Muon is Gauge Fixing

deriving the "Muon" orthogonalization

in terms of the renormalization group

charles@calculatedcontent.com

What is Muon ?: Momentum Update Orthogonalization

- For each layer, form the momentum the standard W update

$$\mathbf{M}_t = \beta \mathbf{M}_{t-1} - \eta \nabla_{\mathbf{W}} L$$

$$\Delta \mathbf{W}_t = \mathbf{M}_t$$

- Orthogonalized, all eigenvalues are one

$$\Delta \mathbf{W}_t \longrightarrow \widetilde{\Delta \mathbf{W}}_t = \mathrm{Orth}(\Delta \mathbf{W}_t)$$

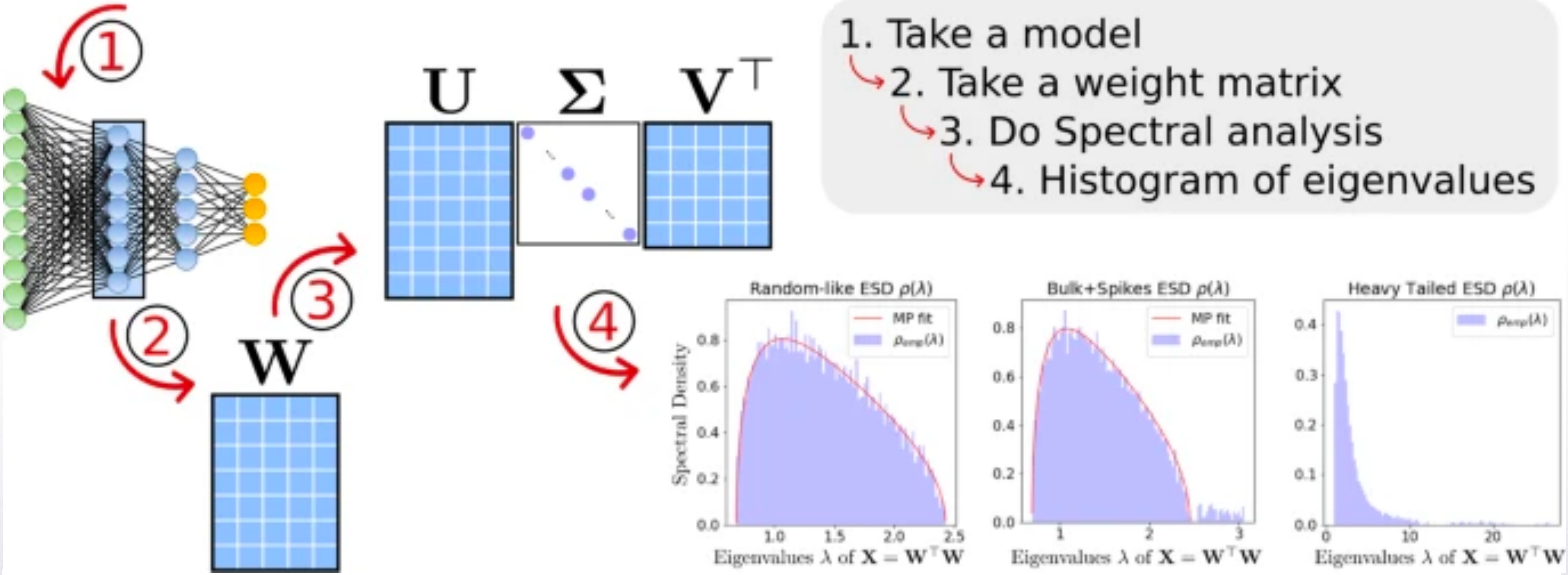
$$\widetilde{\Delta \mathbf{W}}_t^\top \widetilde{\Delta \mathbf{W}}_t = \mathbf{I}$$

$$\boxed{\lambda_i = 1}$$

w|w weight|watcher.ai HOME USAGE RESEARCH PRESENTATIONS CODE EXAMPLES COMMUNITY ABOUT

Analyzing DNN Weight matrices with WeightWatcher

1. Take a model
2. Take a weight matrix
3. Do Spectral analysis
4. Histogram of eigenvalues



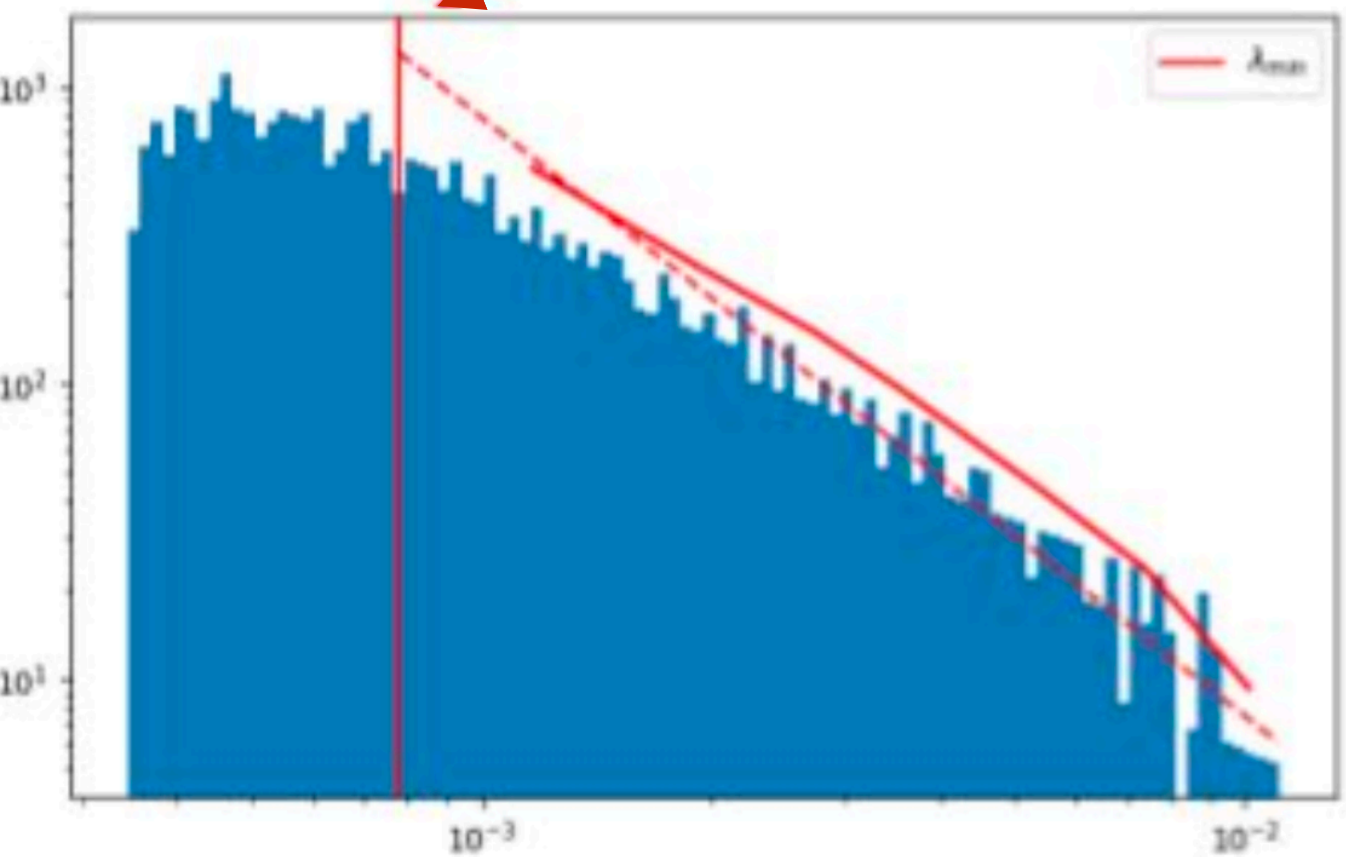
$$\rho_{emp}(\lambda) \sim \lambda^{-\alpha}.$$

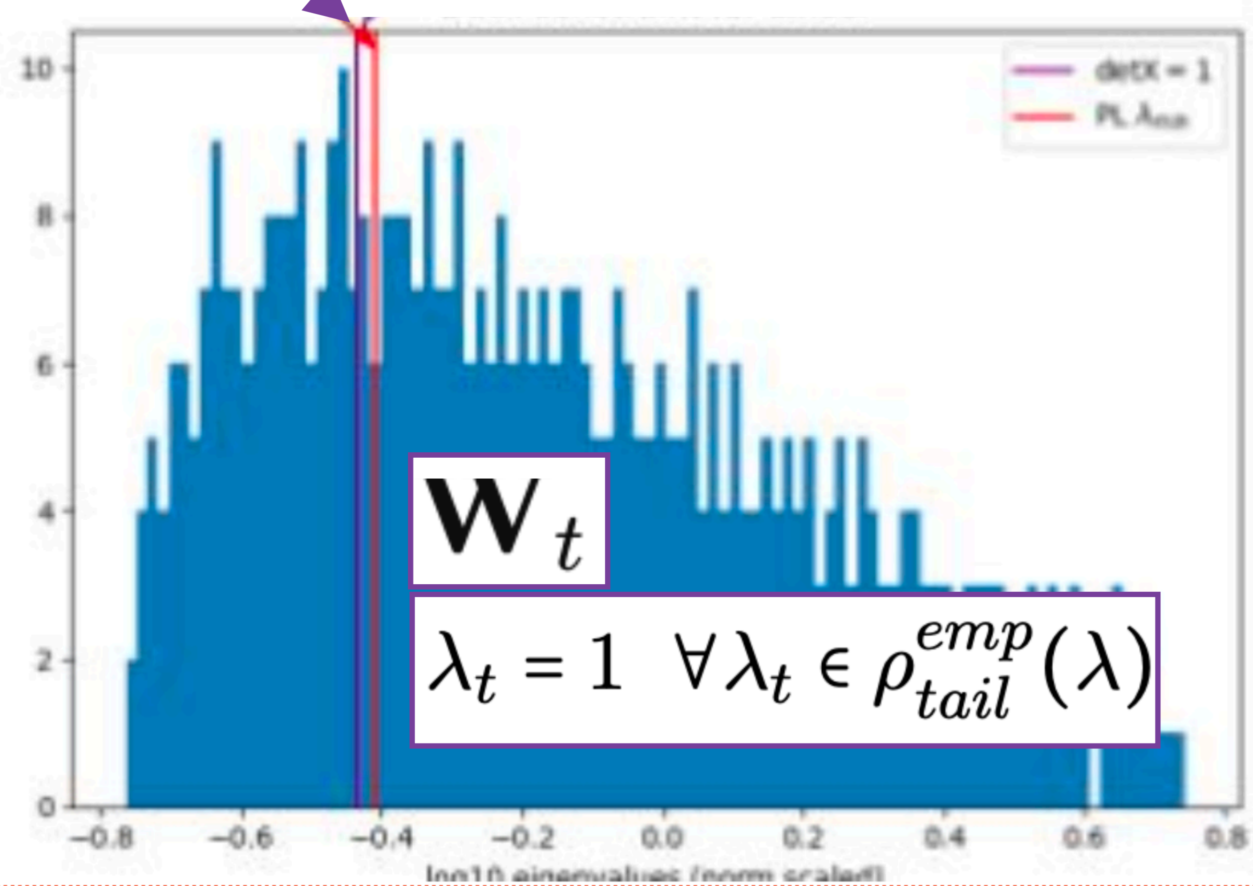
Theory-based: Layer Quality metrics

HTSR (2021) $\rho_{tail}(\lambda) \sim \lambda^{-2}$ $\leftrightarrow$ $\sum_{\lambda_i \in tail} \log \lambda_i = 0$ SETOL (2025)

Start of Power Law Tail (alpha=2.0)

Start of tail giving Volume Preserving Transformation





How is Muon related: to Renormalization Group ?

non-Trivial: the RG fixed point

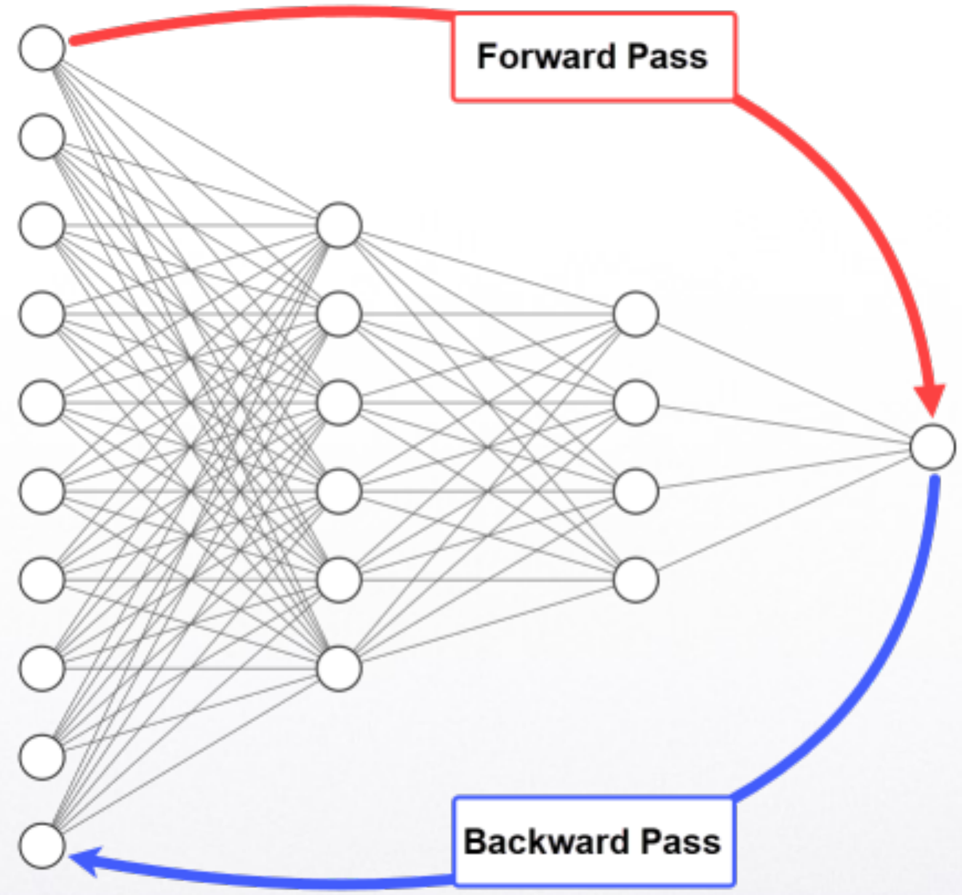
$$\lambda_t = 1 \;\; \forall \lambda_t \in \rho^{emp}_{tail}(\lambda)$$

Trivial: fixes the gauge of the RG flow

$$\lambda_t = 1 \;\; \forall \lambda_t \in \rho^{emp}(\lambda)$$

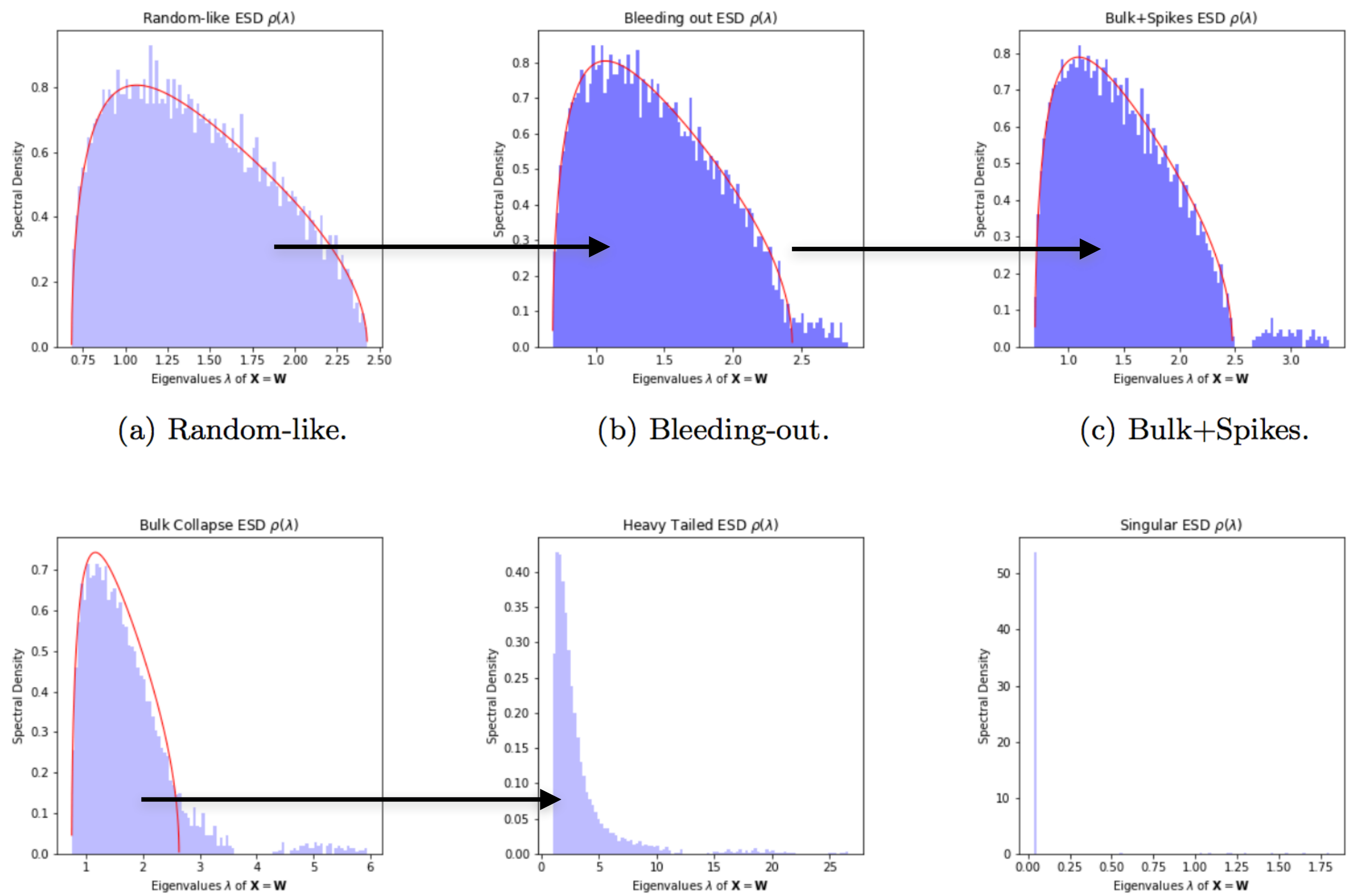
What does this mean ?

Why w|w works ? Fix Point Equation



W → W → W → W → w

HTSR Theory: 5+1 Phases of Training



(a) Random-like. (b) Bleeding-out. (c) Bulk+Spikes.

(d) Bulk-decay. (e) Heavy-tailed. (f) Singularity.

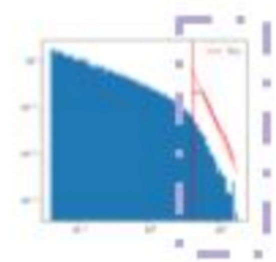
Implicit Self-Regularization in Deep Neural Networks: Evidence from Random Matrix Theory and Implications for Learning

Charles H. Martin, Michael W. Mahoney; JMLR 22(165):1–73, 2021.

Why w|w works ? Renormalization Group (RG) theory

W

X



$$\ln \int d\mu(\mathbf{S}) e^{N\beta \operatorname{Tr}[\mathbf{H}_{\bar{\mathcal{Q}}^2}]} \xrightarrow{RG} \lim_{N\gg 1} \ln \int d\mu(\tilde{\mathbf{A}}) e^{N\beta \operatorname{Tr}[\mathbf{H}^{ECS}_{\bar{\mathcal{Q}}^2}]}$$

Volume Preserving or Scale-Invariant Transformation

$$\det\left(\tilde{\mathbf{A}}\right) = 1 \quad \text{or} \quad \operatorname{Tr}\left[\ln \tilde{\mathbf{A}}\right] = 0,$$

$$\prod_t \lambda_t = 1 \;\; \forall \lambda_t \in \rho^{emp}_{tail}(\lambda),$$

SETOL: SemiEmpirical Theory of (Deep) Learning

Charles H. Martin & Christopher Hinrichs (arxiv, 2025).

ERG Condition: non-Trivial & Trivial solutions

non-Trivial: the RG fixed point

$$\lambda_t = 1 \;\; \forall \lambda_t \in \rho^{emp}_{tail}(\lambda)$$

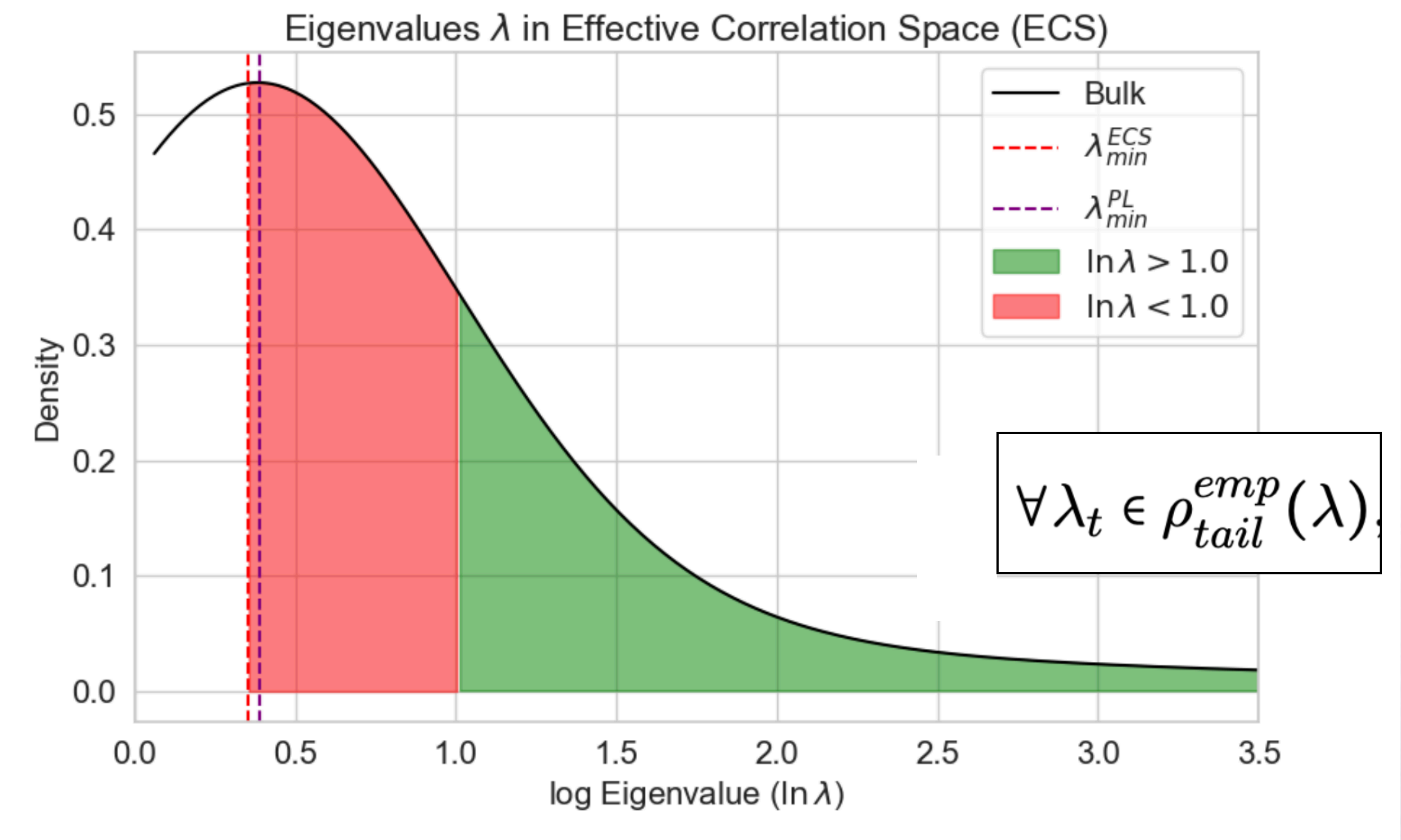
Trivial: fixes the gauge of the RG flow

$$\lambda_t = 1 \;\; \forall \lambda_t \in \rho^{emp}(\lambda)$$

(TM)

New Principle of Ideal Learning:

Wilson Exact Renormalization Group (ERG) Condition



Non-Trivial Solution to ERG Trace-Log Condition $\alpha = 2$

ALPHA

Muon: Removes Redundant RG Directions

Unrestricted flow allows dilations $\mathbf{W}_{t+1} = \mathbf{W}_t + \Delta\mathbf{W}_t$

$$\mathbf{W} \to \mathbf{W}e^{\mathbf{S}} \qquad \mathrm{GL}(n) = O(n) \times \mathrm{SPD}(n)$$

The update lives in the tangent space of the flow

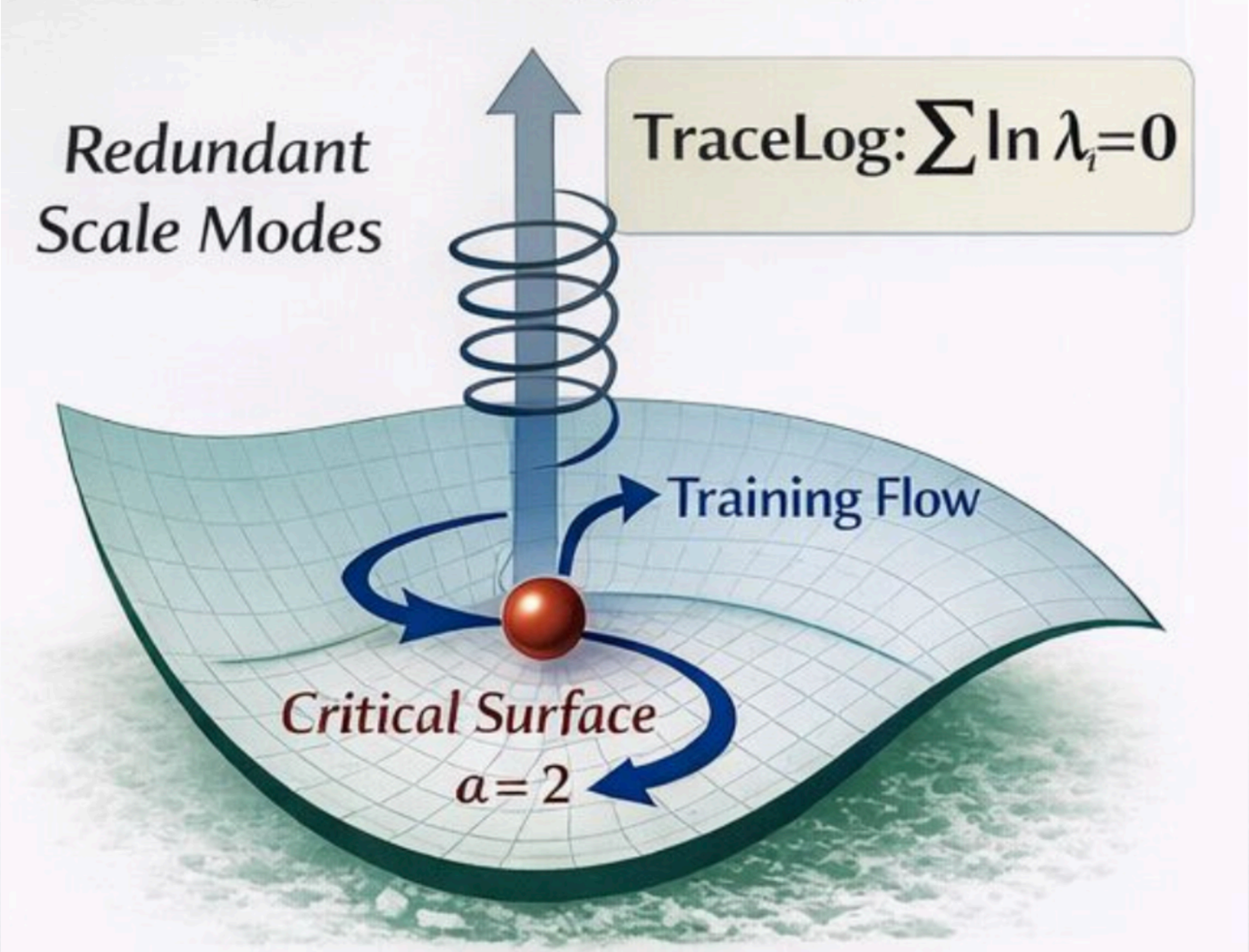
$$\Delta\mathbf{W}_t \approx \eta\dot{\mathbf{W}} \in T_{\mathbf{W}_t}\mathrm{GL}(n)$$

$$\Delta\mathbf{W}_t \longrightarrow \Delta\mathbf{W}_t + (\text{redundant direction})$$

Muon removes the dilation subgroup in the RG flow

$$\mathrm{GL}(n) \longrightarrow O(n)$$

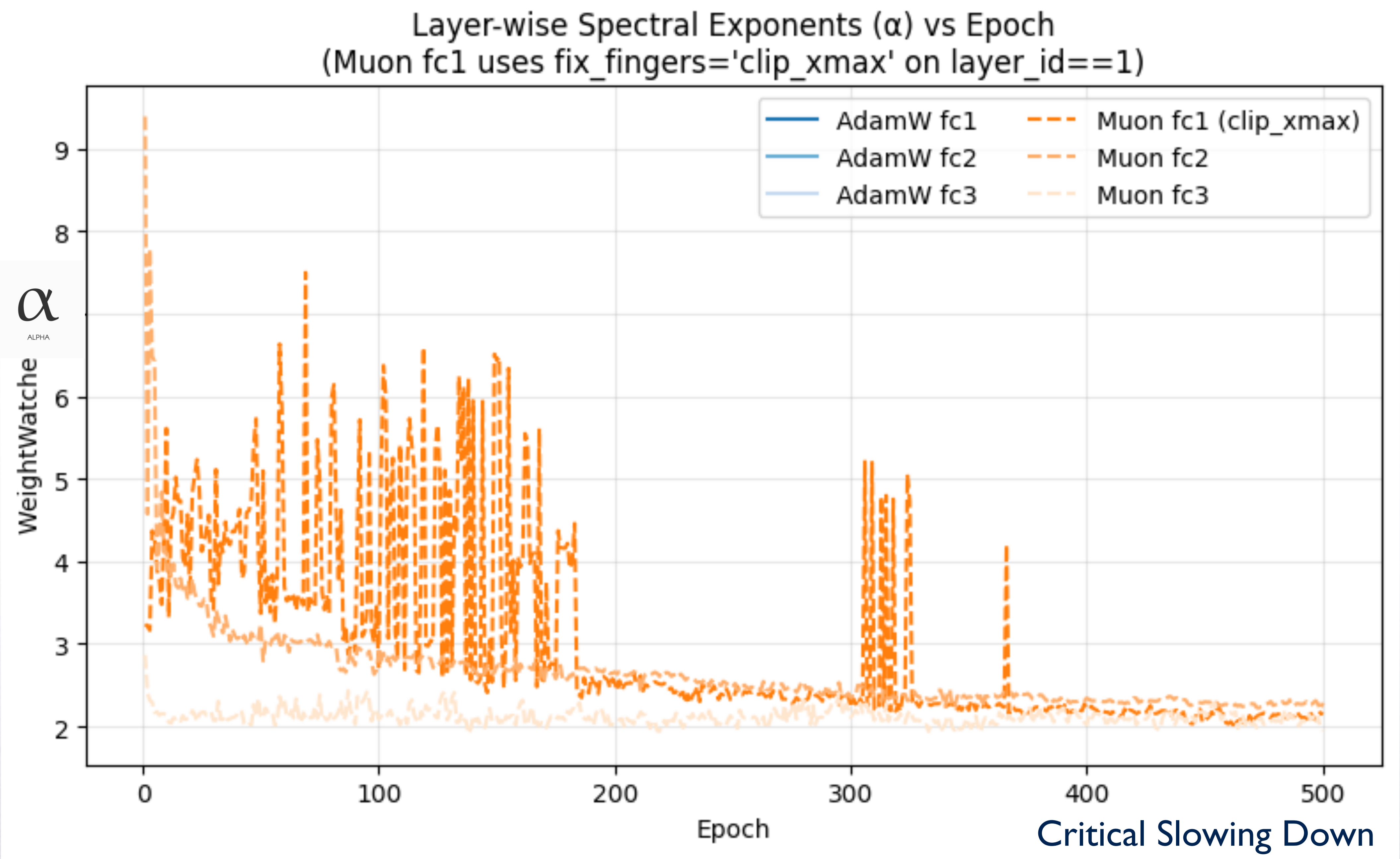
Muon: Removes Redundant RG Directions



$$\mathbf{W} \rightarrow \mathbf{W}e^{\mathbf{S}}$$

$$\Delta\mathbf{W}_t \approx \eta\dot{\mathbf{W}} \in T_{\mathbf{W}_t}\mathrm{GL}(n)$$

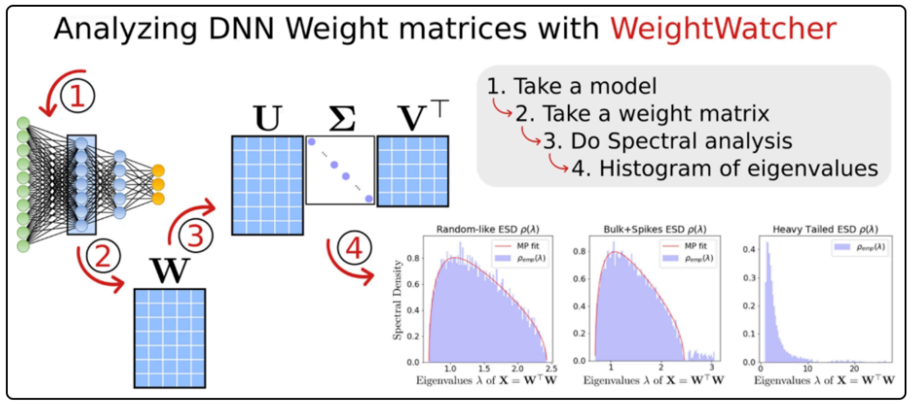
Muon: Removes Redundant Directions



(TM)

w|w weight|watcher.ai

HOME USAGE RESEARCH PRESENTATIONS CODE EXAMPLES COMMUNITY ABOUT



weight | watcher

Data-Free Diagnostics for Deep Learning

WeightWatcher (w|w) is an open-source, diagnostic tool for analyzing Deep Neural Networks (DNN), without needing access to training or even test data. It is based on theoretical research into Why Deep Learning Works, using the new Theory of Heavy-Tailed Self-Regularization (HT-SR), published in JMLR and Nature Communications.

WeightWatcher is a one-of-a-kind must-have tool for anyone training, deploying, or monitoring Deep Neural Networks (DNNs).

pip install weightwatcher

and check out our latest LLM Leaderboard!

https://weightwatcher.ai

175K+ downloads; 17K+ stars

downloads 104k | release v0.7.1.5 | license Apache-2.0 | Published in Nature | Video Tutorial | discord 14 online | LinkedIn | Blog

We are looking for early adopters and collaborators

charles@calculatedcontent.com