Novelty Assessment Report

Paper: The Polar Express: Optimal Matrix Sign Methods and their Application to the Muon Algorithm
PDF URL: https://openreview.net/pdf?id=yRtgZ1K8hO
Venue: ICLR 2026 Conference Submission
Year: 2026
Report Generated: 2025-12-27

Abstract

Computing the polar decomposition and the related matrix sign function has been a well-studied problem in numerical analysis for decades. Recently, it has emerged as an important subroutine within the Muon algorithm for training deep neural networks. However, the requirements of this application differ sharply from classical settings: deep learning demands GPU-friendly algorithms that prioritize high throughput over high precision. We introduce Polar Express, a new method for computing the polar decomposition. Like Newton–Schulz and other classical polynomial methods, our approach uses only matrix-matrix multiplications, making it very efficient on GPUs. Inspired by earlier work of Chen \& Chow and Nakatsukasa \& Freund, Polar Express adapts the update rule at each iteration by solving a minimax optimization problem. We prove that this strategy minimizes error in a worst-case sense, allowing Polar Express to converge as rapidly as possible both in the early iterations and asymptotically. We also address finite-precision issues, making it practical to use in `bfloat16` . When integrated into Muon, our method yields consistent improvements in validation loss for a GPT-2 model on one to ten billion tokens from the FineWeb dataset, outperforming recent alternatives across a range of learning rates.

Core Task Landscape

This paper addresses: **Computing the Polar Decomposition for Neural Network Optimization**

A total of **19 papers** were analyzed and organized into a taxonomy with **15 categories**.

Taxonomy Overview

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

- **Polar Decomposition Algorithms and Theory**
- **Neural Network Optimization with Orthogonality Constraints**
- **Statistical Estimation and Structured Optimization**
- **Spatial Transformation and Equivariance**
- **SAR Polarimetry and Remote Sensing Applications**

Complete Taxonomy Tree

- Computing the Polar Decomposition for Neural Network Optimization Survey Taxonomy
- Polar Decomposition Algorithms and Theory
  - Classical Numerical Methods (1 papers)
  - [5] Computing the polar decompositionâ with applications (j., 1986) View paper
  - Adaptive Minimax Optimization Methods ★ (1 papers)
  - [0] The Polar Express: Optimal Matrix Sign Methods and their Application to the Muon Algorithm (Anon et al., 2026) View paper
  - Continuous-Time and Discrete-Time Neural Network Models (1 papers)
  - [17] Time-Varying Polar Decomposition by Continuous-Time Model and Discrete-Time Algorithm of Zeroing Neural Network Using Zhang Time Discretization (ZTD) (Zanyu Tang, 2021) View paper
  - Theoretical Foundations and Applications (3 papers)
  - [8] On a Neural Implementation of Brenier's Polar Factorization (Vesseron, 2024) View paper
  - [18] Applied Linear Algebra and Big Data Course Book (Kabir K. Gandhi, 2019) View paper
  - [19] The Neural Solids; For optimization problems (Giansalvo Cirrincione, 2001) View paper
- Neural Network Optimization with Orthogonality Constraints
  - Low-Rank Adaptation and Fine-Tuning (1 papers)
  - [2] PoLAR: Polar-Decomposed Low-Rank Adapter Representation (Zhang Liang, 2025) View paper
  - Orthogonal Training Frameworks (1 papers)
  - [7] TAOTF: A Two-stage Approximately Orthogonal Training Framework in Deep Neural Networks (Taoyong Cui, 2022) View paper
  - Graph Neural Network Orthogonalization (2 papers)
  - [1] Simple orthogonal graph representation learning (student abstract) (Cui, 2024) View paper
  - [6] Expressive 1-lipschitz neural networks for robust multiple graph learning against adversarial attacks (Zhao Xin, 2021) View paper
  - Matrix-Gradient Preconditioning (1 papers)
  - [10] PolarGrad: A Class of Matrix-Gradient Optimizers from a Unifying Preconditioning Perspective (Lau, 2025) View paper
- Statistical Estimation and Structured Optimization
  - Group Synchronization (1 papers)
  - [11] Optimal orthogonal group synchronization and rotation group synchronization (Gao Chao, 2022) View paper
  - Atomic Decomposition via Polar Alignment (1 papers)
  - [3] Atomic decomposition via polar alignment: The geometry of structured optimization (Zhenan Fan, 2020) View paper
  - Quantum Singular Value Decomposition (1 papers)

    - [4] Variational quantum singular value decomposition (Wang Xin, 2021) View paper
- Spatial Transformation and Equivariance (1 papers)
    - [9] Polar transformer networks (Carlos Esteves, 2017) View paper
- SAR Polarimetry and Remote Sensing Applications
    - Crop and Land Cover Mapping (3 papers)
    - [12] Crop Type Mapping Based on Polarization Information of Time Series Sentinel-1 Images Using Patch-Based Neural Network (Yuying Liu, 2023) View paper
    - [14] Water-Body Detection From Spaceborne SAR Images With DBO-CNN (Qi-ming Yuan, 2023) View paper
    - [16] Characterizing Ancient Channel of the Yellow River From Spaceborne SAR: Case Study of Chinese Gaofen-3 Satellite (Ning Li, 2021) View paper
    - Target Detection and Identification (1 papers)
    - [13] Anti-Corner Reflector Array Method Based on Pauli Polarization Decomposition and BP Neural Network (Liang Ziyao, 2021) View paper
    - Medical Polarimetry Imaging (1 papers)
    - [15] Detecting cervical intraepithelial neoplasia using polarimetry parameters and multichannel convolutional neural network (Yang Dong, 2021) View paper

Narrative

Core task: Computing the polar decomposition for neural network optimization. The field encompasses several distinct branches that reflect both foundational algorithmic concerns and diverse application domains. At its center, Polar Decomposition Algorithms and Theory addresses the numerical methods and convergence guarantees needed to factorize matrices into orthogonal and positive-definite components—a classical problem dating back to foundational work such as Computing Polar Decomposition[5]. Adjacent to this, Neural Network Optimization with Orthogonality Constraints explores how enforcing or learning orthogonal weight matrices can improve training stability and generalization, with recent efforts like Simple Orthogonal Graph[1] and PoLAR[2] demonstrating practical benefits. A third branch, Statistical Estimation and Structured Optimization, examines problems such as group synchronization and optimal transport where polar-like factorizations arise naturally. Meanwhile, Spatial Transformation and Equivariance investigates geometric transformations in vision tasks, and SAR Polarimetry and Remote Sensing Applications focuses on radar signal processing for earth observation, illustrating the breadth of contexts in which polar decompositions prove useful.

Within the algorithmic core, a particularly active line of work concerns adaptive minimax optimization methods that balance computational efficiency with numerical stability. Polar Express[0] sits squarely in this space, proposing a novel optimizer that leverages polar decomposition to handle non-convex landscapes more robustly. Its emphasis on adaptive step-size rules and minimax formulations contrasts with earlier approaches like Polar Alignment[3], which focused on aligning learned representations through polar factorization, and PolarGrad[10], which integrates polar updates directly into gradient descent. These neighboring works share a common interest in exploiting orthogonal structure, yet they differ in whether the decomposition is computed explicitly at each iteration or approximated via cheaper surrogates. Open questions remain about the trade-offs between exact polar updates and scalable heuristics, as well as how these methods generalize across different network architectures and loss surfaces.

Related Works in Same Category

No sibling papers were found in the same taxonomy leaf. A taxonomy-subtopic-level comparison will be produced instead.

Taxonomy-Level Summary

The original leaf on Adaptive Minimax Optimization Methods focuses on algorithms that dynamically adjust update rules through minimax optimization to achieve optimal convergence rates in neural network training. The sibling subtopics cover complementary aspects: Classical Numerical Methods addresses traditional iterative schemes like Newton-based approaches, Continuous-Time and Discrete-Time Neural Network Models handles time-varying polar decomposition through zeroing neural networks, and Theoretical Foundations explores the mathematical underpinnings and generalizations of polar decomposition itself.

Similarities: - All subtopics relate to computing polar decomposition, which is central to neural network optimization - Each category addresses convergence and computational efficiency, though through different mechanisms - All exclude overlapping methods to maintain clear boundaries (e.g., adaptive methods excluded from classical, static methods excluded from continuous-time)

Differences: - Adaptive Minimax methods focus on learning-based update rule adaptation, while Classical Numerical Methods use fixed iterative schemes - The original leaf targets optimization convergence rates through minimax formulations, whereas Continuous-Time Models address time-varying decomposition problems - Adaptive Minimax is application-oriented toward neural network training, while Theoretical Foundations emphasizes mathematical theory and generalizations - Classical methods employ acceleration techniques for polynomial schemes, contrasting with the adaptive rule modification in minimax approaches

Suggested Search Directions: - Hybrid approaches combining adaptive minimax strategies with classical acceleration techniques - Connections between continuous-time neural network models and discrete adaptive optimization - Theoretical convergence guarantees for adaptive minimax methods compared to classical bounds

Sibling Subtopics

- **Classical Numerical Methods** (leaves: 1, papers: 1)
- Scope: Newton-based and polynomial iterative schemes for polar decomposition with acceleration techniques.
- Exclude: Adaptive minimax methods and neural network-specific algorithms belong to other categories.
- **Continuous-Time and Discrete-Time Neural Network Models** (leaves: 1, papers: 1)
- Scope: Zeroing neural network approaches for time-varying polar decomposition with discretization schemes.
- Exclude: Static matrix methods and deep learning optimization applications belong elsewhere.
- **Theoretical Foundations and Applications** (leaves: 1, papers: 3)
- Scope: Mathematical theory of polar decomposition including generalizations and educational treatments.
- Exclude: Computational algorithms and specific optimization implementations belong to other categories.

Contributions Analysis

Overall novelty summary. The paper introduces Polar Express, an adaptive minimax method for computing polar decomposition tailored to GPU-based neural network training. According to the taxonomy, it resides in the 'Adaptive Minimax Optimization Methods' leaf under 'Polar Decomposition Algorithms and Theory'. Notably, this leaf contains no sibling papers in the current taxonomy, suggesting it occupies a relatively sparse research direction. The taxonomy distinguishes this category from classical fixed-rule methods and continuous-time models, positioning Polar Express as a specialized approach that adapts update rules via minimax optimization for optimal convergence.

The taxonomy reveals that neighboring leaves include 'Classical Numerical Methods' (Newton-based and polynomial schemes), 'Continuous-Time and Discrete-Time Neural Network Models' (zeroing neural networks for time-varying decomposition), and 'Theoretical

Foundations and Applications' (mathematical theory and generalizations). The scope notes clarify that Polar Express diverges from classical fixed-rule methods by incorporating adaptive minimax optimization, and from continuous-time models by focusing on discrete iterative updates. The broader 'Neural Network Optimization with Orthogonality Constraints' branch addresses orthogonal training frameworks and low-rank adaptation, but these methods typically enforce constraints rather than compute decompositions as a subroutine.

Among 21 candidates examined across three contributions, none were flagged as clearly refutable. The core algorithm examined 5 candidates with 0 refutable matches; the optimality proof examined 6 candidates with 0 refutable matches; and the finite-precision modifications examined 10 candidates with 0 refutable matches. This suggests that within the limited search scope—top-K semantic matches plus citation expansion—no prior work was found that directly overlaps with the combination of adaptive minimax optimization, GPU-oriented design, and bfloat16 compatibility. The absence of sibling papers in the taxonomy leaf further indicates that this specific intersection of concerns has received limited prior attention.

Based on the limited literature search (21 candidates), the work appears to occupy a novel position at the intersection of classical polar decomposition theory and modern deep learning infrastructure demands. The taxonomy structure and contribution-level statistics suggest that while related methods exist in neighboring leaves, the specific combination of adaptive minimax updates, GPU efficiency, and low-precision arithmetic has not been extensively explored. However, the analysis does not cover exhaustive searches across all numerical linear algebra or optimization venues, leaving open the possibility of relevant work outside the examined scope.

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

Contribution 1: Polar Express algorithm for computing polar decomposition

Description: The authors propose Polar Express, an iterative method that dynamically adapts polynomial update rules at each iteration by solving a minimax optimization problem. This approach minimizes worst-case error and converges super-exponentially while using only GPU-friendly matrix-matrix multiplications.

This contribution was assessed against **5 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. The matrix sign decomposition and its relation to the polar decomposition

URL: View paper

Brief Assessment

Matrix Sign Decomposition[21] focuses on the relationship between matrix sign decomposition and polar decomposition using rational iterations (polynomials in numerator and denominator), not the specific minimax polynomial optimization approach of Polar Express.

---

2. A note on the extension of the polar decomposition for the multidimensional Burgers equation

URL: View paper

Brief Assessment

Burgers Polar Extension[23] addresses polar decomposition for the multidimensional Burgers equation in fluid dynamics, not iterative matrix algorithms for neural network optimization. The mathematical contexts are entirely different.

---

3. SVD for very large matrices: An approach with polar decomposition and polynomial approximation

URL: View paper

Brief Assessment

SVD Large Matrices[20] focuses on accelerating polar decomposition through Chebyshev polynomial approximation to bypass QR decomposition for SVD applications. The original paper's Polar Express uses minimax-optimal polynomial compositions specifically designed for GPU-friendly matrix-matrix multiplications in deep learning contexts, representing a different algorithmic approach and application domain.

---

4. Rotation Matrix and Angles of Rotation in the Polar Decomposition

URL: View paper

Brief Assessment

Rotation Matrix Angles[24] focuses on computing rotation matrices and angles using Newton's and Halley's methods for polar decomposition, not on developing optimal polynomial composition methods for GPU-friendly matrix sign computation as in the original paper.

---

5. A sixth-order iterative method for approximating the polar decomposition of an arbitrary matrix

URL: View paper

Brief Assessment

Sixth-Order Iterative[22] focuses on a fixed sixth-order iterative scheme, whereas the original paper proposes an adaptive method that dynamically changes polynomial update rules at each iteration by solving minimax optimization problems.

Contribution 2: Optimality proof for composition of polynomials

Description: The authors prove that their greedy polynomial selection strategy yields the optimal composition of polynomials for approximating the matrix sign function in the supremum norm. This theoretical result (Theorem 3.1) guarantees that Polar Express achieves the best possible worst-case convergence rate.

This contribution was assessed against **6 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Polynomial approximation of piecewise analytic functions

URL: View paper

Brief Assessment

Piecewise Analytic Approximation[41] focuses on approximating piecewise analytic functions using polynomial sequences with exponential convergence rates. The ORIGINAL paper addresses optimal polynomial composition for matrix sign function approximation in the supremum norm with worst-case convergence guarantees (Theorem 3.1), which is a different problem domain.

---

2. A PadÃ© family of iterations for the matrix sign function and related problems

URL: View paper

Brief Assessment

Pade Matrix Sign[37] focuses on convergence regions for Padé iterations for matrix sign/sector functions, not on optimality of polynomial compositions in the supremum norm sense claimed by the original paper.

---

3. 2-norm error bounds and estimates for Lanczos approximations to linear systems and rational matrix functions

URL: View paper

Brief Assessment

Lanczos Error Bounds[39] focuses on error estimation for Lanczos approximations to linear systems and rational matrix functions using Gaussian quadrature theory, not on optimal polynomial approximation for the matrix sign function or worst-case convergence rates for polar decomposition methods.

---

4. Computing the matrix sign and absolute value functions

URL: View paper

Brief Assessment

Matrix Sign Function[40] focuses on polynomial iterations for matrix sign computation with known eigenvalue locations, but does not establish optimality of polynomial compositions in the supremum norm sense proven in the original paper's Theorem 3.1.

---

5. Computing via Least Squares Polynomial Approximations

URL: View paper

Brief Assessment

Least Squares Polynomial[38] focuses on approximating matrix functions via least squares polynomial approximations to splines, not on proving optimality of greedy polynomial composition for matrix sign function approximation in the supremum norm.

---

6. A Sixth-Order Iterative Scheme Through Weighted Rational Approximations for Computing the Matrix Sign Function

URL: View paper

Brief Assessment

Weighted Rational Approximations[36] uses rational approximations with weight functions for the matrix sign function, while the original paper proves optimality for polynomial compositions in the supremum norm. These are fundamentally different approximation strategies (rational vs. polynomial).

Contribution 3: Finite-precision modifications for bfloat16 compatibility

Description: The authors develop specific modifications to stabilize the algorithm when working in half-precision arithmetic (bfloat16), including rescaling polynomials and using slightly suboptimal polynomials in early iterations to handle numerical round-off errors.

This contribution was assessed against **10 related papers** from the literature. Papers with potential prior art are analyzed in detail with textual evidence; others receive brief assessments.

---

1. Mixed precision algorithms in numerical linear algebra

URL: View paper

Brief Assessment

Mixed Precision Algorithms[30] is a survey of mixed precision numerical linear algebra algorithms broadly. It does not specifically address the polar decomposition algorithm or the particular finite-precision modifications (rescaling polynomials, using suboptimal polynomials in early iterations) that the original paper develops for their specific polar express method in bfloat16.

---

2. An SMT Formalization of Mixed-Precision Matrix Multiplication: Modeling Three Generations of Tensor Cores

URL: View paper

Brief Assessment

Tensor Cores SMT[28] focuses on formalizing tensor core behavior for matrix multiplication hardware, not on algorithmic modifications for numerical stability in bfloat16 arithmetic for iterative matrix algorithms.

---

3. Probabilistic rounding error analysis for numerical linear algebra

URL: View paper

Brief Assessment

Probabilistic Rounding Error[32] focuses on probabilistic error analysis for numerical linear algebra operations, not on specific algorithmic modifications for stabilizing matrix sign methods in bfloat16 arithmetic as developed in the original paper.

---

4. Evaluation of Bfloat16, Posit, and Takum Arithmetics in Sparse Linear Solvers

URL: View paper

Brief Assessment

Sparse Solver Arithmetics[29] evaluates bfloat16 in sparse linear solvers (LU, QR, GMRES) but does not address the specific modifications needed for matrix sign function computation or polar decomposition algorithms in half-precision arithmetic, which is the focus of the original paper's contribution.

---

5. When precision meets position: Bfloat16 breaks down rope in long-context training

URL: View paper

Brief Assessment

Bfloat16 RoPE[25] addresses numerical stability issues in rotary position embeddings under bfloat16 precision for long-context training, not matrix sign function algorithms. The candidate focuses on positional encoding degradation in transformers, while the original paper develops polynomial rescaling methods for matrix polar decomposition in the Muon optimizer.

---

6. Numerical Performance of the Implicitly Restarted Arnoldi Method in OFP8, Bfloat16, Posit, and Takum Arithmetics

URL: View paper

Brief Assessment

Arnoldi Mixed Precision[27] evaluates the Arnoldi method across various precision formats (OFP8, bfloat16, posit, takum) but does not develop specific algorithmic modifications for stabilizing matrix algorithms in half-precision arithmetic. The original paper's contribution involves developing rescaling polynomials and using suboptimal polynomials in early iterations to handle numerical round-off errors in bfloat16, which is a distinct algorithmic innovation not addressed in the candidate.

---

7. High accuracy matrix computations on neural engines: A study of qr factorization and its applications

URL: View paper

Brief Assessment

Neural Engines QR[34] addresses half-precision (fp16) arithmetic for QR factorization, not the polar decomposition algorithm in bfloat16. The candidate focuses on matrix factorization stability rather than polynomial iteration methods for matrix sign functions.

---

8. Approximate computing in numerical linear algebra: algorithms, analysis, and applications

URL: View paper

Brief Assessment

Approximate Computing[33] is a broad survey on approximate computing in numerical linear algebra. While it discusses finite-precision considerations and low-precision arithmetic (including bfloat16), it does not specifically address the polar decomposition algorithm or the particular stabilization techniques (rescaling polynomials, using suboptimal polynomials in early iterations) described in the original paper's contribution.

---

9. A survey of numerical linear algebra methods utilizing mixed-precision arithmetic

URL: View paper

Brief Assessment

Mixed-Precision Survey[26] focuses on numerical linear algebra methods across various precision formats (fp16, bfloat16, fp32, fp64) but does not specifically address polynomial-based matrix sign function algorithms or the particular stabilization techniques (rescaling polynomials, using suboptimal polynomials in early iterations) described in the original paper's contribution.

---

10. Squeezing a matrix into half precision, with an application to solving linear systems

URL: View paper

Brief Assessment

Half Precision Matrix[31] focuses on converting matrices to half precision for linear system solvers (GMRES-IR), not on stabilizing iterative polynomial methods for computing polar decompositions in neural network optimizers.

Appendix: Text Similarity Detection

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

References

- [0] The Polar Express: Optimal Matrix Sign Methods and their Application to the Muon Algorithm View paper
- [1] Simple orthogonal graph representation learning (student abstract) View paper
- [2] PoLAR: Polar-Decomposed Low-Rank Adapter Representation View paper
- [3] Atomic decomposition via polar alignment: The geometry of structured optimization View paper
- [4] Variational quantum singular value decomposition View paper
- [5] Computing the polar decompositionâ with applications View paper
- [6] Expressive 1-lipschitz neural networks for robust multiple graph learning against adversarial attacks View paper
- [7] TAOTF: A Two-stage Approximately Orthogonal Training Framework in Deep Neural Networks View paper
- [8] On a Neural Implementation of Brenier's Polar Factorization View paper
- [9] Polar transformer networks View paper
- [10] PolarGrad: A Class of Matrix-Gradient Optimizers from a Unifying Preconditioning Perspective View paper
- [11] Optimal orthogonal group synchronization and rotation group synchronization View paper
- [12] Crop Type Mapping Based on Polarization Information of Time Series Sentinel-1 Images Using Patch-Based Neural Network View paper
- [13] Anti-Corner Reflector Array Method Based on Pauli Polarization Decomposition and BP Neural Network View paper
- [14] Water-Body Detection From Spaceborne SAR Images With DBO-CNN View paper
- [15] Detecting cervical intraepithelial neoplasia using polarimetry parameters and multichannel convolutional neural network View paper
- [16] Characterizing Ancient Channel of the Yellow River From Spaceborne SAR: Case Study of Chinese Gaofen-3 Satellite View paper
- [17] Time-Varying Polar Decomposition by Continuous-Time Model and Discrete-Time Algorithm of Zeroing Neural Network Using Zhang Time Discretization (ZTD) View paper
- [18] Applied Linear Algebra and Big Data Course Book View paper
- [19] The Neural Solids; For optimization problems View paper
- [20] SVD for very large matrices: An approach with polar decomposition and polynomial approximation View paper
- [21] The matrix sign decomposition and its relation to the polar decomposition View paper
- [22] A sixth-order iterative method for approximating the polar decomposition of an arbitrary matrix View paper
- [23] A note on the extension of the polar decomposition for the multidimensional Burgers equation View paper
- [24] Rotation Matrix and Angles of Rotation in the Polar Decomposition View paper
- [25] When precision meets position: Bfloat16 breaks down rope in long-context training View paper
- [26] A survey of numerical linear algebra methods utilizing mixed-precision arithmetic View paper
- [27] Numerical Performance of the Implicitly Restarted Arnoldi Method in OFP8, Bfloat16, Posit, and Takum Arithmetics View paper
- [28] An SMT Formalization of Mixed-Precision Matrix Multiplication: Modeling Three Generations of Tensor Cores View paper
- [29] Evaluation of Bfloat16, Posit, and Takum Arithmetics in Sparse Linear Solvers View paper
- [30] Mixed precision algorithms in numerical linear algebra View paper
- [31] Squeezing a matrix into half precision, with an application to solving linear systems View paper
- [32] Probabilistic rounding error analysis for numerical linear algebra View paper
- [33] Approximate computing in numerical linear algebra: algorithms, analysis, and applications View paper
- [34] High accuracy matrix computations on neural engines: A study of qr factorization and its applications View paper
- [35] Computing fundamental matrix decompositions accurately via the matrix sign function in two iterations: The power of Zolotarev's functions View paper
- [36] A Sixth-Order Iterative Scheme Through Weighted Rational Approximations for Computing the Matrix Sign Function View paper
- [37] A PadÃ© family of iterations for the matrix sign function and related problems View paper
- [38] Computing via Least Squares Polynomial Approximations View paper

- [39] 2-norm error bounds and estimates for Lanczos approximations to linear systems and rational matrix functions View paper
- [40] Computing the matrix sign and absolute value functions View paper
- [41] Polynomial approximation of piecewise analytic functions View paper