

Demystifying Manifold Constraints in LLM Pre-training

Kang An^{1*} Jiaxiang Li^{2*} Donald Goldfarb³ Shiqian Ma¹

¹Rice University

²Virginia Tech

³Columbia University

*Equal contribution jasonljx@vt.edu

Introduction

Matrix optimizers normalize momentum. What changes when we also constrain the **weights**?

How do manifold constraints interact with RMSNorm, weight decay, and training precision?

We use **MACRO** to study activation scales, weight rotation, and the update-to-weight ratio.

MACRO: A Single-Loop Constrained Optimizer

For $\mathbf{W} \in \mathbb{R}^{p \times n}$ (p : output; n : input dimension), solve

$$\min_{\mathbf{W} \in \mathcal{M}} \mathcal{L}(\mathbf{W}) = \mathbb{E}_{\xi}[\ell(\mathbf{W}, \xi)].$$

We focus on the Frobenius and spectral constraints:

$$\mathcal{M}_{\text{fro}} = \{\mathbf{W} : \|\mathbf{W}\|_F = R\}, \quad \mathcal{M}_{\text{spec}} = \{\mathbf{W} : \|\mathbf{W}\|_2 = R\}.$$

Msign-Aligned Constrained Riemannian Optimizer:

$$\begin{aligned} \mathbf{G}_t &= \nabla \ell(\mathbf{W}_t, \xi_t), \\ \mathbf{M}_t &= \beta \mathbf{M}_{t-1} + (1 - \beta) \mathbf{G}_t, \\ \Phi_t &= \Pi_{T_{\mathbf{W}_t} \mathcal{M}}(\mathbf{M}_t), \quad \mathbf{O}_t = \text{msign}(\Phi_t), \\ \mathbf{D}_t &= \frac{R \mathbf{O}_t}{\|\mathbf{O}_t\|_{\mathcal{M}} + \epsilon}, \quad \mathbf{W}_{t+1} = P_{\mathcal{M}}(\mathbf{W}_t - \eta_t \mathbf{D}_t). \end{aligned}$$

Here $\Pi_{T_{\mathbf{W}} \mathcal{M}}$ is the tangent projection, $\text{msign}(U \Sigma V^{\top}) = U V^{\top}$, and $\|\cdot\|_{\mathcal{M}}$ is the constraint norm.

One matrix-sign evaluation; no inner bisection. The sign direction need not remain tangent.

η_t : learning rate; β : momentum coefficient; ϵ : numerical stabilizer. Experiments use $P_{\mathcal{M}}(A) = RA/\|A\|_{\mathcal{M}}$; spectral rescaling is not exact Euclidean projection.

Initialization-Based Radius Selection

Use a common multiplier $r > 0$ with norm-specific scaling:

$$R_{\text{fro}} = r \sqrt{p} \quad R_{\text{spec}} = r \sqrt{p/n}.$$

Let $r_X = \|\mathbf{X}\|_F / \sqrt{Nn}$ and $r_Y = \|\mathbf{X} \mathbf{W}_0^{\top}\|_F / \sqrt{Np}$ for N input vectors. Under normalized Gaussian initialization independent of $\mathbf{X}$,

$$\mathbb{E}[r_Y^2 \mid \mathbf{X}] = r^2 r_X^2 \quad (\text{fro}), \quad r_Y \leq r r_X \quad (\text{spec}).$$

Convergence Guarantee

For compact C^3 manifolds (e.g., Frobenius), a lower-bounded smooth loss, unbiased bounded-variance gradients, and the theorem's parameter schedule give

$$\min_{0 \leq t < T} \mathbb{E} \|\nabla_{\mathcal{M}} \mathcal{L}(\mathbf{W}_t)\|_* = O(T^{-1/4}),$$

where $\|\cdot\|_*$ is the nuclear norm. The spectral sphere is nonsmooth at repeated largest singular values.

1. Activation Dynamics: Interplay with RMSNorm

Ablation: remove standard learnable RMSNorm layers from a 330M Qwen3-like model. Parameter-free normalizations, QK-norm, and Norm-Gated SwiGLU are retained (this is *not* a fully normalization-free architecture).

Optimizer	Learning rate				
	0.003	0.005	0.007	0.01	0.03
Muon	2.912	2.849	2.823	2.813	NaN
MACRO-fro	2.901	2.825	2.787	2.758	2.781
MACRO-spec	2.820	2.768	2.747	2.739	2.819

Validation loss after 8.9B tokens; lower is better. MACRO is stable across this tested learning-rate grid.

With standard RMSNorm: larger radii coincide with smaller learned scale parameters γ . At 600M, varying $r \in [1, 5]$ changes validation PPL by at most 0.035 within each MACRO variant.

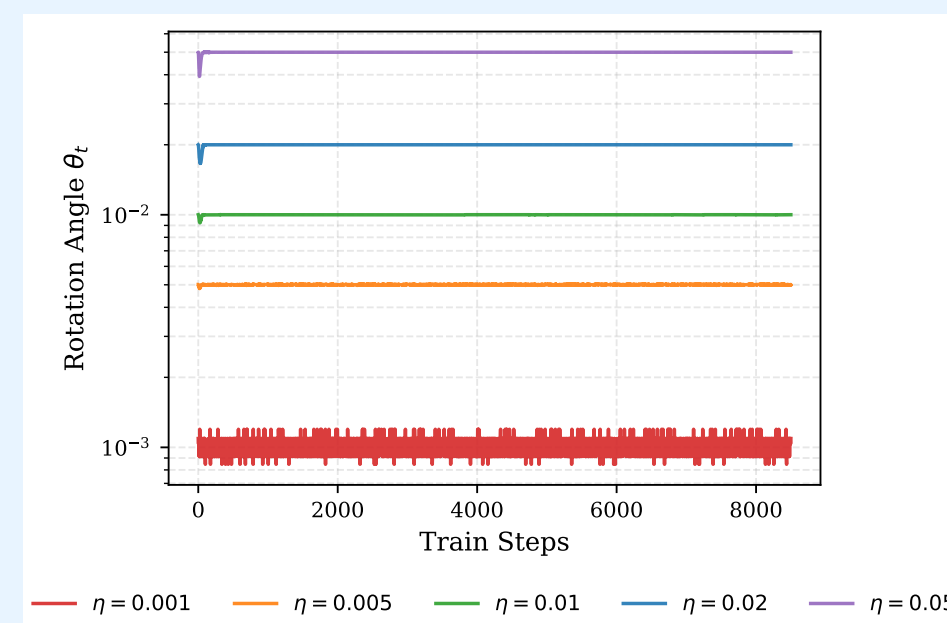
Takeaway: constraints and RMSNorm have aligned roles.

2. Rotational Dynamics: Interplay with Weight Decay

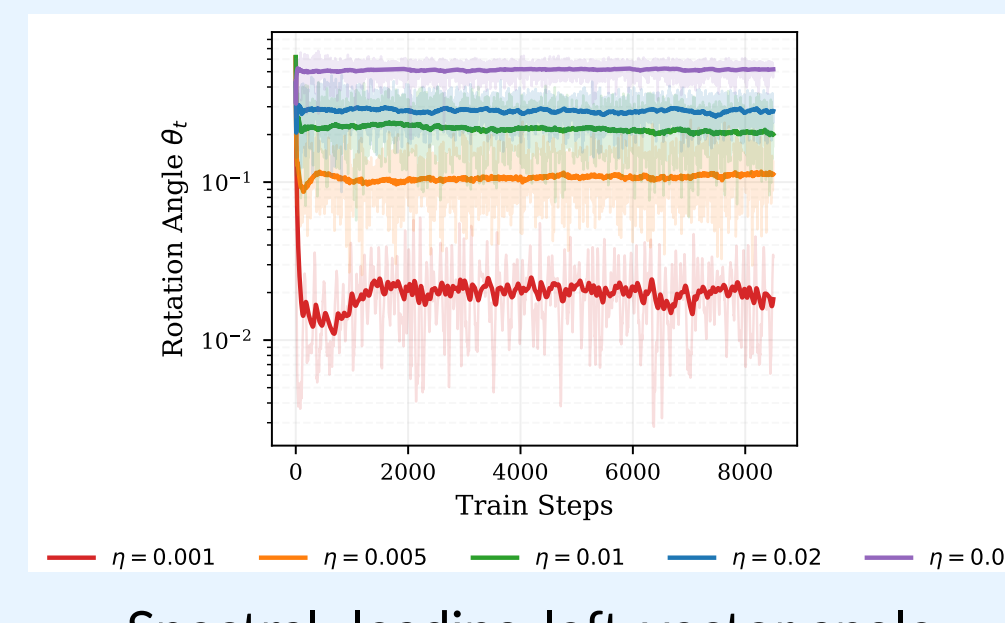
Frobenius sphere. Define the matrix angle $\theta_t = \angle(\mathbf{W}_{t+1}, \mathbf{W}_t)$ and $\alpha_t = \langle \mathbf{W}_t, \mathbf{O}_t \rangle_F / (R \|\mathbf{O}_t\|_F)$. Ignoring the numerical stabilizer, the projected update satisfies

$$\cos \theta_t = \frac{1 - \eta_t \alpha_t}{\sqrt{1 - 2\eta_t \alpha_t + \eta_t^2}}.$$

Thus $\theta_t \approx \eta_t \sqrt{1 - \alpha_t^2} \approx \eta_t$ for small steps and the empirically observed $\alpha_t \approx 0$.



Frobenius: global matrix angle.



Spectral: leading-left-vector angle.

Spectral sphere. Leading singular-vector rotations vary during training; the perturbation analysis depends on a positive local spectral gap.

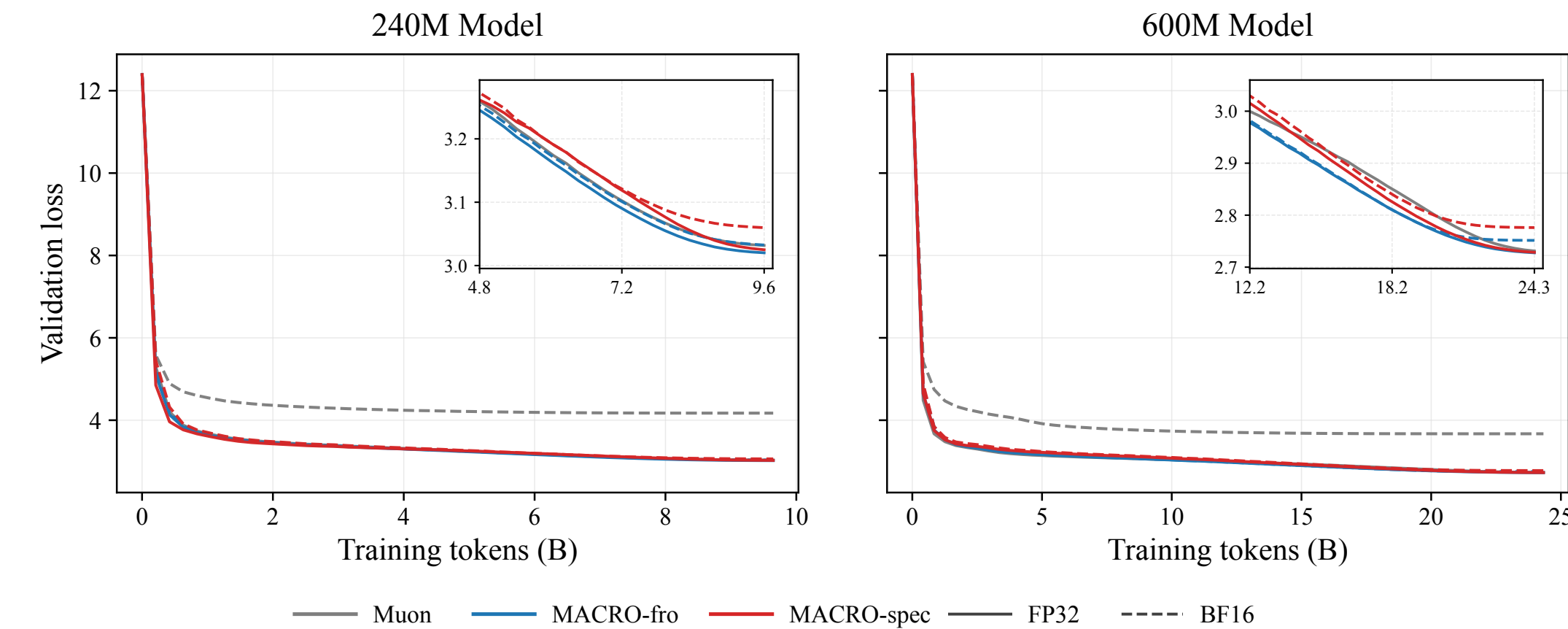
Weight-decay ablation. At 330M, changing weight decay from 0 to 0.1 changes MACRO validation loss by at most 0.001 across both constraints and three learning rates.

Takeaway: no weight decay is needed on constrained weights; adding constraints regulates rotation during training.

3. Relative Updates: Interplay with BF16 Precision

For the proposed update $\Delta \mathbf{W}_t = -\eta_t \mathbf{D}_t$, **before projection and rounding**, MACRO-fro enforces

$$\rho_t := \frac{\|\Delta \mathbf{W}_t\|_F}{\|\mathbf{W}_t\|_F} = \eta_t \quad (\epsilon = 0).$$



Validation loss with FP32 (solid) and BF16 (dashed) parameter storage. Insets show late-training differences.

Final loss increase: BF16 – FP32	240M	600M
Muon	1.1470	0.9394
MACRO-fro	0.0172	0.0239
MACRO-spec	0.0317	0.0476

BF16 parameter, gradient, and state storage; no FP32 master copy. One seed; matched schedules; no stochastic rounding or error compensation.

Interpretation: controlled relative updates help explain reduced cancellation.

Standard Pre-training Results

Standard Qwen3-like models on C4: 40 tokens per parameter, five seeds; using TorchTitan.

Optimizer	240M	600M	1.2B
Muon	20.72	15.37	12.99
MuonH-fro	20.55	15.23	12.92
MuonH-spec	20.63	15.33	12.88
MACRO-fro	20.58	15.29	12.85
MACRO-spec	20.50	15.30	12.90

Validation perplexity (five-seed means); best entries in bold.

Efficiency: MACRO-fro adds 1.6% end-to-end step time versus Muon at 240M, using plain implementations.