Novelty Assessment Report

Paper: Convergence of Muon with Newton-Schulz
PDF URL: https://openreview.net/pdf?id=lJSfxtLpLm
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2026-01-01

Abstract

We analyze Muon as originally proposed and used in practice---using the momentum orthogonalization with a few Newton-Schulz steps. The prior theoretical results replace this key step in Muon with an exact SVD-based polar factor. We prove that Muon with Newton-Schulz converges to a stationary point with the same rate as the SVD-polar idealization, up to a constant factor for given the number of Newton-Schulz steps Q. We further analyze this constant factor, and prove that it converges to 1 doubly exponentially in Q and improves with κ, which is the degree of a polynomial used in Newton-Schulz required when approximating the orthogonalization direction. We also prove that Muon removes the typical square-root-of-rank loss compared to its vector-based counterpart, SGD with momentum. Our results explain why Muon with a few low-degree Newton-Schulz steps matches exact-polar (SVD) behavior at much faster wall-clock time, and explain how much momentum matrix orthogonalization via Newton-Schulz benefits over the vector-based optimizer. Overall, our theory justifies the practical Newton-Schulz design of Muon, narrowing its practice–theory gap.

Core Task Landscape

This paper addresses: **Momentum-Based Matrix Optimization with Approximate Orthogonalization**

A total of **15 papers** were analyzed and organized into a taxonomy with **8 categories**.

Taxonomy Overview

The research landscape has been organized into the following main categories:

- **Convergence Theory and Analysis**
- **Algorithmic Design and Implementation**
- **Applications and Extensions**

Complete Taxonomy Tree

- Momentum-Based Matrix Optimization with Approximate Orthogonalization Survey Taxonomy
- Convergence Theory and Analysis
  - Momentum-Based Optimizers with Approximate Orthogonalization ★ (3 papers)
  - [0] Convergence of Muon with Newton-Schulz (Anon et al., 2026) View paper
  - [4] Beyond the ideal: Analyzing the inexact muon update (Shulgin, 2025) View paper
  - [10] High-dimensional isotropic scaling dynamics of Muon and SGD (G Wang, 2025) View paper
  - Accelerated Methods with Orthogonality Constraints (2 papers)
  - [8] Accelerated Optimization With Orthogonality Constraints (Siegel, 2022) View paper
  - [15] Quadratic Optimization with Orthogonality Constraints: Explicit Lojasiewicz Exponent and Linear Convergence (Wu, 2016) View paper
- Algorithmic Design and Implementation
  - Exact Orthogonalization Methods (2 papers)
  - [1] A feasible method for optimization with orthogonality constraints (Zaiwen Wen, 2013) View paper
  - [2] Simplifying momentum-based positive-definite submanifold optimization with applications to deep learning (Lin Wu, 2023) View paper
  - Approximate and Efficient Orthogonalization (3 papers)
  - [5] NorMuon: Making Muon more efficient and scalable (Li, 2025) View paper
  - [11] AuON: A Linear-time Alternative to Semi-Orthogonal Momentum Updates (Maity, 2025) View paper
  - [13] Improving generalization in DNNs through enhanced orthogonality in momentum-based optimizers (Zhixing Lu, 2025) View paper
  - Parallelizable and Scalable Approaches (1 papers)
  - [3] Parallelizable algorithms for optimization problems with orthogonality constraints (Gao, 2019) View paper
- Applications and Extensions
  - Deep Learning and Neural Network Training (1 papers)
  - [9] Novel Tensor Norm Optimization for Neural Network Training Acceleration (Banik, 2025) View paper
  - Federated and Distributed Learning (1 papers)
  - [7] FedMuon: Accelerating Federated Learning with Matrix Orthogonalization (Liu, 2025) View paper
  - Specialized Matrix Factorization and Statistical Methods (3 papers)
  - [6] Structured Joint Sparse Orthogonal Nonnegative Matrix Factorization for Fault Detection (Xi Zhang, 2023) View paper
  - [12] Incomplete Gaussian elimination as a preconditioning for generalized conjugate gradient acceleration (J. R. Wallis, 1983) View paper
  - [14] Sparse uncorrelated linear discriminant analysis for undersampled problems (Xiaowei Zhang, 2015) View paper

Narrative

Core task: convergence analysis of momentum-based matrix optimization with approximate orthogonalization. The field centers on designing and analyzing optimization algorithms that maintain approximate orthogonality constraints while leveraging momentum to accelerate convergence. The taxonomy divides naturally into three main branches. Convergence Theory and Analysis focuses on establishing rigorous guarantees for momentum-based optimizers that incorporate approximate orthogonalization schemes, examining how iterative projections or corrections affect convergence rates and stability. Algorithmic Design and Implementation explores practical variants and computational strategies, including parallelizable schemes like Parallelizable Orthogonality[3] and adaptive normalization approaches such as NorMuon[5]. Applications and Extensions branch out to specialized domains—ranging from federated learning settings in FedMuon[7] to tensor decompositions in Tensor Norm[9] and sparse factorization problems like Sparse Orthogonal NMF[6]—demonstrating how these core ideas adapt to diverse problem structures.

Recent work has concentrated on refining the interplay between momentum dynamics and orthogonality maintenance, with several studies exploring trade-offs between computational cost and approximation quality. Muon Newton-Schulz[0] sits squarely within the convergence theory branch, providing rigorous analysis of momentum-based optimizers that use Newton-Schulz iterations for approximate orthogonalization. It shares thematic ground with Inexact Muon[4], which examines how inexact orthogonalization steps influence convergence, and with Isotropic Muon[10], which investigates isotropy properties under similar momentum schemes. These works collectively address a central question: how much approximation error can be tolerated in orthogonalization while preserving the benefits of momentum acceleration? By establishing convergence guarantees under relaxed orthogonality conditions, Muon Newton-Schulz[0] contributes to a growing understanding of feasible, scalable optimization on matrix manifolds.

Related Works in Same Category

The following **2 sibling papers** share the same taxonomy leaf node with the original paper:

1. Beyond the ideal: Analyzing the inexact muon update

Authors: Shulgin, Egor, Egor Shulgin, Orabona, Francesco, et al. (10 authors total) | **Year/Venue**: 2025 | **URL**: View paper

Abstract

The Muon optimizer has rapidly emerged as a powerful, geometry-aware alternative to AdamW, demonstrating strong performance in large-scale training of neural networks. However, a critical theory-practice disconnect exists: Muon's efficiency relies on fast, approximate orthogonalization, yet all prior theoretical work analyzes an idealized, computationally intractable version assuming exact SVD-based updates. This work moves beyond the ideal by providing the first analysis of the inexact orthogon...

Relationship Analysis

Both papers belong to the same taxonomy category analyzing momentum-based optimizers with approximate orthogonalization techniques, specifically focusing on Newton-Schulz iterations instead of exact SVD in the MUON optimizer. They overlap in examining the convergence properties of MUON with inexact orthogonalization and quantifying the approximation error from using Newton-Schulz steps. However, the original paper provides explicit convergence rates showing the polar approximation error decays doubly exponentially with Newton-Schulz steps and proves rank-dependent improvements over SGD with momentum, while the candidate paper develops a general Linear Minimization Oracle (LMO) framework with additive error models and reveals the coupling between oracle inexactness and optimal hyperparameters (step size and momentum).

---

2. High-dimensional isotropic scaling dynamics of Muon and SGD

Authors: G Wang, E Paquette, A Agarwala | **Year/Venue**: 2025 | **URL**: View paper

Abstract

â¦ Momentum Orthogonalized by Newton-Schulz (Muon) is a promising algorithm which uses approximate orthogonalization of matrix-â¦ to SGD in a matrix-valued linear regression setting. â¦

Relationship Analysis

Both papers belong to the same taxonomy category analyzing momentum-based optimizers with approximate orthogonalization techniques, specifically focusing on Muon with Newton-Schulz iterations. While the original paper provides nonconvex convergence guarantees for Muon with Newton-Schulz under standard smoothness assumptions and quantifies the polar approximation error's exponential decay, the candidate paper investigates Muon's high-dimensional isotropic scaling dynamics in a matrix-valued linear regression setting using free probability theory. The key difference is that the original paper focuses on general nonconvex optimization with explicit convergence rates and iteration complexity bounds, whereas the candidate paper analyzes specific scaling regimes and normalization schemes in high-dimensional limits, revealing that Muon's default Frobenius normalization may degenerate to normalized SGD at very large dimensions.

Contributions Analysis

Overall novelty summary. The paper establishes convergence guarantees for Muon using Newton-Schulz iterations for approximate orthogonalization, proving that it matches the convergence rate of the idealized SVD-based version up to a constant factor. It resides in the 'Momentum-Based Optimizers with Approximate Orthogonalization' leaf, which contains only three papers total. This is a sparse research direction within the broader taxonomy of fifteen papers, suggesting the specific combination of momentum-based matrix optimization with approximate orthogonalization remains relatively underexplored theoretically.

The taxonomy reveals neighboring work in 'Accelerated Methods with Orthogonality Constraints' focusing on condition number dependence, and in 'Approximate and Efficient Orthogonalization' addressing computational efficiency without momentum dynamics. The paper bridges these areas by analyzing how approximate orthogonalization via Newton-Schulz interacts with momentum acceleration. Its sibling papers examine inexact orthogonalization and isotropy properties, indicating the leaf concentrates on understanding approximation quality trade-offs in momentum schemes rather than exact methods or non-momentum approaches found in adjacent branches.

Among thirteen candidates examined, the first contribution (convergence with Newton-Schulz) shows one refutable candidate from seven examined, while the second contribution (polar approximation error analysis) also has one refutable candidate from three examined. The third contribution (sharper rank dependence) appears more novel with zero refutable candidates among three examined. The limited search scope means these statistics reflect top-K semantic matches rather than exhaustive coverage. The first two contributions face more substantial prior work overlap within this constrained candidate set, while the rank-dependence analysis appears less anticipated by nearby literature.

Based on the limited literature search of thirteen candidates, the work addresses a recognized gap in analyzing practical Newton-Schulz implementations versus idealized SVD assumptions. The sparse taxonomy leaf and modest candidate pool suggest the analysis covers a focused slice of the field rather than comprehensive prior art. The rank-dependence result shows stronger novelty signals within the examined scope, though broader literature may contain additional relevant work not captured by semantic search.

This paper presents **3 main contributions**, each analyzed against relevant prior work:

Contribution 1: First convergence result of MUON with Newton-Schulz

Description: The authors provide the first theoretical convergence analysis for the practical MUON optimizer that uses Newton-Schulz iterations for momentum orthogonalization, rather than the exact SVD-based polar decomposition assumed in prior work. This analysis covers the algorithm as actually implemented and used in practice.

This contribution was assessed against **7 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. ROOT: Robust Orthogonalized Optimizer for Neural Network Training

URL: View paper

Brief Assessment

ROOT[22] focuses on robustness improvements to MUON through adaptive Newton iterations and proximal optimization, but does not provide convergence analysis for MUON with Newton-Schulz iterations. The candidate addresses different technical challenges (dimensional fragility and outlier noise) rather than theoretical convergence guarantees.

---

2. Beyond the ideal: Analyzing the inexact muon update

URL: View paper

Prior Art Analysis

Inexact Muon[4] provides the first theoretical analysis of MUON with inexact orthogonalization, specifically addressing the Newton-Schulz approximation that the original paper claims to be the first to analyze. Both papers claim to be 'the first' to analyze MUON with Newton-Schulz iterations rather than exact SVD. The candidate paper explicitly states it provides 'the first analysis of the inexact orthogonalized update at muon's core' and addresses the 'theory-practice disconnect' where 'muon's efficiency relies on fast, approximate orthogonalization, yet all prior theoretical work analyzes an idealized, computationally intractable version assuming exact svd-based updates.' This directly challenges the novelty claim of being first to analyze Newton-Schulz in MUON.

Evidence

Evidence 1 - **Rationale**: Both papers claim to provide the first analysis of MUON with inexact/Newton-Schulz orthogonalization. The candidate explicitly states that 'all prior theoretical work analyzes an idealized...version assuming exact svd-based updates' and claims to provide 'the first analysis of the inexact orthogonalized update,' directly contradicting the original's claim of being first. - **Original**: we present the first nonconvex convergence guarantees for m uon with a finite number of n ewton -schulz steps (theorem 1), as originally proposed and practically used. It is important to note that even for convex optimization, the convergence of m uon with newton -schulz has not been shown previousl... - **Candidate**: however, a critical theory-practice disconnect exists: muon's efficiency relies on fast, approximate orthogonalization, yet all prior theoretical work analyzes an idealized, computationally intractable version assuming exact svd-based updates. this work moves beyond the ideal by providing the first ...

Evidence 2 - **Rationale**: The original claims its key distinction is not replacing Newton-Schulz with exact SVD. The candidate makes the same claim about analyzing inexact updates rather than idealized SVD-based versions, suggesting both address the same gap in prior work. - **Original**: the key distinction from the existing analyses of muon is that we do not replace newton -schulz in the original muon with the exact polar computed by svd. - **Candidate**: this work moves beyond the ideal by providing the first analysis of the inexact orthogonalized update at muon's core. we develop our analysis within the general framework of linear minimization oracle (lmo)-based optimization, introducing a realistic additive error model to capture the inexactness o...

---

3. MGUP: A Momentum-Gradient Alignment Update Policy for Stochastic Optimization

URL: View paper

Brief Assessment

MGUP[25] focuses on momentum-gradient alignment for selective parameter updates in optimizers like AdamW, Lion, and MUON. It does not analyze Newton-Schulz iterations or provide convergence guarantees for MUON's orthogonalization mechanism.

---

4. Preconditioned Inexact Stochastic ADMM for Deep Model

URL: View paper

Brief Assessment

Preconditioned Stochastic ADMM[21] focuses on ADMM-based optimization with Newton-Schulz as one preconditioning option among several, not on analyzing MUON's convergence properties with Newton-Schulz iterations specifically.

---

5. AuON: A Linear-time Alternative to Orthogonal Momentum Updates

URL: View paper

Brief Assessment

AuON Alternative[23] proposes a different optimizer (AuON) that avoids Newton-Schulz iterations entirely, using normalized nonlinear scaling instead. It does not provide convergence analysis for MUON with Newton-Schulz.

---

6. High-dimensional isotropic scaling dynamics of Muon and SGD

URL: View paper

Brief Assessment

Isotropic Muon[10] analyzes MUON in a matrix-valued linear regression setting with isotropic data, deriving risk recursions rather than nonconvex convergence guarantees. The original paper provides the first nonconvex convergence analysis for practical MUON with Newton-Schulz iterations.

---

7. ANDI: Adaptive Norm-Distribution Interface

URL: View paper

Brief Assessment

ANDI[24] mentions Newton-Schulz iterations and momentum-based updates but does not provide theoretical convergence analysis for MUON with Newton-Schulz. The candidate focuses on a different optimizer (ANDI) rather than analyzing MUON's convergence properties.

Contribution 2: Analysis of polar approximation error and wall-clock convergence

Description: The authors establish that the approximation error from using Newton-Schulz instead of exact SVD vanishes doubly exponentially in the number of Newton-Schulz steps and improves with polynomial degree. This shows that even a few Newton-Schulz steps achieve convergence rates arbitrarily close to the idealized SVD variant while being computationally much cheaper.

This contribution was assessed against **3 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Beyond the ideal: Analyzing the inexact muon update

URL: View paper

Prior Art Analysis

Inexact Muon[4] analyzes how approximation error affects convergence and reveals the coupling between oracle inexactness and optimization parameters. The candidate paper explicitly quantifies 'performance degradation as a function of the lmo inexactness/error' and reveals 'a fundamental coupling between this inexactness and the optimal step size and momentum.' This addresses the same core question as the original paper's analysis of how Newton-Schulz approximation error impacts convergence rates, though potentially from a different theoretical framework (LMO-based vs. direct polar approximation analysis).

Evidence

Evidence 1 - **Rationale**: Both papers analyze how approximation error in the orthogonalization step affects convergence. While the original focuses on polar approximation error decay rates, the candidate provides explicit bounds on performance degradation as a function of inexactness, addressing the same fundamental question of how approximation quality impacts optimization. - **Original**: we prove that the polar approximation error εq due to using n ewton -schulz instead of svd, in the muon optimizer, decays doubly exponentially with the number of n ewton -schulz steps q and decays with the degree κ of the polynomial required to approximate the orthogonalization direction of the mome... - **Candidate**: our analysis yields explicit bounds that quantify performance degradation as a function of the lmo inexactness/error. we reveal a fundamental coupling between this inexactness and the optimal step size and momentum: lower oracle precision requires a smaller step size but larger momentum parameter.

Evidence 2 - **Rationale**: Both papers analyze the practical implications of Newton-Schulz approximation quality on convergence. The original shows few steps suffice for near-SVD performance with better wall-clock time. The candidate reveals that approximation precision must be co-tuned with learning parameters, both addressing how approximation quality affects practical optimization performance. - **Original**: thus, even with a few steps of newton -schulz , the convergence of muon with newton -schulz becomes arbitrarily close to that of the svd-variant in the number of iterations (theorems 2 and 4). hence, given that per-iteration computation is much more efficient for newton -schulz steps compared to svd... - **Candidate**: these findings elevate the approximation procedure (e.g., the number of newton-schulz steps) from an implementation detail to a critical parameter that must be co-tuned with the learning schedule. nanogpt experiments directly confirm the predicted coupling, with optimal learning rates clearly shifti...

---

2. Computing fundamental matrix decompositions accurately via the matrix sign function in two iterations: The power of Zolotarev's functions

URL: View paper

Brief Assessment

Matrix Sign Zolotarev[17] focuses on computing matrix decompositions via Zolotarev's rational approximations to the sign function, not on analyzing Newton-Schulz approximation error in momentum-based optimizers. The paper addresses fundamentally different problems: matrix decomposition algorithms versus optimizer convergence analysis.

---

3. The polar express: Optimal matrix sign methods and their application to the muon algorithm

URL: View paper

Brief Assessment

Polar Express[16] focuses on optimizing the polar decomposition computation itself through minimax optimization, not on analyzing approximation error decay rates in Newton-Schulz versus SVD methods for convergence analysis.

Contribution 3: Sharper rank dependence in MUON with Newton-Schulz

Description: The authors prove that MUON with Newton-Schulz removes the square-root-of-rank factor from the convergence rate compared to SGD with momentum, demonstrating a concrete theoretical advantage of matrix-aware optimization over vector-based methods under the same stationarity metric.

This contribution was assessed against **3 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Low-rank Momentum Factorization for Memory Efficient Training

URL: View paper

Brief Assessment

Low-rank Momentum[18] focuses on memory-efficient training via low-rank momentum factorization for large models, not on proving rank-dependent convergence rate improvements over SGD with momentum under the same stationarity metric.

---

2. Momentum Tracking: Momentum Acceleration for Decentralized Deep Learning on Heterogeneous Data

URL: View paper

Brief Assessment

Momentum Tracking[20] addresses decentralized learning with data heterogeneity, not rank-dependent convergence rates in matrix-aware optimization methods.

---

3. On the O(Ã□d/T1/4) Convergence Rate of RMSProp and Its Momentum Extension Measured by ð□□□l Norm: Better Dependence on the Dimension

URL: View paper

Brief Assessment

RMSProp Convergence[19] analyzes adaptive gradient methods (RMSProp) with dimension-dependent convergence rates, while the original paper studies matrix-aware optimization (MUON) with rank-dependent rates. These are fundamentally different optimization paradigms addressing different structural properties.

Appendix: Text Similarity Detection

No high-similarity text segments were detected across any compared papers.

References

- [0] Convergence of Muon with Newton-Schulz View paper
- [1] A feasible method for optimization with orthogonality constraints View paper
- [2] Simplifying momentum-based positive-definite submanifold optimization with applications to deep learning View paper
- [3] Parallelizable algorithms for optimization problems with orthogonality constraints View paper
- [4] Beyond the ideal: Analyzing the inexact muon update View paper

- [5] NorMuon: Making Muon more efficient and scalable View paper
- [6] Structured Joint Sparse Orthogonal Nonnegative Matrix Factorization for Fault Detection View paper
- [7] FedMuon: Accelerating Federated Learning with Matrix Orthogonalization View paper
- [8] Accelerated Optimization With Orthogonality Constraints View paper
- [9] Novel Tensor Norm Optimization for Neural Network Training Acceleration View paper
- [10] High-dimensional isotropic scaling dynamics of Muon and SGD View paper
- [11] AuON: A Linear-time Alternative to Semi-Orthogonal Momentum Updates View paper
- [12] Incomplete Gaussian elimination as a preconditioning for generalized conjugate gradient acceleration View paper
- [13] Improving generalization in DNNs through enhanced orthogonality in momentum-based optimizers View paper
- [14] Sparse uncorrelated linear discriminant analysis for undersampled problems View paper
- [15] Quadratic Optimization with Orthogonality Constraints: Explicit Lojasiewicz Exponent and Linear Convergence View paper
- [16] The polar express: Optimal matrix sign methods and their application to the muon algorithm View paper
- [17] Computing fundamental matrix decompositions accurately via the matrix sign function in two iterations: The power of Zolotarev's functions View paper
- [18] Low-rank Momentum Factorization for Memory Efficient Training View paper
- [19] On the O(Ãd/T1/4) Convergence Rate of RMSProp and Its Momentum Extension Measured by ð Norm: Better Dependence on the Dimension View paper
- [20] Momentum Tracking: Momentum Acceleration for Decentralized Deep Learning on Heterogeneous Data View paper
- [21] Preconditioned Inexact Stochastic ADMM for Deep Model View paper
- [22] ROOT: Robust Orthogonalized Optimizer for Neural Network Training View paper
- [23] AuON: A Linear-time Alternative to Orthogonal Momentum Updates View paper
- [24] ANDI: Adaptive Norm-Distribution Interface View paper
- [25] MGUP: A Momentum-Gradient Alignment Update Policy for Stochastic Optimization View paper