A Spectral Renormalization-Group View of Learning

Charles H. Martin, PhD
Calculation Consulting, San Francisco, CA
charles@calculationconsulting.com

Jun 26, 2026

Abstract

We study how the layers in deep neural networks (NNs) converge, using Wilsonian Renormalization-Group (RG) ideas to organize the spectral picture. During training, these layers often develop heavy-tailed power-law (PL) spectra, and the strongest checkpoints often appear near the PL exponent $\alpha \simeq 2$. This pattern raises a simple question: why should this value mark a good trained layer? The answer is both statistical and physical. The statistical answer is that $\alpha = 2$ marks the boundary between stable if not optimal learning and overfitting via non-self-averaging. The physical answer is that the same value makes the spectral energy gain scale evenly across logarithmic spectral scales, which is the proposed Wilsonian marginal condition. We then apply the ERG trace-log condition, the proposed SETOL theory, which we justify by taking volume-preserving, scale-invariant transformations to remove arbitrary layer scale along the proposed RG flow. The resulting spectral RG flow connects the HTSR and SETOL theories to optimizers such as Muon, and motivates a new optimizer extension, WW-PGD. Together, these results point toward a spectral Renormalization-Group view of learning, in which related algorithms, in NNs and even traditional machine learning (ML) algorithms, may approach shared spectral structure in the settings studied here.

1 Introduction

Why does Deep Learning Work? Artificial neural networks now power large language models, AlphaGo and AlphaFold-class systems [1, 2], and other foundation models [3, 4, 5, 6], yet we still lack a generally accepted theory that both explains their convergence and tells practitioners how to improve training. Modern networks optimize enormous nonconvex objectives. They memorize parts of their data. They generalize in ways that remain surprising. A theory of learning should explain this coexistence, not merely describe the loss curve.

We don't have a theory of learning; we simply train until the loss stops. But two models can reach similar losses and still build different internal structures. If you just consider the loss, overfitting looks like pre-grokking memorization. Optimizers like Muon whiten or orthogonalize updates without understanding why it helps [28]. These examples point to the same gap: training is not just minimizing the loss; it is motion through internal layer states.

Layer convergence is the missing object. A network trains all layers together, but each dense layer changes its own weight matrix and concentrates information in its own way. Some concentration is necessary because learning compresses data into useful representations. Too much concentration is dangerous because a few dominant tail components can dominate the representation. This is part of the story. But this does not explain how finite trained layers reach correlated spectral endpoints.

HTSR and SETOL provide the starting points. HTSR applies *Random Matrix Theory* (RMT) to trained neural-network weight matrices. It posits $5 + 1$ phases of training by analyzing a layer's *Empirical Spectral Density* (ESD) and identifying signatures of heavy-tailed (HT) universality classes beyond classical *Marchenko–Pastur* RMT. Central to HTSR is the observation that the best layers develop HT spectra fit by a *power law*, with exponent $\alpha \simeq 2$. This motivates treating $\alpha \simeq 2$ not merely as a fitted value, but as a spectral boundary between near optimal self-averaging behavior and overfitting due to non-self-averaging.

SETOL was introduced to explain why this spectral criterion correlates with generalization. Under physically motivated approximations, SETOL expresses a trained layer's contribution to generalization solely as a functional of its empirical covariance spectrum, derived from a layer free energy written as an HCIZ integral. Thus the ESD is the relevant object: the same spectral structure that determines the HTSR exponent α also determines the SETOL energy functional. SETOL further suggests that the optimal regime at $\alpha = 2$ corresponds to approach toward a Wilsonian, scale-invariant, volume-preserving *Exact Renormalization Group* (ERG) fixed point. In this picture, the generalizing components concentrate into a low-rank *Effective Correlation Space* (ECS), whose normalized spectrum satisfies the ERG trace-log gauge-fixing constraint.

Here we thread those two theories into a proposed Wilsonian RG theory of layer convergence. The NN remains a coupled system, but near convergence each layer can be viewed as approaching its own Wilsonian-style spectral fixed point. The layers must remain compatible, but they need not share one global fixed point. The explicit claims are that SETOL derives the trace-log condition, the RG map is operational, and the $\alpha \simeq 2$ claim is empirical and supported by self-averaging and scale-counting arguments. Together, these arguments interpret the HTSR observations and give optimizer design a concrete target: move each layer toward stable correlated structure while preventing overfitting.

2 HTSR: phases of learning

HTSR heavy-tailed phases

RMT-based phenomenology of layer training.

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

- $\alpha > 2$**:** stable, but the tail can be too weak.
- $\alpha \simeq 2$**:** practical optimum for layer convergence.
- $\alpha < 2$**:** overcorrelated / overfit

HTSR is the empirical rule: well-trained layers often sit near the $\alpha \simeq 2$ boundary.

Heavy-Tailed Self-Regularization (HTSR) is a phenomenological theory that describes the phases of training of a NN layer using a unique application of the Marchenko–Pastur (MP) Random Matrix Theory (RMT) in the heavy-tailed (HT) setting, classifying individual NN layers into HT/PL universality classes [8].

The starting point is a single layer. Write the weight matrix as $\mathbf{W} \in \mathbb{R}^{N \times M}$, with $N \geq M$, and define

$$\mathbf{X} = \frac{1}{N}\mathbf{W}^T\mathbf{W}. \tag{1}$$

The operator $\mathbf{X}$ is the layer's weight-correlation operator. Its eigenvectors give input directions in the layer. Its eigenvalues give the spectral quality gain assigned to those eigen-directions. Importantly, HTSR recognizes the elements $\mathbf{W}_{i,j}$ should not be strongly heavy tailed, whereas the eigenvalues λ of $\mathbf{X}$ are, this being a signature of a strongly correlated system near criticality.

HTSR works by fitting the layer ESD to a finite-window power law, where the exponent α places the layer in an RMT/HT universality class. Empirically, strong trained layers often approach the optimal state $\alpha \simeq 2$. Section 4 explains why this boundary is optimal for learning using statistical/statistical mechanics arguments.

The empirical premise comes from observations on hundreds of well-trained models and thousands of layers, using the open-source `WeightWatcher` tool [14]. `WeightWatcher` implements a practical version of the HTSR (and SETOL) diagnostics, and makes these diagnostics reproducible and empirically testable and/or falsifiable.

In many well-trained models, the layers have an ESD with a finite-window tail, starting at λ^+_{PL}, that can be well fit to a power law (PL):

$$\rho(\lambda) \sim \lambda^{-\alpha}, \quad \lambda \geq \lambda^+_{\text{PL}} \tag{2}$$

HTSR uses this tail, together with related power-law metrics, as a data-free diagnostic of learned correlation structure [7, 8, 9]. In practice, the power-law exponent is estimated with the standard Clauset–Shalizi–Newman maximum-likelihood fit which finds the best λ^+_{PL}, as implemented in the open-source WeightWatcher tool [14, 26]. The observed optimum is not simply the smallest exponent. Strong checkpoints often sit near

$$\alpha \simeq 2. \tag{3}$$

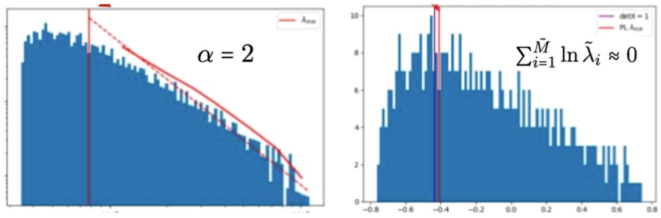

Figure 1. Empirical Spectral Density (ESD) of a well-trained dense layer. Panel (a) shows the ESD of $\mathbf{X} = N^{-1}\mathbf{W}^T\mathbf{W}$ in log-log coordinates. The tail is fit by a finite-window MLE/KS power law with $\alpha \simeq 2$. Panel (b) shows the trace-log-normalized ECS, where $\text{Tr} \log \widetilde{\mathbf{X}}_{\mathcal{R}} \simeq 0$ (where $\mathcal{R}$ denotes the ECS space). The red and purple vertical lines denote the PL tail, $\tilde{\lambda}^+_{\text{PL}}$, and the ECS, $\tilde{\lambda}^+_{\mathcal{R}}$, resp. Plots have been generated using the open-source `WeightWatcher` tool.[14]

Figure 1 shows a representative fitted spectrum.

The diagnostic is practical. The open-source `WeightWatcher` tool estimates the layerwise HTSR PL exponents α from trained weights alone. One can obtain PL fits for both linear and convolutional layers (see [10]). Prior work showed that these data-free spectral metrics can predict model-quality trends without access to training or testing data [9]. Figure 2 shows recent example scans for the OpenAI GPT-OSS 20B and 120B models [6], and the instruct component ($\mathbf{W} := \mathbf{W}_{\text{instruct}} - \mathbf{W}\text{base}$) of the Qwen2.5 instruct models. Each histogram summarizes fitted layerwise α values across all layer weight matrices.

HTSR is phenomenological. It extends the familiar Marchenko-Pastur (MP) Random Matrix Theory (RMT) in a unique way to the layer weight matrices to classify them into several distinct heavy tailed (HT) universality classes. HT RMT was first noted in the theoretical physics literature in the 1990s that when a random matrix is heavy-tailed power-law (HT-PL) elementwise, the ESDs no longer follow the MP law and themselves become HT-PL and fall into different universality classes [20, 21]. HTSR generalized this, and noted that when ESDs themselves are HT-PL, they fall into different universality classes even when the underlying matrix is not HT-PL elementwise. Moreover, HTSR notes that $\alpha \simeq 2.0$ is special, and corresponds to a candidate optimal state of the layer. Roughly, then, for our purposes here, we can reduce the $5+1$ HTSR classes into three phases: $\alpha > 2.0$, $\alpha \simeq 2.0$, and $\alpha < 2.0$.

Still, HTSR is phenomenological. The central question here seeks a more precise but also simple answer: why does $\alpha \simeq 2.0$ correspond to an optimal trained layer with a stable HT-PL empirical spectral density?

Here, we develop an RG argument for $\alpha \simeq 2.0$ that is a posteriori consistent with HTSR (and SETOL), but goes further by integrating HTSR and SETOL; there are two main arguments. First, $\alpha = 2$ is the boundary where generalization is supported by the maximum number of eigencomponents before overfitting. Second, $\alpha = 2$ is what is called marginal in Wilsonian Renormalization Group. But first, to formulate the RG theory, we need to understand its motivation in the SETOL approach.

3 SETOL: the Wilsonian ERG

SETOL is a *Semi-Empirical Theory of Learning*. It connects the layer empirical spectral density (ESD) to multilayer neural-network generalization accuracy using statistical mechanics, providing a rigorous foundation[1] for HTSR. It is semi-empirical because it takes the layer ESD as empirical input and derives a layer-quality functional whose diagnostics can be associated with the HTSR PL metric α. In doing so, it assumes that a NN, near the optimal state of learning, satisfies a volume-preserving, scale-invariant transformation (the trace-log condition) that resembles taking a single step of the Wilsonian Exact Renormalization Group (ERG) [24, 25].

SETOL starts by approximating the average generalization accuracy (one minus the error) of a model, which it calls the average model *Quality* $\overline{Q}_{\rm NN}$.

$$\overline{Q}_{\rm NN} \simeq 1 - \overline{E}^{\rm NN}_{\rm gen}. \tag{4}$$

(where the averages is over all data samples). For a multilayer model, SETOL then approximates total model average quality by a product of layer contributions,

$$\overline{Q}_{\rm NN} \simeq \prod_{\rm layers} \overline{Q}_\ell. \tag{5}$$

Thus a layer quality $\overline{Q} := \overline{Q}_\ell$ estimates how much each layer supports the model's generalization accuracy.

Layer quality has a simple meaning here. It is one layer's contribution to generalization accuracy. A useful layer lowers error. A bad layer adds unstable gain or loses useful capacity. This is the practical meaning of "layer quality" throughout the paper.

SETOL derives the layer quality using the layer-quality (squared) generating function,

$$\begin{aligned} \overline{Q}^2 &= \frac{1}{\beta}\frac{\partial}{\partial n}\left[\lim_{N\gg 1} \beta\Gamma^{\rm IZ}_{\overline{Q}^2}\right] \\ &\simeq_{\text{high-T}} \frac{1}{n}\frac{\partial}{\partial \beta}\left[\lim_{N\gg 1} \beta\Gamma^{\rm IZ}_{\overline{Q}^2}\right]. \end{aligned} \tag{6}$$

[1] At a physics level of rigor.

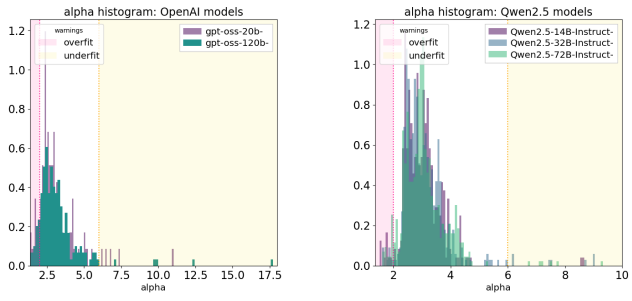

Figure 2. Layerwise HTSR α histograms from linear layers, computed with the open-source `WeightWatcher` tool. Left: GPT-OSS 20B/120B models. Right: Qwen2.5 instruct models. The warning bands are heuristic: small α marks overfit or trap risk, while large α marks weakly correlated or underfit tails. Optimizer notebook details are listed in Appendix A.

Here n is the sample count, and β is the inverse-temperature parameter. This is evaluated in the thermodynamic large-n (and large-N) limit. This generating function $\beta\Gamma^{\rm IZ}_{\overline{Q}^2}$ is essentially an effective layer free energy.

SETOL models each trained layer using a matrix generalization of the classical student–teacher perceptron [19] to matrices. The trained weight matrix plays the role of the fixed teacher, $\mathbf{T} = \mathbf{W}$, while $\mathbf{S}$ denotes a possible student matrix that reproduces the teacher-layer outputs. SETOL then generalizes the ST perceptron free energy in terms of the student correlation operator $\widetilde{\mathbf{A}}$. The compatible matrices $\widetilde{\mathbf{A}}_N$ and $\widetilde{\mathbf{A}}_M$ denote the induced $N \times N$ and $M \times M$ forms of this operator, restricted to the Effective Correlation Space (ECS). In an annealed, high-temperature approximation, the resulting effective free energy, or quality (squared) generating function $\beta\Gamma^{\rm IZ}_{\overline{Q}^2}$, takes the form of a random matrix (HCIZ) integral

$$\begin{aligned} \beta\Gamma^{\rm IZ}_{\overline{Q}^2} &= \frac{1}{N}\log\int d\mu(\widetilde{\mathbf{A}})\exp\{N\mathcal{S}(\widetilde{\mathbf{A}})\}, \\ \mathcal{S}(\widetilde{\mathbf{A}}) &= n\beta\ \mathrm{Tr}\left[\frac{1}{N}\mathbf{T}^T\widetilde{\mathbf{A}}_N\mathbf{T}\right] + \frac{1}{2}\mathrm{Tr}\log(\widetilde{\mathbf{A}}_M). \end{aligned} \tag{7}$$

The first term measures teacher–student alignment: it favors student correlations that match the trained teacher layer. The second term is the log-volume contribution that appears when the theory changes variables from weights to correlations. In the thermodynamic limit, and under the SETOL RG ansatz, this free energy depends only on the empirical spectral density of the teacher covariance. Thus the layer's contribution to generalization accuracy reduces to a functional of the trained weight matrix's ESD.

SETOL free energy and ERG trace-log

Matrix-generalized statistical mechanics student-teacher model of NN generalization accuracy:

$$1 - \overline{E}^{\rm NN}_{\rm gen} \simeq \overline{Q}_{\rm NN} \simeq \prod_\ell \overline{Q}_\ell$$

- **Layer Quality $\overline{Q}_\ell$:** one layer's contribution to total accuracy (1-error). Depends simply on $\mathbf{X}$.
- **Effective Free Energy** $\beta\Gamma^{\rm IZ}_{\overline{Q}^2}$ estimates $\overline{Q}_\ell$ with a random-matrix HCIZ integral
- **Gauge fixing:** remove volume drift with

$$\mathrm{Tr}\log\widetilde{\mathbf{X}}_R = 0 \qquad \Longleftrightarrow \qquad \det\widetilde{\mathbf{X}}_R = 1.$$

SETOL says: relate the HTSR power-law α to the generalization accuracy by evaluating $\beta\Gamma^{\rm IZ}_{\overline{Q}^2}$

The Wilsonian ERG step in SETOL is a change of variables and a change of scale. The bare layer variables are the student weights. The effective variables are the retained correlation modes in $\widetilde{\mathbf{A}}_{\mathcal{R}}$. The SETOL paper describes this step as an Effective-Hamiltonian construction: a bare layer Hamiltonian for $\overline{Q}^2$ is rewritten as a renormalized

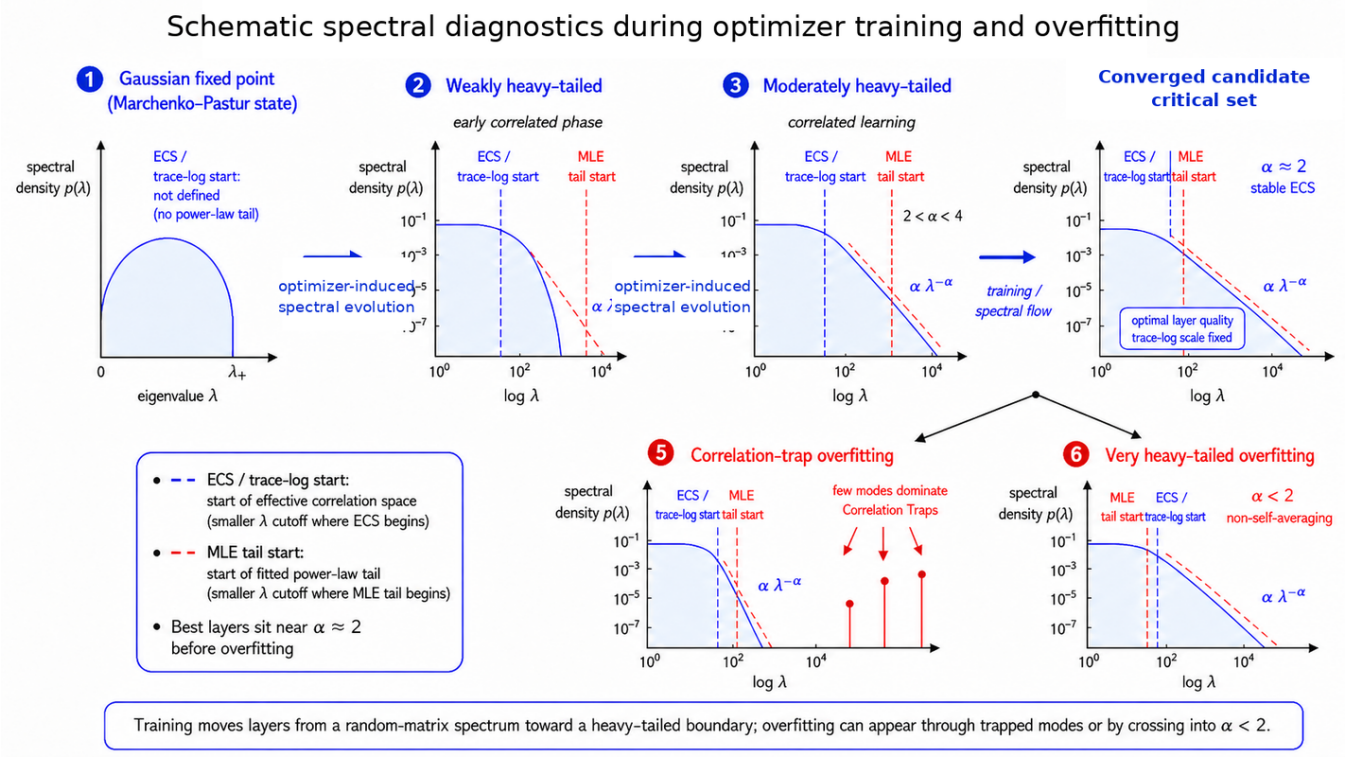

Figure 3. HTSR/SETOL phases during training. The HTSR power-law tail $\mathcal{T}_{\mathrm{PL}}$ and the SETOL Effective Correlation Space $\mathcal{R}_{\mathrm{ECS}}$ are selected independently. During training, they overlap. At convergence, they align $\mathcal{R}^{\star}_{\mathrm{ECS}} \simeq \mathcal{T}^{\star}_{\mathrm{PL}}$, at $\alpha \simeq 2$.

effective Hamiltonian on the ECS [11].

$$d\mu(\mathbf{S})\, e^{n\beta N \operatorname{Tr} \mathcal{H}_{\overline{Q}^2}} \xrightarrow{\text{ERG}} d\mu(\widetilde{\mathbf{A}})\, e^{n\beta N \operatorname{Tr} \mathcal{H}^{\mathrm{ECS}}_{\overline{Q}^2}}. \tag{8}$$

The arrow here is a scale-invariant, volume-preserving transformation, and is effectively taking a single step of the Wilsonian Exact RG (ERG). It is from this convenient construction that the ERG trace-log condition emerges. Here, the optimal ECS flow should approach a stable scale-balanced state near $\alpha \simeq 2$.

We first normalize or recale$\mathbf{X}$ as

$$\operatorname{Tr} \mathbf{X} = M \tag{9}$$

. We call this *Frobenius Rescaling*.

After rescaling, SETOL then defines a volume-preserving, scale-invariant transformation on the student $\mathbf{A}_{\mathcal{R}}$ and teacher $\mathbf{X}_{\mathcal{R}}$ correlation matrices

$$\operatorname{Tr} \log \widetilde{\mathbf{A}}_{\mathcal{R}} = 0 \leftrightarrow \operatorname{Tr} \log \widetilde{\mathbf{X}}_{\mathcal{R}} \simeq 0. \tag{10}$$

. where the tilde denotes that this trace-log condutions holds. (i.e. $\widetilde{\mathbf{X}}_{\mathcal{R}}, \tilde{\lambda}^{+}_{\mathcal{R}}, \tilde{\lambda}$), This is also called *gauge Fixing*, and we may say we are now on the *Gauge Slice*.

The layer quality (squared) is now expressed simply in terms of the tail of the "teacher" ESD using the integrated R-Transform $R_{\widetilde{\mathbf{X}}}(z)$ [23]:

$$\overline{Q}^2 = \sum_{\tilde{\lambda}_i \in \mathcal{R}} \int_{\tilde{\lambda}^{+}_{\mathcal{R}}}^{\tilde{\lambda}_i} R_{\widetilde{\mathbf{X}}}(z)\, dz \tag{11}$$

where $R_{\widetilde{\mathbf{X}}}(z)$ is an analytic function chosen to model the PL tail (and is analogous to a renormalized self-energy) and $\tilde{\lambda}^{+}_{\mathcal{R}}$ is the start of the ECS. This result shows that the layer quality depends only on the ESD (as in HTSR).

The link between HTSR and SETOL is summarized in Figure 3, which should be read as a schematic spectral flow rather than as a static taxonomy. The red vertical line marks the start of the HTSR power-law tail, $\mathcal{T}_{\mathrm{PL}}$, selected by the MLE fit, while the purple vertical line marks the start of the SETOL Effective Correlation Space, $\mathcal{R}_{\mathrm{ECS}}$, selected by the ERG trace-log condition. Training (usually) begins in a Gaussian/Marchenko–Pastur regime where neither boundary is stable, then a weak heavy tail forms, correlated learning develops, and the SETOL ECS boundary and HTSR PL-tail boundary move toward one another. At the convergence point, $\alpha \simeq 2$ and $\operatorname{Tr} \log \widetilde{\mathbf{X}}_{\mathcal{R}} = 0$; the two selected regions then align, $\mathcal{R}^{*}_{\mathrm{ECS}} \simeq \mathcal{T}^{*}_{\mathrm{PL}}$, equivalently $\tilde{\lambda}^{+}_{\mathcal{R}} - \tilde{\lambda}^{+}_{\mathrm{PL}} \simeq 0$. Empirically, this is also the regime in which the truncated-SVD approximation, restricted to the ECS, preserves test/generalization accuracy. The red lower panels show the two failure modes beyond this boundary: a correlation trap, where isolated rank-one perturbations in $\mathbf{W}$ distort the ESD and induce overfitting, and a very-heavy-tailed regime, where $\alpha < 2$ and the system becomes strongly (over) correlated and also overfits. At the boundary is optimal learning and generalization.

SETOL suggests that the best layers in a very well trained model sit at a Wilsonian Exact RG fixed point. The open question is how to formulate the full RG spectral flow, justifying the trace-log condition and $\alpha = 2$ as optimal.

4 The self-averaging boundary

It is a basic principle of statistics that reliable predictions require samples that are representative of the underlying population. The simplest counterexample is estimating the mean of a heavy-tailed power-law distribution. In such a distribution, a small number of extreme observations can dominate the average, so different finite samples may have very different sample means. A more subtle example arises in complex systems near a critical point. There, even when the ensemble mean is well defined, long-range correlations can keep sample-to-sample fluctuations large, so different large samples do not necessarily become representative of one another. In statistical mechanics, these ideas are formalized by the concept of *Self-Averaging*.

Self-averaging versus non-self-averaging

Self-averaging asks if relative sample-to-sample fluctuations $\mathcal{V}_M$ disappear as the system grows.

- $\mathcal{V}_M(\mathcal{O}) \to 0$: a bound can be formed
- $\mathcal{V}_M(\mathcal{O}) \not\to 0$: no useful (variance) bound

The self-averaging boundary is identified here where relative fluctuations $\mathcal{V}_M$ cease to vanish.

Self-averaging means that a macroscopic quantity measured in one large sample becomes close to its average over many samples as the system size M grows (i.e concentrates).. Let $\mathcal{O}_M$ be a macroscopic observable measured in a system of size M, and let the mean over M samples ξ_M be

$$\mu_M := \mathbb{E}[\mathcal{O}_M] = \langle \mathcal{O} \rangle_{\xi_M} \tag{12}$$

We say that $\mathcal{O}_M$ is self-averaging when it concentrates around μ_M as the system size grows. When the second moment is finite, its relative sample-to-sample variance is

$$\mathcal{V}_M(\mathcal{O}) := \frac{\mathrm{Var}(\mathcal{O}_M)}{|\mathbb{E}[\mathcal{O}_M]|^2} = \mathbb{E}\left[\left|\frac{\mathcal{O}_M}{\mu_M} - 1\right|^2\right]. \tag{13}$$

A standard sufficient test for self-averaging is $\mathcal{V}_M(\mathcal{O}) \to 0$ as $M \to \infty$. When $\mathcal{V}_M(\mathcal{O})$ does not vanish, relative fluctuations persist and different large samples can remain measurably different. This is *Non-Self-Averaging*.

In NNs, non-self-averaging arises from two different mechanisms: localized correlation traps, and long-range collective non-vanishing fluctuations. In either case, a model may describe its training sample well while performing poorly on out-of-sample test data.

We propose that the strongest learning occurs at the boundary between self-averaging and non-self-averaging. And this is not specific to NNs but a fundamental property of all related learning algorithms.

But first, it is helpful to understand self-averaging from different theoretical points of view: Statistical Learning Theory (SLT), Statistical Mechanics (StatMech), and Random Matrix Theory (RMT).

4.1 Statistical Learning view

The connection to Statistical Learning Theory (SLT) here is qualitive; it is depicted in Figure 4.

For a fixed tolerance $\epsilon > 0$, consider the probability that one sample differs appreciably from the ensemble mean,

$$p_M(\epsilon) := \Pr\left\{\left|\frac{\mathcal{O}_M}{\mu_M} - 1\right| > \epsilon\right\}. \tag{14}$$

A useful concentration bound has the form

$$p_M(\epsilon) \le B_M(\epsilon), \qquad B_M(\epsilon) \longrightarrow 0 \quad \text{as} \quad M \longrightarrow \infty. \tag{15}$$

Such a bound says that a sufficiently large sample is likely to be close to the ensemble mean.

On the self-averaging side, the relative variance $\mathcal{V}_M(\mathcal{O})$ decreases, so the distribution narrows and the bound can tighten. More data or greater system size then produces a more stable and representative answer. If $\mathcal{V}_M(O) \to 0$ this bound shrinks to zero for every fixed $\epsilon > 0$ since

$$p_M(\epsilon) \le \frac{\mathcal{V}_M(\mathcal{O})}{\epsilon^2}. \tag{16}$$

A sufficiently large sample is then increasingly likely to lie close to the ensemble mean.

On the non-self-averaging side, the relative fluctuations do not vanish and therefor can not gaurentee concentration. There is some fixed tolerance ϵ_0 for which $\mathcal{V}_M(\mathcal{O})$ and $p_M(\epsilon_0)$ remain effectively nonzero as the system grows. Any valid bound must satisfy $p_M(\epsilon_0) \le B_M(\epsilon_0)$ and therefore cannot shrink to zero. One can form a bound and it may still be formally true, but it no longer improves with width. It is flat. In practice, it can become vacuous. Concentration may fail. In such cases, one can describe in-sample data, but the bound no longer supports reliable out-of-sample prediction by itself.

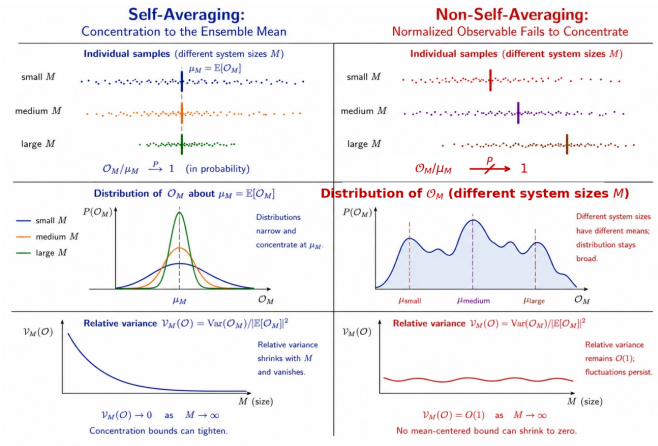

Figure 4. Self-averaging and the failure of a shrinking bound. Here $\mu_M = \mathbb{E}[\mathcal{O}_M]$. In the self-averaging case, the normalized observable $\mathcal{O}_M/\mu_M$ concentrates around one and the relative variance $\mathcal{V}_M(\mathcal{O})$ tends to zero, allowing concentration bounds to tighten. In the non-self-averaging case, the normalized distribution retains finite width and $\mathcal{V}_M(\mathcal{O})$ does not vanish, so no valid mean-centered bound can shrink to zero.

4.2 Statistical Mechanics view

In statistical physics, self-averaging is frequently stated in terms of the data-averaged partition function $\mathcal{Z}_M$ as

$$\langle \log \mathcal{Z}_M \rangle_{\xi_M} \simeq \log \langle \mathcal{Z}_M \rangle_{\xi_M} . \tag{17}$$

Here ξ_M labels one independently prepared sample of size M, and $\langle\cdot\rangle_{\xi_M}$ denotes an average over many such samples. Equation (17) is a strong form of self-averaging: a single large sample gives nearly the same free-energy contribution as the sample average.

To expose the fluctuation correction, define

$$\delta_M := \tfrac{\mathcal{Z}_M - \langle \mathcal{Z}_M \rangle_{\xi_M}}{\langle \mathcal{Z}_M \rangle_{\xi_M}}, \qquad \langle \delta_M \rangle_{\xi_M} = 0, \tag{18}$$

so that

$$\mathcal{V}_M(\mathcal{Z}) := \tfrac{\mathrm{Var}_{\xi_M}(\mathcal{Z}_M)}{\langle \mathcal{Z}_M \rangle^2_{\xi_M}} = \left\langle \delta_M^2 \right\rangle_{\xi_M} . \tag{19}$$

Whenever the expansion is controlled,

$$\begin{aligned} \langle \log \mathcal{Z}_M \rangle_{\xi_M} &= \log \langle \mathcal{Z}_M \rangle_{\xi_M} + \langle \log(1+\delta_M) \rangle_{\xi_M} \\ &= \log \langle \mathcal{Z}_M \rangle_{\xi_M} - \tfrac{1}{2}\mathcal{V}_M(\mathcal{Z}) + \tfrac{1}{3}\left\langle \delta_M^3 \right\rangle_{\xi_M} - \cdots . \end{aligned} \tag{20}$$

Thus the difference between the two sides of Eq. (17) consists of sample-to-sample fluctuation terms. If the relative variance and higher normalized fluctuations vanish, the two expressions agree to leading order. If they remain finite, one large sample need not represent the full collection of samples, thereby failing to generalize.

SETOL and the Spectral Energy. Section 3 gives the corresponding effective free energy of layer quality (squared) generating function $\beta\Gamma\frac{\mathrm{IZ}}{\overline{Q}^2}$, Eq. (6). We use this and ,Eq. (11) to write the ECS-averaged quality (squared) $\tilde{Q}^2 = \overline{Q^2}/M_{\mathcal{R}}$, now averaged over the retained ECS $\mathcal{R}$:

$$\tilde{Q}^2 = \frac{1}{M_{\mathcal{R}}} \sum_{\tilde{\lambda}_i \in \mathcal{R}} G_{\widetilde{\mathbf{X}}}(\tilde{\lambda}_i), \;\; G_{\widetilde{\mathbf{X}}}(\tilde{\lambda}) := \int_{\tilde{\lambda}^+_{\mathcal{R}}}^{\tilde{\lambda}} R_{\widetilde{\mathbf{X}}}(z)\, dz, \tag{21}$$

where $R_{\widetilde{\mathbf{X}}}$ is the R-transform of the retained spectrum. Its local free-cumulant expansion is

$$R_{\mathbf{X}}(z) = \kappa_1^{(f)} + \kappa_2^{(f)} z + \kappa_3^{(f)} z^2 + \cdots . \tag{22}$$

Writing

$$m_p := \frac{1}{M_{\mathcal{R}}} \operatorname{Tr} \widetilde{\mathbf{X}}^p_{\mathcal{R}} = \int \tilde{\lambda}^p_{\mathcal{R}} \widehat{\rho}_{\mathcal{R}}(\tilde{\lambda}_{\mathcal{R}})\, d\tilde{\lambda}_{\mathcal{R}}, \tag{23}$$

To simplify the analysis, let $\tilde{\lambda}^+_{\mathcal{R}} \simeq 0$; evaluating $\hat{Q}^2$ gives

$$\tilde{Q}^2 = \kappa_1^{(f)} m_1 + \tfrac{1}{2}\kappa_2^{(f)} m_2 + \tfrac{1}{3}\kappa_3^{(f)} m_3 + \cdots \tag{24}$$

Since $\kappa_1^{(f)} = m_1$, this gives the ECS-average layer quality as $\tilde{Q} \simeq m_1$. The leading additive spectral quantity is then

$$\widetilde{E}_{\mathcal{R}} := m_1 \simeq \frac{1}{M_{\mathcal{R}}} \operatorname{Tr} \widetilde{\mathbf{X}}_{\mathcal{R}} = \int \tilde{\lambda}_{\mathcal{R}} \widehat{\rho}_{\mathcal{R}}(\tilde{\lambda}_{\mathcal{R}})\, d\tilde{\lambda}_{\mathcal{R}}. \tag{25}$$

Notice that we choose $\operatorname{Tr} \mathbf{X} = M$, but once we are on the gauge slice and have imposed the trace-log condition, $\operatorname{Tr} \widetilde{\mathbf{X}}_{\mathcal{R}} \neq M$ and can flow. This is the *Spectral Energy*; it is the relevant macroscopic observable for our RG power-counting analysis.

Ising-model analogy and the self-averaging test. A similar structure appears in the Ising model. Let $\mathcal{M}$ denote the total magnetization and let a field h enter through $\mathcal{H}(h) = \mathcal{H}_0 - h\mathcal{M}$. Then $\log \mathcal{Z}_M(h)$ is the cumulant-generating function for $\mathcal{M}$:

$$\log \frac{\mathcal{Z}_M(h)}{\mathcal{Z}_M(0)} = \sum_{n \geq 1} \frac{(\beta h)^n}{n!}\, \kappa_n(\mathcal{M}). \tag{26}$$

The first field-dependent term is $\beta h\, \kappa_1(\mathcal{M}) = \beta h \langle \mathcal{M} \rangle$, and the magnetization density is

$$m_M := \frac{\langle \mathcal{M} \rangle_\beta}{M} = \frac{1}{\beta M} \frac{\partial}{\partial h} \log \mathcal{Z}_M(h). \tag{27}$$

where the brackets $\langle \rangle_\beta$ denote the ordinary thermal average within one fixed sample. Thus the total magnetization is the first cumulant, and m_M is its leading additive density generated from the Ising free energy. In the same sense, $\widetilde{E}_{\mathcal{R}} = m_1$ is the first additive spectral quantity generated by the SETOL expansion. It is therefore the first quantity to test for concentration:

$$\mathcal{V}_{M_{\mathcal{R}}}(\widetilde{E}_{\mathcal{R}}) \underset{?}{\longrightarrow} 0. \tag{28}$$

In the disordered Ising model and related random critical systems, the sample-to-sample relative width of the magnetization and susceptibility does not necessarily narrow at the critical point. Instead it can remain finite as the system grows, so these observables are non-self-averaging [35, 36, 37]. The analogy here is that if $\widetilde{E}_{\mathcal{R}}$ does not self-average, then the leading SETOL contribution already depends on the particular training sample or run.

Why criticality defeats averaging. The sample-to-sample fluctuations above compare independently prepared systems; they are not the ordinary thermal fluctuations inside one fixed sample. Away from criticality, the correlation length ℓ_{corr} is finite, so a d-dimensional system of linear size L contains roughly $(L/\ell_{\mathrm{corr}})^d$ nearly independent correlated regions, whose differences average away. At a disordered critical point, $\ell_{\mathrm{corr}} \sim L$, so the effective number of independent regions can remain $O(1)$. Increasing the system size then need not reduce relative fluctuations. This is one reason why magnetization and other singular observables can remain non-self-averaging at criticality.

For the retained spectral energy,

$$\mathrm{Var}_{\xi_M}(\widetilde{E}_{\mathcal{R}}) = \frac{1}{M^2_{\mathcal{R}}} \sum_{i,j \in \mathcal{R}} \mathrm{Cov}_{\xi_M}(\tilde{\lambda}_{\mathcal{R},i}, \tilde{\lambda}_{\mathcal{R},j}). \tag{29}$$

If the eigenvalue contributions fluctuate nearly independently, the off-diagonal terms are small and the average can concentrate. If many modes fluctuate coherently, the covariance terms add rather than cancel, and the relative variance can remain finite even when the spectrum contains many modes. $\widetilde{E}_{\mathcal{R}}$ plays a similar diagnostic role as magnetization for critical, non-self-averaging behavior.

Correlation traps, however, provide a second, different route to non-self-averaging, even at criticality; we discuss this in the next subsection.

4.3 Random Matrix Theory view

Random Matrix Theory gives a direct finite-matrix test of the self-averaging boundary in the form of *Correlation Traps*. Starting from a trained layer, we randomize its entries elementwise,

$$\mathbf{W} \longrightarrow \mathbf{W}^{\pi} := \mathrm{rand}(\mathbf{W}), \tag{30}$$

by flattening the matrix, randomly permuting all NM entries, and reshaping it. This preserves the exact weight multiset and its one-point statistics, but removes the learned arrangement of those weights. For the classical spectral test, $\mathbf{W}^{\pi}$ is therefore uncorrelated and effectively i.i.d:

$$W_{ij}^{\pi} \overset{\text{eff. i.i.d.}}{\sim} \widehat{\rho}_{W}. \tag{31}$$

The observable is the full ESD of $\mathbf{X}^{\pi} = N^{-1}(\mathbf{W}^{\pi})^{T}\mathbf{W}^{\pi}$. For a large finite-variance matrix, this ESD approaches the deterministic Marchenko–Pastur (MP) law [16]; MP is the self-averaging baseline. It is the large-N limit ($Q = \frac{N}{M}$ fixed) of the mean distribution $\rho^{\pi}(\lambda)$ for the ESD of $\mathbf{W}^{\pi}$. For a typical i.i.d. rectangular matrix, we expect all the eigenvalues to lie within the MP bulk, to within finite-size Tracy–Widom (TW) fluctuations [17, 18].

If an eigenvalue λ_{trap} appears beyond the MP bulk edge $\lambda_{\mathrm{MP}}^{+}$ plus the TW scale Δ_{TW}, we call these outliers *Correlation Traps*.

$$\lambda_{\mathrm{trap}} > \lambda_{\mathrm{MP}}^{+} + \Delta_{\mathrm{TW}} \tag{32}$$

Figure 5 illustrates several correlation traps.

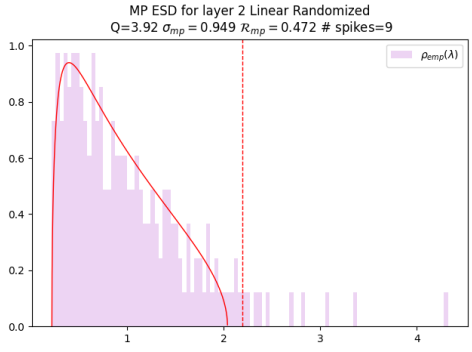

Figure 5. The randomized MP test. After $\mathbf{W} \to \mathrm{rand}(\mathbf{W})$, the bulk follows the MP fit. The dashed line marks the far bulk plus TW edge ($\lambda_{\mathrm{MP}}^{+} + \Delta_{\mathrm{TW}}$); persistent eigenvalues to its right are counted as correlation traps.

A finite number of traps may leave the limiting MP bulk intact, since each outlier carries mass $1/M$. The violation is instead localized at the edge: the largest λ are no longer explained by ordinary finite-size fluctuations.

Figure 6 (top panel) depicts results from a long-horizon grokking experiment, where a three-layer MLP is trained on a subset of MNIST until it induces pre-grokking memorization, then good performance (grokking), and finally overfitting (anti-grokking)[12, 13]. The red line denotes

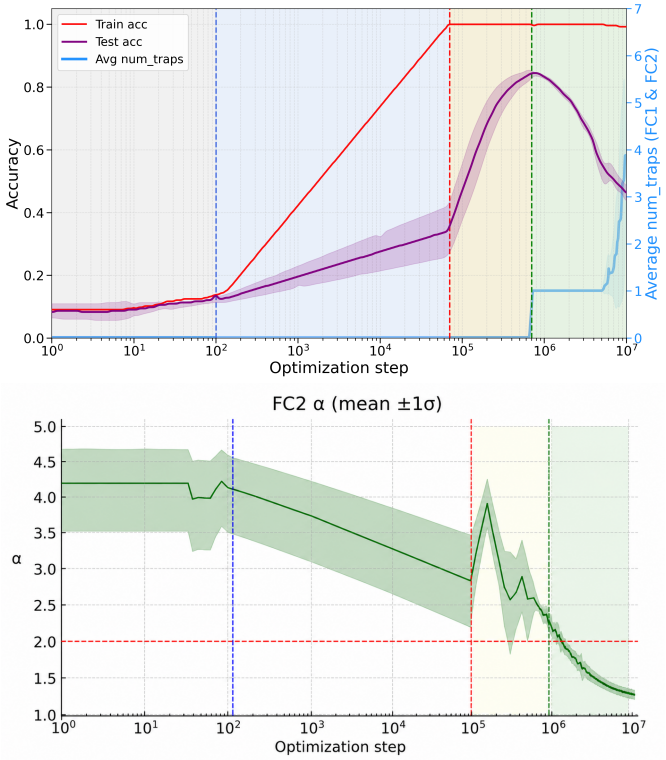

Figure 6. Long-horizon grokking. Top: test accuracy falls as the randomized-spectrum trap count rises. Bottom: the fitted exponent subsequently crosses below $\alpha = 2$. The aligned horizontal axes show the order of the warnings [12, 13].

the training accuracy, which becomes perfect at about 10^5 training steps. The purple line denotes the test accuracy, which increases to optimal at about 10^6 steps, and then drops sharply. The blue line counts the number of correlation traps in the system (in all three layers). The traps become non-zero as soon as the system test accuracy drops at 10^6 steps, and increase with decreasing test accuracy.

The bottom panel displays the average layer α for the FC1 and FC2 layers. Notice that $\alpha \to 2$ as the performance increases, and $\alpha < 2$ after first correlation trap forms and the test error drops. The drop in average α is late because the traps distort the ESD and the PL fits.

These results indicate that conventional overfitting leaves a spectral record in the weights. Training accuracy remains high, but held-out accuracy degrades as persistent MP-edge anomalies accumulate. Traps are often seen with aggressive learning rates or prolonged training. Early stopping and weight decay can help, but neither directly checks whether the randomized ESD has returned to MP.

The trap test is different from HTSR. Traps are measured on rand($\mathbf{W}$); HTSR is measured on the original correlated $\mathbf{W}$ and treats its broad non-MP ESD as learned structure [8]. A trap can distort the upper ESD and damage the power-law fit, sometimes contributing to an apparent $\alpha < 2$. Conversely, a smooth heavy-tailed ESD can cross below two without an isolated trap. Thus collective heavy-tailed fluctuations and localized correlation traps are distinct mechanisms of non-self-averaging.

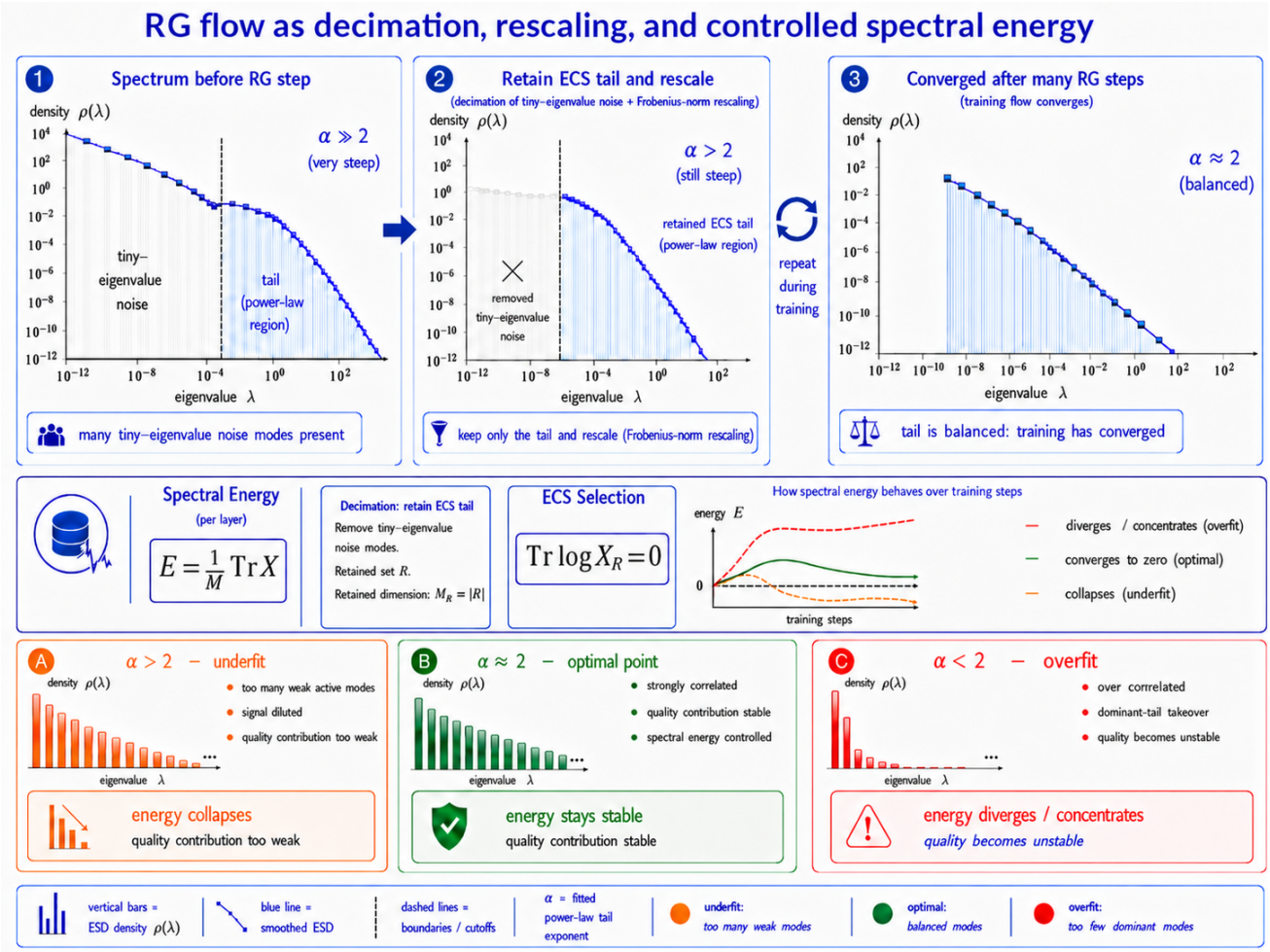

Figure 7. Operational RG flow for one dense layer. The step first decimates tiny-eigenvalue noise and retains an ECS tail $\mathcal{R}$. The SETOL-normalized retained eigenvalues are $\tilde{\lambda}_{\mathcal{R},i}$, and the retained dimension is $M_{\mathcal{R}} = |\mathcal{R}|$. The ERG trace-log condition fixes the product scale, $M_{\mathcal{R}}^{-1}\sum_{i\in\mathcal{R}}\log(\tilde{\lambda}_{\mathcal{R},i}) = 0$. The remaining spectral energy gain is then tested for stability.

5 Wilsonian RG Flow

Optimizer-induced RG flow

Wilsonian RG describes the flow of the layer ESD(s) induced by the optimizer steps $\mathcal{O}_t[\circ]$

$$\mathbf{W}_{t+1} = \mathcal{O}_t[\mathbf{W}_t], \qquad \mathcal{RG}[\rho(\lambda)] \longrightarrow \rho^*_{\text{tail}}(\lambda)$$

The best layers approach an RG fixed point $\mathfrak{F}_{\text{ERG}}$.

This section describes a Wilsonian RG construction for the spectral density $\rho(\lambda)$ as the heavy-tailed power-law density forms and approaches the HTSR $\alpha = 2$ candidate optimal state under the SETOL ERG trace-log rescaling. It is the shape, not the scale, that regulates the final layer quality and therefore the final model generalization accuracy. The retained layer should help accuracy improve, or equivalently help error decrease, without making the layer contribution explode or collapse. ERG trace-log normalization removes arbitrary scale. The RG construction shows when the spectral shape collapses, diverges, or converges.

The goal here is to identify an effective description of a trained layer–the Effective Correlation Space (ECS)–by repeatedly removing small-scale spectral components, retaining the dominant correlation structure, and eliminating redundant overall scale. Rather than develop a formal RG map, we describe the construction operationally using a truncated-SVD picture together with explicit ERG trace-log renormalization. The resulting flow acts on the retained spectral density (in the ECS) and motivates the HTSR $\alpha = 2$ as the boundary between learning and overfitting.

Figure 7 depicts the RG operational training flow which is generated by the training dynamics of the underlying optimizer on each step: $\mathbf{W}_{t+1} = \mathcal{O}_t[\mathbf{W}_t]$. In practice, the RG flow can be seen as a series of truncated-SVD steps, called decimations. Each step keeps the dominant spectral directions that contribute to the generalization accuracy, while dropping directions that do not matter at the current scale. Each retained layer is rescaled according to the trace-log law. The best performing models have layers that approach the SETOL ERG fixed point $\mathfrak{F}_{\text{ERG}}$, converging to an ESD with HTSR PL exponent $\alpha \simeq 2$, neither overfitting ($\alpha < 2$) nor underfitting ($\alpha > 2$) their training data.

5.1 HTSR phases and RG

HTSR provides the empirical starting point for the RG analysis. Section 2 described trained layer spectra using HTSR heavy-tailed universality classes. Here we use those classes as a coarse description of training flow. We do not start with a microscopic model of the data, the optimizer, or the architecture. We start with the observed fact that trained layer ESDs often develop heavy-tailed power-law spectra.

Table 1. HTSR classes interpreted as an RG flow.

Phase	Exponent	RG meaning
Random-like / Bulk+Spikes	no PL fit	Gaussian branch
Weakly HT	$\alpha \gg 2$	weak-energy sector
Moderately HT	$\alpha > 2$	approach to boundary
Strongly HT	$\alpha = 2$	balanced boundary
Very HT / traps	$\alpha < 2$	dominant-tail sector

Table 1 labels the HTSR phases, and Figure 7 shows how a layer moves through them. Training starts with many tiny-eigenvalue modes and a weak or steep tail, then retains an ECS tail and moves the ESD toward a more structured heavy-tailed spectrum. The spectral energy $\widetilde{E}$ from Eq. (25) gives this motion its learning meaning: for $\alpha > 2$ the retained energy is weak and can collapse (orange panel), near $\alpha \simeq 2$ it remains balanced across scales (green panel), and for $\alpha < 2$ the tail modes can carry too much energy and drive the overfit regime (red panel) where the energy diverges.

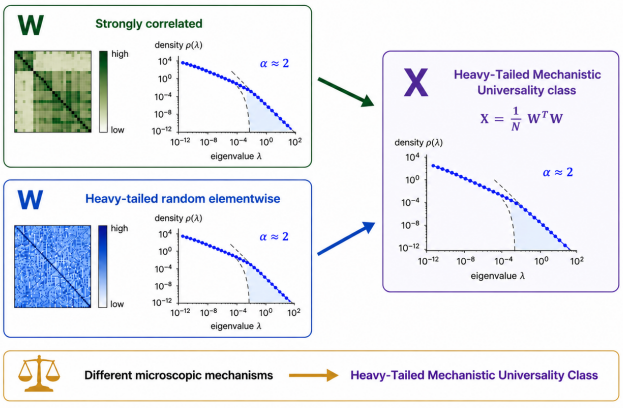

Figure 8. Heavy-Tailed Mechanistic Universality.

Figure 8 illustrates what HTSR calls *Mechanistic Universality*. Different architectures, datasets, and optimizers can produce the same heavy-tailed spectral endpoint. This universality is spectral, not elementwise. It is a property of the ESD of the correlation matrix $\mathbf{X} = N^{-1}\mathbf{W}^T\mathbf{W}$, not a consequence of the distribution of entries W_{ij}. Indeed, the entries are generally not HT (unless the layer is overfit); learned correlations can make the spectrum heavy-tailed. Here we interpret this mechanistic universality as a preview of RG universality; SETOL provides the bridge.

5.2 SETOL and RG

SETOL contributes the scale convention for the RG analysis. Section 3 introduced the layer quality generating function, an effective free energy $\beta\Gamma^{\mathrm{IZ}}_{\overline{Q}^2}$ and showed that the Wilsonian ERG construction leads to the ERG trace-log condition. The present paper does not need to evaluate the full SETOL quality $\overline{Q}$. Instead, we use the ERG trace-log condition to define the scale, and we study the flow of the retained spectral energy $\widetilde{E}$ on that scale.

SETOL → RG energy flow

Spectral Energy:

$$\widetilde{E} = \frac{1}{M_{\mathcal{R}}} \operatorname{Tr} \widetilde{\mathbf{X}}_{\mathcal{R}} = \int \tilde{\lambda}_{\mathcal{R}} \hat{\rho}_{\mathcal{R}}(\tilde{\lambda}_{\mathcal{R}})\, d\tilde{\lambda}_{\mathcal{R}}.$$

Gauge fixing:

$$\operatorname{Tr} \log \widetilde{\mathbf{X}}_{\mathcal{R}} = 0.$$

Truncated SVD is exact

$$\mathbf{W} \to \mathbf{W}_{\mathcal{R}} = \mathbf{U}\mathbf{\Sigma}_{\mathcal{R}}\mathbf{V}^{\top}$$

$$\mathbf{\Sigma}_{\mathcal{R}} := \operatorname{diag}[0, \cdots, 0, \sigma_1, \sigma_2, \cdots \sigma_{M_{\mathcal{R}}}], \quad \sigma_i \sim \sqrt{\lambda_i}$$

RG studies how the spectral components concentrate.

Figure 7 shows RG decimation, rescaling, and the resulting energy flow. The ERG trace-log is the scale-fixing operation; it is not itself the flowing energy. The quantity whose flow we study is the retained spectral energy $\widetilde{E}$, rescaled as needed. Eq. (25) defined this energy as the first moment of the retained ECS. The RG question is therefore concrete: after the trace-log removes arbitrary product scale, how does $\widetilde{E}$ flow across spectral scales?

We answer this question by using the trace-log to measure covariance volume. On the retained positive spectrum,

$$\operatorname{Tr} \log \mathbf{X}_{\mathcal{R}} = \log \det \mathbf{X}_{\mathcal{R}} = \sum_{i \in \mathcal{R}} \log \lambda_i. \tag{33}$$

Thus the trace-log measures the logarithm of the retained covariance volume. It does not add eigenvalue gains directly; it measures their product scale. Once fixed, the natural comparison for the RG flow will be across fixed-ratio bands $[\Lambda, b\Lambda]$, or equivalently across equal-width intervals in $\log \tilde{\lambda}_{\mathcal{R}}$; this lets us define a *Spectral Band Energy*, $G_E(\Lambda, b)$, below.

At the proposed Wilsonian ERG fixed point, the layer weight matrix $\mathbf{W}$ can be replaced by its Truncated SVD form, $\mathbf{W}_{\mathcal{R}} = \mathbf{W}_{\mathcal{R}} = \mathbf{U}\mathbf{\Sigma}_{\mathcal{R}}\mathbf{V}^{\top}$, where $\mathbf{\Sigma}_{\mathcal{R}}$ is the diagonal operator of $M_{\mathcal{R}}$ singular values (σ_i) (properly normalized) defined on the retained ECS, and zero for the other $M - M_{\mathcal{R}}$ ones. Doing this, SETOL predicts the model can faithfully reproduce the test/generalization accuracy. That is, the layer generalizing components fully concentrate into the ECS.

The next subsection shows that the HTSR power-law tail $\rho_{\mathrm{tail}}(\tilde{\lambda})$ makes the energy gain stable at $\alpha = 2$. That is, we illustrate that α is analogous to an RG universal critical exponent.

5.3 Operational view of the RG flow

Figure 7 depicts the spectral RG flow for one layer. The flow turns the ideas from Sections 5.1 and 5.2 into a repeatable operation: decimate tiny-eigenvalue noise, retain an ECS tail, impose the SETOL ERG trace-log gauge, and compare the remaining spectral shape. Section 5.1 supplies the HTSR input: the retained ESD should approach a power-law tail, with the strongest checkpoints often near $\alpha \simeq 2$. Section 5.2 supplies the SETOL input: generalization accuracy is controlled by the layer ESD through the ECS and the trace-log condition. Section 4 supplies the risk test: on the SETOL-normalized retained ECS, the retained spectral energy should self-average instead of concentrating in a few dominant tail modes. This subsection makes that control flow operational.

Figure 7 separates the RG step into four actions. First, the step removes tiny-eigenvalue noise that does not contribute to the retained correlation structure. Second, it keeps the ECS tail and represents the layer by the truncated matrix $\mathbf{W}_{\mathcal{R}}$. Third, it evaluates the retained operator on the Frobenius rescaling $\operatorname{Tr}\widetilde{\mathbf{X}}_{\mathcal{R}} = M$ and checks the ERG trace-log gauge. When the trace-log is exact, the ECS is exact. Fourth, it tests the remaining spectral energy. The point is not to reproduce SGD or any microscopic optimizer. The point is to take an RG view of the training flow and describe the learning process as a coarse-graining induced by the optimizer trajectories by considering the linearized spectral view at each layer.

The three regimes in Figure 7 give the operational meaning of underfit, balanced, and overfit. In the orange regime, $\alpha > 2$ and the tail is steep. The retained layer may self-average, but the spectral energy can be too weak to carry enough learned correlation. In the green regime, $\alpha \simeq 2$ and the retained spectral energy stays balanced across logarithmic spectral scales. This is the stable layer-quality regime. In the red regime, $\alpha < 2$ or correlation traps appear. The spectral energy diverges in the RG flow, the layer quality can become strongly non-self-averaging, and the generalization accuracy degrades.

SETOL suggests that the RG flow induces a truncated layer $\mathbf{W}$. SETOL maps layer quality, and therefore generalization accuracy, to the layer ESD. Once quality depends on the retained spectrum, not on every microscopic weight $W_{i,j}$, an effective description can replace the full layer by a retained ECS matrix. At the fixed point, SETOL predicts that the ECS aligns with the HTSR power-law tail and that the truncated-SVD layer $\mathbf{W}_{\mathcal{R}}$ preserves the test/generalization contribution of the original layer. This is the reason the operational RG step in Figure 7 is a truncated-SVD step.

The RG flow acts on retained spaces and retained operators together. At step (n), the effective object is

$$\left(\mathcal{R}^{(n)}, \mathbf{X}_{\mathcal{R}}^{(n)}\right), \qquad M_{\mathcal{R}}^{(n)} = |\mathcal{R}^{(n)}|. \tag{34}$$

A decimation step changes $\mathcal{R}^{(n)}$ by removing eigencomponents outside the current correlated description. the ECS is selected by imposing the trace-log gauge condition on the Frobenius-normalized retained correlation operator: $\operatorname{Tr}\log\widetilde{\mathbf{X}}^{(n)} = 0$;

SETOL then selects the $\tilde{\lambda}_{\mathcal{R}}^{+}$, the start of the ECS, based on the trace-log condition

$$\sum_{\tilde{\lambda}_{\mathcal{R}}} \log\tilde{\lambda}_{\mathcal{R}} = 0, \quad \tilde{\lambda}_{\mathcal{R}} \geq \tilde{\lambda}_{\mathcal{R}}^{+} \tag{35}$$

At optimality the ECS starts at the PL tail, $\tilde{\lambda}_{\mathcal{R}}^{+} = \tilde{\lambda}_{\mathrm{PL}}^{+}$. When $\alpha < 2$, the ECS can be smaller than the PL, with $\tilde{\lambda}_{\mathcal{R}}^{+} > PLstart$. This indicates that a few tail eigencomponents can dominate. Simultaneously, the PL tail can extend deep into previous bulk region, indicating that the spectral fluctuations may not concentrate (as in Eq. 29). These two mechanisms are central to understanding why $\alpha < 2$ is a signature of overfitting.

The ECS arises implicitly during the optimization step as the learning mechanism effectively compresses the model at each layer, causing the generalizing components of each $\mathbf{W}$ to flow into a lower-rank representation. But it is unclear how much compression is enough, and how much is too much. But because the typical optimizer is not aware of the RG flow, the ECS may be too weak and the spectral energy may collapse. Or, the layer may be overfit to the training data, causing the spectral energy to diverge because it concentrated in a few dominant modes, or many strongly fluctuating modes.

The retained ESD is the shape variable defined by the RG step. Formally, we can write

$$\widehat{\rho}_{\mathcal{R}}(\tilde{\lambda}_{\mathcal{R}}) = \frac{1}{M_{\mathcal{R}}}\sum_{i\in\mathcal{R}}\delta(\tilde{\lambda}_{\mathcal{R}} - \tilde{\lambda}_{\mathcal{R},i}), \qquad \int \widehat{\rho}_{\mathcal{R}}(\tilde{\lambda}_{\mathcal{R}})\, d\tilde{\lambda}_{\mathcal{R}} = 1. \tag{36}$$

In practice, however, we use the discrete retained spectral energy (on the gauge slice)

$$\widetilde{E}_{\mathcal{R}} = \frac{1}{M_{\mathcal{R}}}\operatorname{Tr}\widetilde{\mathbf{X}}_{\mathcal{R}} = \int \tilde{\lambda}_{\mathcal{R}}\, \widehat{\rho}_{\mathcal{R}}(\tilde{\lambda}_{\mathcal{R}})\, d\tilde{\lambda}_{\mathcal{R}}. \tag{37}$$

Section 4 showed why this energy matters: it provides a layer-level quantity that can self-average when many weakly correlated retained eigencomponents contribute, but (again) can fail to self-average in two ways. First, a small number of dominant modes can control the energy, as in correlation traps. Second, and more importantly here, collective correlations among the weights can keep the relative fluctuations of the spectral energy at order one across the entire ECS. This behavior is marginal at $\alpha = 2$ and becomes increasingly tail-dominated for $\alpha < 2$, where the largest eigencomponents can carry an $O(1)$ fraction of the total spectral energy. Although the boundary at $\alpha = 2$ may therefore be marginally non-self-averaging, the spectral energy grows only logarithmically with system size. This is shown below, however, first we would like to show that the SETOL trace-log condition naturally emerges through a scale-invariant RG flow defined by the unit Jacobian, the physically relevant RG transformation.

5.4 Scale Invariant Flow

An RG flow requires a series of scale-invariant decimation maps which here are characterized by the trace-log condition on the retained ECS space during the training flow. The trace-log comes from the Jacobian of the retained coordinate map. Figure 9 isolates the retained coordinate map used in the ERG trace-log gauge step.

In this subsection, $\mathbf{W}_{\mathcal{R}}$ denotes

$$\mathbf{W}_{\mathcal{R}} := \mathbf{W}\mathbf{V}_{\mathcal{R}}, \tag{38}$$

where $\mathbf{V}_{\mathcal{R}}$ contains the retained right singular vectors, and $\mathbf{W}_{\mathcal{R}}$ is $N \times M_{\mathcal{R}}$ (not the zero-padded).

$$\mathbf{X}_{\mathcal{R}} := \mathbf{V}_{\mathcal{R}}^T \mathbf{X}\mathbf{V}_{\mathcal{R}} = \mathrm{diag}(\lambda_i)_{i\in\mathcal{R}}. \tag{39}$$

After selecting $\mathcal{R}$, let $a \in \mathbb{R}^{M_{\mathcal{R}}}$ be

$$y(a) = \frac{1}{\sqrt{N}} \mathbf{W}_{\mathcal{R}} a. \tag{40}$$

The ordinary coordinate Jacobian is therefore

$$\frac{\partial y}{\partial a} = \frac{1}{\sqrt{N}} \mathbf{W}_{\mathcal{R}}. \tag{41}$$

This Jacobian is not a training-time derivative such as $d\mathbf{W}/dt$ or $d\mathbf{X}/dt$. It is the derivative of retained output coordinates with respect to retained input coordinates.

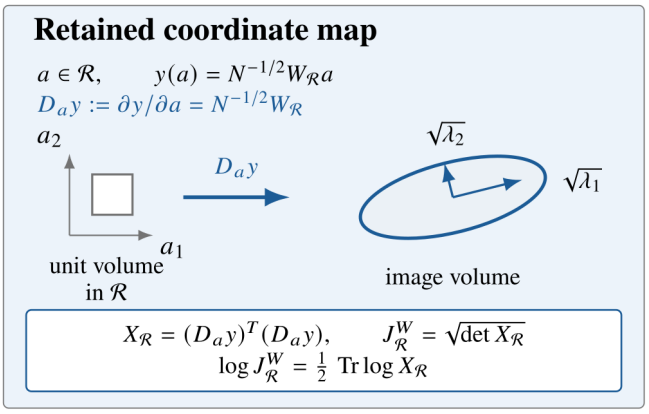

Figure 9. Retained coordinate Jacobian. After the ECS/SVD decimation selects $\mathcal{R}$, the retained layer sends coordinates $a \in \mathbb{R}^{M_{\mathcal{R}}}$ to $y(a) = N^{-1/2}\mathbf{W}_{\mathcal{R}} a$. The ordinary coordinate Jacobian is $\partial y/\partial a = N^{-1/2}\mathbf{W}_{\mathcal{R}}$. Its raw image volume is $J^W_{\mathcal{R}} = \sqrt{\det \mathbf{X}_{\mathcal{R}}}$.

The determinant appears because the retained map changes volume. Since $\mathbf{W}_{\mathcal{R}}$ may be rectangular, we do not take an ordinary determinant of $\mathbf{W}_{\mathcal{R}}$. We take the induced $M_{\mathcal{R}}$-dimensional volume factor. The squared volume factor is the Gram determinant,

$$\det\left[\left(\frac{1}{\sqrt{N}}\mathbf{W}_{\mathcal{R}}\right)^T \left(\frac{1}{\sqrt{N}}\mathbf{W}_{\mathcal{R}}\right)\right] = \det \mathbf{X}_{\mathcal{R}}. \tag{42}$$

Equivalently,

$$\left(\frac{1}{\sqrt{N}}\mathbf{W}_{\mathcal{R}}\right)^T \left(\frac{1}{\sqrt{N}}\mathbf{W}_{\mathcal{R}}\right) = \mathbf{V}_{\mathcal{R}}^T \mathbf{X}\mathbf{V}_{\mathcal{R}} = \mathbf{X}_{\mathcal{R}}. \tag{43}$$

Therefore the raw retained volume Jacobian is

$$J^W_{\mathcal{R}} = \sqrt{\det \mathbf{X}_{\mathcal{R}}} = \prod_{i\in\mathcal{R}} \sqrt{\lambda_i}. \tag{44}$$

Taking logarithms gives

$$\log J^W_{\mathcal{R}} = \frac{1}{2}\log\det \mathbf{X}_{\mathcal{R}} = \frac{1}{2}\mathrm{Tr}\log \mathbf{X}_{\mathcal{R}} = \frac{1}{2}\sum_{i\in\mathcal{R}} \log\lambda_i. \tag{45}$$

Thus the trace-log is the additive log-volume of the retained coordinate map.

The SETOL ERG gauge evaluates this retained volume using the normalized retained operator $\widetilde{\mathbf{X}}_{\mathcal{R}}$ and retained eigenvalues $\tilde{\lambda}_{\mathcal{R},i}$. The normalized retained volume factor is

$$\widetilde{J}_{\mathcal{R}} := \sqrt{\det \widetilde{\mathbf{X}}_{\mathcal{R}}} = \prod_{i\in\mathcal{R}} \sqrt{\tilde{\lambda}_{\mathcal{R},i}}. \tag{46}$$

Taking logarithms gives

$$\log \widetilde{J}_{\mathcal{R}} = \frac{1}{2}\mathrm{Tr}\log \widetilde{\mathbf{X}}_{\mathcal{R}} = \frac{1}{2}\sum_{i\in\mathcal{R}} \log\left(\tilde{\lambda}_{\mathcal{R},i}\right). \tag{47}$$

The scale-invariant ERG condition is

$$\widetilde{J}_{\mathcal{R}} \simeq 1 \qquad \Longleftrightarrow \qquad \mathrm{Tr}\log \widetilde{\mathbf{X}}_{\mathcal{R}} \simeq 0. \tag{48}$$

This is the ERG trace-log gauge-fixing constraint; after the training step $\mathcal{O}_t[\mathbf{W}_t]$, it defines the retained ECS at the next step of the RG flow. The optimizer flow thus both generates and implicitly tracks the RG flow.

The Jacobian argument gives the same trace-log condition as SETOL. SETOL obtains the condition from the HCIZ effective free energy, where the change from weights to correlations produces a log-determinant term. The RG argument obtains the same condition from the retained coordinate map for any selected ECS.

The fixed-point object is a normalized ESD shape, not a raw matrix. The raw operator $\mathbf{X}_{\mathcal{R}}$ can change during training. The operator $\widetilde{\mathbf{X}}_{\mathcal{R}}$ places the retained spectrum on the fixed normalized scale. A fixed point means that this retained spectral shape remains stable after further decimation and trace-log testing. At the proposed fixed point, the SETOL/ECS window and the HTSR power-law window align,

$$\mathcal{R}^\star_{\mathrm{ECS}} \simeq \mathcal{T}^\star_{\mathrm{PL}}, \qquad \rho_{\mathrm{tail}}(\tilde{\lambda}) \sim \tilde{\lambda}^{-2}, \qquad \mathrm{Tr}\log \widetilde{\mathbf{X}}_{\mathcal{R}} \simeq 0. \tag{49}$$

This alignment is not a definition. It is the measurable agreement between the trace-log effective space and the independently fitted HTSR tail.

The truncated-SVD claim is the operational test of the fixed point. SETOL predicts that, at the proposed RG fixed point, one can replace $\mathbf{W}$ with the lower-rank (zero-padded) $\mathbf{W}_{\mathcal{R}}$ without degrading the layer's generalization contribution. This is why the RG step can be described as a truncated-SVD decimation.

The next question is then to understand how $\alpha = 2$ emerges as the boundary between optimal learning and overfitting and is analogous to a critical universal exponent.

Table 2. Scale counting, statistics, and empirical meaning of the three spectral regimes.

Regime	spectral energy per scale	Scale-flow meaning	Statistical meaning	Learning meaning
$\alpha > 2$	$\tilde{\lambda}^2\rho(\lambda) \sim \tilde{\lambda}^{2-\alpha} \to 0$	Spectral band gain decays; weak-tail sector or approach toward the trivial fixed point $\mathfrak{F}_0$.	First moment self-averages; dominant tail components stay controlled.	Stable but potentially weakly correlated or under-trained tail.
$\alpha = 2$	$\tilde{\lambda}^2\rho(\lambda) \sim \tilde{\lambda}^0$	Scale-balanced logarithmic-band state; candidate $\mathfrak{F}_{\mathrm{ERG}}$ when ECS aligns with the fitted tail.	Marginal self-averaging boundary. Universal critical exponent.	Maximal stable structured-variation capture before tail-component domination.
$\alpha < 2$	$\tilde{\lambda}^2\rho(\lambda) \sim \tilde{\lambda}^{2-\alpha} \to \infty$	Spectral band gain grows toward the dominant tail components.	Tail components and fluctuations can dominate the first moment.	Overfit risk; correlation traps; held-out degradation.

5.5 RG Scale counting

RG scale counting asks how the layer-quality contribution changes as we move across spectral scales. In our spectral RG description, the spectrum is not treated as a flat list of eigenvalues. It is grouped into scale bands $[\tilde{\lambda}, b\tilde{\lambda}]$ that are natural for a power-law distribution. Then we ask whether each band contributes a stable amount on the SETOL trace-log gauge, or if the gain collapses/diverges.

RG boundary

RG counts gain by scale.

$$G_E(\Lambda, b) = \int_{\Lambda}^{b\Lambda} \tilde{\lambda}\rho(\lambda)\, d\tilde{\lambda}.$$

$$\rho(\lambda) = C\tilde{\lambda}^{-\alpha}, \qquad \alpha = 2 \Rightarrow G_E(\Lambda, b) = C \log b.$$

- $\alpha > 2$**:** larger bands contribute less and less gain.
- $\alpha = 2$**:** each log band contributes the same gain.
- $\alpha < 2$**:** spectral energy diverges

$\alpha \simeq 2$ is the log-scale-balanced boundary.

One RG step is a spectral operation in this paper. It discards eigencomponents outside the current effective description. It evaluates the retained operator on the fixed SETOL scale $\mathrm{Tr}\,\widetilde{\mathbf{X}}_{\mathcal{R}} = M$. It then compares the retained spectral shape with the previous retained shape. A fixed point means this decimate-and-test operation no longer changes the retained spectral shape or layer-quality gain.

The word "exact" is narrow here. In SETOL, it refers to taking a single scale-invariant transformation, as in Eq. 8. Here, it refers to a proposed Wilsonian map on the retained spectral distribution, not to a closed-form solution of the full network dynamics. We identify the candidate boundary and the candidate fixed-point signature of this map. It does not claim that one finite checkpoint exactly solves the full RG flow, although it is hypothesized.

The three spectral regimes are summarized in Table 2. These are direct analogues to the reduced description of the HTSR classes in Table 1. Each regime arises by considering the behavior of the spectral band energy in the limit of a continuous RG flow along the log-bands.

The fixed point is layerwise. We do not posit one global fixed point for an entire multi-layer network. Rather, each layer has its own spectrum, its own retained ECS, and its own Wilsonian-style spectral fixed point. A strong checkpoint is therefore a coordinated network state in which the matrix-like layers sit near their own SETOL ERG fixed points.

Because the spectral density $\rho(\lambda)$ is heavy-tailed power law for a well-trained layer, the natural RG scale band is logarithmic in $\tilde{\lambda}$, and the natural coordinate is

$$x = \log \tilde{\lambda}. \tag{50}$$

A fixed-width interval in x is a fixed-ratio spectral shell $\tilde{\lambda} \in [\Lambda, b\Lambda]$. This is the spectral analogue of coarse-graining by length scale in ordinary RG. Each logarithmic band asks how much spectral energy is contributed by one multiplicative decade of eigenvalues.

The object being counted is retained spectral energy $\widetilde{E}$ (36)/ Since $d\tilde{\lambda}_{\mathcal{R}} = \tilde{\lambda}_{\mathcal{R}}\, d\log\tilde{\lambda}_{\mathcal{R}}$, we have

$$d\widetilde{E} = \tilde{\lambda}_{\mathcal{R}}\widehat{\rho}_{\mathcal{R}}(\tilde{\lambda}_{\mathcal{R}})\, d\tilde{\lambda}_{\mathcal{R}} = \tilde{\lambda}^2_{\mathcal{R}}\widehat{\rho}_{\mathcal{R}}(\tilde{\lambda}_{\mathcal{R}})\, d\log\tilde{\lambda}_{\mathcal{R}}. \tag{51}$$

Thus the spectral energy per logarithmic spectral scale is

$$\frac{d\widetilde{E}}{d\log\tilde{\lambda}_{\mathcal{R}}} = \tilde{\lambda}^2_{\mathcal{R}}\widehat{\rho}_{\mathcal{R}}(\tilde{\lambda}_{\mathcal{R}}). \tag{52}$$

Substituting $\widehat{\rho}_{\mathcal{R}}(\tilde{\lambda}_{\mathcal{R}}) \simeq C\tilde{\lambda}^{-\alpha}_{\mathcal{R}}$ then gives $C\tilde{\lambda}^{2-\alpha}_{\mathcal{R}}$, so $\alpha = 2$ is precisely the case in which each multiplicative spectral shell contributes the same spectral energy. In the RG sense, we say $\alpha = 2$ *is marginal*.

This quantity answers an interpretable question: how much of the retained layer's quality gain, and therefore how much of its possible accuracy contribution, lives near this multiplicative eigenvalue scale?

A power-law tail makes the answer transparent. Given the retained PL tail, we can evaluate Eq. 52 as

$$\widehat{\rho}_{\mathcal{R}}(\tilde{\lambda}_{\mathcal{R}}) \simeq C\tilde{\lambda}^{-\alpha}_{\mathcal{R}} \to \frac{d\widetilde{E}}{d\log\tilde{\lambda}_{\mathcal{R}}} \sim C\tilde{\lambda}^{2-\alpha}_{\mathcal{R}}. \tag{53}$$

Table 2 summarizes the three RG regimes. When $\alpha > 2$, the high-eigenvalue shells carry less spectral energy with scale. The strongest generalizing components are less and less important. Generalization occurs but is weak. When $\alpha < 2$, high-eigenvalue shells carry more spectral energy with scale. The largest retained directions can dominate and the spectral energy diverges. When $\alpha = 2$, the change in energy is constant for all logarithmic shells; each shell carries comparable spectral energy gain.

Table 3. Special RG notation used in this section.

Term	Notation	Meaning here
RG step	$\mathcal{R}_{\mathrm{ECS}}$	Project to the retained ECS and evaluate the ESD on the SETOL scale.
Fixed point	$\mathcal{R}_{\mathrm{RG}}[\rho^\star] = \rho^\star$	A SETOL-normalized ESD that is unchanged by the RG step.
Weak-tail branch	$\mathfrak{F}_0, \quad g_E$ undefined	The trivial solution HTSR Random-like phase and nearby ESDs.
Correlated branch	$\mathfrak{F}_{\mathrm{ERG}}$	The empirical RG fixed point candidate.
Band Energy	$G_E(\Lambda, b)$	Spectral energy carried by one logarithmic spectral band.
Band gain	$g_E(\tilde{\lambda}_{\mathcal{R}}) := \tilde{\lambda}_{\mathcal{R}}^2 \widehat{\rho}_{\mathcal{R}}(\tilde{\lambda}_{\mathcal{R}})$	Spectral energy per logarithmic spectral scale.
Spectral Beta function	$\beta_E(g_E) = dg_E/d\log\tilde{\lambda}_{\mathcal{R}}$	The scale derivative of the spectral energy. For a PL tail, $\beta_E(g_E) = (2-\alpha)g_E$.
Scaling dimension	$y_E = 2 - \alpha$	The RG exponent controlling energy or spectral energy.
Marginal condition	$y_E = 2 - \alpha = 0$	The logarithmic spectral bands are balanced and carry comparable energy/gain.

We now do the full RG scale-counting calculation of the spectral band energy gain for each regime. Table 3 summarizes the RG notation used below.

The finite-shell calculation says the same thing without derivatives. For a multiplicative shell $[\Lambda, b\Lambda]$,

$$G_E(\Lambda, b) = \int_{\Lambda}^{b\Lambda} \tilde{\lambda}_{\mathcal{R}} \widehat{\rho}_{\mathcal{R}}(\tilde{\lambda}_{\mathcal{R}})\, d\tilde{\lambda}_{\mathcal{R}}. \tag{54}$$

If $\widehat{\rho}_{\mathcal{R}}(\tilde{\lambda}_{\mathcal{R}}) = C\tilde{\lambda}_{\mathcal{R}}^{-\alpha}$, then for $\alpha \neq 2$,

$$G_E(\Lambda, b) = \frac{C\left[(b\Lambda)^{2-\alpha} - \Lambda^{2-\alpha}\right]}{2-\alpha}, \tag{55}$$

while for $\alpha = 2$,

$$G_E(\Lambda, b) = C \log b. \tag{56}$$

The last expression is independent of Λ. That is the cleanest statement of spectral scale balance: a fixed-ratio band carries the same spectral energy no matter where the band is placed in the tail.

We therefore use the spectral shell gain itself as the running coupling,

$$g_E(\tilde{\lambda}_{\mathcal{R}}) := \frac{d\widetilde{E}}{d\log\tilde{\lambda}_{\mathcal{R}}} = \tilde{\lambda}_{\mathcal{R}}^2 \widehat{\rho}_{\mathcal{R}}(\tilde{\lambda}_{\mathcal{R}}). \tag{57}$$

This derivative is taken within one fixed checkpoint while neighboring spectral bands are compared.

The local slope is a scale derivative. It is not a training-time derivative. It compares neighboring eigenvalue bands inside one checkpoint. In RG language, this scale derivative is called a *beta function*. Define

$$\alpha_{\mathrm{loc}}(\tilde{\lambda}_{\mathcal{R}}) = -\frac{d\log\widehat{\rho}_{\mathcal{R}}(\tilde{\lambda}_{\mathcal{R}})}{d\log\tilde{\lambda}_{\mathcal{R}}}. \tag{58}$$

Then

$$s(\tilde{\lambda}_{\mathcal{R}}) = \frac{d}{d\log\tilde{\lambda}_{\mathcal{R}}} \log\left[\tilde{\lambda}_{\mathcal{R}}^2 \widehat{\rho}_{\mathcal{R}}(\tilde{\lambda}_{\mathcal{R}})\right] = 2 - \alpha_{\mathrm{loc}}(\tilde{\lambda}_{\mathcal{R}}), \tag{59}$$

and equivalently

$$\frac{dg_E}{d\log\tilde{\lambda}_{\mathcal{R}}} = \left[2 - \alpha_{\mathrm{loc}}(\tilde{\lambda}_{\mathcal{R}})\right] g_E(\tilde{\lambda}_{\mathcal{R}}). \tag{60}$$

For an exact power law on the fitting window, we can then define *the leading, or tree-level, RG beta function*

$$\beta_E(g_E) \equiv \frac{dg_E}{d\log\tilde{\lambda}_{\mathcal{R}}} = (2-\alpha)g_E \tag{61}$$

where the associated scaling dimension is

$$y_E = 2 - \alpha. \tag{62}$$

The sign of y_E determines whether spectral energy is relevant, irrelevant, or marginal under rescaling. This scale-counting argument identifies the marginal boundary, but it does not by itself prove a complete Wilsonian fixed point.

The fixed-point language should therefore be read operationally. A fixed point is not a single training checkpoint. It is a SETOL-normalized spectral shape that is stable after the final decimation and trace-log testing.

The weak-tail branch is the trivial solution in which no extensive retained ECS survives the coarse graining:

$$\mathfrak{F}_0: \qquad w_{\mathcal{R}} := \frac{M_{\mathcal{R}}}{M} \longrightarrow 0. \qquad g_E \text{ undefined} \tag{63}$$

The correlated solution is the ERG fixed-point candidate,

$$\mathfrak{F}_{\mathrm{ERG}}: \qquad 0 < g_E^\star < \infty, \quad \alpha \simeq 2, \quad \mathcal{R}_{\mathrm{ECS}}^\star \simeq \mathcal{T}_{\mathrm{PL}}^\star. \tag{64}$$

Power-law scale invariance alone does not uniquely select $\alpha = 2$; what distinguishes $\alpha^\star \simeq 2$ is that the spectral coupling is marginal.

This is the RG critical boundary. The scale exponent $y_E = 2-\alpha$ changes sign there, so $\alpha = 2$ plays the role of the critical exponent for this HTSR spectral universality class. Thus $\alpha = 2$ is the tree-level scale-balance condition for logarithmic bands, while $\mathfrak{F}_{\mathrm{ERG}}$ is the empirical correlated fixed-point candidate.

Returning to the earlier discussion, notice that the HTSR boundary between self-averaging and non-self-averaging phases occurs at the same exponent as the RG scale-balance boundary. That is why $\alpha = 2$ is not merely a fitted number. It is the proposed operational critical point where a layer has learned as much correlation structure as possible before overfitting.

By leveraging this RG spectral view, one can now design optimizers that approximate the RG flow; the next section explores this.

5.6 Removing Redundant RG Directions

A Wilsonian RG flow is a flow on equivalence classes, not on every microscopic coordinate. Field redefinitions, coordinate changes, and normalization conventions can move the representative of a theory without changing its long-distance content. These motions are the *redundant directions* of RG theory. They are distinct from irrelevant directions: an irrelevant perturbation changes the theory but decays under coarse graining, whereas a redundant perturbation changes only its description. In the retained spectral description, the physical state is therefore the trace-log-gauge-fixed ECS modulo transformations that leave its quotient spectral shape unchanged.

For a small displacement $\delta\mathbf{W}$, two redundancies are relevant:

- **Trace-log (volume) drift.** The update changes the retained product scale and leaves the chosen gauge slice, while leaving the normalized spectral shape essentially unchanged.
- **Isospectral coordinate motion.** The update changes retained coordinates by an orthogonal similarity transformation without changing the retained eigenvalue multiset.

The remaining component is the candidate physical RG motion: it can redistribute spectral mass, change the tail burden, or change the fitted exponent α.

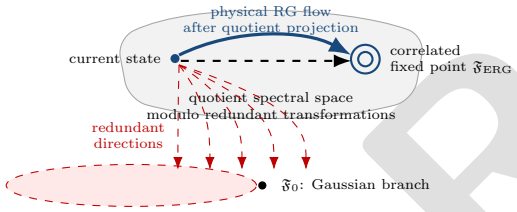

Figure 10. Redundant directions in the spectral RG flow. A raw update contains physical and redundant components. Projection by $\mathbf{I} - P_{\mathrm{red}}$ removes motion inside the redundant fiber and leaves the quotient-space flow toward $\mathfrak{F}_{\mathrm{ERG}}$.

Let the abstract coarse-graining map be

$$\Pi_{\mathrm{RG}} : \mathbf{W} \longmapsto [\widetilde{\mathbf{X}}_{\mathcal{R}}], \tag{65}$$

where brackets denote the equivalence class after gauge fixing. A redundant direction $\mathbf{Z}$ lies in the kernel of the linearized map,

$$D\Pi_{\mathrm{RG}}(\mathbf{W})[\mathbf{Z}] = 0. \tag{66}$$

Thus every infinitesimal update decomposes as

$$\delta\mathbf{W} = \delta\mathbf{W}_{\mathrm{phys}} + \delta\mathbf{W}_{\mathrm{red}}, \qquad \delta\mathbf{W}_{\mathrm{red}} \in \ker D\Pi_{\mathrm{RG}}(\mathbf{W}), \tag{67}$$

and the quotient representative is

$$\delta\mathbf{W}_{\mathrm{RG}} = (\mathbf{I} - P_{\mathrm{red}})\delta\mathbf{W}. \tag{68}$$

By construction, this removal must preserve the physical infinitesimal motion,

$$D\Pi_{\mathrm{RG}}(\mathbf{W})[\delta\mathbf{W}_{\mathrm{RG}}] = D\Pi_{\mathrm{RG}}(\mathbf{W})[\delta\mathbf{W}]. \tag{69}$$

The trace-log condition provides a convenient gauge slice,

$$T_{\mathcal{R}}(\mathbf{W}) := \mathrm{Tr}\log\widetilde{\mathbf{X}}_{\mathcal{R}} = 0. \tag{70}$$

A projected vector must remain tangent to this slice to first order,

$$DT_{\mathcal{R}}(\mathbf{W})[\delta\mathbf{W}_{\mathrm{RG}}] = 0, \tag{71}$$

or equivalently $T_{\mathcal{R}}(\mathbf{W} + \epsilon\delta\mathbf{W}_{\mathrm{RG}}) = O(\epsilon^2)$. This removes scale drift normal to the slice, but it is not the full quotient condition: retained-coordinate transformations can satisfy $DT_{\mathcal{R}} = 0$ and still be redundant. Abstractly, if $\{\mathbf{Z}_\mu\}$ spans the chosen redundant subspace and

$$H_{\mu\nu} = \langle \mathbf{Z}_\mu, \mathbf{Z}_\nu \rangle_F, \tag{72}$$

then the local orthogonal projector is

$$P_{\mathrm{red}}\delta\mathbf{W} = \sum_{\mu,\nu} \mathbf{Z}_\mu (H^\dagger)_{\mu\nu} \langle \mathbf{Z}_\nu, \delta\mathbf{W} \rangle_F. \tag{73}$$

Writing $\mathbf{G}_T := \nabla_{\mathbf{W}} T_{\mathcal{R}}$, the normal trace-log component is

$$\delta\mathbf{W}_{\log} = \frac{\langle \delta\mathbf{W}, \mathbf{G}_T \rangle_F}{\langle \mathbf{G}_T, \mathbf{G}_T \rangle_F} \mathbf{G}_T. \tag{74}$$

Subtracting it produces a vector tangent to the gauge slice. This is necessary but not sufficient. If retained coordinates are rotated by

$$\mathbf{O}(\epsilon) = e^{\epsilon\mathbf{A}}, \qquad \mathbf{A}^\top = -\mathbf{A}, \tag{75}$$

then

$$\widetilde{\mathbf{X}}_{\mathcal{R}} \longmapsto \mathbf{O}(\epsilon)^\top \widetilde{\mathbf{X}}_{\mathcal{R}} \mathbf{O}(\epsilon), \qquad \delta\widetilde{\mathbf{X}}_{\mathcal{R}} = [\widetilde{\mathbf{X}}_{\mathcal{R}}, \mathbf{A}]. \tag{76}$$

This motion preserves both $\mathrm{spec}(\widetilde{\mathbf{X}}_{\mathcal{R}})$ and $T_{\mathcal{R}}$, so it lies inside the gauge slice while remaining redundant. The quotient projector must therefore remove both the normal scale direction and the selected tangent generators of these isospectral orbits.

The interpretation is geometric. The trace-log condition chooses a local section through a bundle of equivalent spectral descriptions; P_{red} removes the vertical component of the raw vector field; and $\delta\mathbf{W}_{\mathrm{RG}}$ is its horizontal representative. Fixed points should consequently be defined in the quotient: a representative may continue to move by a redundant transformation even when its physical RG state has stopped evolving. Conversely, satisfying $T_{\mathcal{R}} = 0$ does not by itself identify a fixed point, because substantial physical shape flow can remain tangent to that slice.

This subsection states only that RG geometry. A finite optimizer step generally leaves the slice at second order and can change the retained support. It therefore requires a retraction, damping, acceptance safeguards, and recomputation of the ECS. Those implementation choices, including how the redundant basis is approximated without overwhelming the supervised optimizer trajectory, are deferred to Section 6.

6 Optimizer Extensions

The base optimizer supplies the supervised training trajectory. The RG construction does not replace that dynamics; it acts on the completed optimizer step and nudges only the retained spectral coordinates associated with redundant RG motion. Any correction should therefore remain local, damped, and subordinate to the loss-driven update.

RG-aware correction

For each proposed optimizer update $\Delta\mathbf{W}_t$, select the working ECS (i.e by the midpoint rule)

$$M_{\mathcal{R}_t} = \left\lfloor \frac{M_{\mathrm{PL},t} + M_{\mathrm{TL},t}}{2} \right\rfloor, \qquad \mathcal{R}_t = \{(1), \ldots, (M_{\mathcal{R}_t})\}.$$

Use the off-gauge residual

$$T_t(\mathbf{W}_t) = \operatorname{Tr}\log \overline{\mathbf{X}}_{\mathcal{R}_t}(\mathbf{W}_t)$$

to nudge the proposed update toward

$$\operatorname{Tr}\log \widetilde{\mathbf{X}}_{\mathcal{R}_t} = 0, \qquad \alpha_{\mathcal{R}_t} = 2,$$

and remove redundent isospectral motion

$$\overline{\mathbf{X}}_{\mathcal{R}_t} \longmapsto \mathbf{O}^\top \overline{\mathbf{X}}_{\mathcal{R}_t} \mathbf{O}, \qquad \mathbf{O}(\epsilon) = e^{\epsilon \mathbf{A}}, \qquad \mathbf{A}^\top = -\mathbf{A}.$$

Nudge the optimizer trajectory toward the RG flow.

The construction summarized above has three parts. First, we select a working retained space between the trace-log ECS boundary and the fitted power-law boundary. The midpoint rule is a conservative finite-size choice: it avoids relying on either noisy boundary alone while retaining modes that are plausibly both correlated and described by the fitted tail. Second, the off-gauge trace-log residual measures motion in the retained log-volume direction. Contraction along this direction drives the correlated sector back toward the weak-tail trivial fixed point $\mathfrak{F}_0$; suppressing that component prevents this redundant return flow, while the spectral shape is nudged toward the marginal state $\alpha = 2$. Finally, one would also like to remove isospectral rotations within the retained space. This step is harder because a rotation that leaves one layer's ESD unchanged need not leave the full network function unchanged unless neighboring layers transform compatibly.

The first operation is a one-dimensional projection fixed by the trace-log. The second is a larger quotient-space problem: an isospectral rotation of one layer is not automatically redundant for the complete network.

6.1 RG-guided optimizer flow

Let the base optimizer propose

$$\widehat{\mathbf{W}}_{t+1} = \mathrm{Proposal}_t(\mathbf{W}_t, \mathcal{H}_t), \qquad \Delta\mathbf{W}_t = \widehat{\mathbf{W}}_{t+1} - \mathbf{W}_t, \tag{77}$$

where $\mathcal{H}_t$ is its internal state. The RG extension acts on this finite displacement after momentum, preconditioning, clipping, weight decay, and the other microscopic optimizer rules have already been applied. Its purpose is not to replace the supervised optimizer, but to nudge the proposed displacement and the resulting optimizer trajectory toward the candidate spectral RG flow.

1. Select a working ECS and measure its off-gauge trace-log. At step t, select a working ECS candidate $\mathcal{R}_t$ from the spectrum of the current layer (or, in a finite-step implementation, from the base proposal), and let $\mathbf{V}_{\mathcal{R}_t}$ contain its retained right singular vectors. This selection identifies the retained space on which the RG diagnostic is evaluated. It does *not* assert that the retained operator already satisfies the trace-log condition. During the local differential calculation, $\mathcal{R}_t$ and $\mathbf{V}_{\mathcal{R}_t}$ are held fixed. The equations below linearize at $\mathbf{W}_t$; a post-proposal implementation may instead evaluate the same residual and gradient at $\widehat{\mathbf{W}}_{t+1}$. Neither choice assumes that the evaluated state lies on the trace-log target.

Define the pre-gauge retained operator

$$\overline{\mathbf{X}}_{\mathcal{R}_t}(\mathbf{W}) := \mathbf{V}_{\mathcal{R}_t}^\top \left(\frac{1}{N} \mathbf{W}^\top \mathbf{W} \right) \mathbf{V}_{\mathcal{R}_t}, \qquad M_{\mathcal{R}_t} = |\mathcal{R}_t|, \tag{78}$$

and its trace-log residual

$$T_t(\mathbf{W}) := \operatorname{Tr}\log \overline{\mathbf{X}}_{\mathcal{R}_t}(\mathbf{W}), \qquad \mathcal{G}_t := \{\mathbf{W} : T_t(\mathbf{W}) = 0\}. \tag{79}$$

The set $\mathcal{G}_t$ is the candidate trace-log target for the fixed working support. The actual retained state generally lies off this target: $T_t(\mathbf{W}_t)$ and $T_t(\widehat{\mathbf{W}}_{t+1})$ need not vanish. Their nonzero values are the control signal used to measure how the external optimizer trajectory departs from the proposed RG flow.

Writing

$$\mathbf{G}_{T,t} := \nabla_{\mathbf{W}} T_t(\mathbf{W}_t), \tag{80}$$

the fixed-support gradient is

$$\mathbf{G}_{T,t} = \frac{2}{N} \mathbf{W}_t \mathbf{V}_{\mathcal{R}_t} \overline{\mathbf{X}}_{\mathcal{R}_t}(\mathbf{W}_t)^{-1} \mathbf{V}_{\mathcal{R}_t}^\top, \tag{81}$$

provided the retained operator is positive definite. Because $\overline{\mathbf{X}}_{\mathcal{R}_t}$ is the pre-gauge operator rather than a pointwise normalized one, T_t is not identically zero and $\mathbf{G}_{T,t}$ is generally nonzero. If $\langle \mathbf{G}_{T,t}, \mathbf{G}_{T,t} \rangle_F$ falls below a numerical tolerance, the correction is skipped or regularized.

2. Nudge the proposed update toward the trace-log target. Let

$$d_t := DT_t(\mathbf{W}_t)[\Delta\mathbf{W}_t] = \langle \mathbf{G}_{T,t}, \Delta\mathbf{W}_t \rangle_F. \tag{82}$$

A tangent projection would impose $DT_t[\Delta\mathbf{W}] = 0$ and would therefore preserve the current residual to first order. That is not the desired operation when $T_t(\mathbf{W}_t) \neq 0$. Instead, choose a tracking rate $0 \leq \gamma_{T,t} \leq 1$ and replace the normal component of the base update by a component

that reduces the current residual:

$$\Delta\mathbf{W}_t^{\text{track}} := \Delta\mathbf{W}_t - \frac{d_t + \gamma_{T,t}T_t(\mathbf{W}_t)}{\langle \mathbf{G}_{T,t}, \mathbf{G}_{T,t}\rangle_F}\mathbf{G}_{T,t},$$
$$DT_t(\mathbf{W}_t)[\Delta\mathbf{W}_t^{\text{track}}] = -\gamma_{T,t}T_t(\mathbf{W}_t). \tag{83}$$

Thus the correction does not assume that the current retained ECS is already on the trace-log surface. It uses the off-gauge residual to steer the optimizer trajectory toward that surface. When $T_t(\mathbf{W}_t) = 0$, Eq. (83) reduces to the ordinary tangent projection as a limiting case.

Early in training it may be preferable to use only a one-sided branch-protection rule rather than active tracking. With

$$a_t = \frac{d_t}{\langle \mathbf{G}_{T,t}, \mathbf{G}_{T,t}\rangle_F}, \qquad a_t^- := \min(a_t, 0), \tag{84}$$

use

$$\Delta\mathbf{W}_t^{\not\to\mathfrak{F}_0} = \Delta\mathbf{W}_t - a_t^-\mathbf{G}_{T,t}. \tag{85}$$

This cancels only first-order contraction of the retained log-volume; updates that expand it are unchanged. It is a safeguard against motion toward $\mathfrak{F}_0$, not an assertion that $T_t = 0$.

3. Remove only validated isospectral coordinate motion. The trace-log correction above is a steering term. It should be distinguished from quotienting genuinely redundant coordinate motion. Let $\mathbf{A}^\top = -\mathbf{A}$ and $\mathbf{O}(\epsilon) = e^{\epsilon\mathbf{A}}$. Then

$$\overline{\mathbf{X}}_{\mathcal{R}_t} \longmapsto \mathbf{O}(\epsilon)^\top\overline{\mathbf{X}}_{\mathcal{R}_t}\mathbf{O}(\epsilon), \qquad \delta\overline{\mathbf{X}}_{\mathcal{R}_t} = [\overline{\mathbf{X}}_{\mathcal{R}_t}, \mathbf{A}]. \tag{86}$$

This motion preserves the retained eigenvalues, the ESD, and T_t, whether or not T_t is zero. It is therefore tangent to a fixed trace-log level set. Removing it requires tangent generators in addition to the scalar trace-log tracking correction.

For a selected weight-space basis $\{\mathbf{Z}_{A_j,t}\}_{j=1}^{m_t}$ of candidate isospectral generators, define

$$H_{ij,t} = \langle \mathbf{Z}_{A_i,t}, \mathbf{Z}_{A_j,t}\rangle_F,$$
$$\mathbf{P}_{\text{iso},t}\Delta\mathbf{W} = \sum_{i,j=1}^{m_t}\mathbf{Z}_{A_i,t}(H_t^+)_{ij}\langle \mathbf{Z}_{A_j,t}, \Delta\mathbf{W}\rangle_F. \tag{87}$$

The RG-nudged update is then

$$\Delta\mathbf{W}_t^{\text{RG}} = (\mathbf{I} - \mathbf{P}_{\text{iso},t})\Delta\mathbf{W}_t^{\text{track}}. \tag{88}$$

Because an exact isospectral generator satisfies $DT_t(\mathbf{W}_t)[\mathbf{Z}_{A_j,t}] = 0$, this quotient step preserves the first-order trace-log tracking condition in Eq. (83).

This construction is potentially implementable, but harder for three reasons. The full isospectral tangent space has dimension $M_{\mathcal{R}_t}(M_{\mathcal{R}_t} - 1)/2$; its generators can be nearly dependent; and rotating the retained coordinates of one layer generally changes the inputs presented to the next layer. Such a direction is redundant for the network only when neighboring layers transform compatibly or a direct functional test shows that the network output is unchanged. A practical approximation therefore samples only a small, well-conditioned family with

$$m_t \ll M_{\mathcal{R}_t}^2, \tag{89}$$

and retains only generators whose effect is stable under resampling.

The tangent part needs an additional network-level test. An ESD regards retained rotations as equivalent, but the full network generally does not: a right-coordinate rotation of one layer changes the coordinates passed to the next. A sampled generator should therefore be paired with a compensating neighboring-layer transformation or rejected unless, on a small calibration batch $\{x\}$,

$$\frac{\|f_{\mathbf{W}+\epsilon\mathbf{Z}_{A,t}}(x) - f_{\mathbf{W}}(x)\|}{\epsilon\|\mathbf{Z}_{A,t}\|_F} \le \varepsilon_{\text{func}}. \tag{90}$$

Only generators that pass this functional check should enter the sampled projector.

4. Take a finite step, measure the residual, and optionally retract partway. A damped finite candidate is

$$\mathbf{W}_{t+1}^{\text{cand}} = \mathbf{W}_t + \rho_t\Delta\mathbf{W}_t^{\text{RG}}, \qquad 0 < \rho_t \le 1. \tag{91}$$

For the fixed working support, define the base and corrected residuals

$$\tau_t^{\text{base}} := T_t(\widehat{\mathbf{W}}_{t+1}), \qquad \tau_t^{\text{cand}} := T_t(\mathbf{W}_{t+1}^{\text{cand}}). \tag{92}$$

The corrected residual is generally nonzero. Indeed, that is expected away from the endpoint: the purpose of the extension is to nudge the optimizer trajectory toward the RG target, not to assume that every intermediate retained ECS already lies on it. At fixed support, the first-order relation is

$$T_t(\mathbf{W}_{t+1}^{\text{cand}}) = \big(1 - \rho_t\gamma_{T,t}\big)T_t(\mathbf{W}_t) + \mathcal{O}(\rho_t^2). \tag{93}$$

If a finite spectral restoration is desired, let $\overline{\lambda}_i$ denote the retained eigenvalues of $\overline{\mathbf{X}}_{\mathcal{R}_t}(\mathbf{W}_{t+1}^{\text{cand}})$ and define

$$\mathbf{X}_{\mathcal{R}_t}^{(\kappa_t)} := \exp\left(-\kappa_t\frac{\tau_t^{\text{cand}}}{M_{\mathcal{R}_t}}\right)\overline{\mathbf{X}}_{\mathcal{R}_t}(\mathbf{W}_{t+1}^{\text{cand}}),$$
$$\lambda_i^{(\kappa_t)} := \overline{\lambda}_i\exp\left(-\kappa_t\frac{\tau_t^{\text{cand}}}{M_{\mathcal{R}_t}}\right), \qquad 0 \le \kappa_t \le 1. \tag{94}$$

Then

$$\operatorname{Tr}\log\mathbf{X}_{\mathcal{R}_t}^{(\kappa_t)} = (1 - \kappa_t)\tau_t^{\text{cand}}. \tag{95}$$

Thus $\kappa_t < 1$ is a partial nudge and does not place the actual retained operator on the trace-log surface. Only the exact retraction $\kappa_t = 1$ produces the gauge-fixed representative, for which one may write

$$\widetilde{\mathbf{X}}_{\mathcal{R}_t} := \mathbf{X}_{\mathcal{R}_t}^{(1)}, \qquad \operatorname{Tr}\log\widetilde{\mathbf{X}}_{\mathcal{R}_t} = 0. \tag{96}$$

The tilde is therefore reserved for the result of an exact retraction (or the limiting fixed-point representative), not

for the pre-gauge operator that is differentiated during training.

After the finite correction, the ECS is recomputed. Define the post-correction raw residual on the recomputed support by

$$\tau_t^+ := \operatorname{Tr}\log \overline{\mathbf{X}}_{\mathcal{R}_t^+}(\mathbf{W}_{t+1}^{\mathrm{cand}}). \tag{97}$$

The correction should be accepted only when it remains local. For minibatch loss L_t and recomputed support $\mathcal{R}_t^+$, useful checks are

$$\begin{aligned} &L_t(\mathbf{W}_{t+1}^{\mathrm{cand}}) \le L_t(\widehat{\mathbf{W}}_{t+1}) + \delta_{L,t}, \\ &\frac{|\mathcal{R}_t^+ \cap \mathcal{R}_t|}{|\mathcal{R}_t^+ \cup \mathcal{R}_t|} \ge s_{\min}, \qquad |\tau_t^+| \le |\tau_t^{\mathrm{base}}| + \delta_{T,t}. \end{aligned} \tag{98}$$

Failure of any test reduces ρ_t, $\gamma_{T,t}$, or κ_t, or skips the RG correction. These safeguards prevent the spectral nudge from overwhelming supervised descent or jumping across a moving ECS boundary.

Operationally, the proposed extension is

$$\begin{aligned} \text{Propose} \to \text{select } \mathcal{R}_t \to \text{measure } T_t, \\ \to \text{nudge} \to \text{validated quotient} \to \text{recompute}. \end{aligned} \tag{99}$$

The target surface must not be confused with the actual retained ECS encountered along training. At a generic optimizer step,

$$\operatorname{Tr}\log \overline{\mathbf{X}}_{\mathcal{R}_t}(\mathbf{W}_t) \neq 0 \qquad \text{in general.} \tag{100}$$

That nonzero residual is the point of the construction: it tells the RG extension how to alter $\Delta\mathbf{W}_t$ so that the external optimizer trajectory moves toward, rather than being assumed to lie on, the candidate RG flow. The zero trace-log condition is imposed only by an exact retraction or reached asymptotically at the proposed fixed point.

A staged implementation would monitor the raw trace-log residual, the overlap of successive ECS supports, the supervised loss, and the functional residual of each sampled generator. Early in training, only the one-sided contraction safeguard may be used. After the retained support stabilizes, ρ_t, $\gamma_{T,t}$, and κ_t may be increased gradually and a small isospectral basis may be sampled. If any diagnostic deteriorates, the method falls back to the raw optimizer proposal. Muon and WW-PGD below are therefore partial approximations, not implementations of the complete trace-log-tracking plus validated-isospectral quotient construction.

6.2 Muon: a crude approximation

Muon is a matrix-aware optimizer for hidden-layer weights [28, 38, 29]. It maintains a momentum-like matrix update and approximately orthogonalizes it with a short Newton–Schulz iteration before applying the learning rate; non-matrix parameters are normally handled by AdamW. The motivation in Keller Jordan's blog is spectral: ordinary adaptive updates can drive hidden-layer matrices toward undesirable low-rank structure, so Muon replaces the anisotropic singular spectrum of the raw update by an approximately isotropic one. This is effective empirically, but it is not derived from the RG quotient geometry and does not identify which directions are actually redundant.

Let the momentum-like update have thin SVD

$$\mathbf{G}_t = \mathbf{U}_t \mathbf{\Sigma}_t \mathbf{V}_t^\top. \tag{101}$$

The polar idealization of Muon replaces it by

$$\Delta\mathbf{W}_t = \operatorname{Ortho}(\mathbf{G}_t) \simeq \mathbf{U}_t \mathbf{V}_t^\top. \tag{102}$$

Consequently,

$$\begin{aligned} \Delta\mathbf{W}_t^\top \Delta\mathbf{W}_t &\simeq \mathbf{V}_t \mathbf{V}_t^\top, \\ \operatorname{spec}_+(\Delta\mathbf{W}_t^\top \Delta\mathbf{W}_t) &= \{1, \dots, 1\}. \end{aligned} \tag{103}$$

so the active update singular values share one scale and

$$\operatorname{Tr}_+ \log(\Delta\mathbf{W}_t^\top \Delta\mathbf{W}_t) \simeq 0. \tag{104}$$

The learning rate remains outside this normalization,

$$\mathbf{W}_{t+1} = \mathbf{W}_t - \eta_t \Delta\mathbf{W}_t. \tag{105}$$

Muon therefore removes anisotropic scale and conditioning from the *update*, which may suppress excessive motion along a few singular directions. But it does not select an ECS, compute the trace-log normal $\mathbf{G}_T$, identify motion toward $\mathfrak{F}_0$, remove tangent isospectral generators, or restore the trace-log of the layer state. It can remove some redundant motion, but it also changes physical components because it does not know the quotient decomposition.

Muon should thus be interpreted as a crude operational test: broad update isotropization can change the spectral route followed during training, but it is not an implementation of the sampled isospectral quotient in Eq. (87) and does not establish exact RG projection or critical slowing down.

6.3 WW-PGD: Projected Gradient Descent

WeightWatcher-guided Projected Gradient Dynamics (WW-PGD) is the first-pass implementation designed to test the operational RG idea more directly. It does not compute the sampled isospectral quotient of Eq. (87), and it does not attempt to remove the complicated isospectral directions tangent to fixed trace-log level sets. Instead, the base optimizer proposes a finite layer matrix, after which WW-PGD modifies a selected spectral window. The method removes or restores one scalar retained-volume coordinate and simultaneously nudges the remaining shape coordinates toward the marginal $\alpha \simeq 2$ rank-order form.

In the branch picture, this scalar direction is the trace-log drift toward the chosen trace-log target. One sign of that drift reduces retained correlated volume and points back toward the weak-tail or trivial branch $\mathfrak{F}_0$; the opposite sign can overconcentrate the retained spectrum. WW-PGD controls this one-dimensional volume motion while leaving the harder isospectral quotient problem for future work. It is therefore a partial test of the proposal, not a realization of the complete RG optimizer.

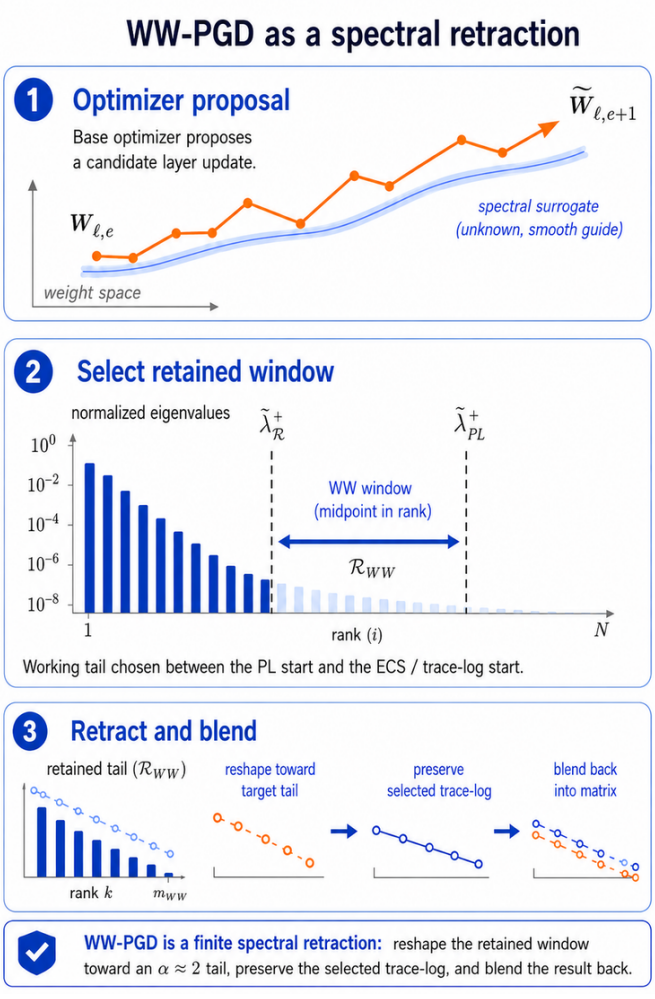

Figure 11. WW-PGD as an approximate finite spectral correction. The base optimizer proposes a candidate layer. WW-PGD selects a window between the fitted PL start and the ECS/trace-log start, reshapes it toward an $\alpha \simeq 2$ target, restores the selected trace-log before blending, and mixes the result back into the optimizer trajectory.

For epoch or outer step e, write

$$\widehat{\mathbf{W}}_{\ell,e+1} = \mathcal{O}^{(e)}_{\text{base}}(\mathbf{W}_{\ell,e}), \qquad \mathbf{W}_{\ell,e+1} = \mathcal{R}_{\text{WW}}(\widehat{\mathbf{W}}_{\ell,e+1}). \tag{106}$$

The SVD and spectral diagnostics of the proposed layer are then evaluated using the conventions already established earlier in the paper. Let m_{PL} and $m_{\mathcal{R}}$ denote the numbers of eigenvalues retained by the fitted PL tail and by the trace-log ECS. WW-PGD chooses the midpoint in rank space,

$$m_{\text{WW}} = \left\lfloor \frac{m_{\text{PL}} + m_{\mathcal{R}}}{2} \right\rfloor, \qquad \mathcal{R}_{\text{WW}} = \{(1), \ldots, (m_{\text{WW}})\}. \tag{107}$$

The correction is skipped when this window is too small or unstable.

Inside $\mathcal{R}_{\text{WW}}$, rank k is assigned the target

$$\mu_k = k^{-q}, \qquad q = 1, \qquad \alpha_{\text{target}} = 1 + q^{-1} = 2. \tag{108}$$

The target $\tilde{\lambda}^{\star}_k = A\mu_k$ uses an amplitude that preserves the current selected trace-log,

$$A = \exp\left[\frac{\sum_{i\in\mathcal{R}_{\text{WW}}} \log \tilde{\lambda}_i - \sum_{k=1}^{m_{\text{WW}}} \log \mu_k}{m_{\text{WW}}} \right]. \tag{109}$$

Thus the target changes the rank-order shape without independently imposing a new product scale.

The move is deliberately soft. With homotopy $h_e \in [0,1]$, define

$$\begin{aligned} g_k &= \log \tilde{\lambda}_k - \log \tilde{\lambda}^{\star}_k, \\ r_k &= \text{clip}\left(\frac{1 - \eta_C h_e g_k}{1 + \eta_C h_e g_k}, r_{\min}, r_{\max} \right), \end{aligned} \tag{110}$$

with provisional values

$$\hat{\tilde{\lambda}}_k = \tilde{\lambda}_k r_k. \tag{111}$$

The selected trace-log is then restored before reconstruction,

$$\tilde{\lambda}^{\text{ret}}_k = \hat{\tilde{\lambda}}_k \exp\left[\frac{T_{\text{WW}} - \sum_{j=1}^{m_{\text{WW}}} \log \hat{\tilde{\lambda}}_j}{m_{\text{WW}}} \right], \tag{112}$$

$$T_{\text{WW}} = \sum_{i\in\mathcal{R}_{\text{WW}}} \log \tilde{\lambda}_i. \tag{113}$$

In log-eigenvalue coordinates, Eq. (112) removes exactly one scalar component: the uniform retained log-volume direction. The remaining $m_{\text{WW}} - 1$ coordinates describe spectral shape. This is the direct connection to the redundant-direction discussion: WW-PGD controls one normal gauge

coordinate but does not project the tangent isospectral orbit.

The retained singular values are reconstructed through their scale-free ratios, producing $\mathbf{W}^{\text{shaped}}$. The final state is blended with the base proposal,

$$\mathbf{W}^{+} = (1 - \eta_B h_e)\widehat{\mathbf{W}} + \eta_B h_e \mathbf{W}^{\text{shaped}}. \tag{114}$$

The restoration in Eq. (112) is exact on the selected shaped spectrum before this final blend. The complete WW-PGD update is intentionally approximate: after blending, the selected trace-log generally has a residual. This is a feature of the first-pass design. The method nudges the optimizer toward the proposed RG trajectory without forcing every intermediate checkpoint exactly onto the final trace-log slice.

A useful diagnostic is the post-blend residual

$$\varepsilon_T^{+} = \left| \sum_{i \in \mathcal{R}_{\text{WW}}} \log \tilde{\lambda}_i(\mathbf{W}^{+}) - T_{\text{WW}} \right|. \tag{115}$$

Together with the supervised loss, the stability of the selected window, and the movement of the PL and ECS boundaries, this residual can be used to adapt η_B, η_C, and the homotopy schedule. If the correction damages the optimizer trajectory, it should be weakened or skipped. Near a stable endpoint, it may be tightened.

WW-PGD is therefore the first operational experiment on the proposed RG extension. It tests whether removing one retained-volume degree of freedom and steering the remaining spectral shape toward the marginal target can improve training without imposing the full quotient geometry. A more complete optimizer would add the local update-space quotient of Eq. (87), including carefully selected isospectral generators, while retaining the same safeguards against trajectory distortion.

7 Experimental Results

7.1 Muon test

The Muon test asks whether update geometry changes the spectral route. The controlled model is a three-layer MNIST MLP, described in the Appendix A. The comparison holds the data, model, and diagnostics fixed. Only the matrix-valued update rule changes: AdamW [27] is compared to Muon. At checkpoints, `WeightWatcher` fits the HTSR exponent of each fully connected layer

$$\alpha_\ell(t), \qquad \ell \in \{\mathrm{FC1}, \mathrm{FC2}, \mathrm{FC3}\}. \tag{116}$$

Figure 12 compares results for Muon against AdamW [27]; it depicts the test and training accuracies and the layerwise $\alpha_\ell(t)$ per epoch. The top panels compare AdamW and Muon over the first twenty (20) epochs, while the bottom panels track Muon for five-hundred (500) epochs.

Consider the top panels first. AdamW reaches high train and test accuracy more slowly than Muon, but reaches a comparable range after twenty (20) epochs. The AdamW and Muon layerwise $\alpha_\ell(t)$ curves behave very differently over these early epochs. For AdamW, all layer $\alpha_\ell(t)$ values drop quickly. FC2 and FC3 approach the optimal value $\alpha = 2$, but FC1 immediately collapses below 2.0 and remains trapped in the proposed overfitting regime, even as accuracy increases. In contrast, only Muon's FC3 layer approaches $\alpha = 2$ quickly, while FC1 and FC2 fluctuate at first, and then remain near $\alpha = 4$.

The bottom panels show Muon over very long-horizon runs. Neither Muon training accuracy nor test accuracy drops; both increase slowly and asymptotically. Muon FC1 and FC2 track each other, first fluctuating strongly with $\alpha_\ell(t) \in [3, 6]$. Eventually, however, all layer values satisfy $\alpha_\ell(t) \to 2$, again very slowly.

This comparison is intentionally untuned. We made no attempt to optimize learning rates or make Muon win the MNIST benchmark.

In summary, AdamW overfits very quickly, while Muon approaches the RG fixed point asymptotically slowly. This is consistent with Muon removing the RG redundant scale directions, and could be associated with RG *critical slowing down*.

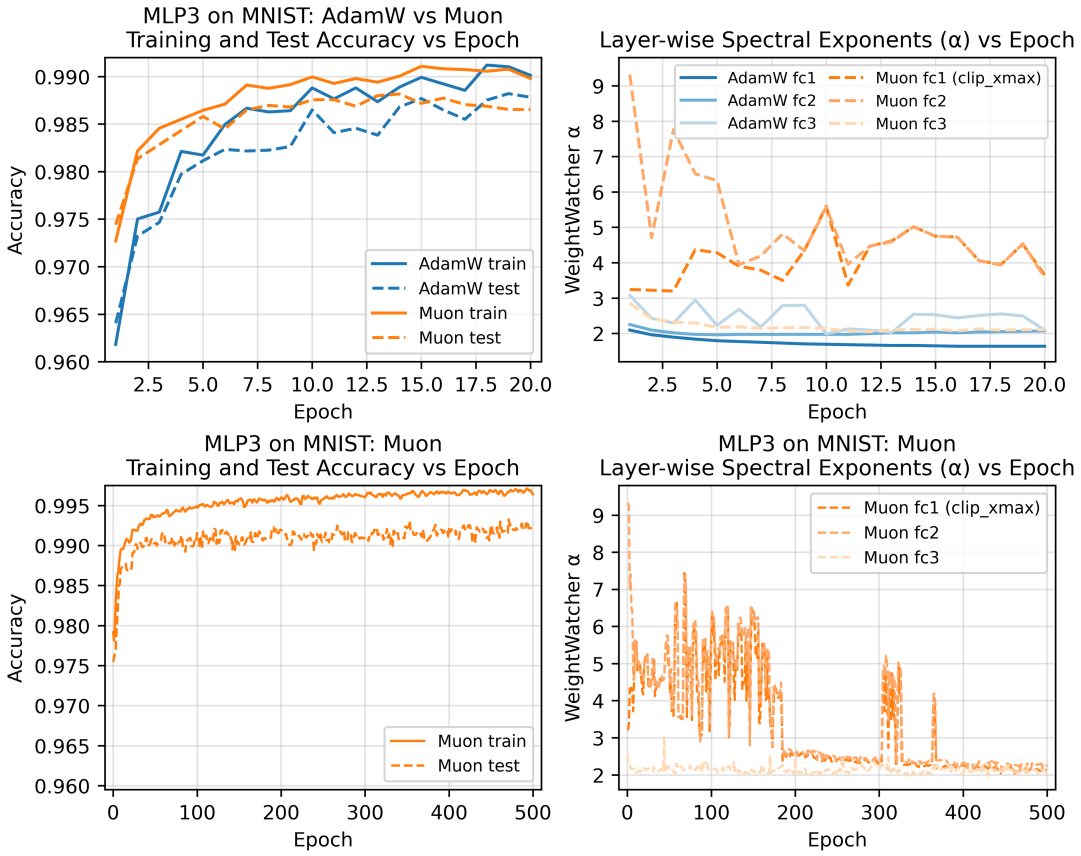

Figure 12. MNIST optimizer comparison: AdamW versus Muon. Same three-layer MNIST MLP trained with AdamW and Muon. Short-horizon panel compares early epochs; long-horizon panel tracks the Muon $\alpha_\ell(t) \to 2$.

7.2 WW-PGD test

Test 1: Fixing AdamW runs. The WW-PGD test asks whether the optimizer can reduce AdamW overfitting by preventing the layer exponents $\alpha_\ell(t)$ from dropping below 2, while simultaneously helping all layers approach the RG fixed point $\alpha \simeq 2$. The experiment is similar to the Muon test (Section 7.1), and is described in Appendix A.

The WW-PGD condition uses the same AdamW mini-batch optimizer, but applies the epoch-boundary retraction in Eq. (106). The experiment runs for 35 epochs and five independent runs. The WW-PGD hyperparameters are $(q = 1;\ \eta_B = 0.5;\ \eta_C = 0.25)$ with zero warmup and a five-epoch homotopy ramp.

Figure 13 shows the main effect. At each epoch, `WeightWatcher` logs the HTSR exponent $\alpha_\ell(t)$. The plotted curves report the mean and standard deviation for each layer across five runs (FC1, FC2, FC3).

Figure 13 shows the layerwise results. AdamW alone lets the $\alpha_\ell(t)$ exponents separate by layer. As in the Muon comparison, the FC1 exponent satisfies $\alpha_\ell(t) < 2$ throughout most of training, while the FC2 exponent satisfies $\alpha_\ell(t) \geq 2$. FC3 satisfies $\alpha_\ell(t) \gg 2$; however, this does not materially change the results.

WW-PGD improves the layer $\alpha_\ell(t)$ for all layers, while keeping the training and test accuracies close to, or only slightly below, the baseline AdamW results.

The convergence to the RG optimal boundary is much faster than with Muon. This is the key takeaway. This makes WW-PGD a direct finite-step test of the spectral-RG optimizer hypothesis.

Test 2: Fixing an anti-grokking checkpoint. As a further stress test, we test whether WW-PGD can repair a deeply overfit checkpoint. Starting from an anti-grokking checkpoint shown in Figure 6 [13], we resume training using AdamW, Muon, and WW-PGD. Figure 14 depicts preliminary experiments [15], showing that the WW-PGD optimizer can recover the test accuracy even after long-horizon overfitting.

Together, these two experiments provide preliminary evidence that WW-PGD repairs the overfit checkpoints tested here while driving all layer $\alpha_\ell(t) \simeq 2$.

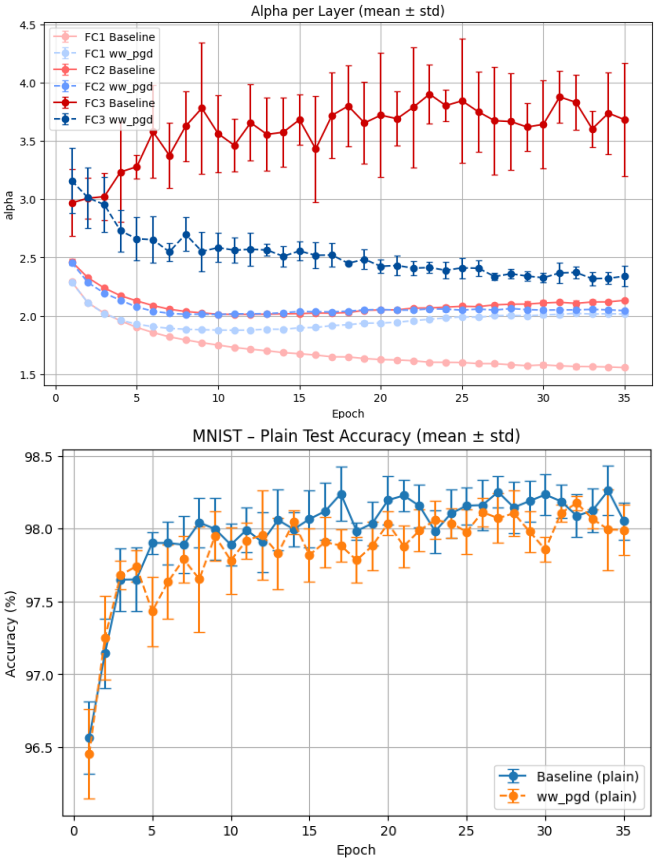

Figure 13. WW-PGD stabilizes layerwise HTSR exponents while preserving test accuracy. A three-layer MNIST MLP is trained with AdamW alone or AdamW plus WW-PGD. Top: layerwise fitted HTSR exponents for FC1–FC3. Bottom: ordinary MNIST test accuracy. Curves show mean ± standard deviation across five runs.

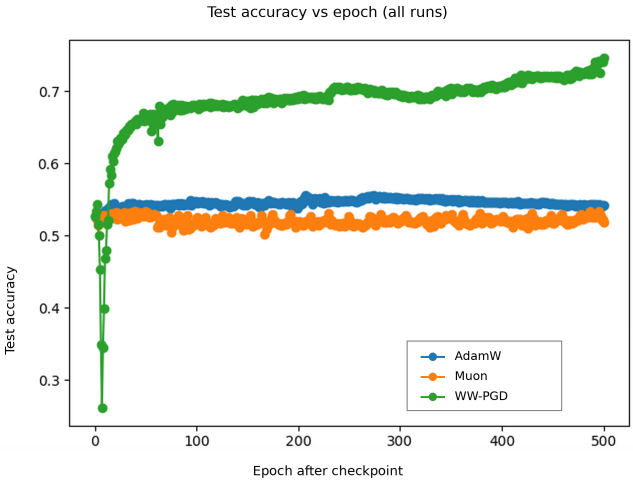

Figure 14. WW-PGD corrects overfit layers. Extending training on an anti-grokking checkpoint from Figure 6. [15]

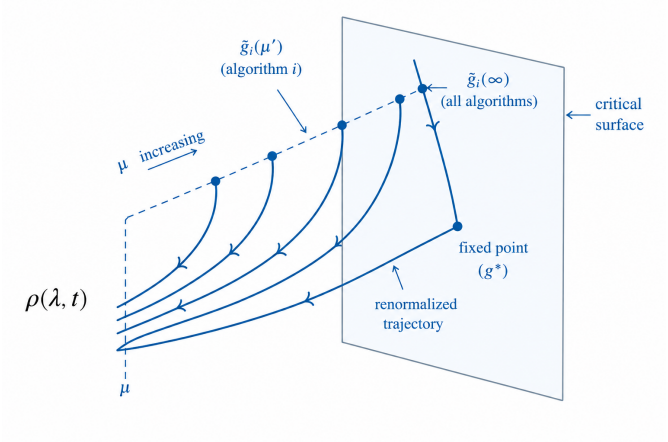

Figure 15. RG flow in the space of effective correlations. Different nonlinear learning algorithms follow different microscopic trajectories, but coarse-graining can suppress irrelevant directions and carry retained spectral structure toward a candidate fixed point. Candidate universal quantities are read off near the fixed point, not from the microscopic update rule.

8 Implications of Universality

It is conjectured that neural networks, boosted trees, and related methods can be grouped under one universality story. They use different update rules, but each can induce an effective correlation structure on the data [22]. Figure 15 represents that structure by effective couplings $\widetilde{g}_i(\mu)$ that flow with RG scale μ. The RG question is whether these spectra approach a common critical surface and candidate fixed point $g^\star$.

Universality of learning algorithms

Universality asks whether different nonlinear learners can flow to the same coarse-grained candidate spectral fixed point.

- Nonlinear learning algorithms can be linearized near their fixed points.
- Spectra, not parameters, may characterize the optimal conditions.
- Self-averaging bulk modes support transfer.
- Non-self-averaging modes reveal dominant atypical structure that causes overfitting.
- XGBoost is a conjectural test case.

RG universality means different algorithms may share a candidate fixed-point spectrum.

The RG claim is local. A learner can be nonlinear away from convergence. Near a stable endpoint, its coarse-grained flow can be linearized in a few spectral coordinates $\widetilde{g}_i$. In Figure 15, the renormalized trajectories bend toward the same critical surface. The stable directions define the universality class, and the fixed point $g^\star$ defines the universal spectral observables.

Near convergence, the invariant object is the spectrum of the retained correlation operator. Parameters change with architecture, but the proposed ERG point may be universal. A non-neural learner needs its own effective correlation operator. Universality becomes testable only after that operator and its $\widetilde{g}_i$ coordinates are defined.

Self-averaging is another proposed natural universal property. The bulk spectral averages should stabilize across samples, initializations, and checkpoints. Non-self-averaging modes should concentrate into a few dominant elements in the spectral tail. Tail modes are not automatically bad. They become risky when a few directions carry too much spectral energy gain, causing correlation traps to form and/or $\alpha < 2$.

Neural networks give the direct case. Different depths, widths, initializations, and optimizers can produce similar endpoint spectra. The shared object is the normalized effective correlation spectrum. This view also reframes optimizer design. A good optimizer should reduce loss while steering the spectrum toward a self-averaging marginal boundary between optimality and overfitting, with $\alpha \simeq 2$ as a candidate universal spectral exponent.

XGBoost gives a plausible extension [34]. Boosting has no dense layer weight matrix, so the current layer-specific theory does not apply directly. Still, each tree changes the residuals and the prediction function. Those updates induce a data-data correlation structure in prediction space. A practical test would form a centered Gram operator from tree contributions or prediction increments, normalize its scale, and track its spectrum across boosting rounds. Early work (not shown) suggests this is possible. Future work will test this construction on XGBoost models.

Universality does not say that all algorithms compute the same function. It says that successful learners may share the same coarse-grained spectral organization. Generalization would then depend less on update details and more on self-averaging, marginality, and controlling the dominant modes to prevent overfitting or collapse. More broadly, universality suggests transferability. In semi-empirical electronic-structure theory, transferable parameters work because different molecules can share the same effective valence-space structure. Martin and Freed tested this idea by deriving semiempirical π-electron parameters from ab initio effective valence-shell Hamiltonians and studying when those parameters transfer across related molecules [31, 32, 33]. The analogy is illustrative. The SETOL effective free energy $\beta\Gamma\frac{\mathrm{IZ}}{Q^2}$ is evaluated by taking a single step of the Wilsonian Exact RG (ERG) and is expressed in terms of the R-Transform of the layer, which is akin to a renormalized self-energy and analogous to the more general effective Hamiltonian approach. Moreover, if learned correlation spectra have candidate universal fixed-point structure, then spectral diagnostics may transfer across related models and tasks. Transferability would then come from the effective correlation space, not from the microscopic implementation.

9 Conclusions

This paper proposes an RG theory of NN layer convergence, integrating the older phenomenological HTSR approach and the more recently developed SETOL theory. HTSR supplies the empirical fact: strong layers often develop heavy-tailed spectra near $\alpha \simeq 2$. SETOL supplies the trace-log condition: remove arbitrary layer scale and study the retained effective correlation space. The main claim is that these two pictures meet at the same spectral endpoint.

The value $\alpha = 2$ has two meanings. Statistically, it marks the boundary between stable bulk-supported learning and domination by a few extreme eigencomponents. RG scale counting gives the same boundary. At $\alpha = 2$, each logarithmic spectral shell contributes comparable spectral energy gain. This agreement makes $\alpha \simeq 2$ a critical boundary, not just a fitted number.

The fixed point is layerwise. Each dense layer has its own retained spectrum, trace-log scale, and dominant-tail burden. A trained network is therefore not one monolithic fixed point. It is a coordinated state in which many layers approach compatible spectral endpoints.

The practical diagnostic is spectral. A good layer should keep useful correlation gain spread across many retained eigencomponents. A bad layer can concentrate too much gain in a few dominant tail modes. The trace-log condition removes scale. The power-law tail measures shape. The dominant-tail eigencomponents test whether the layer remains self-averaging.

The optimizer message is nuanced. Raw optimizers can spend update norm on scale-only motion. A better optimizer should reduce loss while steering the retained spectrum toward the correlated, self-averaging branch. This view suggests why matrix-aware optimizers such as Muon are relevant. It also suggests new optimizer designs based on trace-log geometry.

The `WeightWatcher` message becomes sharper. The HTSR exponent α, the trace-log coordinate, ECS–tail alignment, trap counts, and the tail burden are not isolated heuristics. They can be read as candidate RG observables of the same layerwise spectral flow. They require the weights and the weights alone, without labels, gradients, or test or even training data.

The empirical program is direct. Track these observables across checkpoints. Compare them across optimizers. Ask whether held-out performance improves when the spectrum approaches the marginal boundary and degrades when $\alpha < 2$ and/or traps form. This turns the RG picture into a falsifiable diagnostic program.

The limitations are also direct. This paper does not prove convergence of SGD, AdamW, or Muon to the spectral RG fixed point. It proposes an operational RG interpretation of retained layer spectra and connects that interpretation to HTSR and SETOL theory and validated empirical diagnostics. The evidence supports this picture on the models studied here, but it does not establish universality across all architectures, datasets, optimizers, or nonlinear learners. This remains to be explored.

The broader implication is transferability. Semi-empirical electronic-structure theory uses transferable parameters because related molecules share effective valence-space structure. The analogy here is direct. If learned correlation spectra have candidate universal fixed-point structure, then useful spectral parameters may transfer across related models, tasks, optimizers, and algorithm families.

The final message is simple. Learning is not only loss minimization. Learning is spectral organization. In this picture, strong layers move toward a RG correlated fixed point where scale is removed, the tail is balanced, and the weights remain self-averaging.

The theory in one pass

Dense layers learn by moving their spectra.

- HTSR observes heavy-tailed spectra near $\alpha \simeq 2$.
- SETOL connects heavy-tailed layer spectra to model accuracy (called layer quality), and yields the ERG trace-log laws.
- Self-averaging helps explain why $\alpha = 2$ is a useful boundary.
- RG scale counting explains why $\alpha = 2$ is marginal, suggesting a candidate universal spectral exponent.
- Optimizers should reduce scale-only motion and move toward the ERG fixed point.
- The WW-PGD optimizer extension can repair overfit models.
- The theory may be extensible to XGBoost and related algorithms: $\alpha = 2$ and $\mathrm{Tr} \log \tilde{\mathbf{X}} = 0$.

Keep the learned tail, remove its scale, and test whether the retained layer stays balanced.

Table 4 summarizes the notation and conventions used throughout.

Acknowledgements

The author gratefully acknowledges Hari K. Prakash for valuable discussions, constructive feedback, guidance throughout this work, and for his help in developing the WW-PGD optimizer.

Table 4. Notation and conventions used in this report.

Group	Symbol	Meaning
Layer	$\mathbf{W} \in \mathbb{R}^{N\times M}$, $N \geq M$	Dense-layer weights.
Layer	$\mathbf{X} = N^{-1}\mathbf{W}^T\mathbf{W}$	Layer correlation operator.
Layer	$\lambda_i > 0$, $\mathcal{S}^+ = \{i : \lambda_i > 0\}$	Positive spectral support.
Layer	$\sigma(\mathbf{W}) = \sqrt{N\lambda(\mathbf{X})}$,	Singular values of $\mathbf{W}$ (pre-rescaling).
Layer	$\operatorname{Tr}\mathbf{X} = M$,	Frobenius rescaling.
PL tail	$\rho_{\text{tail}}(\lambda) \sim \lambda^{-\alpha}$	HTSR tail; fitted exponent α.
PL tail	$\mathcal{T}_{\text{PL}}$, $\mathcal{T}^\star_{\text{PL}}$	CSN-selected PL window.
ECS	$\mathcal{R} \subseteq \mathcal{S}^+$, $M_{\mathcal{R}} = \lvert\mathcal{R}\rvert$	Retained ECS and size.
ECS	$\mathcal{R}^\star_{\text{ECS}}$	Stabilized SETOL ECS.
Alignment	$\mathcal{R}^\star_{\text{ECS}} \simeq \mathcal{T}^\star_{\text{PL}}$	SETOL–HTSR window match.
Trace-log	$\mathbf{X}_{\mathcal{R}} = \mathbf{U}^T_{\mathcal{R}}\mathbf{X}\mathbf{U}_{\mathcal{R}}$	ECS-restricted operator.
Trace-log	$\operatorname{Tr}\log\widetilde{\mathbf{X}}_{\mathcal{R}} = \sum_{i\in\mathcal{R}}\log\tilde{\lambda}_{\mathcal{R},i} = 0$	Scale-invariant condition.
Energy	$E_+ = M^{-1}\operatorname{Tr}\widetilde{\mathbf{X}}$	Retained spectral energy
Energy	$\widetilde{E} = M_{\mathcal{R}}^{-1}\sum_{i\in\mathcal{R}}\tilde{\lambda}_{\mathcal{R},i} = \int \tilde{\lambda}_{\mathcal{R}}\widehat{\rho}_{\mathcal{R}}(\tilde{\lambda}_{\mathcal{R}})d\tilde{\lambda}_{\mathcal{R}}$	Retained spectral energy.
Tail	$\phi_k^R = \sum_{j=1}^k \tilde{\lambda}_{\mathcal{R},R(j)} / \sum_{i\in\mathcal{R}}\tilde{\lambda}_{\mathcal{R},i}$	Top-k gain fraction.
Tail	$M^R_{\text{tr}} = (\sum_{i\in\mathcal{R}}\tilde{\lambda}_{\mathcal{R},i})^2 / \sum_{i\in\mathcal{R}}\tilde{\lambda}^2_{\mathcal{R},i}$	Effective contributor count.
Scale	$x = \log\tilde{\lambda}_{\mathcal{R}}$, $[\Lambda, b\Lambda]$	Log coordinate and shell.
Scale	$G_E(\Lambda, b) = \int_\Lambda^{b\Lambda}\tilde{\lambda}_{\mathcal{R}}\widehat{\rho}_{\mathcal{R}}(\tilde{\lambda}_{\mathcal{R}})d\tilde{\lambda}_{\mathcal{R}}$	Band energy gain.
Scale	$g_E(\tilde{\lambda}_{\mathcal{R}}) = \dfrac{d\widetilde{E}}{d\log\tilde{\lambda}_{\mathcal{R}}} = \tilde{\lambda}^2_{\mathcal{R}}\widehat{\rho}_{\mathcal{R}}(\tilde{\lambda}_{\mathcal{R}})$;	Spectral energy gain per logarithmic scale.
Beta	$\beta_E(g_E) = dg_E/d\log\tilde{\lambda}_{\mathcal{R}} = (2-\alpha)g_E$	Leading PL scale derivative.
Marginal	$y_E = 2 - \alpha = 0$	Log-band balance.
Fixed point	$\mathcal{R}_{\text{RG}}[\rho^\star] = \rho^\star$	Normalized spectral fixed point.
Branch	$\mathfrak{F}_0$, $\mathfrak{F}_{\text{ERG}}$	RG spectral fixed points or branches.
Branch	$\mathfrak{F}_0 : g_E$ undefined	Trivial/random-like branch.
Branch	$\mathfrak{F}_{\text{ERG}} : 0 < g^\star_E < \infty,\ \alpha \simeq 2,\ \mathcal{R}^\star_{\text{ECS}} \simeq \mathcal{T}^\star_{\text{PL}}$	Empirical SETOL ERG fixed point or branch.
Quality	$\overline{Q}_{\text{NN}}$, $\overline{Q}_\ell$	Model and layer quality.
Optimizer	$A_{\mathcal{R}} = \mathbf{U}_{\mathcal{R}}\mathbf{X}^{-1}_{\mathcal{R}}\mathbf{U}^T_{\mathcal{R}}$, $G_T = \nabla_W T_{\mathcal{R}}$	Trace-log gradient objects.
Optimizer	$\Delta\mathbf{W}_\perp$	Trace-log projected update.
WW-PGD	$\mathcal{T}$, $m = \lvert\mathcal{T}\rvert$	Selected tail and size.
WW-PGD	$\tilde{\lambda}^\star_k = Ak^{-q}$, $q = 1$	Rank-order tail target.
WW-PGD	η_B, η_C, h_e	Blend, Cayley, homotopy factors.

A Optimizer experiment details

This appendix records the optimizer protocols. Both tests use MNIST and the same three-layer MLP, $784 \to 512 \to 512 \to 10$, with ReLU nonlinearities. The linear layers are logged as FC1–FC3. MNIST inputs use mean 0.1307 and standard deviation 0.3081. At checkpoints, `WeightWatcher` records the fitted HTSR exponent α and the tail diagnostics for each fully connected layer.

Reproduction notebooks. The Muon panel uses the public notebooks AdamWvsMuon.ipynb, MLP3-MNIST-AdamW.ipynb, and MLP3-MNIST-Muon.ipynb. The WW-PGD wrapper is reproduced from WW_PGD_QuickStart.ipynb. The plotted curves use the settings below.

Muon test. The Muon test compares AdamW and Muon on the same model and data. AdamW gives the baseline trajectory. Muon is applied to matrix-valued hidden-layer updates. Scalar, bias, and other non-matrix parameters are handled by the associated base optimizer. The short-horizon panel tracks the early optimizer separation. The long-horizon panel tracks the slower Muon approach over hundreds of epochs. Training and test accuracy are logged with the layerwise curves $\alpha(t)$, $\ell \in \{\mathrm{FC1}, \mathrm{FC2}, \mathrm{FC3}\}$.

WW-PGD test. The WW-PGD test compares AdamW with AdamW+WW-PGD. The minibatch optimizer is fixed. The batch size is 128. AdamW uses learning rate 10^{-3}. Gradients are clipped to norm 1.0. The main run uses 35 epochs and five independent runs. At every epoch, `WeightWatcher` analyzes the model with `detX=True`, `randomize=False`, and `plot=False`. WW-PGD applies one epoch-boundary spectral retraction with `min_tail`=5, $q = 1$, $\eta_B = 0.5$, $\eta_C = 0.25$, `use_detx=True`, `warmup_epochs`=0, `ramp_epochs`=5, and `enable_trap_pgd=False`.

WW-PGD correction. For each eligible layer, WW-PGD forms $\lambda_i = s_i^2$ and selects the retained tail from `WeightWatcher` diagnostics for $\tilde{\lambda}^{+}_{\mathrm{PL}}$ and $\tilde{\lambda}^{+}_{\mathcal{R}}$. It sorts the tail eigenvalues in decreasing order and uses the target $\tilde{\lambda}^{\star}_{k} = Ak^{-1}$. The amplitude A preserves the retained tail trace-log. The moved tail is recentered, reconstructed with the original singular vectors, and blended back into the post-AdamW layer. The standard test loader uses ordinary MNIST. The perturbed loader uses random rotation by 7°, random translation by 3%, small Gaussian blur, and no horizontal flip.

References

[1] D. Silver et al., "Mastering the game of Go with deep neural networks and tree search," *Nature* **529**, 484–489, 2016. DOI: 10.1038/nature16961.

[2] J. Jumper et al., "Highly accurate protein structure prediction with AlphaFold," *Nature* **596**, 583–589, 2021. DOI: 10.1038/s41586-021-03819-2.

[3] R. Bommasani et al., "On the Opportunities and Risks of Foundation Models," arXiv:2108.07258, 2021.

[4] A. Yang et al., "Qwen2.5 Technical Report," arXiv:2412.15115, 2025.

[5] OpenAI et al., "gpt-oss-120b & gpt-oss-20b Model Card," arXiv:2508.10925, 2025.

[6] OpenAI, "GPT-4 Technical Report," arXiv:2303.08774, 2023.

[7] M. W. Mahoney and C. H. Martin, "Traditional and heavy-tailed self-regularization in neural network models," *Proceedings of the 36th International Conference on Machine Learning*, PMLR 97:4284–4293, 2019; arXiv:1901.08276.

[8] C. H. Martin and M. W. Mahoney, "Implicit self-regularization in deep neural networks: Evidence from random matrix theory and implications for learning," *Journal of Machine Learning Research* 22(165):1–73, 2021. `https://jmlr.org/papers/v22/20-410.html`.

[9] C. H. Martin, T. S. Peng, and M. W. Mahoney, "Predicting trends in the quality of state-of-the-art neural networks without access to training or testing data," *Nature Communications* 12, 4122, 2021. DOI: 10.1038/s41467-021-24025-8.

[10] C. H. Martin and M. W. Mahoney, "Heavy-Tailed Universality Predicts Trends in Test Accuracies for Very Large Pre-Trained Deep Neural Networks," in *Proceedings of the 2020 SIAM International Conference on Data Mining*, SIAM, 2020, pp. 505–513. doi: `10.1137/1.9781611976236.57`.

[11] C. H. Martin and C. Hinrichs, "SETOL: A semi-empirical theory of (deep) learning," arXiv:2507.17912, 2025.

[12] H. K. Prakash and C. H. Martin, "Detecting overfitting in neural networks during long-horizon grokking using random matrix theory," arXiv:2605.12394, 2026.

[13] H. K. Prakash and C. H. Martin, "Late-stage generalization collapse in grokking: Detecting anti-grokking with WeightWatcher," arXiv:2602.02859, 2026.

[14] C. H. Martin, "WeightWatcher: An open-source diagnostic tool for deep neural networks," GitHub repository, `https://github.com/CalculatedContent/WeightWatcher`.

[15] H. K. Prakash and C. H. Martin, "WW-PGD Repair of Overfit and Anti-Grokking Models," work in progress, 2026.

[16] V. A. Marchenko and L. A. Pastur, "Distribution of eigenvalues for some sets of random matrices," *Mathematics of the USSR-Sbornik* 1, 457–483, 1967. DOI: 10.1070/SM1967v001n04ABEH001994.

[17] C. A. Tracy and H. Widom, "On orthogonal and symplectic matrix ensembles," *Communications in Mathematical Physics* 177, 727–754, 1996. DOI: 10.1007/BF02099545; arXiv:solv-int/9509007.

[18] J. Baik, G. Ben Arous, and S. Péché, "Phase transition of the largest eigenvalue for nonnull complex sample covariance matrices," *The Annals of Probability* **33**(5), 1643–1697, 2005. DOI: 10.1214/009117905000000233.

[19] H. S. Seung, H. Sompolinsky, and N. Tishby, "Statistical mechanics of learning from examples," *Physical Review A* 45, 6056–6091, 1992. DOI: 10.1103/PhysRevA.45.6056.

[20] P. Cizeau and J.-P. Bouchaud, "Theory of Lévy matrices," *Physical Review E* 50, 1810–1822, 1994. DOI: 10.1103/PhysRevE.50.1810.

[21] M. Potters and J.-P. Bouchaud, *A First Course in Random Matrix Theory: For Physicists, Engineers and Data Scientists*, Cambridge University Press, 2020.

[22] S. Galluccio, J.-P. Bouchaud, and M. Potters, "Rational decisions, random matrices and spin glasses," *Physica A* 259, 449–456, 1998. DOI: 10.1016/S0378-4371(98)00332-X; arXiv:cond-mat/9801209.

[23] A. Zee, "Law of addition in random matrix theory," *Nuclear Physics B* 474, 726–744, 1996. DOI: 10.1016/0550-3213(96)00276-3; arXiv:cond-mat/9602146.

[24] K. G. Wilson and J. B. Kogut, "The renormalization group and the epsilon expansion," *Physics Reports* 12, 75–199, 1974. DOI: 10.1016/0370-1573(74)90023-4.

[25] J. Polchinski, "Renormalization and effective Lagrangians," *Nuclear Physics B* 231, 269–295, 1984. DOI: 10.1016/0550-3213(84)90287-6.

[26] A. Clauset, C. R. Shalizi, and M. E. J. Newman, "Power-law distributions in empirical data," *SIAM Review* 51, 661–703, 2009. DOI: 10.1137/070710111.

[27] I. Loshchilov and F. Hutter, "Decoupled weight decay regularization," *International Conference on Learning Representations*, 2019; arXiv:1711.05101.

[28] K. Jordan, Y. Jin, V. Boza, J. You, F. Cesista, L. Newhouse, and J. Bernstein, "Muon: An optimizer for hidden layers in neural networks," online technical note, 2024, URL: `https://kellerjordan.github.io/posts/muon/`.

[29] J. Liu et al., "Muon is scalable for LLM training," arXiv:2502.16982, 2025.

[30] J. Bernstein and L. Newhouse, "Modular Duality in Deep Learning," in *Proceedings of the 42nd International Conference on Machine Learning*, PMLR 267:3920–3930, 2025; arXiv:2410.21265.

[31] C. H. Martin and K. F. Freed, "Ab initio computation of semiempirical π-electron methods. I. Constrained, transferable valence spaces in $\mathcal{H}^{\nu}$ calculations," *Journal of Chemical Physics* **100**, 7454–7470, 1994. DOI: 10.1063/1.466889.

[32] C. H. Martin, "Highly accurate ab initio π-electron Hamiltonians for small protonated Schiff bases," *Journal of Physical Chemistry* **100**, 14310–14315, 1996. DOI: 10.1021/jp960617w.

[33] C. H. Martin, "Redesigning semiempirical-like π-electron theory with second order effective valence shell Hamiltonian (H^{ν}) theory: Application to large protonated Schiff bases," *Chemical Physics Letters* **257**, 229–237, 1996. DOI: 10.1016/0009-2614(96)00548-9.

[34] T. Chen and C. Guestrin, "XGBoost: A scalable tree boosting system," in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785–794, 2016. DOI: 10.1145/2939672.2939785; arXiv:1603.02754.

[35] S. Wiseman and E. Domany, "Lack of self-averaging in critical disordered systems," *Physical Review E* **52**, 3469–3484, 1995. DOI: 10.1103/PhysRevE.52.3469.

[36] A. Aharony and A. B. Harris, "Absence of Self-Averaging and Universal Fluctuations in Random Systems near Critical Points," *Physical Review Letters* **77**, 3700–3703 (1996). doi:10.1103/PhysRevLett.77.3700.

[37] S. Wiseman and E. Domany, "Finite-Size Scaling and Lack of Self-Averaging in Critical Disordered Systems," *Physical Review Letters* **81**, 22–25 (1998). doi:10.1103/PhysRevLett.81.22.

[38] K. Jordan, "Muon: An optimizer for hidden layers in neural networks," Keller Jordan blog, Dec. 8, 2024. [Online]. Available: `https://kellerjordan.github.io/posts/muon/`.

[39] F. J. Wegner, "Some invariance properties of the renormalization group," *Journal of Physics C: Solid State Physics* **7**, no. 12, 2098–2108 (1974). doi:10.1088/0022-3719/7/12/004.

[40] J. Polchinski, "Renormalization and effective lagrangians," *Nuclear Physics B* **231**, no. 2, 269–295 (1984). doi:10.1016/0550-

3213(84)90287-6.