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

Euijin Hong EHONG@ANDREW.CMU.EDU
Carnegie Mellon University

Abstract

We propose the *total training energy*, the integral squared error of the back-propagated signal, as a trajectory-level diagnostic for deep learning optimizers, viewing every first-order method as a discrete-time feedback controller closing the loop around the loss landscape. On quadratic losses with a Kronecker-factored Gauss–Newton Hessian, the total training energy equals the squared $\mathcal{H}_2$ norm of the closed loop and decomposes *exactly* into per-mode energy contributions. Jensen's inequality implies that uniform per-mode rates strictly minimize the total training energy among controllers with a common weighted-mean contraction rate, which Muon's polar step implements to the extent its update directions align with the Kronecker eigenbasis. On MNIST with one-, two-, and three-layer MLPs trained with SGD, AdamW, and Muon, Muon attains the smallest Gini coefficient of the per-mode energy distribution on *every* monitored hidden matrix and the smallest cumulative energy on two of three, with the advantage concentrated in the low-curvature tail of the spectrum. The framework recasts "orthogonalizing momentum helps" as a measurable, mechanism-level claim about the geometry of the Hessian spectrum.

1. Introduction

Deep learning optimizers are usually analyzed through their *terminal* loss $L(\boldsymbol{W}_T)$, producing convergence-rate bounds under convexity or Polyak–Łojasiewicz assumptions [5, 8]. This perspective leaves two practically important questions underspecified. First, why do structurally different update rules — Adam's diagonal rescaling [3] versus Muon's matrix polar decomposition [2] — produce strikingly different *transient dynamics* even when their asymptotic rates are comparable? Second, what scalar, measurable diagnostic captures optimizer quality along the entire trajectory rather than only at convergence?

We approach both questions through discrete-time linear control. Every first-order optimizer can be written as a feedback controller acting on a plant whose output is the gradient; the closed loop is a dynamical system whose trajectory is the training history (Figure 1). In this view the natural trajectory cost is the $\mathcal{H}_2$ norm of the closed-loop transfer function, equivalently the integral squared error (ISE) of the back-propagated signal $\boldsymbol{\delta}_t = \partial L_t / \partial \boldsymbol{y}_t$. We call this the *total training energy* $\mathcal{E} \triangleq \sum_{t=0}^{T} \|\boldsymbol{\delta}_t\|_2^2$.

$\mathcal{E}$ is *exact and free* to measure ($\|\boldsymbol{\delta}_t\|^2$ is already produced by back-propagation), admits a closed-form decomposition over Hessian eigenmodes whose per-mode contribution $\mathcal{E}_{ij} \propto 1/(1-\rho_{ij}^2)$ is strictly convex in the per-mode contraction rate ρ_{ij}. By Jensen's inequality, an optimizer that

. An LLM was used as a drafting and editing aid; all theoretical claims, derivations, and experiments are the authors' own and were human-verified.

© E. Hong.

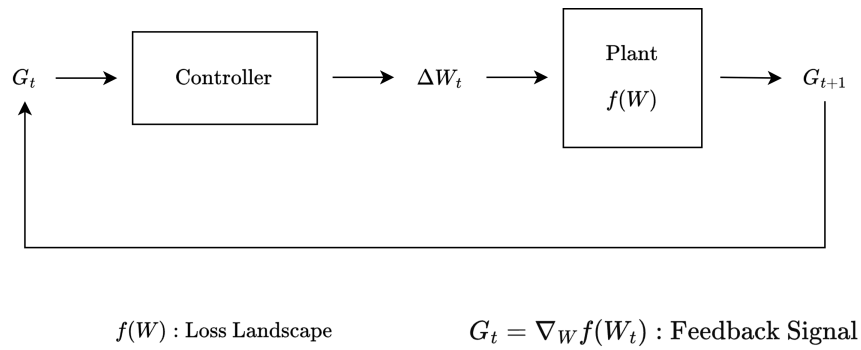

Figure 1: Neural-network training as a discrete-time feedback loop. The optimizer (CONTROLLER) consumes the gradient $\boldsymbol{G}_t = \nabla_{\boldsymbol{W}} f(\boldsymbol{W}_t)$ and emits a weight update $\Delta \boldsymbol{W}_t$ that drives the loss landscape (PLANT $f(\boldsymbol{W})$). The total training energy $\mathcal{E} = \sum_t \|\boldsymbol{\delta}_t\|^2$ is the $\mathcal{H}_2$ norm (ISE) of the closed loop and serves as a trajectory-level diagnostic.

equalizes ρ_{ij} across modes strictly dominates any optimizer with the same mean rate but uneven spread. Muon's polar factorization $\boldsymbol{M}_t = U\Sigma V^\top \mapsto UV^\top$ is exactly such an equalizer — the control-theoretic origin of its empirical advantage over diagonal preconditioners.

Contributions. (1) We formalize the optimizer-as-controller correspondence on the matrix-Gauss–Newton quadratic and show that total energy is the squared $\mathcal{H}_2$ norm of the closed loop (§2). (2) We derive the per-mode decomposition $\mathcal{E} = \sum_{ij} \lambda_i \mu_j \, a_{ij}(0)^2/(1-\rho_{ij}^2)$ and prove via Jensen's inequality that uniform contraction rates minimize $\mathcal{E}$, giving a control-theoretic justification of Muon's polar step (§3). (3) We validate the predicted orderings on MNIST with three MLP architectures and SGD, AdamW, and the official Muon optimizer, finding that Muon attains the lowest Gini coefficient of $\{\mathcal{E}_{ij}\}$ on every monitored hidden matrix and the lowest cumulative energy on two of three (§4). The Jensen lower bound (5) predicts that the cost of non-equalization scales with the product $\kappa(\Sigma_x)\kappa(\Sigma_\delta)$ rather than its average, suggesting that optimizer choice should enter neural scaling laws multiplicatively in the Kronecker factor condition numbers.

2. Optimizer-as-Controller Framework

Feedback loop. The first-order update $\Delta \boldsymbol{W}_t = -\alpha\, \mathcal{C}(\boldsymbol{G}_{0:t})$ is a feedback law in which the controller $\mathcal{C}$ (the optimizer, with internal state given by momentum, second-moment estimate, or preconditioner) observes the plant output $\boldsymbol{G}_t = \nabla_{\boldsymbol{W}} f(\boldsymbol{W}_t)$ and produces the control input. For an LTI closed loop with linear layer $\boldsymbol{y}_t = \boldsymbol{W}_t \boldsymbol{x}_t$ and squared-loss back-propagated error $\boldsymbol{\delta}_t$, the *total training energy*

$$\mathcal{E} \triangleq \sum_{t=0}^{T} \|\boldsymbol{\delta}_t\|_2^2 \tag{1}$$

equals the squared $\mathcal{H}_2$ norm of the transfer function from weight error to output error [9]; lower $\mathcal{E}$ corresponds to a controller that extinguishes error faster and with less residual accumulation along the entire trajectory — a strictly stronger criterion than terminal-loss convergence.

Kronecker structure of the Hessian. For a layer with squared loss $f(\boldsymbol{W}) = \frac{1}{2}\|\boldsymbol{AWB} - \boldsymbol{C}\|_F^2$, the Hessian of the vectorized parameter satisfies

$$\mathcal{H} = \nabla^2_{\mathrm{vec}\,\boldsymbol{W}} f \;=\; (\boldsymbol{B}\boldsymbol{B}^\top) \otimes (\boldsymbol{A}^\top \boldsymbol{A}), \tag{2}$$

with eigenvalues that are pairwise products of the two factors' eigenvalues and eigenvectors of Kronecker form. For a general layer $\boldsymbol{y} = \boldsymbol{W}\boldsymbol{x}$, the Gauss–Newton Hessian has the analogous form $\mathcal{H} \approx \Sigma_x \otimes \Sigma_\delta$ with $\Sigma_x = \mathbb{E}[\boldsymbol{x}\boldsymbol{x}^\top]$, $\Sigma_\delta = \mathbb{E}[\boldsymbol{\delta}\boldsymbol{\delta}^\top]$ [7]. We fix one convention for the entire paper. Write the eigendecompositions of the two Kronecker factors as $\Sigma_x = \sum_i \lambda_i\, v_i v_i^\top$ (input factor, eigenpairs (λ_i, v_i) indexed by i) and $\Sigma_\delta = \sum_j \mu_j\, u_j u_j^\top$ (back-propagated-error factor, eigenpairs (μ_j, u_j) indexed by j), so that $\mathcal{H} = \Sigma_x \otimes \Sigma_\delta$ has eigenvalues $\{\lambda_i \mu_j\}$ and eigenvectors $\{v_i \otimes u_j\}$. This factorization is the mathematical hinge for the per-mode analysis, since the condition number factorizes as $\kappa(\mathcal{H}) = \kappa(\Sigma_x)\kappa(\Sigma_\delta)$, so spectral anisotropy compounds multiplicatively along the two factors.

3. Per-Mode Decomposition and Spectral Equalization

Modal energy and the $1/(1-\rho^2)$ law. Substituting $\boldsymbol{\delta}_t = \Delta\boldsymbol{W}_t\, \boldsymbol{x}_t$ (valid locally for squared loss near $\boldsymbol{W}^\star$) and taking expectation over $\boldsymbol{x}_t$, $\mathbb{E}_x[\|\boldsymbol{\delta}_t\|^2] = \mathrm{tr}(\Delta\boldsymbol{W}_t^\top \Delta\boldsymbol{W}_t \Sigma_x)$. Projecting the weight error onto the Kronecker eigenbasis fixed in §2 through the *mode amplitudes* $a_{ij}(t) \triangleq u_j^\top \Delta\boldsymbol{W}_t\, v_i$ (the component of $\Delta\boldsymbol{W}_t$ along the eigenvector $v_i \otimes u_j$, so that i indexes the Σ_x factor and j the Σ_δ factor), the total energy decomposes exactly as

$$\mathcal{E} \;=\; \sum_{t=0}^{T} \sum_{i,j} \lambda_i \mu_j\, a_{ij}(t)^2. \tag{3}$$

Under a geometric contraction model $|a_{ij}(t)| = \rho_{ij}^t |a_{ij}(0)|$ with per-mode rate $\rho_{ij} \in [0,1)$, summing the geometric series gives the central object of the framework,

$$\boxed{\mathcal{E}_{ij} \;=\; \frac{\lambda_i \mu_j\, a_{ij}(0)^2}{1-\rho_{ij}^2}.} \tag{4}$$

Total energy is dominated not by the average mode but by the *slowest* ones, because $1/(1-\rho^2) \to \infty$ as $\rho \to 1$. Ill-conditioned spectra pay a disproportionate price along their high-ρ tail.

Jensen's inequality and spectral equalization. Let $w_{ij} = \lambda_i \mu_j\, a_{ij}(0)^2 \geq 0$ be mode weights with total $W = \sum_{ij} w_{ij}$, and let $\bar{\rho} = \sum_{ij}(w_{ij}/W)\rho_{ij}$ be the weighted-mean contraction rate. The function $\phi(\rho) = 1/(1-\rho^2)$ is strictly convex on $[0,1)$, so Jensen's inequality yields

$$\mathcal{E} \;=\; \sum_{ij} w_{ij}\, \phi(\rho_{ij}) \;\geq\; \frac{W}{1-\bar{\rho}^2}, \tag{5}$$

with equality *iff* all ρ_{ij} are equal. Among controllers with a common weighted-mean contraction rate, the optimizer that equalizes the per-mode rates strictly minimizes total energy. The non-negative Jensen gap $\Delta_{\text{Jensen}} \triangleq \mathcal{E} - W/(1-\bar{\rho}^2)$ is an optimizer-quality defect that vanishes only for an ideal spectral equalizer.

Proposition 1 (Muon is a spectral equalizer under eigenbasis alignment) *Let $\boldsymbol{M}_t = \beta \boldsymbol{M}_{t-1} + \boldsymbol{G}_t$ accumulate momentum and let Muon apply the polar map $\mathrm{polar}(\boldsymbol{M}_t) = UV^\top$ from $\boldsymbol{M}_t = U\Sigma V^\top$ before the step [2], so that the resulting update has all singular values equal to one and the per-step progress is identical along every left/right singular direction of $\boldsymbol{M}_t$. Assume the* ***eigenbasis-alignment*** *condition, that near a local quadratic minimum with Kronecker Hessian (2) the dominant left and right singular directions of $\boldsymbol{M}_t$ coincide with the Kronecker eigenbasis $\{v_i \otimes u_j\}$ of $\mathcal{H}$. Then the equalization of singular values transfers to equalization of the per-mode rates ρ_{ij}, and by (5) Muon satisfies the optimality condition for trajectory energy among uniform-rate controllers.*

Remark 2 (Plausibility of the alignment assumption) *The eigenbasis-alignment condition is the crux of the argument and is not automatic. The buffer $\boldsymbol{M}_t$ accumulates raw gradients, and its singular vectors need not coincide with the eigenvectors of $\mathcal{H}$ in general. The condition is most plausible near convergence, where the gradient signal is dominated by curvature directions and the momentum buffer inherits the eigenstructure of $\mathcal{H}$; it degrades when $\boldsymbol{M}_t$ is dominated by transient non-curvature gradient components, in which case singular-value equalization need not equalize $\{\rho_{ij}\}$. The equalization claim should therefore be read as holding to the extent the alignment holds.*

In contrast, SGD-with-momentum inherits the anisotropy of $\mathcal{H}$ multiplicatively, with $\rho_{ij} \approx 1 - \alpha\lambda_i\mu_j$, producing a wide spread. AdamW's diagonal second-moment rescaling partially compensates along the coordinate axes but cannot correct the inter-parameter anisotropy captured by the off-diagonal Kronecker structure of $\mathcal{H}$, so a residual non-zero Jensen gap remains. Recent analyses of Adam and Shampoo via spectral steepest descent [1] are consistent with this view, since spectral-norm-constrained steps are an alternative description of the same equalization mechanism.

Why $\{\mathcal{E}_{ij}\}$ stays nonuniform under rate equalization. The per-mode energy (4) factors as $\mathcal{E}_{ij} = w_{ij} \cdot 1/(1-\rho_{ij}^2)$ into a *mode-weight* part $w_{ij} = \lambda_i\mu_j\, a_{ij}(0)^2$ and a *rate* part $1/(1-\rho_{ij}^2)$. Muon acts only on the rate part; the weights inherit the anisotropic Kronecker spectrum of $\mathcal{H}$, which no first-order optimizer alters. Consequently $\{\mathcal{E}_{ij}\}$ remains highly nonuniform even under perfect rate equalization, so a Gini coefficient of $\{\mathcal{E}_{ij}\}$ close to one is *expected* for every optimizer, Muon included, rather than a sign that equalization failed. The quantity that isolates the mechanism is the dispersion of the rates $\{\rho_{ij}\}$ themselves, or equivalently the Jensen gap $\Delta_{\text{Jensen}} = \mathcal{E} - W/(1-\bar{\rho}^2)$; the Gini of $\{\mathcal{E}_{ij}\}$ convolves this rate dispersion with the fixed weight anisotropy, which is why its absolute value stays near one. That Muon nonetheless attains the smallest Gini of $\{\mathcal{E}_{ij}\}$ on every monitored layer is consistent with its acting on the only channel available to it, the rates, with the residual set by the weight anisotropy that no optimizer removes.

4. Experimental Validation

Setup. We instantiate three MNIST MLPs [4] of increasing depth, spanning zero to two Muon-eligible hidden weight matrices: MLP-1 (784→10); MLP-2 (784→256→10; hidden matrix `fc1`); MLP-3 (784→256→128→10; hidden matrices `fc1`, `fc2`). All layers use ReLU and no batch normalization. We compare **SGD** (Nesterov, β=0.9), **AdamW** (β_1=0.9, β_2=0.999, weight decay 0.01; Loshchilov and Hutter 6), and the official **Muon** implementation [2] (Newton–Schulz coefficients $(3.4445, -4.7750, 2.0315)$, five iterations in bfloat16, Nesterov momentum), routing hidden weight matrices to Muon and biases / output layer to the auxiliary AdamW as prescribed. Learning

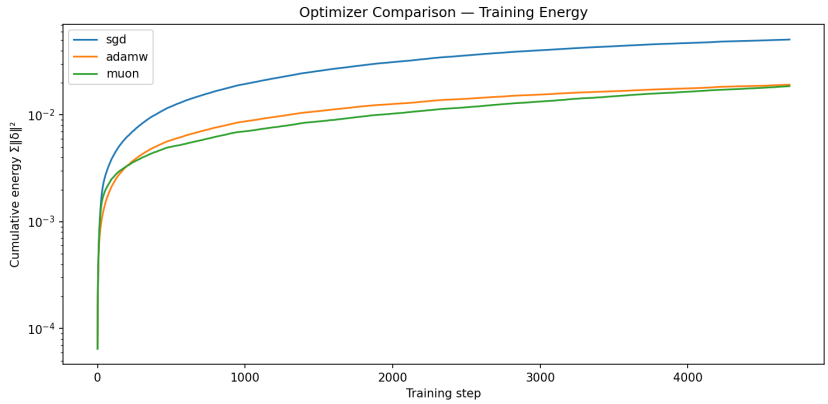

Figure 2: Cumulative training energy $\sum_{s\le t}\|\boldsymbol{\delta}_s\|^2$ on MLP-3 (log scale). The ordering $\mathcal{E}_{\text{SGD}}>\mathcal{E}_{\text{AdamW}}>\mathcal{E}_{\text{Muon}}$ emerges in the early transient and is preserved throughout training, with Muon's advantage over AdamW concentrated in the steady-state plateau.

Table 1: Optimizer comparison on Muon-eligible hidden layers (lower is better in both metrics). **Bold** marks the minimum per row and per block. The Gini block reports the concentration of $\{\mathcal{E}_{ij}\}$, whose near-unit values for all optimizers are explained in §3.

		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**

rates were tuned once on MLP-2 and reused. We train ten epochs at batch size 128 under cross-entropy loss, recording per-step $\|\boldsymbol{\delta}_t\|^2$ via forward / backward hooks; a float64 exponential moving estimator maintains $\widehat{\Sigma_x}, \widehat{\Sigma_\delta}$ and projects the weight error onto the Kronecker eigenbasis once per epoch to compute $\{\mathcal{E}_{ij}\}$ and its Gini coefficient.

Total energy and Gini. Table 1 reports both metrics; Figure 2 plots the cumulative-energy trajectory on MLP-3. The predicted ordering $\mathcal{E}_{\text{Muon}}<\mathcal{E}_{\text{AdamW}}<\mathcal{E}_{\text{SGD}}$ holds on two of three monitored layers; on MLP-3 `fc1`, AdamW edges Muon by a small margin, which the Jensen bound (5) permits whenever AdamW reaches a smaller weighted-mean rate $\bar{\rho}$, since cumulative energy orders optimizers only at a *common* $\bar{\rho}$ and is therefore rate-confounded. Across all three layers Muon attains the smallest Gini of $\{\mathcal{E}_{ij}\}$, consistent with Proposition 1; as §3 explains, these values stay near one because the mode-weights carry the Hessian's fixed anisotropy, so the informative signal is the ordering across optimizers rather than the distance from zero. AdamW reduces Gini relative to SGD via its diagonal preconditioner but cannot flatten the inter-parameter anisotropy in the off-diagonal Kronecker structure, leaving a residual gap.

Task-level metrics. For MLP-1 (no Muon-eligible matrix, so Muon collapses to AdamW), the three optimizers are statistically indistinguishable on test accuracy (92.39–92.52%), validating that Muon's gains come from the matrix-polar step rather than the auxiliary AdamW. For MLP-2, Muon achieves the lowest test loss and highest accuracy ($0.0011/98.13\%$ vs. $0.0307/98.05\%$ for SGD and $0.0154/98.02\%$ for AdamW). For MLP-3, accuracy ordering reverses (97.46% vs. 97.83–97.88%) although the energy ordering on Muon-eligible layers is preserved, reflecting that trajectory energy is a *training-side* criterion that need not predict generalization.

5. Discussion

What the framework explains, and what it does not. A single mechanism, spectral equalization of per-mode contraction rates, simultaneously explains Muon's smaller Gini of $\{\mathcal{E}_{ij}\}$ and, on most layers, its lower cumulative energy under anisotropic Kronecker curvature; it also predicts that Muon's advantage shrinks as the Hessian becomes more diagonal, matching our MLP-1 result. The decomposition (3) assumes a quadratic Kronecker loss; on cross-entropy with ReLU it is a local approximation, and stochastic gradient noise contributes a time-constant floor to $\|\boldsymbol{\delta}_t\|^2$. The framework is silent about generalization.

High-dimensional scaling implications. The per-mode law (4) predicts that the Hessian condition number $\kappa(\mathcal{H}) = \kappa(\Sigma_x)\kappa(\Sigma_\delta)$ *multiplies* along the two Kronecker factors, so non-equalizing optimizers pay a cost that grows multiplicatively in the per-factor condition numbers. This yields a falsifiable scaling prediction, that the Jensen gap should scale with the *product* $\kappa(\Sigma_x)\kappa(\Sigma_\delta)$ rather than with the average $\frac{1}{2}(\kappa(\Sigma_x) + \kappa(\Sigma_\delta))$. We emphasize that this prediction is *untested* by the present work. Our experiments are small MNIST MLPs, the opposite regime from the large-scale setting in which a Muon-versus-Adam scaling gap would be measured, so the prediction is offered as future work and not as a claim supported by these results. A natural route to test it is the $\mathcal{H}_2$-optimal controller obtained from the discrete algebraic Riccati equation, which replaces Muon's uniform unit gain with curvature-adaptive per-mode gains. This does not contradict the Jensen result. The Jensen bound (5) establishes that uniform rates are optimal *at a fixed* weighted-mean rate $\bar{\rho}$, whereas the Riccati controller solves a different problem, spending control effort to lower $\bar{\rho}$ itself, so its non-uniform gains pursue a smaller mean rate rather than a more uniform spread.

References

[1] Jeremy Bernstein and Laker Newhouse. Old optimizer, new norm: An anthology. *arXiv preprint arXiv:2409.20325*, 2024.

[2] Keller Jordan. Muon: An optimizer for hidden layers in neural networks. Blog post, 2024. URL https://kellerjordan.github.io/posts/muon/.

[3] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[4] Yann LeCun. The mnist database of handwritten digits. *http://yann. lecun. com/exdb/mnist/*, 1998.

[5] Stanislas Łojasiewicz. Sur la géométrie semi-et sous-analytique. In *Annales de l'institut Fourier*, volume 43, pages 1575–1595, 1993.

[6] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. *arXiv preprint arXiv:1711.05101*, 2017.

[7] James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pages 2408–2417. PMLR, 2015.

[8] Boris T Polyak. Gradient methods for the minimisation of functionals. *USSR Computational Mathematics and Mathematical Physics*, 3(4):864–878, 1963.

[9] Kemin Zhou, John Comstock Doyle, Keith Glover, et al. *Robust and optimal control*, volume 40. Prentice hall New Jersey, 1996.