sum_{i=1}^{N} d_i = \operatorname{tr}(B) = \sum_{i=1}^{N} \lambda_i, \tag{23}$$

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

$$\operatorname{diag}(Q^\top B Q) = d. \tag{24}$$

Let

$$Q = \begin{bmatrix} u_1 & u_2 & \cdots & u_N \end{bmatrix}. \tag{25}$$

Then

$$u_i^\top B u_i = d_i. \tag{26}$$

Moreover,

$$\frac{\operatorname{tr}(B)}{n} = \frac{\|m\|_F^2}{n}. \tag{27}$$

Define

$$\alpha = \frac{\|m\|_F}{\sqrt{n}}. \tag{28}$$

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

$$\sqrt{d_i} = \alpha. \tag{29}$$

Let

$$I = \{i : d_i \neq 0\}. \tag{30}$$

Then $|I| = n$.

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

$$u_i^\top m^\top m u_i = d_i. \tag{31}$$

Hence

$$\|mu_i\|_2 = \sqrt{d_i} = \alpha. \tag{32}$$

Define

$$v_i = \frac{mu_i}{\alpha}. \tag{33}$$

Then

$$\|v_i\|_2 = 1. \tag{34}$$

Moreover,

$$mu_i = \alpha v_i. \tag{35}$$

Define

$$M_i = v_i u_i^\top, \qquad i \in I. \tag{36}$$

Then each M_i is rank one. Also,

$$\langle M_i, M_j \rangle_F = \operatorname{tr}(M_i^\top M_j) = \operatorname{tr}(u_i v_i^\top v_j u_j^\top) = (v_i^\top v_j)(u_j^\top u_i). \tag{37}$$

Since the vectors u_i are orthonormal, we obtain

$$\langle M_i, M_j \rangle_F = \delta_{ij}. \tag{38}$$

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

Furthermore,

$$\alpha \sum_{i \in I} M_i = \alpha \sum_{i \in I} v_i u_i^\top = \sum_{i \in I} m u_i u_i^\top. \tag{39}$$

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

$$\|mu_i\|_2^2 = u_i^\top m^\top m u_i = u_i^\top B u_i = d_i = 0. \tag{40}$$

Thus

$$mu_i = 0 \qquad \text{for } i \notin I. \tag{41}$$

Therefore,

$$\sum_{i \in I} m u_i u_i^\top = \sum_{i=1}^{N} m u_i u_i^\top. \tag{42}$$

Since Q is orthogonal,

$$\sum_{i=1}^{N} u_i u_i^\top = QQ^\top = I. \tag{43}$$

Hence

$$\alpha \sum_{i \in I} M_i = m \sum_{i=1}^{N} u_i u_i^\top = mQQ^\top = m. \tag{44}$$

Thus,

$$m = \alpha \sum_{i \in I} M_i. \tag{45}$$

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

$$u_i^\top mm^\top u_i = d_i. \tag{46}$$

Hence

$$\|m^\top u_i\|_2 = \sqrt{d_i} = \alpha. \tag{47}$$

Define

$$v_i = \frac{m^\top u_i}{\alpha}. \tag{48}$$

Then

$$\|v_i\|_2 = 1. \tag{49}$$

Moreover,

$$m^\top u_i = \alpha v_i. \tag{50}$$

Define

$$M_i = u_i v_i^\top, \qquad i \in I. \tag{51}$$

Then each M_i is rank one. Also,

$$\langle M_i, M_j \rangle_F = \operatorname{tr}(M_i^\top M_j) = \operatorname{tr}(v_i u_i^\top u_j v_j^\top) = (u_i^\top u_j)(v_j^\top v_i). \tag{52}$$

Since the vectors u_i are orthonormal, we obtain

$$\langle M_i, M_j \rangle_F = \delta_{ij}. \tag{53}$$

Furthermore,

$$\alpha \sum_{i \in I} M_i = \alpha \sum_{i \in I} u_i v_i^\top = \sum_{i \in I} u_i u_i^\top m. \tag{54}$$

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

$$\|m^\top u_i\|_2^2 = u_i^\top mm^\top u_i = u_i^\top B u_i = d_i = 0. \tag{55}$$

Thus

$$m^\top u_i = 0 \qquad \text{for } i \notin I. \tag{56}$$

Therefore,

$$\sum_{i \in I} u_i u_i^\top m = \sum_{i=1}^{N} u_i u_i^\top m. \tag{57}$$

Since Q is orthogonal,

$$\sum_{i=1}^{N} u_i u_i^\top = QQ^\top = I. \tag{58}$$

Hence

$$\alpha \sum_{i \in I} M_i = \left(\sum_{i=1}^{N} u_i u_i^\top \right) m = QQ^\top m = m. \tag{59}$$

Thus,

$$m = \alpha \sum_{i \in I} M_i. \tag{60}$$

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

$$m = \alpha \sum_{i \in I} M_i, \qquad \alpha = \frac{\|m\|_F}{\sqrt{n}}, \tag{61}$$

where

$$\langle M_i, M_j \rangle_F = \delta_{ij}, \qquad \operatorname{rank}(M_i) = 1. \tag{62}$$

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

Suppose

$$m = \sum_i a_i p_i q_i^\top. \tag{63}$$

Let

$$I := \{i : \ a_i \neq 0\} \tag{64}$$

be the active index set, and define

$$M_i = p_i q_i^\top, \qquad i \in I. \tag{65}$$

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

$$\langle p_i, p_j \rangle = \delta_{ij}, \qquad \langle q_i, q_j \rangle = \delta_{ij}. \tag{66}$$

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

$$\begin{aligned}
\langle M_i, M_j \rangle_F &= \left\langle p_i q_i^\top , \, p_j q_j^\top \right\rangle_F && (67)\\
&= \operatorname{tr}\Big((p_i q_i^\top)^\top p_j q_j^\top \Big) && (68)\\
&= \operatorname{tr}\Big(q_i p_i^\top p_j q_j^\top \Big) && (69)\\
&= \langle p_i, p_j \rangle \langle q_i, q_j \rangle && (70)\\
&= \delta_{ij}. && (71)
\end{aligned}$$

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

$$\widetilde{p}_i = \operatorname{sign}(a_i) p_i. \tag{72}$$

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

$$m = \sum_{i \in I} |a_i| \, \widetilde{p}_i q_i^\top. \tag{73}$$

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

$$mq_j = \sum_{i\in I} |a_i|\, \widetilde{p}_i q_i^\top q_j \tag{74}$$

$$= \sum_{i\in I} |a_i|\, \widetilde{p}_i \langle q_i, q_j\rangle \tag{75}$$

$$= |a_j|\, \widetilde{p}_j. \tag{76}$$

Similarly,

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

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

$$= |a_j|\, q_j. \tag{79}$$

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

$$m^\top m q_j = |a_j|^2 q_j. \tag{80}$$

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

$$\|m\|_* = \sum_{i\in I} |a_i| = \sum_i |a_i|. \tag{81}$$

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

$$K(m) = \sum_i |a_i| \tag{82}$$

coincides with the nuclear norm:

$$K(m) = \|m\|_*. \tag{83}$$

A.3. Zero Explicit Time Contribution of Standard Examples

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

$$\langle V_i(t), V_j(t)\rangle = \delta_{ij}. \tag{84}$$

Then, for $h > 0$,

$$\begin{aligned}
\langle V_i(t+h), V_j(t+h)\rangle &= \left\langle V_i(t) + h\frac{V_i(t+h)-V_i(t)}{h}, V_j(t) + h\frac{V_j(t+h)-V_j(t)}{h}\right\rangle \\
&= \langle V_i(t), V_j(t)\rangle + h\left\langle V_i(t), \frac{V_j(t+h)-V_j(t)}{h}\right\rangle \\
&\quad + h\left\langle V_j(t), \frac{V_i(t+h)-V_i(t)}{h}\right\rangle \\
&\quad + h^2\left\langle \frac{V_i(t+h)-V_i(t)}{h}, \frac{V_j(t+h)-V_j(t)}{h}\right\rangle.
\end{aligned} \tag{85}$$

Since

$$\langle V_i(t+h), V_j(t+h)\rangle = \langle V_i(t), V_j(t)\rangle = \delta_{ij}, \tag{86}$$

we obtain

$$\begin{aligned} 0 = h\left[\left\langle V_i(t), \frac{V_j(t+h)-V_j(t)}{h}\right\rangle + \left\langle V_j(t), \frac{V_i(t+h)-V_i(t)}{h}\right\rangle\right] \\ + h^2\left\langle \frac{V_i(t+h)-V_i(t)}{h}, \frac{V_j(t+h)-V_j(t)}{h}\right\rangle. \end{aligned} \tag{87}$$

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

$$\langle V_i(t), \dot{V}_j(t)\rangle + \langle V_j(t), \dot{V}_i(t)\rangle = 0. \tag{88}$$

Equivalently,

$$\langle \dot{V}_i(t), V_j(t)\rangle + \langle V_i(t), \dot{V}_j(t)\rangle = 0. \tag{89}$$

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

$$\langle \dot{V}_i(t), V_i(t)\rangle = 0. \tag{90}$$

SGD. Suppose that

$$m = \alpha \sum_i M_i, \qquad \alpha > 0, \tag{91}$$

where

$$\langle M_i, M_j\rangle_F = \delta_{ij}. \tag{92}$$

For the coefficient-wise potential

$$K_t(m) = \sum_i |a_i|, \tag{93}$$

the explicit time-variation term is

$$\partial_t K_t(m) = \sum_i \operatorname{sign}(a_i)\langle \dot{M}_i, m\rangle_F. \tag{94}$$

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

$$\operatorname{sign}(a_i) = 1. \tag{95}$$

Therefore,

$$\begin{aligned} \partial_t K_t(m) &= \sum_i \langle \dot{M}_i, m\rangle_F \\ &= \sum_i \left\langle \dot{M}_i, \alpha \sum_j M_j \right\rangle_F \\ &= \alpha \sum_i \sum_j \langle \dot{M}_i, M_j\rangle_F. \end{aligned} \tag{96}$$

By applying (89) to the orthonormal family $\{M_i(t)\}_i$, we have

$$\langle \dot{M}_i, M_j\rangle_F + \langle M_i, \dot{M}_j\rangle_F = 0. \tag{97}$$

Hence,

$$\begin{aligned}\sum_i \sum_j \langle \dot{M}_i, M_j \rangle_F &= -\sum_i \sum_j \langle M_i, \dot{M}_j \rangle_F \\ &= -\sum_j \sum_i \langle \dot{M}_j, M_i \rangle_F \\ &= -\sum_i \sum_j \langle \dot{M}_i, M_j \rangle_F.\end{aligned} \tag{98}$$

Therefore,

$$\sum_i \sum_j \langle \dot{M}_i, M_j \rangle_F = 0. \tag{99}$$

Hence

$$\partial_t K_t(m) = 0. \tag{100}$$

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

$$\{M_i\}_i, \tag{101}$$

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

$$m = \sum_i a_i M_i. \tag{102}$$

The coefficient-wise potential is defined by

$$K_t(m) = \sum_i |a_i|. \tag{103}$$

Since the basis is fixed, we have

$$\dot{M}_i = 0 \tag{104}$$

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

$$\begin{aligned}\partial_t K_t(m) &= \sum_i \operatorname{sign}(a_i) \langle \dot{M}_i, m \rangle_F \\ &= 0.\end{aligned} \tag{105}$$

Hence,

$$\partial_t K_t(m) = 0. \tag{106}$$

Muon. Suppose that

$$m = \sum_i \sigma_i M_i, \qquad \sigma_i > 0, \tag{107}$$

where

$$M_i = u_i v_i^\top. \tag{108}$$

Assume

$$\langle u_i, u_j \rangle = \delta_{ij}, \qquad \langle v_i, v_j \rangle = \delta_{ij}. \tag{109}$$

Then

$$\dot{M}_i = \dot{u}_i v_i^\top + u_i \dot{v}_i^\top. \tag{110}$$

The explicit time-variation term is

$$\partial_t K_t(m) = \sum_i \operatorname{sign}(\sigma_i) \langle \dot{M}_i, m \rangle_F. \tag{111}$$

Since

$$\sigma_i > 0, \tag{112}$$

we have

$$\operatorname{sign}(\sigma_i) = 1. \tag{113}$$

Thus,

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

Substituting

$$m = \sum_j \sigma_j M_j, \tag{115}$$

we get

$$\begin{aligned} \partial_t K_t(m) &= \sum_i \left\langle \dot{M}_i, \sum_j \sigma_j M_j \right\rangle_F \\ &= \sum_i \sum_j \sigma_j \langle \dot{M}_i, M_j \rangle_F. \end{aligned} \tag{116}$$

Since

$$M_j = u_j v_j^\top, \tag{117}$$

we have

$$\begin{aligned} \langle \dot{M}_i, M_j \rangle_F &= \left\langle \dot{u}_i v_i^\top + u_i \dot{v}_i^\top, u_j v_j^\top \right\rangle_F \\ &= \langle \dot{u}_i, u_j \rangle \langle v_i, v_j \rangle + \langle u_i, u_j \rangle \langle \dot{v}_i, v_j \rangle. \end{aligned} \tag{118}$$

Therefore,

$$\begin{aligned} \partial_t K_t(m) &= \sum_i \sum_j \sigma_j \left[\langle \dot{u}_i, u_j \rangle \langle v_i, v_j \rangle + \langle u_i, u_j \rangle \langle \dot{v}_i, v_j \rangle \right] \\ &= \sum_i \sigma_i \left[\langle \dot{u}_i, u_i \rangle + \langle \dot{v}_i, v_i \rangle \right]. \end{aligned} \tag{119}$$

By applying (90) to the orthonormal families $\{u_i(t)\}_i$ and $\{v_i(t)\}_i$, respectively, we have

$$\langle \dot{u}_i, u_i \rangle = 0, \qquad \langle \dot{v}_i, v_i \rangle = 0. \tag{120}$$

Hence

$$\partial_t K_t(m) = 0. \tag{121}$$

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:

$$\partial_t K_t(m) = 0. \tag{122}$$

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,

$$K_t(m) = K(m) \qquad \text{for all } m. \tag{123}$$

Consequently, its Fenchel conjugate is also time-independent:

$$\begin{aligned} K_t^*(y) &= \sup_m \{\langle y, m\rangle_F - K_t(m)\} \\ &= \sup_m \{\langle y, m\rangle_F - K(m)\} \\ &= K^*(y). \end{aligned} \tag{124}$$

Hence,

$$\partial_t K_t^*(y) = 0 \qquad \text{for all } y. \tag{125}$$

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

$$\partial_t K_t(m) = 0, \qquad \partial_t K_t^*(y) = 0. \tag{126}$$

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

$$\chi_{\mathcal{C}}(y) = \begin{cases} 0, & y \in \mathcal{C}, \\ +\infty, & y \notin \mathcal{C}. \end{cases} \tag{127}$$

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

$$K_{\mathrm{SGD}}(m) = \sqrt{n}\|m\|_F. \tag{128}$$

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

$$K^*_{\mathrm{SGD}}(y) = \chi_{\{\|y\|_F \le \sqrt{n}\}}(y). \tag{129}$$

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

$$\min_X F(X) \qquad \text{s.t.} \qquad \|\lambda X\|_F \le \sqrt{n}. \tag{130}$$

Equivalently,

$$\|X\|_F \le \frac{\sqrt{n}}{\lambda}. \tag{131}$$

For Lion, the coordinate-wise sign structure corresponds to

$$K_{\mathrm{Lion}}(m) = \|m\|_1, \qquad \nabla K_{\mathrm{Lion}}(m) = \mathrm{sign}(m). \tag{132}$$

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

$$K^*_{\mathrm{Lion}}(y) = \chi_{\{\|y\|_\infty \le 1\}}(y). \tag{133}$$

Therefore, the corresponding constrained problem is

$$\min_X F(X) \qquad \text{s.t.} \qquad \|\lambda X\|_\infty \le 1, \tag{134}$$

or equivalently,

$$\|X\|_\infty \le \frac{1}{\lambda}. \tag{135}$$

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

$$K_{\mathrm{Muon}}(m) = \|m\|_{\mathrm{tr}}. \tag{136}$$

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

$$K^*_{\mathrm{Muon}}(y) = \chi_{\{\|y\|_{\mathrm{op}} \le 1\}}(y). \tag{137}$$

Thus, the corresponding constrained problem is

$$\min_X F(X) \qquad \text{s.t.} \qquad \|\lambda X\|_{\mathrm{op}} \le 1, \tag{138}$$

or equivalently,

$$\|X\|_{\mathrm{op}} \le \frac{1}{\lambda}. \tag{139}$$

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

$$W_{t+1} = W_t - \eta_t \xi_t, \qquad g_t = \nabla \mathcal{L}(W_t), \tag{140}$$

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,

$$\langle M_i(t), M_j(t)\rangle_F = \delta_{ij}, \quad rank(M_i) = 1 \qquad \text{for all } i,j \in \{1,\dots,N\}. \tag{141}$$

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

$$g_t = \sum_{i=1}^N \langle g_t, M_i(t)\rangle_F M_i(t). \tag{142}$$

Define the time-dependent potential

$$K_t(g) = \sum_{i=1}^N |\langle g, M_i(t)\rangle_F| \,. \tag{143}$$

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

$$\xi_t \in \partial K_t(g_t) \tag{144}$$

as

$$\xi_t = \sum_{i\in I_t} \operatorname{sign}\left(\langle g_t, M_i(t)\rangle_F\right) M_i(t), \tag{145}$$

where

$$I_t = \{i : \langle g_t, M_i(t)\rangle_F \neq 0\}, \qquad N_t = |I_t|. \tag{146}$$

By construction,

$$\langle g_t, \xi_t\rangle_F = \left\langle g_t, \sum_{i\in I_t} \operatorname{sign}\left(\langle g_t, M_i(t)\rangle_F\right) M_i(t) \right\rangle_F \tag{147}$$

$$= \sum_{i\in I_t} |\langle g_t, M_i(t)\rangle_F| \,. \tag{148}$$

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

$$\|\xi_t\|_F^2 = \left\| \sum_{i\in I_t} \operatorname{sign}\left(\langle g_t, M_i(t)\rangle_F\right) M_i(t) \right\|_F^2 \tag{149}$$

$$= \sum_{i\in I_t} 1 = N_t. \tag{150}$$

Define

$$\mu_t = \frac{1}{N_t} \sum_{i \in I_t} |\langle g_t, M_i(t) \rangle_F| , \tag{151}$$

whenever $N_t > 0$. Then

$$\langle g_t, \xi_t \rangle_F = N_t \mu_t, \qquad \|\xi_t\|_F^2 = N_t. \tag{152}$$

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

$$\mathcal{L}(W_{t+1}) \leq \mathcal{L}(W_t) + \langle \nabla \mathcal{L}(W_t), W_{t+1} - W_t \rangle_F + \frac{L}{2} \|W_{t+1} - W_t\|_F^2 \tag{153}$$

$$= \mathcal{L}(W_t) - \eta_t \langle g_t, \xi_t \rangle_F + \frac{L \eta_t^2}{2} \|\xi_t\|_F^2 \tag{154}$$

$$= \mathcal{L}(W_t) - \eta_t N_t \mu_t + \frac{L \eta_t^2}{2} N_t \tag{155}$$

$$= \mathcal{L}(W_t) - \eta_t N_t \left(\mu_t - \frac{L \eta_t}{2} \right). \tag{156}$$

Therefore, if

$$0 < \forall \eta_t < \frac{2 \mu_t}{L}, \tag{157}$$

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

$$\eta_t = \frac{\mu_t}{L} \tag{158}$$

gives

$$\mathcal{L}(W_{t+1}) \leq \mathcal{L}(W_t) - \frac{N_t \mu_t^2}{2L}. \tag{159}$$

Define

$$D_t = \frac{N_t \mu_t^2}{2L}. \tag{160}$$

Then

$$\mathcal{L}(W_{t+1}) \leq \mathcal{L}(W_t) - D_t. \tag{161}$$

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

$$\sum_{t=0}^{T-1} D_t \leq \mathcal{L}(W_0) - \mathcal{L}(W_T). \tag{162}$$

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

$$\sum_{t=0}^{T-1} D_t \leq \mathcal{L}(W_0) - \mathcal{L}(W^*) < \infty. \tag{163}$$

Hence,

$$\sum_{t=0}^{\infty} D_t < \infty, \qquad D_t \to 0. \tag{164}$$

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

$$D_t = \frac{1}{2LN_t} \left(\sum_{i \in I_t} |\langle g_t, M_i(t) \rangle_F| \right)^2 . \tag{165}$$

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

$$D_t \geq \frac{1}{2LN_t} \sum_{i \in I_t} |\langle g_t, M_i(t) \rangle_F|^2 . \tag{166}$$

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

$$\sum_{i \in I_t} |\langle g_t, M_i(t) \rangle_F|^2 = \|g_t\|_F^2. \tag{167}$$

Therefore,

$$D_t \geq \frac{\|g_t\|_F^2}{2LN_t}. \tag{168}$$

Since $N_t \leq N$, we further obtain

$$D_t \geq \frac{\|g_t\|_F^2}{2LN}. \tag{169}$$

Because $D_t \to 0$, it follows that

$$\|g_t\|_F^2 \to 0. \tag{170}$$

Equivalently,

$$\nabla \mathcal{L}(W_t) \to 0. \tag{171}$$

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. Continuous-Time Convergence Analysis for Momentum Dynamics with Time-Dependent Potentials

This appendix follows the static Lion-$\mathcal{K}$/Muon Lyapunov argument and isolates the additional terms that arise when the static potential K is replaced by a time-dependent potential K_t. We therefore do not attempt to rederive every step of the standard static-potential proof; instead, we focus on the explicit time-variation terms of K_t and K_t^*, and on the assumptions under which these terms can be controlled.

C.1. 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

$$K_t(m) := \sum_{i=1}^N |\langle m, M_i(t) \rangle| . \tag{172}$$

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

$$K_t^*(y) = \sup_{m \in \mathcal{X}} \{\langle y, m \rangle - K_t(m)\} . \tag{173}$$

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

$$m = \sum_{i=1}^N a_i M_i(t), \qquad a_i = \langle m, M_i(t) \rangle. \tag{174}$$

Hence, by (172),

$$K_t(m) = \sum_{i=1}^N |a_i|. \tag{175}$$

Define

$$\mathcal{C}_t := \left\{ y \in \mathcal{X} : \max_{1 \le i \le N} |\langle y, M_i(t) \rangle| \le 1 \right\} . \tag{176}$$

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

$$\chi_{\mathcal{C}_t}(y) := \begin{cases} 0, & y \in \mathcal{C}_t, \\ +\infty, & y \notin \mathcal{C}_t. \end{cases} \tag{177}$$

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

$$|\langle y, M_i(t) \rangle| \le 1, \qquad i = 1, \dots, N. \tag{178}$$

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

$$\begin{aligned} \langle y, m \rangle &= \sum_{i=1}^N a_i \langle y, M_i(t) \rangle && (179) \\ &\le \sum_{i=1}^N |a_i| \, |\langle y, M_i(t) \rangle| && (180) \\ &\le \sum_{i=1}^N |a_i| && (181) \\ &= K_t(m). && (182) \end{aligned}$$

Therefore,

$$\langle y, m\rangle - K_t(m) \le 0, \qquad \forall m \in \mathcal{X}. \tag{183}$$

Since equality is attained at $m = 0$, it follows from (173) that

$$K_t^*(y) = 0. \tag{184}$$

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

$$|\langle y, M_j(t)\rangle| > 1. \tag{185}$$

For $s > 0$, choose

$$m_s = s \ \mathrm{sign}\left(\langle y, M_j(t)\rangle\right) M_j(t). \tag{186}$$

Then

$$K_t(m_s) = s, \tag{187}$$

and

$$\langle y, m_s\rangle = s\,|\langle y, M_j(t)\rangle|\,. \tag{188}$$

Hence

$$\langle y, m_s\rangle - K_t(m_s) = s\left(|\langle y, M_j(t)\rangle| - 1\right) \tag{189}$$

$$\to +\infty \qquad \text{as } s \to \infty. \tag{190}$$

Thus,

$$K_t^*(y) = +\infty. \tag{191}$$

Combining the two cases, we obtain

$$K_t^*(y) = \begin{cases} 0, & \max\limits_{1\le i\le N} |\langle y, M_i(t)\rangle| \le 1, \\ +\infty, & \text{otherwise.} \end{cases} \tag{192}$$

Equivalently,

$$K_t^*(y) = \chi_{\mathcal{C}_t}(y). \tag{193}$$

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,

$$|\langle y, M_i(t)\rangle| \le \|y\|\,\|M_i(t)\| = \|y\|. \tag{194}$$

Hence,

$$\|y\| \le 1 \quad \Longrightarrow \quad y \in \mathcal{C}_t, \qquad \forall t \ge 0. \tag{195}$$

By (193), this implies

$$\|y\| \le 1 \quad \Longrightarrow \quad K_t^*(y) = 0, \qquad \forall t \ge 0. \tag{196}$$

Assume that the trajectory is bounded:

$$\sup_{t\ge 0} \|x_t\| \le R \tag{197}$$

for some $R > 0$. If

$$\lambda \leq \frac{1}{R}, \tag{198}$$

then

$$\|\lambda x_t\| \leq \lambda R \leq 1, \qquad \forall t \geq 0. \tag{199}$$

Therefore,

$$K_t^*(\lambda x_t) = 0, \qquad \forall t \geq 0. \tag{200}$$

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

$$\|\lambda x_t\| \leq 1. \tag{201}$$

Hence, by (196),

$$K_s^*(\lambda x_t) = 0, \qquad \forall s \geq 0. \tag{202}$$

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

$$\partial_s K_s^*(\lambda x_t)\big|_{s=t} = 0, \qquad \forall t \geq 0. \tag{203}$$

Equivalently, we write

$$\partial_t K_t^*(\lambda x_t) = 0, \qquad \forall t \geq 0. \tag{204}$$

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

C.2. Momentum convergence under time-varying potentials

We now analyze the momentum dynamics under the time-dependent Lion-$\mathcal{K}_t$-type continuous-time dynamics

$$\dot{m}_t = -\nabla F(x_t) - m_t, \tag{205}$$

$$\dot{x}_t = \nabla K_t(m_t^f) - \lambda x_t, \tag{206}$$

where

$$m_t^f := m_t - \epsilon(\nabla F(x_t) + m_t), \qquad \epsilon \in (0, 1]. \tag{207}$$

Here K_t is a time-dependent convex potential and K_t^* denotes its Fenchel conjugate.

We assume throughout this subsection that

$$F^\star := \inf_x F(x) > -\infty, \tag{208}$$

$$\sup_{t \geq 0} \|x_t\| \leq R \quad \text{and} \quad \lambda \leq \frac{1}{R}, \tag{209}$$

$$K_t \longrightarrow K_\infty \quad \text{as } t \to \infty. \tag{210}$$

Since K_t may be nonsmooth, the explicit time-variation term is interpreted in the a.e. sense: for a.e. t, the map $s \mapsto K_s(m_t)$ is differentiable at $s = t$, and

$$\partial_s K_s(m_t)|_{s=t} := \frac{d}{ds} K_s(m_t)\bigg|_{s=t}.$$

We assume that this explicit time-variation has nonpositive cumulative drift:

$$\int_0^T \partial_s K_s(m_t)|_{s=t}\, dt \le 0, \qquad \forall T \ge 0. \tag{211}$$

The convergence in (210) is understood in a sense strong enough to pass to the limit in the subgradient graph: if $m_t \to m_\infty$, $y_t \to y_\infty$, and $y_t \in \partial K_t(m_t)$, then

$$y_\infty \in \partial K_\infty(m_\infty). \tag{212}$$

Motivated by the Lion-$\mathcal{K}$ Lyapunov function, define

$$H_t(x,m) := F(x) + \frac{1}{\lambda} K_t^*(\lambda x) + c\left(K_t^*(\lambda x) + K_t(m) - \langle m, \lambda x\rangle\right), \tag{213}$$

where

$$c := \frac{1-\epsilon}{1+\epsilon\lambda} \ge 0. \tag{214}$$

The last term in (213) is the Fenchel gap between m and λx, and hence

$$K_t^*(\lambda x) + K_t(m) - \langle m, \lambda x\rangle \ge 0. \tag{215}$$

By the result of the previous subsection, the boundedness assumption (209) implies

$$K_t^*(\lambda x_t) = 0, \qquad \partial_t K_t^*(\lambda x_t) = 0, \qquad \forall t \ge 0. \tag{216}$$

Consequently, along the trajectory,

$$H_t(x_t, m_t) \ge F(x_t) \ge F^\star. \tag{217}$$

For a fixed potential K_t, the standard Lion-$\mathcal{K}$ Lyapunov calculation can be written explicitly as a sum of monotonicity gaps. To see this, freeze K_t and write

$$m_t^f = m_t - \epsilon(\nabla F(x_t) + m_t). \tag{218}$$

For notational simplicity, choose

$$z_t \in \partial K_t^*(\lambda x_t).$$

In the smooth case, this is simply $z_t = \nabla K_t^*(\lambda x_t)$. Then the frozen-$K_t$ Lyapunov calculation gives

$$\begin{aligned}\frac{d}{dt} H_t(x_t, m_t)\Big|_{K_t \text{ frozen}} \le &-\frac{\lambda+1}{1+\epsilon\lambda}\left\langle \nabla K_t(m_t^f) - \lambda x_t,\, m_t^f - z_t\right\rangle \\ &- \frac{1-\epsilon}{\epsilon(1+\epsilon\lambda)}\left\langle \nabla K_t(m_t^f) - \nabla K_t(m_t),\, m_t^f - m_t\right\rangle.\end{aligned} \tag{219}$$

For notational clarity, define the two monotonicity gaps

$$\Gamma_t := \left\langle \nabla K_t(m_t^f) - \lambda x_t,\, m_t^f - z_t\right\rangle, \tag{220}$$

$$\Delta_t := \left\langle \nabla K_t(m_t^f) - \nabla K_t(m_t),\, m_t^f - m_t\right\rangle. \tag{221}$$

Since $z_t \in \partial K_t^*(\lambda x_t)$, Fenchel conjugacy gives

$$\lambda x_t \in \partial K_t(z_t). \tag{222}$$

Together with

$$\nabla K_t(m_t^f) \in \partial K_t(m_t^f),$$

the monotonicity of ∂K_t yields

$$\Gamma_t \geq 0. \tag{223}$$

Similarly, since

$$\nabla K_t(m_t^f) \in \partial K_t(m_t^f), \qquad \nabla K_t(m_t) \in \partial K_t(m_t),$$

the monotonicity of ∂K_t gives

$$\Delta_t \geq 0. \tag{224}$$

Now define

$$D_t := a\,\Gamma_t + b\,\Delta_t, \qquad a := \frac{\lambda+1}{1+\epsilon\lambda} > 0, \qquad b := \frac{1-\epsilon}{\epsilon(1+\epsilon\lambda)} \geq 0. \tag{225}$$

Equivalently,

$$\begin{aligned} D_t = {} & \frac{\lambda+1}{1+\epsilon\lambda}\left\langle \nabla K_t(m_t^f) - \lambda x_t,\, m_t^f - z_t \right\rangle \\ & + \frac{1-\epsilon}{\epsilon(1+\epsilon\lambda)}\left\langle \nabla K_t(m_t^f) - \nabla K_t(m_t),\, m_t^f - m_t \right\rangle. \end{aligned} \tag{226}$$

Therefore,

$$D_t \geq 0. \tag{227}$$

and hence

$$\frac{d}{dt} H_t(x_t, m_t)\Big|_{K_t \text{ frozen}} \leq -D_t. \tag{228}$$

At times where the relevant a.e. derivatives exist, the total derivative of $H_t(x_t, m_t)$ can be decomposed into the frozen-K_t contribution and the explicit time-variation of the potentials:

$$\frac{d}{dt} H_t(x_t, m_t) = \frac{d}{dt} H_t(x_t, m_t)\Big|_{K_t \text{ frozen}} + \left(\frac{1}{\lambda} + c\right) \partial_s K_s^*(\lambda x_t)|_{s=t} + c\,\partial_s K_s(m_t)|_{s=t}\,. \tag{229}$$

Using the frozen-K_t Lyapunov estimate, we obtain, for a.e. $t \geq 0$,

$$\frac{d}{dt} H_t(x_t, m_t) \leq -D_t + \left(\frac{1}{\lambda} + c\right) \partial_s K_s^*(\lambda x_t)|_{s=t} + c\,\partial_s K_s(m_t)|_{s=t}\,. \tag{230}$$

By (203), we have

$$\partial_s K_s^*(\lambda x_t)|_{s=t} = 0, \qquad \text{for a.e. } t \geq 0. \tag{231}$$

Therefore,

$$\frac{d}{dt} H_t(x_t, m_t) \leq -D_t + c\,\partial_s K_s(m_t)|_{s=t}\,, \qquad \text{for a.e. } t \geq 0. \tag{232}$$

Using the cumulative nonpositive-drift assumption (211), integrating (232) from 0 to T, we obtain

$$\begin{aligned} H_T(x_T, m_T) - H_0(x_0, m_0) &\le -\int_0^T D_t\, dt + c\int_0^T \partial_s K_s(m_t)|_{s=t}\, dt \\ &\le -\int_0^T D_t\, dt, \end{aligned} \tag{233}$$

where the second inequality follows from (211). Hence

$$H_T(x_T, m_T) + \int_0^T D_t\, dt \le H_0(x_0, m_0). \tag{234}$$

Since $H_T(x_T, m_T) \ge F^\star$, it follows that

$$\int_0^\infty D_t\, dt \le H_0(x_0, m_0) - F^\star < \infty. \tag{235}$$

Assume that D_t is uniformly continuous along the trajectory. Then, by Barbalat's lemma, since $D_t \ge 0$ and $\int_0^\infty D_t\, dt < \infty$, we have

$$D_t \longrightarrow 0 \qquad \text{as } t \to \infty. \tag{236}$$

Since

$$D_t = a\Gamma_t + b\Delta_t, \qquad a > 0, \quad b \ge 0,$$

and $\Gamma_t, \Delta_t \ge 0$, the convergence $D_t \to 0$ implies

$$\Gamma_t \longrightarrow 0. \tag{237}$$

If $b > 0$, then it also implies

$$\Delta_t \longrightarrow 0. \tag{238}$$

Define the Fenchel–Young gap between m_t^f and λx_t by

$$\mathcal{G}_t := K_t(m_t^f) + K_t^*(\lambda x_t) - \langle \lambda x_t, m_t^f \rangle. \tag{239}$$

By the Fenchel–Young inequality,

$$\mathcal{G}_t \ge 0. \tag{240}$$

Moreover, since $z_t \in \partial K_t^*(\lambda x_t)$, Fenchel conjugacy gives

$$K_t(z_t) + K_t^*(\lambda x_t) = \langle \lambda x_t, z_t \rangle. \tag{241}$$

Therefore,

$$\begin{aligned} \mathcal{G}_t &= K_t(m_t^f) - K_t(z_t) - \langle \lambda x_t, m_t^f - z_t \rangle \\ &\le \Big\langle \nabla K_t(m_t^f), m_t^f - z_t \Big\rangle - \Big\langle \lambda x_t, m_t^f - z_t \Big\rangle \\ &= \Big\langle \nabla K_t(m_t^f) - \lambda x_t,\ m_t^f - z_t \Big\rangle = \Gamma_t. \end{aligned} \tag{242}$$

Combining the preceding inequalities, we obtain

$$0 \le \mathcal{G}_t \le \Gamma_t. \tag{243}$$

Hence $\Gamma_t \to 0$ implies

$$\mathcal{G}_t \longrightarrow 0. \tag{244}$$

Using $K_t^*(\lambda x_t) = 0$, this gives

$$K_t(m_t^f) - \langle \lambda x_t, m_t^f \rangle \longrightarrow 0. \tag{245}$$

Thus the Fenchel–Young inequality becomes asymptotically tight between m_t^f and λx_t.

Invoking the Lion-$\mathcal{K}$ momentum-tracking estimate used in Lemma 3 and in the proof of Theorem 6 of [2],

$$m_t^f + \nabla F(x_t) \to 0,$$

the preceding Fenchel-gap convergence can be converted into the limiting KKT condition by the same gap-to-KKT argument as in Proposition 2 and Section 7.4 of [2]. In particular, since $K_t \to K_\infty$ in the subgradient-graph sense assumed above, the asymptotic Fenchel consistency between m_t^f and λx_t passes to the limiting potential K_∞. Hence every limiting point satisfies the KKT-type condition associated with the induced limiting constrained problem. This completes the reduction of the time-dependent potential case to the standard Lion-$\mathcal{K}$/Muon gap-to-KKT convergence argument.

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

This appendix is a discrete-time analogue of the preceding time-dependent Lyapunov analysis. As in the static Lion-$\mathcal{K}$/Muon proof, we use the same Lyapunov structure, but replace the static potential K by a time-dependent potential K_t. The purpose is not to rederive the full static-potential argument, but to isolate the additional discrete variation terms caused by the change from K_t to K_{t+1}, and to state conditions under which these terms do not contribute positively to the averaged residual bound.

We use the discrete-time Lion-$\mathcal{K}$ Lyapunov function from [2], replacing the static potential K by the time-dependent potential K_t.

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

The time-dependent Lyapunov function is

$$\begin{aligned} \mathcal{H}_t &:= \mathcal{F}(X_t) - \mathcal{F}^\star + \frac{1}{\lambda} K_t^*(\lambda X_t) + \frac{c_t}{\lambda}\left(K_t^*(\lambda X_t) + K_t(M_t) - \langle \lambda X_t, M_t \rangle\right) \\ &= \mathcal{F}(X_t) - \mathcal{F}^\star + \frac{a_t}{\lambda} K_t^*(\lambda X_t) + \frac{c_t}{\lambda} K_t(M_t) - c_t \langle X_t, M_t \rangle. \end{aligned} \tag{246}$$

The main difference from the static-K case is that the actual next Lyapunov value $\mathcal{H}_{t+1}$ is defined with K_{t+1}, whereas the one-step static Lion-$\mathcal{K}$ estimate applies with a fixed potential during

the step $t \to t+1$. Therefore, we introduce the frozen next Lyapunov value

$$\widehat{\mathcal{H}}_{t+1}^{(t)} := \mathcal{F}(X_{t+1}) - \mathcal{F}^\star + \frac{1}{\lambda} K_t^*(\lambda X_{t+1}) + \frac{c_{t+1}}{\lambda} \left(K_t^*(\lambda X_{t+1}) + K_t(M_{t+1}) - \langle \lambda X_{t+1}, M_{t+1} \rangle \right). \tag{247}$$

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

$$\mathcal{H}_{t+1} := \mathcal{F}(X_{t+1}) - \mathcal{F}^\star + \frac{1}{\lambda} K_{t+1}^*(\lambda X_{t+1}) + \frac{c_{t+1}}{\lambda} \left(K_{t+1}^*(\lambda X_{t+1}) + K_{t+1}(M_{t+1}) - \langle \lambda X_{t+1}, M_{t+1} \rangle \right). \tag{248}$$

With the potential frozen at K_t, we obtain

$$\eta_t \left(a_t \Gamma_t^{(t)} + b_t \Delta_t^{(t)} \right) \le \mathcal{H}_t - \widehat{\mathcal{H}}_{t+1}^{(t)} + \frac{L}{2} \eta_t^2 \left\| \nabla K_t(\widetilde{M}_{t+1}) - \lambda X_{t+1} \right\|_F^2. \tag{249}$$

Equation (249) is obtained by applying the static Lion-$\mathcal{K}$ one-step Lyapunov estimate of [2] with the potential frozen at K_t. The new issue in the time-dependent setting is that the actual next Lyapunov value is defined using K_{t+1}, whereas $\widehat{\mathcal{H}}_{t+1}^{(t)}$ in (249) is computed using K_t. Therefore, compared with the static proof, the only additional term to be controlled is the potential-shift discrepancy $\mathcal{H}_{t+1} - \widehat{\mathcal{H}}_{t+1}^{(t)}$.

Here the residual terms are exactly the static-K residuals, but computed with K_t:

$$\Gamma_t^{(t)} := \left\langle \nabla K_t(\widetilde{M}_{t+1}) - \lambda X_{t+1}, \widetilde{M}_{t+1} - \nabla K_t^*(\lambda X_{t+1}) \right\rangle, \tag{250}$$

$$\Delta_t^{(t)} := \left\langle \nabla K_t(\widetilde{M}_{t+1}) - \nabla K_t(M_{t+1}), \widetilde{M}_{t+1} - M_{t+1} \right\rangle. \tag{251}$$

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 (247) from (248), we have

$$\begin{aligned} \mathcal{H}_{t+1} - \widehat{\mathcal{H}}_{t+1}^{(t)} &= \frac{1}{\lambda} \left[K_{t+1}^*(\lambda X_{t+1}) - K_t^*(\lambda X_{t+1}) \right] \\ &\quad + \frac{c_{t+1}}{\lambda} \left[K_{t+1}^*(\lambda X_{t+1}) - K_t^*(\lambda X_{t+1}) \right] \\ &\quad + \frac{c_{t+1}}{\lambda} \left[K_{t+1}(M_{t+1}) - K_t(M_{t+1}) \right] \\ &= \frac{a_{t+1}}{\lambda} \left[K_{t+1}^*(\lambda X_{t+1}) - K_t^*(\lambda X_{t+1}) \right] \\ &\quad + \frac{c_{t+1}}{\lambda} \left[K_{t+1}(M_{t+1}) - K_t(M_{t+1}) \right]. \end{aligned} \tag{252}$$

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 (252) into (249) yields the time-dependent one-step inequality

$$\begin{aligned} \eta_t \left(a_t \Gamma_t^{(t)} + b_t \Delta_t^{(t)} \right) &\le \mathcal{H}_t - \mathcal{H}_{t+1} + \frac{L}{2} \eta_t^2 \left\| \nabla K_t(\widetilde{M}_{t+1}) - \lambda X_{t+1} \right\|_F^2 \\ &\quad + \frac{a_{t+1}}{\lambda} \left[K_{t+1}^*(\lambda X_{t+1}) - K_t^*(\lambda X_{t+1}) \right] \\ &\quad + \frac{c_{t+1}}{\lambda} \left[K_{t+1}(M_{t+1}) - K_t(M_{t+1}) \right]. \end{aligned} \tag{253}$$

We now impose the following assumptions.

First, as in the continuous-time analysis, we assume that λ is chosen sufficiently small so that

$$K_t^*(\lambda X_{t+1}) = 0, \qquad K_{t+1}^*(\lambda X_{t+1}) = 0$$

along the trajectory. Under this assumption, the dual variation term in (253) vanishes:

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

Second, we assume cumulative nonpositive primal variation:

$$\sum_{t=0}^{T-1}\left[K_{t+1}(M_{t+1}) - K_t(M_{t+1})\right] \le 0, \qquad \forall T \ge 1. \tag{254}$$

Let

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

Then (254) says that

$$S_T \le 0, \qquad \forall T \ge 1.$$

Third, assume that the learning-rate schedule is nonincreasing:

$$\eta_{t+1} \le \eta_t.$$

Since

$$c_t = \frac{\eta_t \lambda \beta_1}{\eta_t \lambda(1-\beta_1) + (1-\beta_2)}$$

is an increasing function of η_t, we have

$$0 < c_{t+1} \le c_t.$$

Define

$$w_t := \frac{c_{t+1}}{\lambda}.$$

Then $w_t > 0$ and $w_{t+1} \le w_t$. By Abel summation,

$$\sum_{t=0}^{T-1} w_t V_t = w_{T-1} S_T + \sum_{t=0}^{T-2} (w_t - w_{t+1}) S_{t+1}. \tag{255}$$

Since $S_T \le 0$, $S_{t+1} \le 0$, $w_{T-1} > 0$, and $w_t - w_{t+1} \ge 0$, every term on the right-hand side is nonpositive. Therefore,

$$\sum_{t=0}^{T-1} \frac{c_{t+1}}{\lambda}\left[K_{t+1}(M_{t+1}) - K_t(M_{t+1})\right] \le 0. \tag{256}$$

Thus the explicit primal variation term in the summed version of (253) does not contribute positively.

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

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

Then, at differentiable points,

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

and hence

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

The same bound holds for any subgradient at nondifferentiable points.

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

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

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

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

Therefore,

$$\begin{aligned} \left\| \nabla K_t(\widetilde{M}_{t+1}) - \lambda X_{t+1} \right\|_F^2 &\le 2\|\nabla K_t(\widetilde{M}_{t+1})\|_F^2 + 2\|\lambda X_{t+1}\|_F^2 \\ &\le 2N + 2N = 4N. \end{aligned} \tag{257}$$

Summing (253) from $t = 0$ to $T - 1$, and using the vanishing of the dual variation term, we obtain

$$\begin{aligned} \sum_{t=0}^{T-1} \eta_t \left(a_t \Gamma_t^{(t)} + b_t \Delta_t^{(t)} \right) \le \mathcal{H}_0 - \mathcal{H}_T + \frac{L}{2} \sum_{t=0}^{T-1} \eta_t^2 \left\| \nabla K_t(\widetilde{M}_{t+1}) - \lambda X_{t+1} \right\|_F^2 \\ + \sum_{t=0}^{T-1} \frac{c_{t+1}}{\lambda} \left[K_{t+1}(M_{t+1}) - K_t(M_{t+1}) \right]. \end{aligned} \tag{258}$$

By the weighted variation estimate (256),

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

Therefore, the explicit primal variation term does not contribute positively, and (258) gives

$$\begin{aligned} \sum_{t=0}^{T-1} \eta_t \left(a_t \Gamma_t^{(t)} + b_t \Delta_t^{(t)} \right) &\le \mathcal{H}_0 - \mathcal{H}_T + \frac{L}{2} \sum_{t=0}^{T-1} \eta_t^2 \left\| \nabla K_t(\widetilde{M}_{t+1}) - \lambda X_{t+1} \right\|_F^2 \\ &\le \mathcal{H}_0 - \mathcal{H}_T + 2LN \sum_{t=0}^{T-1} \eta_t^2, \end{aligned} \tag{259}$$

where the last inequality follows from (257). Equivalently,

$$\frac{1}{T}\sum_{t=0}^{T-1}\eta_t\left(a_t\Gamma_t^{(t)}+b_t\Delta_t^{(t)}\right)\le\frac{\mathcal{H}_0-\mathcal{H}_T}{T}+\frac{2LN}{T}\sum_{t=0}^{T-1}\eta_t^2. \tag{260}$$

Thus, if $\mathcal{H}_T$ is bounded below and

$$\frac{1}{T}\sum_{t=0}^{T-1}\eta_t^2\to 0,$$

then

$$\frac{1}{T}\sum_{t=0}^{T-1}\eta_t\left(a_t\Gamma_t^{(t)}+b_t\Delta_t^{(t)}\right)\to 0.$$

This is the time-dependent-potential analogue of the averaged residual bound in the discrete Lion-$\mathcal{K}$ analysis of [2]. After this estimate is established, the remaining gap-to-KKT argument follows in the same sense as [2], with K, Γ_t, and Δ_t replaced by K_t, $\Gamma_t^{(t)}$, and $\Delta_t^{(t)}$, respectively. Thus the conclusion should be interpreted as a K_t-dependent stationarity conclusion unless an additional limiting-potential assumption is imposed.

A fixed limiting KKT interpretation requires an additional stabilization assumption, such as $K_t\to K_\infty$ along the generated trajectory. Under this assumption, the K_t-dependent residuals pass to the KKT residuals associated with the limiting potential K_∞, by the same subgradient-graph convergence argument as in Appendix C.2.

D.1. Application to Random Orthogonal-Basis Normalization

We now explain how the preceding time-dependent-potential analysis applies to Random Orthogonal-Basis Normalization (ROBN). In ROBN, the potential K_t is defined by a randomly sampled Frobenius-orthonormal basis $\{M_i(t)\}_{i=1}^N$:

$$K_t(M)=\sum_{i=1}^{N}|\langle M,M_i(t)\rangle_F|. \tag{261}$$

To rewrite this in vector form, let

$$q_i(t):=\operatorname{vec}(M_i(t)),\qquad Q_t:=\begin{bmatrix}q_1(t) & q_2(t) & \cdots & q_N(t)\end{bmatrix}. \tag{262}$$

Since the Frobenius inner product is preserved under vectorization, we have

$$\langle M,M_i(t)\rangle_F=q_i(t)^\top\operatorname{vec}(M). \tag{263}$$

Moreover, because $\{M_i(t)\}_{i=1}^N$ is Frobenius-orthonormal, $\{q_i(t)\}_{i=1}^N$ is an orthonormal basis after vectorization. Hence Q_t is an orthogonal matrix. Therefore,

$$Q_t^\top \operatorname{vec}(M) = \begin{bmatrix} q_1(t)^\top \operatorname{vec}(M) \\ q_2(t)^\top \operatorname{vec}(M) \\ \vdots \\ q_N(t)^\top \operatorname{vec}(M) \end{bmatrix} = \begin{bmatrix} \langle M, M_1(t)\rangle_F \\ \langle M, M_2(t)\rangle_F \\ \vdots \\ \langle M, M_N(t)\rangle_F \end{bmatrix}. \tag{264}$$

Taking the ℓ_1-norm of this coefficient vector gives

$$K_t(M) = \sum_{i=1}^N |\langle M, M_i(t)\rangle_F| = \|Q_t^\top \operatorname{vec}(M)\|_1. \tag{265}$$

The deterministic analysis above used a cumulative nonpositive primal variation condition to remove the explicit time-variation term. For ROBN, such a pathwise monotonicity condition is generally too strong, because the orthonormal basis is resampled over time. Hence the instantaneous difference

$$K_{t+1}(M_{t+1}) - K_t(M_{t+1})$$

need not be nonpositive for every realization.

Instead, ROBN is naturally associated with an averaged cancellation property. For the idealized ROBN analysis, we model the sampled orthogonal matrices $\{Q_t\}$ as independent Haar-random orthogonal matrices. This is natural because ROBN starts from a Gaussian random matrix and applies an orthogonalization procedure. In the actual implementation, a finite number of Newton–Schulz iterations is used, so the resulting matrix should be understood as an approximately orthogonal random matrix; the Haar model is an idealized rotation-invariant model of this construction.

We now state the stochastic cancellation condition for the basis-variation term. Let

$$V_t := K_{t+1}(M_{t+1}) - K_t(M_{t+1}).$$

Here K_t is the potential induced by the random orthogonal basis Q_t.

We denote by $\mathcal{F}_t$ the sigma-field generated by the history available after the current basis Q_t has been sampled and the next momentum state M_{t+1} has been formed, but before the next basis Q_{t+1} is sampled. In particular, $\mathcal{F}_t$ contains the past trajectory, the sampled bases up to the current one, and the current momentum state M_{t+1}, but it does not contain the freshly resampled next basis Q_{t+1}. For example, one may take

$$\mathcal{F}_t := \sigma(M_0, \dots, M_{t+1},\, Q_0, \dots, Q_t,\, \text{stochastic gradients and randomness used up to the formation of } M_{t+1})\,.$$

Thus, $\mathcal{F}_t$ represents the information available after M_{t+1} and Q_t are known, but before the next Haar-like random basis Q_{t+1} is drawn.

With this choice of filtration, M_{t+1} and Q_t are fixed under conditioning on $\mathcal{F}_t$, while Q_{t+1} is freshly resampled and is independent of $\mathcal{F}_t$. Let

$$x := \mathrm{vec}(M_{t+1}) \in \mathbb{R}^N.$$

Conditional on $\mathcal{F}_t$, the vector x is fixed. Since Q_{t+1} is modeled as a Haar-random orthogonal matrix, rotation invariance gives

$$Q_{t+1}^\top x \overset{d}{=} \|x\|_2 U,$$

where U is uniformly distributed on the unit sphere $\mathbb{S}^{N-1}$. Hence

$$\mathbb{E}\Big[\|Q_{t+1}^\top x\|_1 \,\Big|\, \mathcal{F}_t\Big] = \|x\|_2\, \mathbb{E}[\|U\|_1].$$

Define

$$\alpha_N := \mathbb{E}[\|U\|_1] > 0.$$

Since $\|x\|_2 = \|M_{t+1}\|_F$, we obtain

$$\mathbb{E}[K_{t+1}(M_{t+1}) \,|\, \mathcal{F}_t] = \alpha_N \|M_{t+1}\|_F.$$

Equivalently, since the coordinates of U are exchangeable,

$$\alpha_N = N\, \mathbb{E}[|U_1|],$$

which depends only on the dimension N.

On the other hand, Q_t and M_{t+1} are already contained in $\mathcal{F}_t$, so $K_t(M_{t+1})$ is $\mathcal{F}_t$-measurable. Therefore,

$$\mathbb{E}[V_t \mid \mathcal{F}_t] = \alpha_N \|M_{t+1}\|_F - K_t(M_{t+1}).$$

In the high-momentum regime considered here, for example with $\beta = 0.95$, the new momentum state M_{t+1} is largely determined by the accumulated past momentum. Although M_{t+1} may depend on the current sampled basis Q_t, we assume that this dependence does not create a systematic alignment bias between M_{t+1} and the coordinate directions induced by Q_t. We formalize this weak-alignment assumption by requiring that the current coefficient norm is close to its Haar-averaged value:

$$K_t(M_{t+1}) \approx \alpha_N \|M_{t+1}\|_F.$$

Equivalently,

$$\mathbb{E}[V_t \mid \mathcal{F}_t] \approx 0.$$

For the formal convergence argument, we impose the corresponding idealized zero-bias condition

$$\mathbb{E}[V_t \mid \mathcal{F}_t] = 0.$$

Equivalently,

$$K_t(M_{t+1}) = \alpha_N \|M_{t+1}\|_F.$$

This condition should be understood as the idealized exact version of the weak-alignment approximation

$$K_t(M_{t+1}) \approx \alpha_N \|M_{t+1}\|_F,$$

which is expected to be accurate when the Q_t-dependent fresh update is small relative to the accumulated momentum. It is a stochastic regularity assumption motivated by the Haar-like basis model and the high-momentum regime, rather than a consequence of Haar sampling alone.

We also assume that the basis-variation term has a uniformly bounded conditional second moment:

$$\mathbb{E}[V_t^2 \mid \mathcal{F}_t] \le C$$

for some constant $C < \infty$ and all t. This condition is mild in the coefficient-norm setting. Indeed, if

$$K_t(M) = \|Q_t^\top \operatorname{vec}(M)\|_1,$$

where $Q_t \in \mathbb{R}^{N\times N}$ is orthogonal, then

$$K_t(M) \le \sqrt{N}\|Q_t^\top \operatorname{vec}(M)\|_2 = \sqrt{N}\|M\|_F.$$

Hence

$$|V_t| \le K_{t+1}(M_{t+1}) + K_t(M_{t+1}) \le 2\sqrt{N}\|M_{t+1}\|_F.$$

Therefore, if the momentum sequence is uniformly bounded, namely

$$\|M_{t+1}\|_F \le R$$

along the generated trajectory, then

$$V_t^2 \le 4NR^2.$$

In particular,

$$\mathbb{E}[V_t^2 \mid \mathcal{F}_t] \le 4NR^2.$$

Under these assumptions, the unweighted basis-variation average vanishes almost surely. Let

$$S_T := \sum_{t=0}^{T-1} V_t.$$

With the above indexing, V_t is measurable with respect to $\mathcal{F}_{t+1}$, since the next basis Q_{t+1} has been sampled by that time, while the idealized zero-bias condition gives

$$\mathbb{E}[V_t \mid \mathcal{F}_t] = 0.$$

Thus $\{V_t\}$ is a martingale-difference sequence with respect to the filtration $\{\mathcal{F}_{t+1}\}$, equivalently after the standard one-step index shift. Moreover, the bounded conditional second-moment assumption implies

$$\mathbb{E}[V_t^2] = \mathbb{E}\big[\mathbb{E}[V_t^2 \mid \mathcal{F}_t]\big] \le C.$$

Therefore,

$$\sum_{t=1}^{\infty} \frac{\mathbb{E}[V_t^2]}{t^2} \le C \sum_{t=1}^{\infty} \frac{1}{t^2} < \infty.$$

Hence, by the strong law of large numbers for martingale differences,

$$\frac{1}{T} S_T = \frac{1}{T} \sum_{t=0}^{T-1} V_t \to 0 \quad \text{almost surely.}$$

Equivalently,

$$\frac{1}{T}\sum_{t=0}^{T-1}\left[K_{t+1}(M_{t+1})-K_t(M_{t+1})\right]\to 0 \quad \text{almost surely.}$$

This gives the unweighted basis-variation condition used below.

As a diagnostic sanity check, Appendix H reports a controlled MLP experiment in which the empirical basis-variation average is observed to be close to zero.

Although the mean-zero fluctuation property naturally motivates the unweighted averaged cancellation of V_t, the time-dependent Lyapunov inequality contains the weighted variation term

$$\frac{c_{t+1}}{\lambda}V_t.$$

Thus, for the averaged residual estimate, it is sufficient to verify the weighted Cesàro condition

$$\frac{1}{T}\sum_{t=0}^{T-1}\frac{c_{t+1}}{\lambda}\left[K_{t+1}(M_{t+1})-K_t(M_{t+1})\right]\to 0. \tag{266}$$

Equivalently, since $\lambda>0$ is fixed, it is enough to show

$$\frac{1}{T}\sum_{t=0}^{T-1}c_{t+1}\left[K_{t+1}(M_{t+1})-K_t(M_{t+1})\right]\to 0. \tag{267}$$

Below, we show that this weighted Cesàro condition follows from the unweighted Cesàro cancellation under a mild boundedness condition on the variation terms.

To derive the weighted Cesàro condition, recall that

$$V_t := K_{t+1}(M_{t+1})-K_t(M_{t+1}).$$

Assume the unweighted Cesàro cancellation motivated above:

$$\frac{1}{T}\sum_{t=0}^{T-1}V_t\to 0 \quad \text{almost surely.}$$

We show that this condition implies the weighted Cesàro condition required in the Lyapunov bound.

Since c_t is nonincreasing and bounded below, there exists $c_\infty\ge 0$ such that $c_t\to c_\infty$. Moreover,

$$c_{t+1}-c_\infty\ge 0$$

for all t. Hence

$$\frac{1}{T}\sum_{t=0}^{T-1}c_{t+1}V_t = c_\infty\frac{1}{T}\sum_{t=0}^{T-1}V_t+\frac{1}{T}\sum_{t=0}^{T-1}(c_{t+1}-c_\infty)V_t.$$

The first term converges to zero by the unweighted averaged cancellation. For the second term, assume that $\{V_t\}$ is bounded, i.e.,

$$|V_t|\le B$$

for some $B > 0$. Then

$$\left|\frac{1}{T}\sum_{t=0}^{T-1}(c_{t+1}-c_\infty)V_t\right| \le \frac{1}{T}\sum_{t=0}^{T-1}(c_{t+1}-c_\infty)|V_t| \le B\frac{1}{T}\sum_{t=0}^{T-1}(c_{t+1}-c_\infty).$$

Since $c_{t+1} - c_\infty \to 0$, its Cesàro average also converges to zero:

$$\frac{1}{T}\sum_{t=0}^{T-1}(c_{t+1}-c_\infty) \to 0.$$

Therefore,

$$\frac{1}{T}\sum_{t=0}^{T-1}(c_{t+1}-c_\infty)V_t \to 0.$$

Consequently,

$$\frac{1}{T}\sum_{t=0}^{T-1}c_{t+1}V_t \to 0.$$

That is,

$$\frac{1}{T}\sum_{t=0}^{T-1}c_{t+1}\left[K_{t+1}(M_{t+1}) - K_t(M_{t+1})\right] \to 0. \tag{268}$$

The boundedness of V_t follows, for example, if the momentum sequence $\{M_{t+1}\}$ is bounded. Indeed, since K_t is an ℓ_1-type coefficient norm induced by a Frobenius-orthonormal basis,

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

Hence

$$|V_t| \le K_{t+1}(M_{t+1}) + K_t(M_{t+1}) \le 2\sqrt{N}\|M_{t+1}\|_F.$$

Therefore, if $\sup_t \|M_{t+1}\|_F < \infty$, then $\sup_t |V_t| < \infty$.

Substituting the time-dependent one-step inequality and summing from $t = 0$ to $T-1$, we have

$$\sum_{t=0}^{T-1}\eta_t\left(a_t\Gamma_t^{(t)} + b_t\Delta_t^{(t)}\right) \le \mathcal{H}_0 - \mathcal{H}_T + 2LN\sum_{t=0}^{T-1}\eta_t^2 + \sum_{t=0}^{T-1}\frac{c_{t+1}}{\lambda}\left[K_{t+1}(M_{t+1}) - K_t(M_{t+1})\right], \tag{269}$$

where we used the discrete smoothness remainder bound from (257). Dividing by T, we obtain

$$\frac{1}{T}\sum_{t=0}^{T-1}\eta_t\left(a_t\Gamma_t^{(t)} + b_t\Delta_t^{(t)}\right) \le \frac{\mathcal{H}_0 - \mathcal{H}_T}{T} + \frac{2LN}{T}\sum_{t=0}^{T-1}\eta_t^2 + \frac{1}{T}\sum_{t=0}^{T-1}\frac{c_{t+1}}{\lambda}\left[K_{t+1}(M_{t+1}) - K_t(M_{t+1})\right]. \tag{270}$$

Therefore, if $\mathcal{H}_T$ is bounded below,

$$\frac{1}{T}\sum_{t=0}^{T-1}\eta_t^2 \to 0, \tag{271}$$

and the ROBN basis-variation term satisfies the weighted Cesàro condition (266), which follows from (268), then

$$\frac{1}{T}\sum_{t=0}^{T-1}\eta_t\left(a_t\Gamma_t^{(t)} + b_t\Delta_t^{(t)}\right) \to 0. \tag{272}$$

Thus, ROBN inherits the averaged residual convergence mechanism of the time-dependent Lion-$\mathcal{K}$ analysis, up to an asymptotically vanishing basis-variation term. The resulting conclusion should be interpreted as an averaged K_t-dependent stationarity result. If, in addition, the sampled potentials stabilize along the generated trajectory, for example $K_t \to K_\infty$, then the same limiting-potential argument yields a stationarity interpretation with respect to the limiting potential K_∞.

D.2. A Simple Orthogonal Construction Satisfying Nonpositive Variation

We give a simple construction showing that the nonpositive primal-variation assumption is not vacuous even when the potentials are induced by orthogonal coefficient maps. Let

$$x_{t+1} := \mathrm{vec}(M_{t+1}),$$

and define

$$K_t(M) := \|Q_t^\top \mathrm{vec}(M)\|_1,$$

where Q_t is the orthogonal matrix whose columns are the vectorized orthonormal basis elements of the matrix space.

For each t, choose Q_{t+1} so that its first column is aligned with x_{t+1}. If $x_{t+1} \neq 0$, set

$$q_1(t+1) = \frac{x_{t+1}}{\|x_{t+1}\|_2}.$$

Then choose the remaining columns as an orthonormal basis of $x_{t+1}^\perp$. Such a choice always exists: take any basis of

$$x_{t+1}^\perp = \{z \in \mathbb{R}^N : \langle z, x_{t+1}\rangle = 0\}$$

and apply the Gram–Schmidt procedure. Hence

$$Q_{t+1} = \begin{bmatrix} q_1(t+1) & q_2(t+1) & \cdots & q_N(t+1)\end{bmatrix}$$

is an orthogonal matrix.

By construction,

$$q_1(t+1)^\top x_{t+1} = \|x_{t+1}\|_2,$$

and, for $i = 2, \ldots, N$,

$$q_i(t+1)^\top x_{t+1} = 0.$$

Therefore,

$$Q_{t+1}^\top x_{t+1} = \begin{bmatrix} \|x_{t+1}\|_2 \\ 0 \\ \vdots \\ 0 \end{bmatrix}.$$

It follows that

$$K_{t+1}(M_{t+1}) = \|Q_{t+1}^\top x_{t+1}\|_1 = \|x_{t+1}\|_2.$$

On the other hand, for any orthogonal Q_t,

$$\|x_{t+1}\|_2 = \|Q_t^\top x_{t+1}\|_2 \le \|Q_t^\top x_{t+1}\|_1 = K_t(M_{t+1}).$$

Hence

$$K_{t+1}(M_{t+1}) \le K_t(M_{t+1}),$$

and therefore

$$K_{t+1}(M_{t+1}) - K_t(M_{t+1}) \le 0.$$

Summing over t, we obtain

$$\sum_{t=0}^{T-1} [K_{t+1}(M_{t+1}) - K_t(M_{t+1})] \le 0, \qquad \forall T \ge 1.$$

This construction is trajectory-adaptive and is not intended to describe the random resampling mechanism of ROBN. Rather, it provides a concrete orthogonal-basis selection rule showing that the nonpositive primal-variation assumption can be satisfied in a non-vacuous way.

Appendix E. 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

$$\widetilde{\Delta}_t = A_t^\top \Delta_t, \tag{273}$$

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

$$D_t = A_t\widehat{\Delta}_t = \sum_i p_i(t)\widehat{\Delta}_{t,i:}, \qquad \widehat{\Delta}_{t,i:} = \frac{\widetilde{\Delta}_{t,i:}}{\|\widetilde{\Delta}_{t,i:}\|_2 + \varepsilon}. \tag{274}$$

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

$$K_t(\Delta) = \sum_i \left\|(A_t^\top \Delta)_{i:}\right\|_2. \tag{275}$$

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

$$A_t\widehat{\Delta}_t \in \partial_\Delta K_t(\Delta_t), \qquad \widehat{\Delta}_{t,i:} = \frac{\widetilde{\Delta}_{t,i:}}{\|\widetilde{\Delta}_{t,i:}\|_2}. \tag{276}$$

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

$$K_{t+1}(\Delta) \neq K_t(\Delta). \tag{277}$$

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

$$K_t(\Delta) = \sum_{j=1}^{n} \left\|(\Delta A_t^\top)_{:j}\right\|_2, \tag{278}$$

and the ROBN direction satisfies

$$\widehat{\Delta}_t A_t \in \partial_\Delta K_t(\Delta_t) \tag{279}$$

for nonzero columns when $\varepsilon = 0$.

Appendix F. Experimental Details and Hyperparameter Selection

F.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

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

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.

F.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 G. 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,

$$G = \begin{bmatrix} G_{\mathrm{top}} \\ G_{\mathrm{bottom}} \end{bmatrix}, \qquad G_{\mathrm{top}}, G_{\mathrm{bottom}} \in \mathbb{R}^{n\times m}. \tag{280}$$

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

$$\Delta = \begin{bmatrix} \mathrm{NS}(G_{\mathrm{top}}) \\ \mathrm{NS}(G_{\mathrm{bottom}}) \end{bmatrix}. \tag{281}$$

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

$$G = \begin{bmatrix} G_{\mathrm{top}} \\ G_{\mathrm{bottom}} \end{bmatrix}, \qquad G_{\mathrm{top}}, G_{\mathrm{bottom}} \in \mathbb{R}^{n\times m}. \tag{282}$$

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

$$\mathrm{NS}(G_{\mathrm{top}}) = \sum_i \sigma_i^{\mathrm{top}} u_i^{\mathrm{top}} v_i^{\mathrm{top}\top}, \qquad \mathrm{NS}(G_{\mathrm{bottom}}) = \sum_j \sigma_j^{\mathrm{bottom}} u_j^{\mathrm{bottom}} v_j^{\mathrm{bottom}\top} \tag{283}$$

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

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

$$p_i^{\mathrm{top}} = \begin{bmatrix} u_i^{\mathrm{top}} \\ 0 \end{bmatrix}, \qquad q_i^{\mathrm{top}} = v_i^{\mathrm{top}}, \tag{284}$$

and

$$p_j^{\mathrm{bottom}} = \begin{bmatrix} 0 \\ u_j^{\mathrm{bottom}} \end{bmatrix}, \qquad q_j^{\mathrm{bottom}} = v_j^{\mathrm{bottom}}. \tag{285}$$

Then the SNSU update can be written as

$$\Delta = \sum_i \sigma_i^{\mathrm{top}} p_i^{\mathrm{top}} q_i^{\mathrm{top}\top} + \sum_j \sigma_j^{\mathrm{bottom}} p_j^{\mathrm{bottom}} q_j^{\mathrm{bottom}\top}. \tag{286}$$

Equivalently, setting

$$M_i^{\mathrm{top}} = p_i^{\mathrm{top}} q_i^{\mathrm{top}\top}, \qquad M_j^{\mathrm{bottom}} = p_j^{\mathrm{bottom}} q_j^{\mathrm{bottom}\top}, \tag{287}$$

we obtain

$$\Delta = \sum_i a_i^{\mathrm{top}} M_i^{\mathrm{top}} + \sum_j a_j^{\mathrm{bottom}} M_j^{\mathrm{bottom}}, \qquad a_i^{\mathrm{top}} = \sigma_i^{\mathrm{top}}, \quad a_j^{\mathrm{bottom}} = \sigma_j^{\mathrm{bottom}}. \tag{288}$$

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

$$\{p_k(t)\}_k = \{p_i^{\mathrm{top}}\}_i \cup \{p_j^{\mathrm{bottom}}\}_j, \qquad \{q_k(t)\}_k = \{q_i^{\mathrm{top}}\}_i \cup \{q_j^{\mathrm{bottom}}\}_j, \tag{289}$$

and

$$\{a_k(t)\}_k = \{a_i^{\mathrm{top}}\}_i \cup \{a_j^{\mathrm{bottom}}\}_j = \{\sigma_i^{\mathrm{top}}\}_i \cup \{\sigma_j^{\mathrm{bottom}}\}_j. \tag{290}$$

With this notation,

$$\Delta = \sum_k a_k(t) p_k(t) q_k(t)^\top. \tag{291}$$

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

$$\left\langle p_i^{\mathrm{top}}, p_j^{\mathrm{bottom}} \right\rangle = 0 \qquad \text{for all } i, j. \tag{292}$$

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

$$I_t := \{k : a_k(t) \neq 0\}, \tag{293}$$

we have

$$\langle p_k(t), p_\ell(t) \rangle = \delta_{k\ell}, \qquad k, \ell \in I_t. \tag{294}$$

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:

$$\partial_t K_t(m_t) = 0. \tag{295}$$

We compare SNSU with Muon under the same ViT-B/16 ImageNet-1K setting as Appendix F.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 F.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.

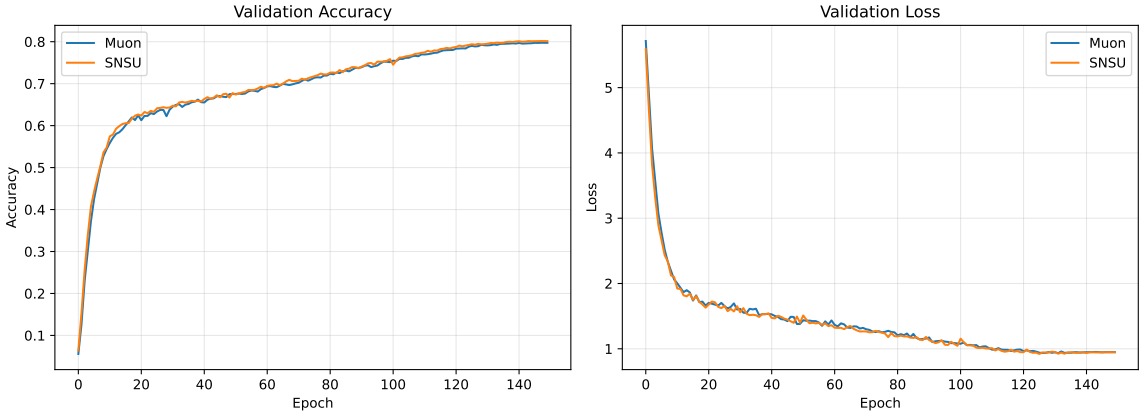


Figure 1: Auxiliary single-seed validation curves for Muon and SNSU on ImageNet-1K using a ViT-B/16 backbone.

Appendix H. Additional Diagnostic for ROBN Basis-Variation Cancellation

We provide an additional diagnostic check for the averaged basis-variation cancellation assumption used in Appendix D.1. The goal of this experiment is not to validate the zero-bias assumption in full generality, but to check whether the empirical basis-variation average remains close to zero 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 empirical basis-variation average

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

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 de-

generate one. The measured basis-variation statistics are

$$\sum_{t=0}^{T-1} V_t = 0.006165, \qquad T = 19954, \qquad \frac{1}{T}\sum_{t=0}^{T-1} V_t = 3.089 \times 10^{-7}.$$

Thus, the empirical average of the basis-variation term is essentially zero in this run. This diagnostic does not prove the idealized zero-bias condition, but it provides empirical support that the ROBN basis-variation fluctuations can be approximately centered in the tested high-momentum 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.