Spectral Equalization Minimizes Total Training Energy: A Control-Theoretic Account of Muon's Advantage

Euijin Hong[1]

[1]Carnegie Mellon University

Motivation

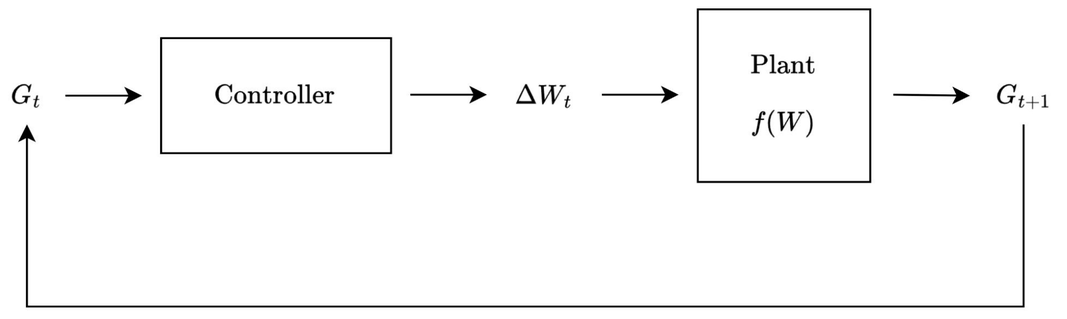
- Q1) Structurally different update rules (Adam's diagonal rescaling, Muon's matrix polar step) show strikingly different transients at comparable asymptotic rates. Why?
- Q2) What scalar, measurable diagnostic captures optimizer quality along the entire trajectory, not only at convergence?
- **Answer.** Treat the optimizer as a discrete-time feedback controller. The trajectory cost is the closed-loop H2 energy, and Muon wins by equalizing per-mode contraction rates.

Contribution

- **Framework.** Optimizer-as-controller correspondence on the Hessian Gauss–Newton quadratic form. Total training energy equals the squared H2 norm of the closed loop.
- **Theory.** Exact decomposition: $\mathcal{E} = \Sigma\, \lambda_i\, \mu_j\, a_{ij}(t)^2$. Under geometric contraction, $\mathcal{E}_{ij} = \lambda_i\, \mu_j\, a_{ij}(0)^2/(1-\rho_{ij}^2)$. Uniform contraction rates strictly minimize total training energy $\mathcal{E}$ among controllers sharing a weighted-mean rate (Jensen's inequality). Muon's polar step attains this optimum under eigenbasis alignment.
- **Validation.** On three MNIST MLPs against SGD and AdamW, Muon attains the lowest Gini of $\{\mathcal{E}_{ij}\}$ on every monitored hidden layer and the lowest cumulative energy on two of three.

Core Methodology

C1. Control-theoretic framework of neural network optimization



$f(W)$: Loss Landscape $G_t = \nabla_W f(W_t)$: Feedback Signal

Neural network training as a discrete-time feedback loop.
The optimizer is the controller, the loss landscape is the plant, the gradient is the feedback signal.

$$\Delta \boldsymbol{W}_t = -\alpha\, \mathcal{C}(\boldsymbol{G}_{0:t}), \quad \boldsymbol{G}_t = \nabla_{\boldsymbol{W}} f(\boldsymbol{W}_t)$$

$$\mathcal{E} = \sum_{t=0}^{T} \|\delta_t\|_2^2 = \|\text{closed loop}\|_{\mathcal{H}_2}^2, \quad \delta_t = \frac{\partial \mathcal{L}_t}{\partial y_t}.$$

$$\mathcal{H} \approx \Sigma_x \otimes \Sigma_\delta, \quad \text{spec}(\mathcal{H}) = \{\lambda_i \mu_j\}, \quad \kappa(\mathcal{H}) = \kappa(\Sigma_x)\, \kappa(\Sigma_\delta).$$

C2. Per-mode decomposition and Spectral Equalization

$$a_{ij}(t) = u_j^\top \Delta W_t v_i \quad \Longrightarrow \quad \mathcal{E} = \sum_t \sum_{i,j} \lambda_i \mu_j\, a_{ij}(t)^2.$$

$$\mathcal{E}_{ij} = \underbrace{\lambda_i\, \mu_j\, a_{ij}(0)^2}_{\text{mode weight } w_{ij}} \cdot \underbrace{\frac{1}{1-\rho_{ij}^2}}_{\text{rate factor}}$$

curvature × init, optimizer-independent — the optimizer's only channel

under $|a_{ij}(t)| = \rho_{ij}^t\, |a_{ij}(0)|$, $0 \le \rho_{ij} < 1$.

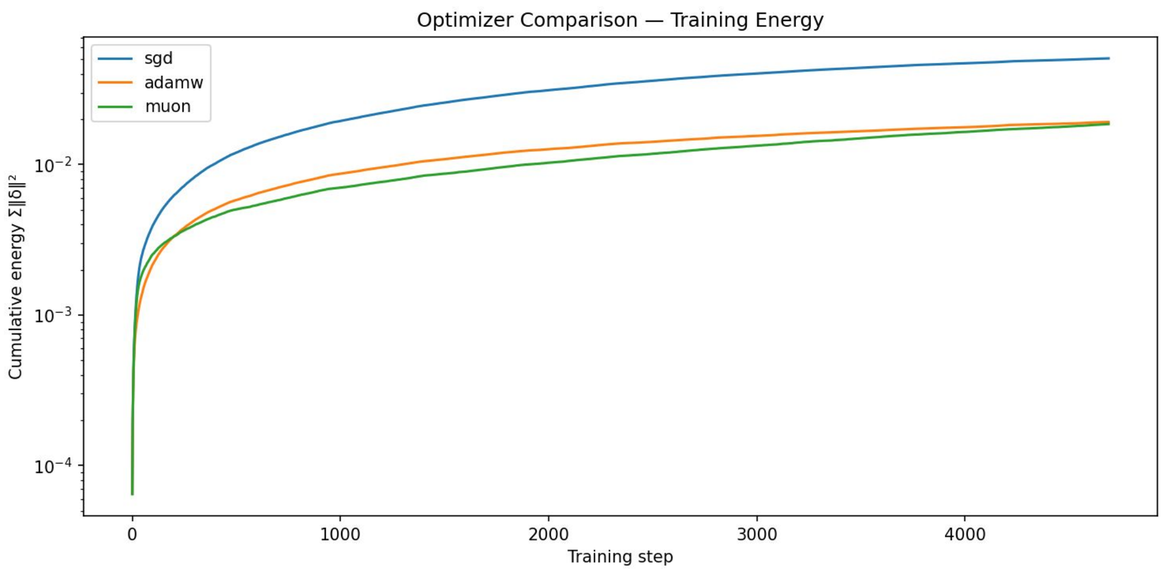
$$\mathcal{E} = \sum_{ij} w_{ij}\, \varphi(\rho_{ij}) \;\ge\; \frac{W}{1-\bar{\rho}^2}, \quad \varphi(\rho) = \frac{1}{1-\rho^2}, \quad \bar{\rho} = \sum_{ij} \frac{w_{ij}}{W}\, \rho_{ij}, \quad W = \sum_{ij} w_{ij}.$$

➢ **Equality holds iff all ρ_{ij} equal.**

$$M_t = \beta M_{t-1} + G_t, \qquad \text{polar}(M_t) = UV^\top \quad \text{from} \quad M_t = U\Sigma V^\top.$$

Empirical Results

Cumulative training energy of MLP-3 model



Optimizer comparison on Muon-eligible hidden layers

		Cumulative energy $\mathcal{E}$			Gini of $\{\mathcal{E}_{ij}\}$		
Model	Layer	SGD	AdamW	Muon	SGD	AdamW	Muon
MLP-2	`fc1`	0.0163	0.0051	**0.0011**	0.9986	0.9954	**0.9528**
MLP-3	`fc1`	0.0210	**0.0064**	0.0097	0.9985	0.9952	**0.9655**
MLP-3	`fc2`	0.0171	0.0040	**0.0014**	0.9974	0.9905	**0.8184**

Key Takeaways

Spectral equalization of per-mode contraction rates, not orthogonality per se, is the measurable mechanism behind Muon's trajectory advantage.

- **Diagnostic.** Total training energy exactly equals the squared H2 norm of the closed loop on Kronecker quadratics.
- **Mechanism.** Per-mode energy $\mathcal{E}_{ij}$ is strictly convex in ρ_{ij}. Thus, at a common mean rate, uniform rates strictly minimize total energy. Muon's polar step implements this to the extent its singular directions align with the Kronecker eigenbasis.
- **Evidence.** Lowest Gini of $\{\mathcal{E}_{ij}\}$ on all monitored hidden matrices, lowest cumulative energy on two of three. The single exception (MLP-3 fc1) is permitted by the bound, since cumulative energy is rate-confounded and orders optimizers only at a common $\bar{\rho}$.
- **Prediction.** The cost of non-equalization scales with the product $\kappa(\Sigma_x)\kappa(\Sigma_\delta)$, not the average, so optimizer choice should enter scaling laws multiplicatively in the Kronecker condition numbers. Validation remains as future work.

Paper Link: