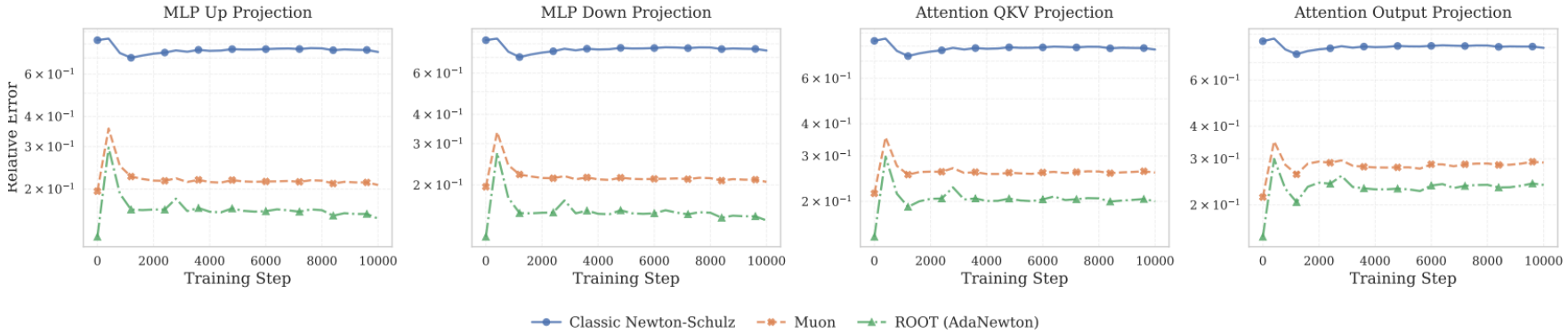
ROOT: Robust Orthogonalized Optimizer for Neural Network Training

251202

가짜연구소 허의주

ROOT: Robust Orthogonalized Optimizer for Neural Network Training

차원적 부정확성과 Outlier라는 두 가지 주요 한계를 해결한 Optimizer



Method	HellaSwag	BoolQ	PIQA	ARC-e	ARC-c	OBQA	SciQ	Wino	WSC	Avg.
AdamW	44.24	**62.60**	72.69	71.63	**37.80**	27.20	89.80	58.09	67.40	59.05
Muon	44.83	61.16	73.07	**74.12**	37.12	29.80	89.50	59.67	67.03	59.59
ROOT	**45.37**	62.08	**73.12**	72.14	36.86	**31.20**	**90.40**	**60.30**	**69.60**	**60.12**

1. Motivation

- Optimizer의 선택은 **수렴 속도**와 **최종 성능** 뿐만 아니라 모델 개발의 **막대한 경제적 비용에 깊은 영향**을 미침
- 기존 Optimizer의 **수치적 불안정성, 차원적 취약성, 이상치 민감도**를 해결하는 새로운 Optimizer 제안

2. Core Methodology

1. 알고리즘적 Robustness: ROOT는 Muon의 고정 계수 뉴턴-슐츠 반복을 차원 별 계수를 사용하는 **적응형 뉴턴 반복**으로 대체
   → 다양한 아키텍처 구성에서 일관된 정밀도 보장, **행렬 차원 변화에 강건**하게 만듦
2. 최적화 Robustness: 데이터 수준 노이즈에 대한 최적화 강건성을 위한 **근접 최적화 (proximal optimization) 프레임워크** 도입
   → **Soft-Thresholding**을 통해 이상치 유발 Gradient noise를 억제, 훈련을 안정화

3. Experimental Results

- Muon 및 Adam 기반 Optimizer 대비 **더 빠르고 안정적인 수렴**을 달성
- 특히 노이즈가 많고 비대칭형 차원에서도 **우수한 최종 성능** 입증
- ROOT의 AdaNewton은 Muon의 고정 계수 방식보다 **직교화 정밀도를 크게 향상**
- ROOT의 Soft-threshold 기반 Outlier 억제 알고리즘은 LLM 및 비전 작업 모두에서 **일반화 성능을 향상**시킴

ROOT: Robust Orthogonalized Optimizer for Neural Network Training

차원적 부정확성과 Outlier라는 두 가지 주요 한계를 해결한 Optimizer

❖ Introduction

- 대규모 언어 모델(LLMs)을 사전 훈련 하는 데 필요한 **계산 요구의 증가**로 최적화 알고리즘 설계는 중요 연구 분야로 자리잡음
- Optimizer의 선택은 **수렴 속도**와 **최종 성능** 뿐만 아니라 모델 개발의 **막대한 경제적 비용**에 깊은 영향을 미침
- Adam 및 AdamW는 모델 크기가 수십억 개의 매개변수로 급증함에 따라 **수치적 불안정성**과 같은 적응형 방법의 내재된 한계가 분명해짐
- 이러한 문제들을 해결하기 위해 Muon은 가중치 행렬을 전체적인 개체로 간주해 **모멘텀 직교화**하는 **아키텍처적 변화를 가져옴**
- 하지만 **차원적 취약성**과 **이상치 민감도**라는 두 가지 중요한 한계가 대두되었고, 본 논문에서는 **두 가지 측면을 해결한 ROOT를 제안**

- **차별점: ROOT의 이중 Robustness 강화 메커니즘**

1. **알고리즘적 Robustness**: ROOT는 Muon의 고정 계수 뉴턴-슐츠 반복을 차원 별 계수를 사용하는 **적응형 뉴턴 반복**으로 대체
   다양한 아키텍처 구성에서 일관된 정밀도 보장, 행렬 차원 변화에 강건하게 만듦
2. **최적화 Robustness**: 데이터 수준 노이즈에 대한 최적화 강건성을 위한 **근접 최적화(proximal optimization)** 프레임워크 도입
   Soft-Thresholding을 통해 이상치 유발 Gradient noise를 억제, 훈련을 안정화

- **주요 성과**: ROOT가 기존 옵티마이저 대비 상당히 개선된 강건성, 더 빠른 수렴, 우수한 최종 성능을 달성

ROOT: Robust Orthogonalized Optimizer for Neural Network Training

차원적 부정확성과 Outlier라는 두 가지 주요 한계를 해결한 Optimizer

❖ Related Works

1. 행렬 인지 옵티마이저 (Matrix-Aware Optimizers)

- **배경**: 딥 뉴럴 네트워크 훈련은 전통적으로 SGD나 AdamW와 같은 표준 옵티마이저에 의존
- **한계**: 모델의 매개변수를 독립적인 벡터로 취급, 계산적으론 효율적이나 가중치 행렬에 내재된 **구조적 상관관계를 간과**
- **Muon의 등장**: Muon Optimizer는 명시적 곡률 근사 대신 **뉴턴-슐츠 반복을 통한 모멘텀 행렬 직교화 적용**
  → O(N) 복잡도를 달성함과 동시에 매개변수 행렬 전반에 걸친 일관된 업데이트 촉진

2. Muon 이후

- **동향**: Muon의 성공 이후 다양한 차원에서의 기능 확장 등장 - **(1) 효율성 및 확장성, (2) 적응성 및 정밀도**
- **특징**: Muon에 부족한 **계산 효율성을 증진**시키고 **Element-Wise Adaptivity를 통합**하려는 시도
- 모델: Dion, LiMuon, Drop-Muon, AdaGO, AdaMuon, CANS 등

ROOT: Robust Orthogonalized Optimizer for Neural Network Training

차원적 부정확성과 Outlier라는 두 가지 주요 한계를 해결한 Optimizer

❖ Approach

- Preliminaries
  - ROOT는 Muon과 같은 직교화 기반 옵티마이저에서 시작, 행렬 구조 매개변수의 최적화 수행

$$M_t = \mu M_{t-1} + \nabla \mathcal{L}(\theta_{t-1})$$
$$M'_t = \text{Newton-Schulz}(M_t)$$
$$\theta_t = \theta_{t-1} - \eta_t M'_t$$

  : Newton-Schulz(NS) 반복을 통해 기울기 모멘텀을 직교화한 후 Parameter 업데이트 수행

- Adaptive Newton Iteration with Fine-grained Coefficients
  - Muon의 한계
    : 고정 계수를 이용한 NS 반복은 **행렬 차원에 상당한 민감도**를 보임
    → 뉴럴 네트워크의 다양한 계층 전반에 걸쳐 **일관적이지 않은 직교화 품질** 발생
  - 행렬 차원 변화에 대한 Robustness 향상
    : ROOT는 세밀한, 차원별 계수를 갖는 적응형 뉴턴-슐츠 반복을 제안,
    이 계수들은 네트워크 아키텍처의 고유한 행렬 크기 각각에 맞춰 학습됨
    → Neural Network의 모든 계층에서 **일관된 고정밀 직교화**를 포장, 행렬 차원 변화에 강건

$$X_k = a^{(m,n)} X_{k-1} + b^{(m,n)} X_{k-1}(X_{k-1}^T X_{k-1}) + c^{(m,n)} X_{k-1}(X_{k-1}^T X_{k-1})^2$$

Algorithm 1 Muon Optimizer (Jordan et al., 2024)

Require: Learning rate η, momentum μ

1: Initialize $M_0 \leftarrow 0$
2: **for** $t = 1, \ldots$ **do**
3: Compute gradient $G_t \leftarrow \nabla_\theta \mathcal{L}_t(\theta_{t-1})$
4: $M_t \leftarrow \mu M_{t-1} + G_t$
5: $M'_t \leftarrow \text{NewtonSchulz5}(M_t)$
6: Update parameters $\theta_t \leftarrow \theta_{t-1} - \eta M'_t$
7: **end for**
8: **Return** θ_t

Table 1. Orthogonalization error reveals dimensional fragility of fixed-coefficient Newton-Schulz iteration.

n	2048	4096	8192	2048	2048	2048	2048
m	2048	4096	8192	3072	4096	8192	16384
a	3.4445	3.4445	3.4445	3.4445	3.4445	3.4445	3.4445
b	-4.7750	-4.7750	-4.7750	-4.7750	-4.7750	-4.7750	-4.7750
c	2.0315	2.0315	2.0315	2.0315	2.0315	2.0315	2.0315
MSE	0.0499	0.0637	0.0761	0.0338	0.0362	0.0340	0.0425
a	3.3334	3.3732	3.3886	2.9091	2.7739	2.5925	2.7045
b	-4.2591	-4.6134	-4.9026	-3.8108	-3.4911	-3.0010	-3.2335
c	1.7791	2.0576	2.3105	1.8600	1.6992	1.4138	1.5336
MSE	0.0352	0.0470	0.0587	0.0024	0.0010	0.0003	0.0003

ROOT: Robust Orthogonalized Optimizer for Neural Network Training

차원적 부정확성과 Outlier라는 두 가지 주요 한계를 해결한 Optimizer

❖ Approach

- Adaptive Newton Iteration with Fine-grained Coefficients

▶ 이상치 억제를 통한 Robust Optimization

- Outlier-Induced Gradient Noise에 대한 Robustness 향상
  : 비정상적으로 큰 크기를 가진 Outlier가 NS 초기 정규화 단계 왜곡할 수 있으며, NS 반복의 다항식적 특성에 의해 증폭될 수 있음
  → ROOT는 **Soft-Thresholding**을 통합한 **인접 부분 최적화**를 통해 이를 해결
- 이상치 억제를 위한 소프트 임계값
  : ROOT는 모멘텀을 2가지 요소로 모델링 - 기본 구성 요소와 Outlier

$$M_t = B_t + O_t$$

최적화 목표: L1-norm 페널티를 이용, Outlier를 명시적으로 제어,

$$\min_{B_t, O_t} \|M_t - B_t - O_t\|_F^2 + \lambda \|O_t\|_1 \quad \text{subject to} \quad \|B_t\| \leq \tau$$

Soft-thresholding으로 Outlier 분리

$$\mathcal{T}_\varepsilon[x]_i = \text{sign}(x_i) \cdot \max(|x_i| - \varepsilon, 0)$$

- AdaNewton은 오직 Robust한 기본 구성 요소인 B_t에만 적용됨

Algorithm 2 ROOT Optimizer

Require: Learning rate η, momentum μ, threshold ε

Initialize $M_0 \leftarrow 0$
for $t = 1, \ldots$ **do**
  Compute gradient $G_t \leftarrow \nabla_\theta \mathcal{L}(\theta_{t-1})$
  $M_t \leftarrow \mu M_{t-1} + G_t$ # Momentum accumulation
  $O_t \leftarrow \mathcal{T}_\varepsilon[M_t]$ # Outlier separation via soft-thresholding
  $B_t \leftarrow M_t - O_t$ # Clipped base components
  $B_t^{\text{orth}} \leftarrow \text{AdaNewton}(B_t)$ # Robust orthogonalization
  Update parameters $\theta_t \leftarrow \theta_{t-1} - \eta B_t^{\text{orth}}$
end for
Return θ_t

ROOT: Robust Orthogonalized Optimizer for Neural Network Training

차원적 부정확성과 Outlier라는 두 가지 주요 한계를 해결한 Optimizer

❖ Experiments

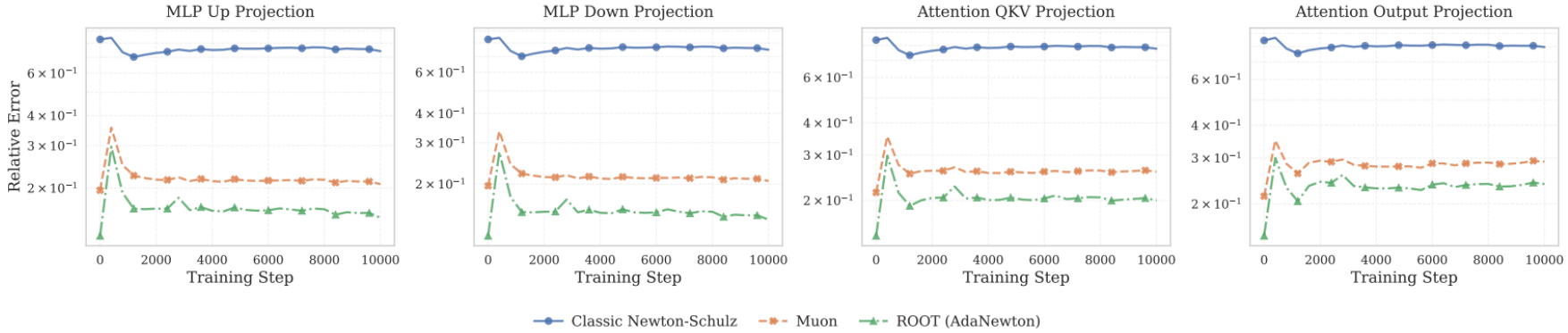
- Experimental Setup
  - ROOT Optimizer 접근 방식 검증을 위해 FineWeb-Edu 데이터셋을 사용, 사전 훈련 수행
  - Ablation studies를 위해 10B 토큰의 하위 집합을, 주요 실험에는 100B 토큰 샘플을 활용
  - 제안된 최적화 개선 사항 평가를 위한 1B Transformer 모델을 훈련
  - 단일 epoch 동안 사전 훈련, 2,000단계의 웜업 단계를 거쳐 최고 학습률의 10%까지 감소하는 코사인 학습률 스케줄을 사용
  - 하이퍼파라미터는 Muon 옵티마이저의 기본 하이퍼파라미터를 채택, AdamW와의 업데이트 RMS를 맞추기 위해 0.2 스케일링 인자를 적용
  - HellaSwag, ARC-easy (ARC-e), ARC-challenge (ARC-c), BoolQ, PIQA, SciQ, WINO, OBQA, WSC를 포함한 포괄적인 학술 벤치마크 세트에서 평가함
  - 모든 평가는 lm-evaluation-harness 프레임워크를 사용하여 **제로샷(zero-shot)**으로 수행

ROOT: Robust Orthogonalized Optimizer for Neural Network Training

차원적 부정확성과 Outlier라는 두 가지 주요 한계를 해결한 Optimizer

❖ Experiments

- Experimental Result



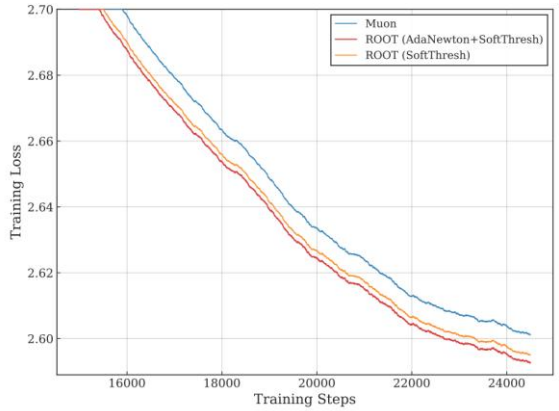


Figure 3. Training loss curves for 10B tokens. ROOT variants demonstrate faster convergence and lower final loss compared to Muon baseline, with full ROOT achieving the best performance.

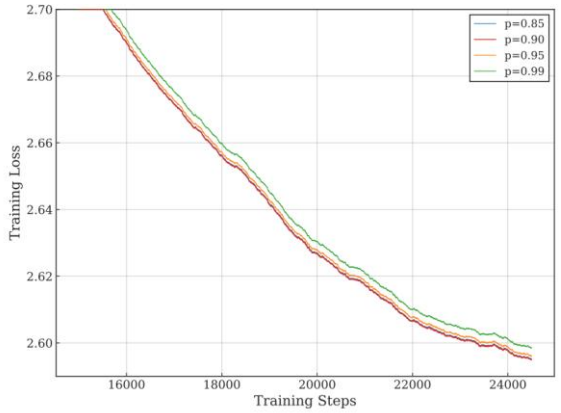


Figure 4. Ablation on the quantile hyperparameter p. The curve with $p = 0.90$ demonstrates the optimal equilibrium between suppressing gradient noise and preserving informative gradient signals.

Table 3. Top-1 Test Accuracy on CIFAR-10 (ViT, 6.3 M, trained from scratch). Fixed params: Muon LR $= 0.02$, AdamW LR $= 0.001$, WD $= 5 \times 10^{-5}$.

METHOD	p (QUANTILE)	ACC (%)
MUON	-	84.67
ROOT	0.95	85.75
ROOT	0.90	86.58
ROOT	0.85	**88.44**

ROOT: Robust Orthogonalized Optimizer for Neural Network Training

차원적 부정확성과 Outlier라는 두 가지 주요 한계를 해결한 Optimizer

❖ Conclusion

- LLMs 훈련에서 나타나는 **두 가지 중요한 한계**를 해결하는 **강건하면서도 직교화된 ROOT Optimizer** 제시
  - 1) dimension-robust orthogonalization
  - 2) new paradigm for stable and efficient neural network optimization
- 이러한 이중 Robustness 메커니즘을 통해 **안정적이고 효율적인 신경망 최적화**를 위한 **새로운 패러다임**을 확립
- 광범위한 실험 검증을 통해,
  - **노이즈가 많고 non-convex scenarios**에서 **우수한 성능을 달성**함을 입증
  - ROOT가 **이론적 보장**과 **실질적인 이점** 모두를 제공함을 확인