

A Spectral Transport Mechanism Underlying the Training Dynamics of Large Language Models

Yizhou Zhang¹, Weichen Wu², Lun Du³, Zhengjie Miao⁴

¹ Variational AI, ² The Voleon Group, ³ Independent Researcher, ⁴ Simon Fraser University
 {zyizhou96, weichen.wu.1996, dulun2834}@gmail.com, zhengjie@sfu.ca

Abstract

Large language models exhibit remarkably regular loss trajectories during training, often following approximate power-law decay over long optimization intervals. The dynamical origin of this phenomenon remains poorly understood. We derive an exact transport–dissipation equation governing the spectral distribution of prediction error under gradient flow, with the irreducible entropy cleanly separated via a KL-divergence-based error observable. The equation is exact; what requires empirical input is the behavior of the spectral flux $\mathcal{F}$ during training. Analyzing public LLM checkpoints via stochastic Lanczos probes, we find that $\mathcal{F}$ is small throughout training: approximately zero in an extended quasi-static phase, and small but nonzero and directional thereafter. Because $\mathcal{F}$ remains weak, perturbation theory yields power-law loss decay as the leading-order solution over any finite window. The first-order correction predicts a systematic, monotonically growing undershoot relative to any early power-law extrapolation. We confirm this across three Pythia scales (410M–7B), where the deviation onset aligns in token count rather than parameter count, and at the loss level on OLMo-2 (7B), which differs in architecture and data. Preliminary results further indicate that the same spectral and loss signatures appear under a Warmup-Stable-Decay schedule, with the drift onset falling at the cooldown.

1 Introduction

Large language models (LLM) exhibit remarkably regular loss trajectories during training, often following an approximate power-law decay over long optimization intervals (Kaplan et al., 2020; Hestness et al., 2017; Hoffmann et al., 2022). Yet their dynamical origin remains poorly understood. Despite its empirical ubiquity, a fundamental question persists: why does gradient descent on a complex, high-dimensional, nonlinearly parameterized system produce such predictable macroscopic behavior?

A growing line of theoretical work addresses this question through analytical models of learning dynamics (Jacot et al., 2018; Lee et al., 2019; Bordelon et al., 2024; Vyas et al., 2023; Tian, 2025; Bordelon & Pehlevan, 2024). These approaches achieve tractability by imposing closure conditions on the spectral structure of the learning operator, assumptions chosen for analytical convenience rather than derived from or validated against realistic LLM training.

We show that approximate power-law loss decay is the leading-order solution of an exact transport–dissipation equation governing the spectral distribution of prediction error under gradient flow:

$$\partial_t \varepsilon(\lambda, t) + \partial_\lambda \mathcal{F}(\lambda, t) = -2\lambda \varepsilon(\lambda, t), \quad (1)$$

where t is training time (measured in tokens), λ ranges over the eigenvalues of the learning operator $M(t)$ defined via Jacobian operators on error function, $\varepsilon(\lambda, t)$ is the spectral energy density of prediction error at eigenvalue λ , and $\mathcal{F}$ is the spectral flux. When $\mathcal{F} = 0$, each spectral mode decays independently, and power-law initial data yields a power-law loss decay. When $\mathcal{F}$ is small but nonzero and directional, the regime we measure empirically throughout training (see Section 4), the first-order correction is an accumulated drift term

that grows monotonically with training time. This yields a falsifiable prediction: the true loss must fall systematically below any early power-law extrapolation, with the undershoot growing monotonically.

The measured undershoot is small in absolute terms, 3–5% by the end of training, which may explain why it has gone unnoticed despite qualitative hints in prior work (Hoffmann et al., 2022). But it is statistically unambiguous and grows monotonically as the framework requires. The deviation onset is aligned in token count rather than parameter count across three Pythia scales spanning 410M to 7B parameters, confirming that the mechanism is governed by the data budget, not model capacity; the same one-sided undershoot, though without spectral-probe evidence, persists on OLMo-2 (7B).

The equation is derived exactly from gradient flow, with the irreducible entropy cleanly separated via a KL-divergence-based error observable; no closure condition is imposed. What requires empirical input is the behavior of $\mathcal{F}$. Analyzing publicly released LLM checkpoints at three scales via stochastic Lanczos probes (Lanczos, 1950), a standard spectral probe in various areas such as quantum physics (Dagotto, 1994), stochastic trace estimation (Ubaru et al., 2017), and neural network Hessian analysis (Ghorbani et al., 2019), we find that $\mathcal{F}$ is small throughout training, with the spectral profile evolving continuously from an early phase in which no net transport is visible, through the onset of a small directional drift, to systematic long-range deviation.

The central contribution is threefold. First, we derived an exact transport–dissipation equation governing the spectral distribution of prediction error under gradient flow, with the irreducible entropy cleanly separated. Second, we track the total spectral flux $\mathcal{F}$ from LLM checkpoints via the shoulder diagnostic, finding it small throughout training, ≈ 0 in an extended quasi-static phase, with a small directional component thereafter; the slope’s stationarity further suggests the coupling contribution is negligible. The accumulated drift $\bar{V}(t)$ entering the loss correction is inferred from the loss rather than integrated from the probe, and closing this loop is left to future work. Third, we propose a falsifiable prediction derived from perturbation theory, systematic, monotonically growing undershoot relative to any early power-law extrapolation, confirmed across three Pythia scales and on OLMo-2 (7B). Whether the same boundary conditions hold beyond the models studied here is an open question; the equation itself is exact.

2 Related Work

Scaling laws in deep learning and LLMs. Empirical scaling laws have established that training loss and downstream performance follow predictable power-law trends across model size, data, and compute (Kaplan et al., 2020; Hoffmann et al., 2022; Hestness et al., 2017; Henighan et al., 2020). These findings have become central to the study of large language models (Wei et al., 2022; Hernandez et al., 2021). Our goal is different: rather than characterizing scaling laws empirically, we seek a dynamical mechanism that explains why approximate power-law loss decay emerges during training.

Kernel and feature-learning perspectives. A major line of theory explains neural optimization through kernel or near-kernel dynamics, where the learning operator remains approximately fixed throughout training (Jacot et al., 2018; Lee et al., 2019; Bordelon et al., 2024). They treat training as kernel regression with a fixed learning operator, yielding tractable dynamics under the assumption that feature evolution is negligible (Bordelon et al., 2024). While these models are analytically tractable, they do not capture the feature evolution observed in realistic large-scale training (Woodworth et al., 2020; Allen-Zhu & Li, 2019). Complementary work studies feature learning in simplified architectures and shows how representation evolution can alter scaling behavior beyond the kernel regime (Bordelon & Pehlevan, 2024; Vyas et al., 2023; Tian, 2025; Yang & Hu, 2021). In general, they introduce structural assumptions about the feature learning process, such as prescribed spectral decay or simplified architectural constraints, motivated by intuition rather than derived from or validated against realistic training.

Our framework shares the same technical foundation as these works, gradient flow and the learning operator $\mathbf{M}(t) = \mathbf{J}\mathbf{J}^*$ as the natural object of study. The key distinction is twofold. First, by defining the error observable through the KL divergence between the data distribution and the model, we make the MSE operator structure faithful to the true optimization objective of LLMs, with the irreducible entropy cleanly separated rather than appearing as an uncontrolled floor. Second, the condition that closes our equation, $\mathcal{F} \approx 0$ in the quasi-static phase, is an empirical finding measured directly from LLM checkpoints, rather than a structural assumption.

Physics-inspired explanations of power laws. Recent work has interpreted neural scaling through analogies to thermodynamics, quantization, and other physical mechanisms (Michaud et al., 2023; Liu et al., 2025b;a; Baek et al., 2025). These approaches provide valuable intuition but are typically phenomenological: the key dynamical quantities are introduced conceptually rather than derived from the training process or defined as directly measurable observables. Our framework instead derives the transport equation from gradient flow operator dynamics and measures the quasi-static condition directly.

Spectral structure of learning dynamics. Our work is also related to analyses of spectral bias and the evolution of the neural tangent kernel during training (Rahaman et al., 2019; Bietti & Mairal, 2019; Canatar & Pehlevan, 2022; Wang et al., 2023). We use spectral structure not as an end point but as the state variable of a transport equation, connecting operator drift, spectral flux, and power-law loss decay within a single framework.

3 Spectral Shell Dynamics in a Drifting Eigenbasis

A physical picture. The dynamics below are those of a dissipative many-body system. Diagonalizing the learning operator $M(t)$ resolves the error into modes, one per eigendirection ϕ_u , a quasi-particle carrying energy $g_u^2 = \langle e, \phi_u \rangle^2$, whose eigenvalue λ_u is a cooling rate: its energy dissipates as $e^{-2\lambda_u t}$, so sharp (large- λ) modes cool fast. As the representation drifts the eigenbasis rotates and modes exchange energy; what matters for the loss is the *net* transport across λ -shells, the spectral flux $\mathcal{F}$. Like a glass of still water whose molecules move violently yet carry no net current, the quasi-static regime has $\mathcal{F} \approx 0$: energy is traded between modes but none crosses the spectrum, and the loss follows a clean power law. When the exchanges stop cancelling and $\mathcal{F}$ carries a persistent net energy toward larger λ , the loss falls below the power law. This net drift is what the paper measures.

3.1 Notation and Preliminaries

In this paper, we study learning as an evolution in a Hilbert function space. Our central object is the prediction error $e(t)$ and how gradient descent redistributes its energy across spectral resolutions. We work in the real separable Hilbert space $\mathcal{H} := L^2(p)$, the space of square-integrable functions with respect to the data distribution $p(x)$. The inner product and norm are defined by

$$\langle f, g \rangle := \mathbb{E}_{x \sim p}[f(x)g(x)], \quad \|f\|^2 = \langle f, f \rangle. \quad (2)$$

Error observable and the KL divergence. Let $p(y|x)$ denote the true data-generating distribution, $q_\theta(y|x)$ the model prediction, and y^* a token sampled from p . The cross-entropy loss decomposes as

$$\mathcal{L}_{\text{CE}}(\theta) = D_{\text{KL}}(p(\cdot|x) \| q_\theta(\cdot|x)) + H[p(y|x)], \quad (3)$$

where D_{KL} is the KL divergence and $H[p(y|x)]$ is the conditional entropy of the data. Since $H[p]$ is independent of θ , the learnable component of the loss is the KL divergence. We define the prediction error as

$$e(x) := \sqrt{2 D_{\text{KL}}(p(\cdot|x) \| q_\theta(\cdot|x))}. \quad (4)$$

This definition makes the MSE loss equal to the KL divergence:

$$\mathcal{L}_{\text{MSE}} = \frac{1}{2} \|e\|^2 = \mathbb{E}_x[D_{\text{KL}}(p(\cdot|x) \| q_\theta(\cdot|x))] = \mathcal{L}_{\text{CE}} - H[p(y|x)]. \quad (5)$$

This identity is exact for all θ . The MSE operator framework thus governs precisely the learnable component of the cross-entropy loss, with the irreducible entropy $H[p]$ cleanly separated as a constant invisible to the gradient flow: $e(x) = 0$ at optimality, and $H[p]$ need not be estimated or subtracted.

Since $\nabla_\theta H[p] = 0$, the induced gradient flow is identical to cross-entropy training:

$$\dot{\theta}(t) = -\nabla_\theta \mathcal{L}_{\text{CE}}(\theta(t)) = -\nabla_\theta \left(\frac{1}{2} \|e\|^2 \right). \quad (6)$$

Proposition 3.1 (Exact error evolution). *Let $J_e : \mathbb{R}^N \rightarrow \mathcal{H}$ denote the Jacobian operator of the error observable,*

$$(J_e v)(x) = \nabla_\theta e(x) \cdot v. \quad (7)$$

Its adjoint $J_e^* : \mathcal{H} \rightarrow \mathbb{R}^N$ satisfies $\nabla_\theta \mathcal{L}(\theta) = J_e^* e$. Define the Gram operator $M(t) : \mathcal{H} \rightarrow \mathcal{H}$ by $M(t) := J_{e(t)} J_{e(t)}^*$. Then by the chain rule, under gradient flow we have

$$\dot{e}(t) = J_{e(t)} \dot{\theta}(t) = -J_{e(t)} J_{e(t)}^* e(t) = -M(t) e(t). \quad (8)$$

where $M(t)$ is self-adjoint and positive semidefinite on $\mathcal{H}$.

Thus gradient descent induces a linear, time-dependent operator evolution in function space. All nonlinear effects of parameter optimization, including the nonlinearity of the error observable itself, are encoded in the time dependence of $M(t)$. To understand how error energy distributes across resolutions, we diagonalize $M(t)$ at each time. Let $\{(\lambda_u(t), \phi_u(t))\}_{u=1}^d$ denote an orthonormal eigendecomposition

$$M(t) \phi_u(t) = \lambda_u(t) \phi_u(t). \quad (9)$$

Define modal coordinates

$$g_u(t) := \langle e(t), \phi_u(t) \rangle. \quad (10)$$

Thus, understanding training amounts to understanding how the amplitudes $g_u(t)$ and eigenvalues $\lambda_u(t)$ co-evolve.

3.2 Exact Mode ODEs in a Drifting Eigenbasis

A key complication arises because $M(t)$ evolves in time. Its eigenvalues drift and its eigenvectors rotate. Therefore the modal coefficients $g_u(t)$ do not evolve independently:

Theorem 3.2 (Exact modal dynamics¹). *Assume simple spectrum for clarity (i.e. $\lambda_u \neq \lambda_v$ for $u \neq v$). Then the modal coefficients satisfy*

$$\dot{g}_u = -\lambda_u g_u + \sum_{v \neq u} \omega_{v \rightarrow u}(t) g_v, \quad (11)$$

where the coupling coefficients are given by

$$\omega_{v \rightarrow u} = \frac{\Omega_{v \rightarrow u}}{\lambda_u - \lambda_v}, \quad \Omega_{v \rightarrow u} := \langle \phi_u, \dot{M} \phi_v \rangle, \quad u \neq v. \quad (12)$$

The first term represents pure dissipation along each eigenmode. The second term arises entirely from eigenbasis rotation. It describes energy transfer between modes induced by representation drift. The expression for $\omega_{v \rightarrow u}$ follows from Kato perturbation theory applied to the symmetric operator $M(t)$. A full proof is provided in Appendix C.

At this microscopic level, the system is highly coupled: every mode interacts with every other mode, and each eigenvalue $\lambda_u(t)$ also evolves in time. Direct analysis of these coupled ODEs is intractable. To make progress, we adopt a statistical-mechanical perspective and coarse-grain the spectrum.

¹The derivation uses only $\mathcal{L} = \frac{1}{2} \|e\|^2$, $\dot{\theta} = -\nabla_\theta \mathcal{L}$, and $M = J_e J_e^*$ positive semidefinite, and thus applies to any differentiable error observable $e(x) = h(f_\theta(x))$.

3.3 Shell Aggregation and Energy Balance

The exact mode ODEs in Theorem 3.2 are high-dimensional and strongly coupled. Following a statistical-mechanical viewpoint, we coarse-grain spectral modes into *shells* indexed by α and track only macroscopic quadratic energy within each shell.

Shell partition. Fix a spectral range of interest $[\lambda_{\min}, \lambda_{\max}]$ and choose boundaries

$$\lambda_{\min} = b_0 < b_1 < \dots < b_K = \lambda_{\max}. \quad (13)$$

Define shell intervals $I_\alpha := [b_{\alpha-1}, b_\alpha]$ for $\alpha \in [K] := \{1, \dots, K\}$. No restriction is imposed on the spacing: shells may be uniform in λ , uniform in $\log \lambda$, or any adaptive discretization.

Theorem 3.3 (Exact shell energy balance). *At each time t , we define the shell quadratic energy $E_\alpha(t)$ and inter-shell exchange $\mathcal{F}_{\beta \rightarrow \alpha}(t)$ as follows:*

$$E_\alpha(t) := \frac{1}{2} \sum_{u \in \mathcal{S}_\alpha(t)} g_u^2(t), \quad \mathcal{F}_{\beta \rightarrow \alpha}(t) := \sum_{\substack{u \in \mathcal{S}_\alpha(t) \\ v \in \mathcal{S}_\beta(t)}} g_u(t) g_v(t) \omega_{v \rightarrow u}(t). \quad (14)$$

where $\mathcal{S}_\alpha(t) := \{u \in [d] : \lambda_u(t) \in I_\alpha\}$ denotes the set of mode indices contained in shell α at time t . Then for each shell $\alpha \in [K]$, the shell energy satisfies

$$\frac{d}{dt} E_\alpha(t) = - \sum_{u \in \mathcal{S}_\alpha(t)} \lambda_u(t) g_u^2(t) + \sum_{\beta \neq \alpha} \mathcal{F}_{\beta \rightarrow \alpha}(t) + B_\alpha(t), \quad (15)$$

where the first term is pure shell dissipation, the second term is pure inter-shell exchange, and $B_\alpha(t)$ is the boundary-crossing term induced by eigenvalue drift across shell boundaries.

3.4 Coarse-Grained Spectral Density and Flux

The shell balance in Theorem 3.3 provides an exact discrete conservation structure. To connect this discrete bookkeeping to a continuous spectral description, we introduce a coarse-grained energy density and flux over the spectral coordinate λ .

Coarse-grained definition. Fix a shell partition $\{I_\alpha\}_{\alpha=1}^K$ with boundaries $b_0 < b_1 < \dots < b_K$ and $I_\alpha = [b_{\alpha-1}, b_\alpha]$. When the shell spacing is sufficiently fine over a resolution range, we define a (shell-smoothed) energy density $\varepsilon(\lambda, t)$ and a transport function $\mathcal{F}(\lambda, t)$ to be any pair of functions satisfying, for each α ,

$$E_\alpha(t) \approx \int_{b_{\alpha-1}}^{b_\alpha} \varepsilon(\lambda, t) d\lambda, \quad \mathcal{F}(b_\alpha, t) - \mathcal{F}(b_{\alpha-1}, t) \approx - \left[\sum_{\beta \neq \alpha} \mathcal{F}_{\beta \rightarrow \alpha}(t) + B_\alpha(t) \right]. \quad (16)$$

In words, ε is the density whose integral recovers shell energy, and $\mathcal{F}(\lambda, t)$ is the flux whose boundary difference recovers the net transport out of the shell.

Theorem 3.4 (Exact microscopic realization via δ and H). *Let $\{(\lambda_u(t), \phi_u(t))\}$ be an instantaneous eigenbasis of $M(t)$ and $g_u(t) = \langle e(t), \phi_u(t) \rangle$ the modal coordinates. Define the distribution-valued energy density and flux by*

$$\varepsilon(\lambda, t) = \frac{1}{2} \sum_u g_u^2(t) \delta(\lambda_u(t) - \lambda), \quad (17)$$

and

$$\mathcal{F}(\lambda, t) = \frac{1}{2} \sum_u g_u^2(t) \dot{\lambda}_u(t) \delta(\lambda_u(t) - \lambda) \quad (18)$$

$$+ \sum_{u,v} h(\lambda_u(t) - \lambda) [1 - h(\lambda_v(t) - \lambda)] g_u(t) g_v(t) \omega_{v \rightarrow u}(t), \quad (19)$$

where δ is Dirac's delta distribution and h is the Heaviside step function.

Then:

- (i) The coarse-grained relations (16) hold exactly when interpreted in the sense of distributions (i.e., integrating (17) over a shell recovers E_α , and the corresponding flux bookkeeping recovers the inter-shell transport terms).
- (ii) In the sense of distributions, $(\varepsilon, \mathcal{F})$ satisfy the exact conservation form

$$\partial_t \varepsilon(\lambda, t) + \partial_\lambda \mathcal{F}(\lambda, t) = -2\lambda \varepsilon(\lambda, t). \quad (20)$$

Theorem 3.4 follows by differentiating (17), using the exact modal ODEs (Theorem 3.2), and applying $\partial_\lambda h(\lambda_u - \lambda) = -\delta(\lambda_u - \lambda)$. A complete derivation is given in Appendix E.

4 Empirical Spectral Probing of Learning Operators

We analyze publicly released Pythia checkpoints (Biderman et al., 2023) (410M, 1B, 7B) to provide direct empirical evidence for the quasi-static phase and its consequences.

4.1 Experimental Setup

Spectral estimation via stochastic Lanczos. The operator $M = J_e J_e^*$ governing error dynamics acts on a high-dimensional function space, making explicit eigendecomposition infeasible for modern LLMs. To estimate shell-resolved spectral quantities at scale, we employ a stochastic Lanczos method (Lanczos, 1950), a standard spectral probe in quantum many-body physics (Dagotto, 1994), stochastic trace estimation (Ubaru et al., 2017), and neural network Hessian analysis (Ghorbani et al., 2019), probing the spectrum of M via operator–vector products without explicitly forming M . The expectation is approximated using fixed mini-batches from the validation set; matrix–vector products Mv are computed via VJP–JVP composition, where the Jacobian is that of the error observable defined in (21). We apply a single-vector Lanczos algorithm with $k = 100$ steps. Implementation details and pseudocode are in Appendix F.

Error observable: binary KL approximation. The full $D_{\text{KL}}(p(\cdot|x) \| q_\theta(\cdot|x))$ requires the true distribution $p(y|x)$ and a vocabulary-wide sum, both infeasible at scale. We adopt a *binary coarsening*: reducing each token prediction to a correct-vs-rest event,

$$D_{\text{bin}}(x) = r \log \frac{r}{s} + (1-r) \log \frac{1-r}{1-s}, \quad e(x) := \sqrt{2 D_{\text{bin}}(x)}, \quad (21)$$

where $r \approx q_{\text{ref}}(y^*|x)$ is a reference model approximation and $s = q_\theta(y^*|x)$; only q_θ is differentiated. By the data processing inequality, $D_{\text{bin}} \leq D_{\text{KL}}$, preserving the mathematical structure of the framework. Learnable and irreducible components are automatically separated by the Lanczos procedure (Appendix F).

4.2 Spectral Evolution Across Training

Figure 1 presents the shell-resolved spectral energy profiles for Pythia-410M, 1B, and 7B, measured at successive checkpoints spanning the full training trajectory. Each row is divided into four panels covering consecutive token intervals, providing a systematic view of how the spectral error distribution evolves as training progresses. The key diagnostic throughout is the position of the left shoulder of the spectral profile: under the pure dissipation equation $\partial_t \varepsilon = -2\lambda \varepsilon$, each shell decays independently at rate 2λ , which naturally drives the spectral peak rightward over time. This peak migration is an intrinsic consequence of dissipation dynamics and should not be conflated with spectral transport. The left shoulder, by contrast, is sensitive to the net spectral flux $\mathcal{F}$: coincident shoulders across checkpoints indicate $\mathcal{F} \approx 0$, while systematic misalignment signals that transport is active.

Early training (panel a, 2.1B–8.4B tokens). The left shoulders of successive checkpoints are in close agreement, providing direct evidence for $\mathcal{F} \approx 0$. The spectral peak migrates rightward while the overall amplitude decreases uniformly, both consistent with pure dissipation, confirming the system is well within the quasi-static phase.

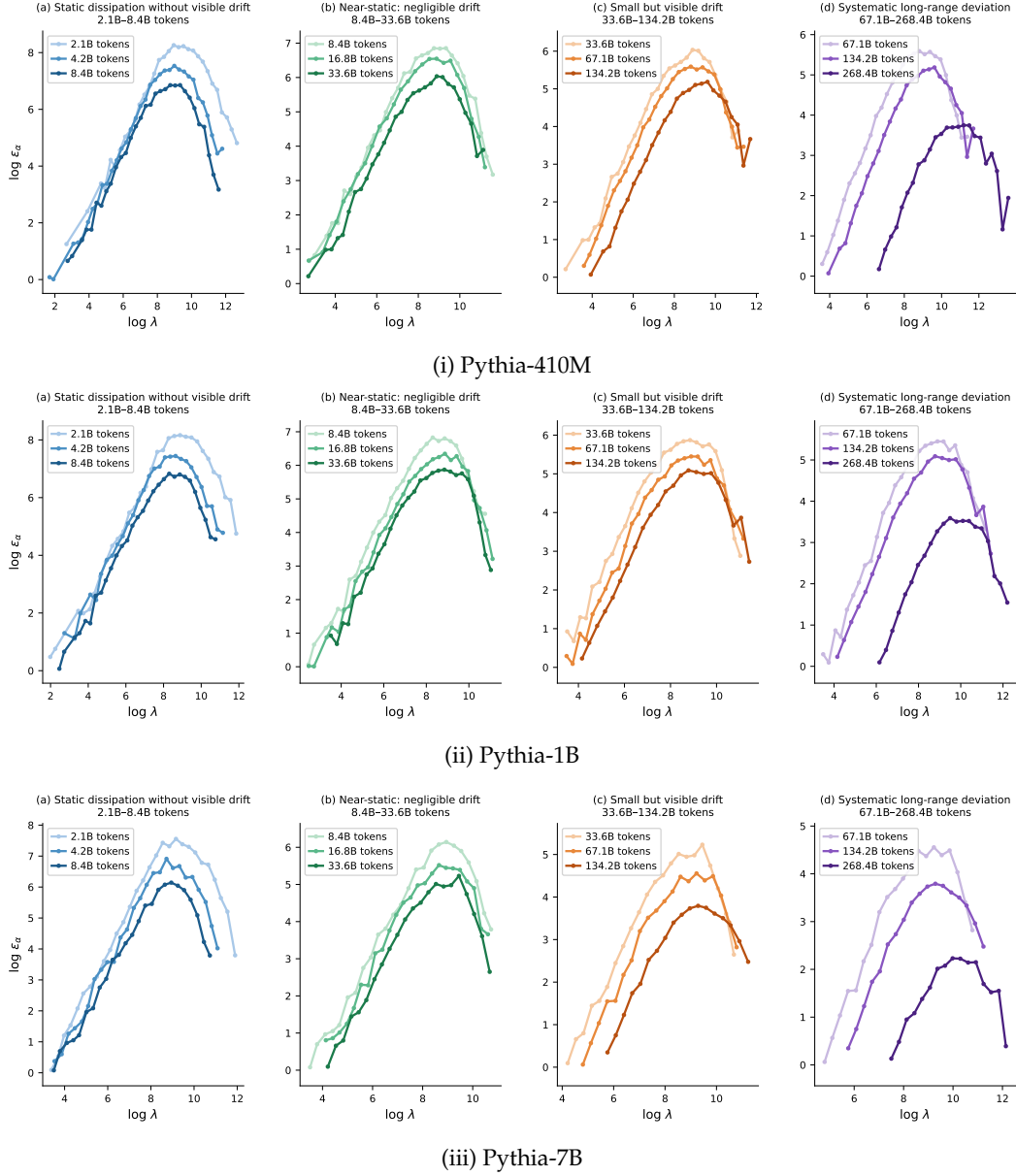


Figure 1: Shell-resolved spectral energy profiles $\log \varepsilon_\alpha$ vs. $\log \lambda$ at successive checkpoints for Pythia-410M, 1B, and 7B. Four consecutive training intervals are shown per model. The left shoulder serves as the primary diagnostic for $\mathcal{F} \approx 0$: coincident shoulders indicate vanishing spectral flux; progressive misalignment signals drift accumulation. Rightward peak migration reflects pure dissipation and does not constitute evidence of transport. The four panels show one continuous drift over four consecutive token intervals; the pattern is consistent across all three scales.

Middle training (panel b and c, 8.4B–134.2B tokens). The quasi-static condition begins to weaken. In panel b, the shoulders remain approximately coincident; in panel c, systematic misalignment first emerges, marking the onset of non-negligible $2\bar{V}(t)$ in Equation (24) of Section 5 and the beginning of deviation from the early-window power-law extrapolation.

Late training (panel d, 67.1B–268.4B tokens). Left-shoulder misalignment becomes pronounced and the high-eigenvalue tail accumulates relatively more energy. The accumulated

drift $\bar{V}(t)$ is no longer negligible; by Theorem 5.1 (see Section 5), the power-law exponent from any earlier window can no longer be extrapolated to the current phase.

Consistency across model scales. We emphasize that the four panels do not correspond to four discrete regimes: they show a single continuous drift over four consecutive token intervals, and the interval boundaries are a presentation choice rather than sharp transitions. With this understanding, the progression, from no visible net transport, through drift onset, to systematic deviation, is strikingly consistent across Pythia-410M, 1B, and 7B despite a nearly twenty-fold difference in parameter count. Phase transitions are aligned in token count rather than parameter count, indicating that the mechanism is governed by the training data budget, not model capacity.²

5 Power-Law Loss Decay and Its Long-Range Breakdown

The experiments in Section 4 track the total spectral flux $\mathcal{F}$ via the left shoulder: $\mathcal{F}$ is ≈ 0 while the shoulder is stationary, and small but directional once it migrates. At the level of the decomposition we can offer only an indication: the profile translates without an appreciable slope change, suggesting the coupling flux $\mathcal{F}_{\text{coup}}$ is negligible and the net flux is drift-dominated; this is inferred from the profile’s geometry, not measured independently. We now derive the consequences of a small net $\mathcal{F}$.

When the net spectral flux vanishes, $\mathcal{F} \approx 0$, Equation (20) reduces to a pure dissipation equation:

$$\partial_t \varepsilon(\lambda, t) = -2\lambda \varepsilon(\lambda, t), \quad (22)$$

with explicit solution $\varepsilon(\lambda, t) = \varepsilon(\lambda, 0) \cdot e^{-2\lambda t}$. Under power-law initial data this yields power-law loss decay (see Appendix G for the derivation). The key question is what happens when $\mathcal{F}$ is small but nonzero.

Defining $v(\lambda, t) = \frac{\mathcal{F}(\lambda, t)}{\varepsilon(\lambda, t)}$, the accumulated drift up to time t is

$$\bar{V}(t) := \int_0^t (t-s) \bar{v}(s) ds, \quad (23)$$

where $\bar{v}(s)$ is the spectrally weighted average of $v(\lambda, s)$ over the surviving error distribution $\varepsilon(\lambda, 0)e^{-2\lambda t}$ (defined precisely in Appendix H). Treating v as a small perturbation and expanding to first order yields the corrected loss

$$L(t) = L_{\text{PL}}(t) [1 - 2\bar{V}(t) + O(\bar{V}^2)], \quad (24)$$

where $L_{\text{PL}}(t)$ is the power-law trajectory under $\mathcal{F} = 0$.

Theorem 5.1 (Local power-law decay with systematic long-range undershoot). *Suppose the net spectral flux satisfies $\mathcal{F} = v\varepsilon$ with v small and $v(\lambda, s) \geq v_{\min} > 0$, that is, the net drift transports error energy toward larger λ where the per-mode dissipation rate 2λ is stronger. Then:*

1. (Local power-law.) *Over any finite window in which $2\bar{V}(t) \ll 1$, the loss is well approximated by a power law: $L(t) \approx L_{\text{PL}}(t)$.*
2. (Systematic undershoot.) *As $\bar{V}(t)$ accumulates, $2\bar{V}(t)$ grows and the correction in Equation (24) becomes significant. The loss falls systematically below the power-law extrapolation from any earlier finite window: $L(t) < L_{\text{PL}}(t)$.*
3. (Irreversibility of undershoot.) *Once $2\bar{V}(t)$ is no longer negligible, no finite power-law fit to the later trajectory will extrapolate correctly to still later times, as the same accumulation mechanism continues to operate.*

²This undershoot is not a shortcoming of the power-law form, a more flexible fit could absorb it, but a directional prediction of the mechanism: net drift toward larger λ forces the loss *below* the extrapolation, with an onset set by the data budget rather than model capacity.

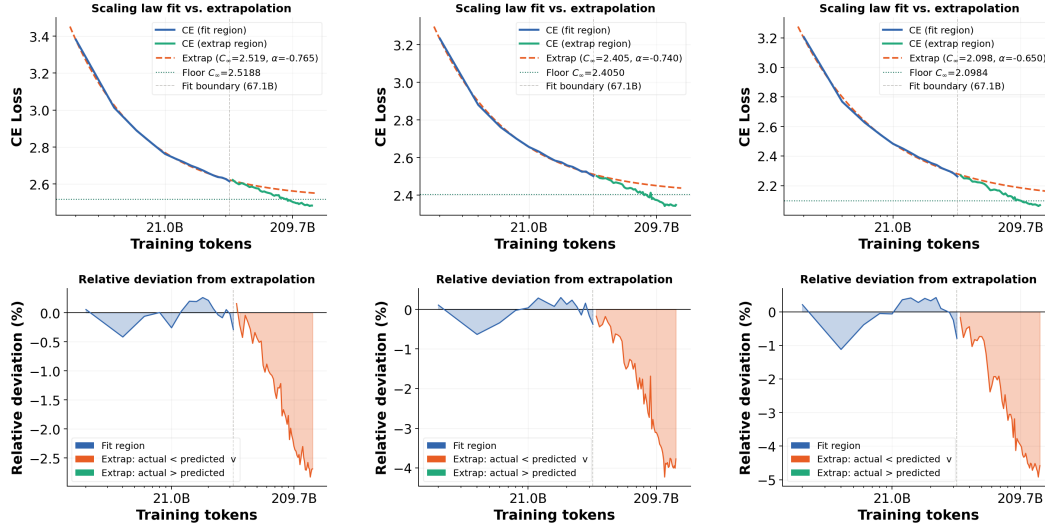


Figure 2: Empirical validation of Theorem 5.1 on Pythia-410M, 1B, and 7B (left to right). **Top row:** CE loss trajectories (solid) against the power-law fit extrapolated from steps 1k–32k (dashed); dotted line shows fitted floor C_∞ . **Bottom row:** relative deviation $(L_{\text{actual}} - L_{\text{pred}}) / L_{\text{pred}}$. In all three models almost every checkpoint beyond the fit boundary falls below the extrapolation, with the undershoot growing monotonically to 3–5%.

Empirical validation. Theorem 5.1 predicts that any power-law fit obtained from an early training window will systematically underestimate the rate of loss reduction at later steps: as the accumulated drift $\tilde{V}(t)$ grows, the correction factor $1 - 2\tilde{V}(t)$ in Equation (24) drives the true loss below the extrapolated trajectory, with the magnitude of the undershoot increasing monotonically in time. We test this prediction directly on Pythia-410M, 1B, and 7B, the same model scales analyzed in Section 4, using dense checkpoints saved every 2,000 steps across the full training run. Following Hoffmann et al. (2022), we fit $L(t) = C_\infty + A \cdot t^\alpha$ to checkpoints with $t \leq 32,000$ steps via grid search over C_∞ with closed-form OLS in log-space (see Appendix I for details).³ The fit quality is uniformly high (RMSE < 0.025 nats); curves are extrapolated without further adjustment.

Results. Figure 2 reports the outcome across all three model scales. In all three models and at almost every checkpoint beyond the fit boundary, the observed loss lies below the extrapolated power law, with the undershoot growing monotonically to 3–5%. The undershoot is not a finite-sample artifact, with almost all held-out checkpoints deviating in the same direction, a chance explanation is ruled out. The monotonically growing deviation with no sign of saturation confirms part (iii) of Theorem 5.1: no re-fit to a later window will extrapolate correctly. Critically, the pattern is identical across 410M to 7B, with deviation onset aligned in token count (i.e. data budget), not model capacity.

Together, these results provide quantitative confirmation of parts (ii) and (iii) of Theorem 5.1, and close the empirical loop with Section 4: power-law breakdown in the loss trajectory is the integrated signature of the spectral drift measured via the Lanczos probe.

Beyond Pythia: Cross-Family, Schedule, and Robustness The analysis above is on Pythia under a cosine schedule; three additional checks show the same signatures under a different family and a different schedule. On OLMo-2 7B, differing in architecture, optimizer, and data, the loss undershoots the early power law one-sidedly and monotonically, reaching around -3% (Appendix L). To test a fundamentally different schedule, we trained a small (~ 109 M) Llama-style model under Warmup-Stable-Decay (Hu et al., 2024; Hägele et al.,

³We fit $\mathcal{L}_{\text{CE}}$ rather than the KL divergence; $H[p]$ is absorbed into C_∞ and does not affect the predicted undershoot.

2024) and applied the same probe: the left shoulder is stationary during the constant-LR phase ($\mathcal{F} \approx 0$) and migrates during the cooldown, and the loss tracks the power law until the cooldown onset and then undershoots by a few percent, the same stable-then-cooldown pattern visible in public WSD logs (Hägele et al., 2024; Schaipp et al., 2025) (Appendix K). Finally, the undershoot is insensitive to the fit window (Appendix J), and the shoulder diagnostic is stable to the Lanczos truncation and calibration batch size (Appendix M).

6 Limitations and Future Work

The framework’s empirical support, while now spanning two model families (Pythia 410M–7B, OLMo-2 7B) and two learning-rate schedules (cosine and, preliminarily, WSD), remains narrower than the generality of the transport–dissipation equation itself, which is exact and architecture-agnostic. In particular, the direct spectral-probe evidence (Figure 1) is still primarily from Pythia, and the WSD spectral evidence comes from a single small model; broader cross-family and larger-scale spectral probing is the most important direction for strengthening the empirical case. Besides, several components are approximations. The binary-KL observable is a lower bound on the full KL; the equation is derived under gradient flow rather than real finite-LR Adam training, though its predictions hold on real checkpoints; and both the coupling/drift decomposition of $\mathcal{F}$ and the accumulated drift $\bar{V}(t)$ are inferred from geometry and the loss rather than measured independently.

The framework identifies the spectral slope as the quantity governing the power-law exponent, but does not predict the exponent from first principles. Extending the perturbative analysis beyond the weak- $\mathcal{F}$ regime, characterizing which boundary conditions hold under different training recipes, and predicting the exponent are natural directions for future work.

Furthermore, the theoretic framework also predicts a consequence beyond the loss curve. Because net drift moves error energy toward sharper-curvature directions, the framework predicts declining compressibility with training, consistent with reports that post-training-quantization degradation grows with pretraining tokens (Kumar et al., 2024) and couples to learning-rate dynamics (Catalan-Tatjer et al., 2025), a falsifiable direction for future work.

7 Conclusion

Starting from $\dot{e} = -Me$ with $M = J_e J_e^*$, where the error observable is defined via the KL divergence, cleanly separating the irreducible entropy, we derived an exact transport–dissipation PDE governing the spectral distribution of prediction error under gradient flow. The central empirical finding concerns the total spectral flux: $\mathcal{F}$ is approximately zero in an extended quasi-static phase, with a small directional drift accumulating thereafter. Because $\mathcal{F}$ remains weak, a perturbative analysis yields power-law loss decay as the leading-order solution over any finite window. The first-order correction is an accumulated drift term that grows monotonically, predicting a systematic undershoot relative to any early power-law extrapolation. Experiments on Pythia-410M, 1B, and 7B confirm both: power-law decay holds locally, and the loss falls below the extrapolated trajectory at almost every held-out checkpoint, with deviation onset governed by the data budget rather than model capacity; the same one-sided undershoot appears on OLMo-2 (7B), and preliminary results under a WSD schedule show the same spectral and loss signatures.

The contribution is structural. The framework identifies $\mathcal{F}$ as the governing quantity: where it is negligible, power-law decay holds; where it accumulates, scaling breaks down in a predictable and measurable way. Predicting the precise exponent from first principles, characterizing the boundary conditions across model families and training recipes (beyond the Pythia, OLMo-2, and preliminary WSD evidence reported here), and extending the analysis beyond the weak-flux regime are natural directions for future work.

Acknowledgement

We are thankful for the discussion with Jihao Long and Runyu Zhang about this work. We also appreciate the suggestions from anonymous reviewer. Zhengjie Miao is partially supported by an NSERC Discovery Grant.

References

- Zeyuan Allen-Zhu and Yuanzhi Li. What can resnet learn efficiently, going beyond kernels? *Advances in neural information processing systems*, 32, 2019.
- David D Baek, Ziming Liu, and Max Tegmark. Geneft: Understanding statics and dynamics of model generalization via physics-inspired effective theory. *Physical Review E*, 111(3): 035307, 2025.
- Stella Biderman, Hailey Schoelkopf, Quentin Anthony, Herbie Bradley, Kyle O’Brien, Eric Hallahan, Mohammad Aflah Khan, Shivanshu Purohit, USVSN Sai Prashanth, Edward Raff, Aviya Skowron, Lintang Sutawika, and Oskar van der Wal. Pythia: A suite for analyzing large language models across training and scaling, 2023. URL <https://arxiv.org/abs/2304.01373>.
- Alberto Bietti and Julien Mairal. On the inductive bias of neural tangent kernels. *Advances in Neural Information Processing Systems*, 32, 2019.
- Blake Bordelon and Cengiz Pehlevan. How feature learning can improve neural scaling laws. *arXiv preprint arXiv:2409.17858*, 2024.
- Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. *arXiv preprint arXiv:2402.01092*, 2024.
- Abdulkadir Canatar and Cengiz Pehlevan. A kernel analysis of feature learning in deep neural networks. In *2022 58th Annual Allerton Conference on Communication, Control, and Computing (Allerton)*, pp. 1–8. IEEE, 2022.
- Albert Catalan-Tatjer, Niccolò Ajroldi, and Jonas Geiping. Training dynamics impact post-training quantization robustness. *arXiv preprint arXiv:2510.06213*, 2025.
- Elbio Dagotto. Correlated electrons in high-temperature superconductors. *Reviews of Modern Physics*, 66(3):763, 1994.
- Behrooz Ghorbani, Shankar Krishnan, and Ying Xiao. An investigation into neural net optimization via hessian eigenvalue density. *International Conference on Machine Learning (ICML)*, 2019.
- Alexander Hägele, Elie Bakouch, Atli Kosson, Loubna B Allal, Leandro Von Werra, and Martin Jaggi. Scaling laws and compute-optimal training beyond fixed training durations. *Advances in Neural Information Processing Systems*, 37:76232–76264, 2024.
- Tom Henighan, Jared Kaplan, Mayank Katz, Mark Chen, Christopher Hesse, Jacob Jackson, Heewoo Jun, Tom Brown, Prafulla Dhariwal, Scott Gray, et al. Scaling laws for autoregressive generative modeling. *arXiv preprint arXiv:2010.14701*, 2020.
- Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws and interpretability of learning from diverse data sources. *arXiv preprint arXiv:2109.07437*, 2021.
- Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.
- Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.
- Shengding Hu, Yuge Tu, Xu Han, Chaoqun He, Ganqu Cui, Xiang Long, Zhi Zheng, Yewei Fang, Yuxiang Huang, Weilin Zhao, et al. Minicpm: Unveiling the potential of small language models with scalable training strategies. *arXiv preprint arXiv:2404.06395*, 2024.

- Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in Neural Information Processing Systems*, 31, 2018.
- Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.
- Tanishq Kumar, Zachary Ankner, Benjamin F Spector, Blake Bordelon, Niklas Muennighoff, Mansheej Paul, Cengiz Pehlevan, Christopher Ré, and Aditi Raghunathan. Scaling laws for precision. *arXiv preprint arXiv:2411.04330*, 2024.
- Cornelius Lanczos. An iteration method for the solution of the eigenvalue problem of linear differential and integral operators. *Journal of Research of the National Bureau of Standards*, 45(4):255–282, 1950.
- Jaehoon Lee, Lechao Xiao, Samuel S Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in Neural Information Processing Systems*, 32, 2019.
- Yizhou Liu, Ziming Liu, and Jeff Gore. Superposition yields robust neural scaling. *arXiv preprint arXiv:2505.10465*, 2025a.
- Ziming Liu, Yizhou Liu, Jeff Gore, and Max Tegmark. Neural thermodynamic laws for large language model training. *arXiv preprint arXiv:2505.10559*, 2025b.
- Eric Michaud, Ziming Liu, Uzey Girit, and Max Tegmark. The quantization model of neural scaling. *Advances in Neural Information Processing Systems*, 36:28699–28722, 2023.
- Nasim Rahaman, Aristide Baratin, Devansh Arpit, Felix Draxler, Min Lin, Fred Hamprecht, Yoshua Bengio, and Aaron Courville. On the spectral bias of neural networks. *Proceedings of Machine Learning Research*, 2019.
- Fabian Schaipp, Alexander Hägele, Adrien Taylor, Umut Simsekli, and Francis Bach. The surprising agreement between convex optimization theory and learning-rate scheduling for large model training. *arXiv preprint arXiv:2501.18965*, 2025.
- Zhiqiang Shen, Tianhua Tao, Liqun Ma, Willie Neiswanger, Zhengzhong Liu, Hongyi Wang, Bowen Tan, Joel Hestness, Natalia Vassilieva, Daria Soboleva, et al. Slimpajama-dc: Understanding data combinations for llm training. *arXiv preprint arXiv:2309.10818*, 2023.
- Yuandong Tian. Provable scaling laws of feature emergence from learning dynamics of grokking. *arXiv preprint arXiv:2509.21519*, 2025.
- Shashanka Ubaru, Jie Chen, and Yousef Saad. Fast estimation of $\text{tr}(f(a))$ via stochastic Lanczos quadrature. *SIAM Journal on Matrix Analysis and Applications*, 38(4):1075–1099, 2017.
- Nikhil Vyas, Alexander Atanasov, Blake Bordelon, Depen Morwani, Sabarish Sainathan, and Cengiz Pehlevan. Feature-learning networks are consistent across widths at realistic scales. *Advances in Neural Information Processing Systems*, 36:1036–1060, 2023.
- Zhichao Wang, Andrew Engel, Anand D Sarwate, Ioana Dumitriu, and Tony Chiang. Spectral evolution and invariance in linear-width neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2023. URL <https://arxiv.org/abs/2211.06506>.
- Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*, 2022.
- Blake Woodworth, Suriya Gunasekar, Jason D Lee, Edward Moroshko, Pedro Savarese, Itay Golan, Daniel Soudry, and Nathan Srebro. Kernel and rich regimes in overparametrized models. In *Conference on Learning Theory*, pp. 3635–3673. PMLR, 2020.

Greg Yang and Edward J Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021. URL <https://arxiv.org/abs/2011.14522>.

A Core objects and their physical reading.

Symbol	Meaning (physical reading)
$e(x)$	Prediction error observable; $\frac{1}{2}\ e\ ^2$ is the learnable loss (KL), zero at optimality
$M = J_e J_e^*$	Learning operator; its eigenvalues set the spectral (curvature) scales
λ	Eigenvalue; large λ = sharp direction, cooling rate 2λ
ϕ_u, g_u	Eigenmode and its amplitude $g_u = \langle e, \phi_u \rangle$; g_u^2 is the energy it carries
$\varepsilon(\lambda, t)$	Spectral energy density of the error along λ
$\mathcal{F}(\lambda, t)$	Spectral flux: <i>net</i> energy transport across λ -shells. $\mathcal{F} \approx 0$ means no net transport, not that modes evolve independently
$\bar{V}(t)$	Accumulated net drift; the quantity driving the undershoot

Table 1: Core objects and their physical reading.

B Proof of Proposition 3.1

We restate Proposition 3.1 for convenience.

Proposition B.1 (Exact error evolution). *Let $e(x)$ be a differentiable error observable (the proof applies to any $e(x) = h(f_\theta(x))$); in the main text we use $e(x) = \sqrt{2 D_{\text{bin}}(x)}$). Let $J_e : \mathbb{R}^N \rightarrow \mathcal{H}$ denote its Jacobian operator,*

$$(J_e v)(x) = \nabla_\theta e(x) \cdot v.$$

Its adjoint $J_e^* : \mathcal{H} \rightarrow \mathbb{R}^N$ satisfies

$$\nabla_\theta L(\theta) = J_e^* e.$$

Define the Gram operator

$$M(t) : \mathcal{H} \rightarrow \mathcal{H}, \quad M(t) = J_{e(t)} J_{e(t)}^*.$$

Then under gradient flow,

$$\dot{e}(t) = -M(t)e(t).$$

Moreover, $M(t)$ is self-adjoint and positive semidefinite on $\mathcal{H}$.

Proof. The proof proceeds for any differentiable $e(x) = h(f_\theta(x))$; we use gradient flow $\dot{\theta} = -\nabla_\theta L$ and $M = J_e J_e^*$.

Differentiating with respect to time and applying the chain rule gives

$$\dot{e}(t) = \frac{d}{dt} e(\theta(t)) = \nabla_\theta e(x) \cdot \dot{\theta}(t) = J_{e(t)} \dot{\theta}(t).$$

Since $\mathbb{R}^N$ and $\mathcal{H}$ are Hilbert spaces, J_e admits an adjoint $J_e^* : \mathcal{H} \rightarrow \mathbb{R}^N$ defined by

$$\langle J_e v, h \rangle_{\mathcal{H}} = \langle v, J_e^* h \rangle_{\mathbb{R}^N}, \quad \forall v \in \mathbb{R}^N, h \in \mathcal{H}.$$

Gradient of the loss. The loss is

$$L(\theta) = \frac{1}{2} \|e(\theta)\|_{\mathcal{H}}^2 = \frac{1}{2} \langle e(\theta), e(\theta) \rangle_{\mathcal{H}} = \mathcal{L}_{\text{CE}}(\theta).$$

For any direction $v \in \mathbb{R}^N$, the directional derivative is

$$DL(\theta)[v] = \langle e(\theta), De(\theta)[v] \rangle_{\mathcal{H}} = \langle e(\theta), J_e v \rangle_{\mathcal{H}} = \langle J_e^* e(\theta), v \rangle_{\mathbb{R}^N}.$$

By the Riesz representation of the Euclidean gradient, this implies

$$\nabla_\theta L(\theta) = J_e^* e(\theta).$$

Verification of gradient equivalence. Since $\frac{1}{2}\|e\|^2 = D_{\text{KL}}(p\|q_\theta) = \mathcal{L}_{\text{CE}} - H[p]$ and $H[p]$ is independent of θ ,

$$\nabla_\theta \left(\frac{1}{2}\|e\|^2 \right) = \nabla_\theta D_{\text{KL}} = \nabla_\theta \mathcal{L}_{\text{CE}}.$$

The gradient flow induced by $\frac{1}{2}\|e\|^2$ is identical to the gradient flow induced by the cross-entropy loss, at every θ .

Under gradient flow,

$$\dot{\theta}(t) = -\nabla_\theta L(\theta(t)) = -J_{e(t)}^* e(t).$$

Substituting this into the chain rule identity yields

$$\dot{e}(t) = J_{e(t)} \dot{\theta}(t) = J_{e(t)} (-J_{e(t)}^* e(t)) = -J_{e(t)} J_{e(t)}^* e(t).$$

Thus,

$$\dot{e}(t) = -M(t)e(t), \quad M(t) := J_{e(t)} J_{e(t)}^*.$$

Finally, $M(t)$ is self-adjoint since

$$(J_{e(t)} J_{e(t)}^*)^* = J_{e(t)} J_{e(t)}^*,$$

and positive semidefinite because for any $h \in \mathcal{H}$,

$$\langle h, M(t)h \rangle = \langle h, J_{e(t)} J_{e(t)}^* h \rangle = \|J_{e(t)}^* h\|^2 \geq 0.$$

□

Remark B.2 (Relationship to log-probability Jacobian). Let $f(x) = \log q_\theta(y^*|x)$ and J_f its Jacobian. Since $e = \sqrt{-2f}$, the chain rule gives $J_e = D \circ J_f$ where $D : h(x) \mapsto -h(x)/e(x)$ is a multiplicative operator. The Gram operator therefore satisfies $M = D M_f D$ where $M_f = J_f J_f^*$ is the Gram operator of the log-probability. This conjugation reweights samples by $1/e(x)^2$: tokens with small NLL (already well-predicted) are upweighted relative to tokens with large NLL, reflecting the natural curvature structure of the cross-entropy loss.

C Proof of Theorem 3.2

This section proves Theorem 3.2 in the main text. We work under the simple-spectrum assumption, i.e. $\lambda_u(t) \neq \lambda_v(t)$ for $u \neq v$.

Theorem C.1 (Exact modal dynamics). *Let*

$$M(t)\phi_u(t) = \lambda_u(t)\phi_u(t)$$

be an orthonormal eigendecomposition of the self-adjoint operator $M(t)$, and define

$$g_u(t) := \langle e(t), \phi_u(t) \rangle.$$

Then the modal coefficients satisfy

$$\dot{g}_u(t) = -\lambda_u(t)g_u(t) + \sum_{v \neq u} \omega_{v \rightarrow u}(t) g_v(t),$$

where

$$\omega_{v \rightarrow u}(t) = \frac{\Omega_{v \rightarrow u}(t)}{\lambda_u(t) - \lambda_v(t)}, \quad \Omega_{v \rightarrow u}(t) := \langle \phi_u(t), \dot{M}(t)\phi_v(t) \rangle.$$

Proof. We divide the proof into two steps.

Step 1: Eigenbasis differentiation formula. Differentiate the eigenvalue equation

$$M\phi_v = \lambda_v\phi_v$$

with respect to time:

$$\dot{M}\phi_v + M\dot{\phi}_v = \dot{\lambda}_v\phi_v + \lambda_v\dot{\phi}_v.$$

Taking inner product with ϕ_u for $u \neq v$, and using self-adjointness of M , we obtain

$$\langle \phi_u, \dot{M}\phi_v \rangle + \langle \phi_u, M\dot{\phi}_v \rangle = \dot{\lambda}_v \langle \phi_u, \phi_v \rangle + \lambda_v \langle \phi_u, \dot{\phi}_v \rangle.$$

Since $\langle \phi_u, \phi_v \rangle = 0$ for $u \neq v$ and

$$\langle \phi_u, M\dot{\phi}_v \rangle = \langle M\phi_u, \dot{\phi}_v \rangle = \lambda_u \langle \phi_u, \dot{\phi}_v \rangle,$$

it follows that

$$\langle \phi_u, \dot{M}\phi_v \rangle + \lambda_u \langle \phi_u, \dot{\phi}_v \rangle = \lambda_v \langle \phi_u, \dot{\phi}_v \rangle.$$

Therefore,

$$\langle \phi_u, \dot{\phi}_v \rangle = \frac{\langle \phi_u, \dot{M}\phi_v \rangle}{\lambda_v - \lambda_u}, \quad u \neq v.$$

Define

$$\Omega_{v \rightarrow u} := \langle \phi_u, \dot{M}\phi_v \rangle, \quad \omega_{v \rightarrow u} := \frac{\Omega_{v \rightarrow u}}{\lambda_u - \lambda_v}.$$

Then

$$\langle \phi_u, \dot{\phi}_v \rangle = -\omega_{v \rightarrow u} = \omega_{u \rightarrow v}, \quad u \neq v.$$

For the diagonal term, differentiating the normalization condition

$$\langle \phi_u, \phi_u \rangle = 1$$

gives

$$\langle \dot{\phi}_u, \phi_u \rangle + \langle \phi_u, \dot{\phi}_u \rangle = 0.$$

In a real Hilbert space this implies

$$\langle \phi_u, \dot{\phi}_u \rangle = 0.$$

Step 2: Derivation of the modal ODE. Recall that

$$g_u(t) = \langle e(t), \phi_u(t) \rangle.$$

Differentiating in time gives

$$\dot{g}_u = \langle \dot{e}, \phi_u \rangle + \langle e, \dot{\phi}_u \rangle.$$

By Proposition 3.1, the error evolves according to

$$\dot{e} = -Me.$$

Hence

$$\langle \dot{e}, \phi_u \rangle = -\langle Me, \phi_u \rangle = -\langle e, M\phi_u \rangle = -\lambda_u \langle e, \phi_u \rangle = -\lambda_u g_u.$$

For the second term, expand e in the instantaneous eigenbasis:

$$e = \sum_v g_v \phi_v.$$

Then

$$\langle e, \dot{\phi}_u \rangle = \sum_v g_v \langle \phi_v, \dot{\phi}_u \rangle.$$

The $v = u$ term vanishes because $\langle \phi_u, \dot{\phi}_u \rangle = 0$, so

$$\langle e, \dot{\phi}_u \rangle = \sum_{v \neq u} g_v \langle \phi_v, \dot{\phi}_u \rangle.$$

Using the formula from Step 1,

$$\langle \phi_v, \dot{\phi}_u \rangle = \frac{\langle \phi_v, \dot{M}\phi_u \rangle}{\lambda_u - \lambda_v}.$$

Since $\dot{M}$ is self-adjoint,

$$\langle \phi_v, \dot{M} \phi_u \rangle = \langle \phi_u, \dot{M} \phi_v \rangle = \Omega_{v \rightarrow u}.$$

Therefore,

$$\langle e, \dot{\phi}_u \rangle = \sum_{v \neq u} g_v \frac{\Omega_{v \rightarrow u}}{\lambda_u - \lambda_v} = \sum_{v \neq u} \omega_{v \rightarrow u} g_v.$$

Combining both terms yields

$$\dot{g}_u = -\lambda_u g_u + \sum_{v \neq u} \omega_{v \rightarrow u} g_v.$$

This proves the claim. $\square$

Remark C.2. The coefficients $\omega_{v \rightarrow u}$ arise entirely from the time dependence of the eigenbasis of $M(t)$. If the eigenbasis were fixed in time, then all off-diagonal couplings would vanish and each mode would dissipate independently as $\dot{g}_u = -\lambda_u g_u$.

D Shell Energy Dynamics

This section proves Theorem 3.3 in the main text. The main subtlety is that the shell boundaries are fixed in the Eulerian spectral coordinate, while the eigenvalues $\lambda_u(t)$ drift in time. As a result, the shell membership changes with time and an additional boundary-crossing term appears.

The proof is identical to the case of a linear error observable; only the definition of $M = J_e J_e^*$ differs, and none of the algebraic steps depend on the specific form of J_e . We reproduce the argument for completeness.

D.1 Spectral shells and shell energies

Let $\{I_\alpha\}_{\alpha=1}^K$ be a partition of the spectral range of interest, with

$$I_\alpha = [b_{\alpha-1}, b_\alpha), \quad b_0 < b_1 < \dots < b_K.$$

At each time t , define the shell index set

$$\mathcal{S}_\alpha(t) := \{u : \lambda_u(t) \in I_\alpha\}.$$

The shell quadratic energy is

$$E_\alpha(t) := \frac{1}{2} \sum_{u \in \mathcal{S}_\alpha(t)} g_u^2(t).$$

It is convenient to write this in terms of the indicator function

$$1_\alpha(\lambda) := 1_{[b_{\alpha-1}, b_\alpha)}(\lambda),$$

so that

$$E_\alpha(t) = \frac{1}{2} \sum_u 1_\alpha(\lambda_u(t)) g_u^2(t).$$

D.2 Differentiating the shell energy

Differentiating $E_\alpha(t)$ with respect to time yields

$$\frac{d}{dt} E_\alpha(t) = \sum_u 1_\alpha(\lambda_u(t)) g_u(t) \dot{g}_u(t) + \frac{1}{2} \sum_u g_u^2(t) \frac{d}{dt} 1_\alpha(\lambda_u(t)).$$

Equivalently,

$$\frac{d}{dt} E_\alpha(t) = \sum_{u \in \mathcal{S}_\alpha(t)} g_u(t) \dot{g}_u(t) + B_\alpha(t),$$

where

$$B_\alpha(t) := \frac{1}{2} \sum_u g_u^2(t) \frac{d}{dt} 1_\alpha(\lambda_u(t))$$

is the boundary-crossing contribution induced by eigenvalue drift.

Formally, $B_\alpha(t)$ vanishes unless some $\lambda_u(t)$ crosses one of the boundaries $b_{\alpha-1}$ or b_α . In the sense of distributions,

$$\frac{d}{dt} 1_\alpha(\lambda_u(t)) = \dot{\lambda}_u(t) \left[\delta(\lambda_u(t) - b_{\alpha-1}) - \delta(\lambda_u(t) - b_\alpha) \right].$$

Hence

$$B_\alpha(t) = \frac{1}{2} \sum_u g_u^2(t) \dot{\lambda}_u(t) \left[\delta(\lambda_u(t) - b_{\alpha-1}) - \delta(\lambda_u(t) - b_\alpha) \right].$$

D.3 Substitution of the modal dynamics

By Theorem 3.2, the modal coefficients satisfy

$$\dot{g}_u = -\lambda_u g_u + \sum_{v \neq u} \omega_{v \rightarrow u} g_v.$$

Substituting this into the previous identity gives

$$\frac{d}{dt} E_\alpha(t) = - \sum_{u \in \mathcal{S}_\alpha(t)} \lambda_u(t) g_u^2(t) + \sum_{u \in \mathcal{S}_\alpha(t)} \sum_{v \neq u} g_u(t) g_v(t) \omega_{v \rightarrow u}(t) + B_\alpha(t).$$

D.4 Cancellation of intra-shell terms

We now split the coupling sum into contributions from $v \in \mathcal{S}_\alpha(t)$ and $v \notin \mathcal{S}_\alpha(t)$:

$$\sum_{u \in \mathcal{S}_\alpha} \sum_{v \neq u} g_u g_v \omega_{v \rightarrow u} = \sum_{u \in \mathcal{S}_\alpha} \sum_{\substack{v \in \mathcal{S}_\alpha \\ v \neq u}} g_u g_v \omega_{v \rightarrow u} + \sum_{u \in \mathcal{S}_\alpha} \sum_{v \notin \mathcal{S}_\alpha} g_u g_v \omega_{v \rightarrow u}.$$

The first term vanishes by the antisymmetry

$$\omega_{v \rightarrow u} = -\omega_{u \rightarrow v},$$

since pairwise contributions cancel:

$$g_u g_v \omega_{v \rightarrow u} + g_v g_u \omega_{u \rightarrow v} = 0.$$

Therefore only inter-shell interactions remain:

$$\frac{d}{dt} E_\alpha(t) = - \sum_{u \in \mathcal{S}_\alpha(t)} \lambda_u(t) g_u^2(t) + \sum_{u \in \mathcal{S}_\alpha(t)} \sum_{v \notin \mathcal{S}_\alpha(t)} g_u(t) g_v(t) \omega_{v \rightarrow u}(t) + B_\alpha(t).$$

D.5 Inter-shell exchange form

For $\beta \neq \alpha$, define the inter-shell exchange

$$\mathcal{F}_{\beta \rightarrow \alpha}(t) := \sum_{u \in \mathcal{S}_\alpha(t)} \sum_{v \in \mathcal{S}_\beta(t)} g_u(t) g_v(t) \omega_{v \rightarrow u}(t).$$

Then

$$\sum_{u \in \mathcal{S}_\alpha(t)} \sum_{v \notin \mathcal{S}_\alpha(t)} g_u(t) g_v(t) \omega_{v \rightarrow u}(t) = \sum_{\beta \neq \alpha} \mathcal{F}_{\beta \rightarrow \alpha}(t),$$

and hence

$$\frac{d}{dt} E_\alpha(t) = - \sum_{u \in \mathcal{S}_\alpha(t)} \lambda_u(t) g_u^2(t) + \sum_{\beta \neq \alpha} \mathcal{F}_{\beta \rightarrow \alpha}(t) + B_\alpha(t).$$

This is exactly the shell energy balance stated in Theorem 3.3.

E Exact Microscopic Flux Representation

This section proves Theorem 3.4 in the main text. We show that the distribution-valued energy density and flux defined in Equations (17)–(19) satisfy the exact conservation law

$$\partial_t \varepsilon(\lambda, t) + \partial_\lambda \mathcal{F}(\lambda, t) = -2\lambda \varepsilon(\lambda, t)$$

in the sense of distributions.

E.1 Distribution-valued energy density

Recall that the instantaneous modal energy density is defined by

$$\varepsilon(\lambda, t) = \frac{1}{2} \sum_u g_u^2(t) \delta(\lambda_u(t) - \lambda).$$

For any test function $\psi \in C_c^\infty(\mathbb{R})$, we have

$$\langle \varepsilon(\cdot, t), \psi \rangle = \frac{1}{2} \sum_u g_u^2(t) \psi(\lambda_u(t)).$$

Thus integrating ε over a shell interval $I_\alpha = [b_{\alpha-1}, b_\alpha)$ recovers the shell energy:

$$\int_{I_\alpha} \varepsilon(\lambda, t) d\lambda = \frac{1}{2} \sum_{u: \lambda_u(t) \in I_\alpha} g_u^2(t) = E_\alpha(t).$$

E.2 Time derivative of the energy density

Differentiate $\varepsilon(\lambda, t)$ with respect to time. Using the product rule,

$$\partial_t \varepsilon(\lambda, t) = \sum_u g_u \dot{g}_u \delta(\lambda_u - \lambda) + \frac{1}{2} \sum_u g_u^2 \dot{\lambda}_u \delta'(\lambda_u - \lambda),$$

where δ' denotes the derivative of the Dirac distribution with respect to its argument.

Substituting the modal dynamics

$$\dot{g}_u = -\lambda_u g_u + \sum_{v \neq u} \omega_{v \rightarrow u} g_v,$$

we obtain

$$\partial_t \varepsilon(\lambda, t) = -\sum_u \lambda_u g_u^2 \delta(\lambda_u - \lambda) + \sum_{u,v} g_u g_v \omega_{v \rightarrow u} \delta(\lambda_u - \lambda) + \frac{1}{2} \sum_u g_u^2 \dot{\lambda}_u \delta'(\lambda_u - \lambda).$$

Since

$$\varepsilon(\lambda, t) = \frac{1}{2} \sum_u g_u^2 \delta(\lambda_u - \lambda),$$

the first term is exactly

$$-\sum_u \lambda_u g_u^2 \delta(\lambda_u - \lambda) = -2\lambda \varepsilon(\lambda, t),$$

because the delta distribution enforces $\lambda = \lambda_u$.

It remains to identify the last two terms as $\partial_\lambda \mathcal{F}$.

E.3 The drift part of the flux

Define

$$\mathcal{F}_{\text{drift}}(\lambda, t) = \frac{1}{2} \sum_u g_u^2(t) \dot{\lambda}_u(t) \delta(\lambda_u(t) - \lambda).$$

Then

$$\partial_\lambda \mathcal{F}_{\text{drift}}(\lambda, t) = -\frac{1}{2} \sum_u g_u^2 \dot{\lambda}_u \delta'(\lambda_u - \lambda),$$

since

$$\partial_\lambda \delta(\lambda_u - \lambda) = -\delta'(\lambda_u - \lambda).$$

Thus the drift flux exactly cancels the term coming from the time dependence of the eigenvalues in $\partial_t \varepsilon$.

E.4 The coupling part of the flux

Now define

$$\mathcal{F}_{\text{coup}}(\lambda, t) = \sum_{u,v} h(\lambda_u - \lambda) [1 - h(\lambda_v - \lambda)] g_u g_v \omega_{v \rightarrow u},$$

where h is the Heaviside step function.

We compute its λ -derivative in the sense of distributions. Using

$$\partial_\lambda h(\lambda_u - \lambda) = -\delta(\lambda_u - \lambda), \quad \partial_\lambda [1 - h(\lambda_v - \lambda)] = \delta(\lambda_v - \lambda),$$

we obtain

$$\partial_\lambda \mathcal{F}_{\text{coup}} = - \sum_{u,v} \delta(\lambda_u - \lambda) [1 - h(\lambda_v - \lambda)] g_u g_v \omega_{v \rightarrow u} + \sum_{u,v} h(\lambda_u - \lambda) \delta(\lambda_v - \lambda) g_u g_v \omega_{v \rightarrow u}.$$

In the second sum, interchange the dummy indices u and v :

$$\sum_{u,v} h(\lambda_u - \lambda) \delta(\lambda_v - \lambda) g_u g_v \omega_{v \rightarrow u} = \sum_{u,v} h(\lambda_v - \lambda) \delta(\lambda_u - \lambda) g_u g_v \omega_{u \rightarrow v}.$$

Using the antisymmetry

$$\omega_{u \rightarrow v} = -\omega_{v \rightarrow u},$$

this becomes

$$- \sum_{u,v} h(\lambda_v - \lambda) \delta(\lambda_u - \lambda) g_u g_v \omega_{v \rightarrow u}.$$

Therefore,

$$\partial_\lambda \mathcal{F}_{\text{coup}} = - \sum_{u,v} \delta(\lambda_u - \lambda) [1 - h(\lambda_v - \lambda) + h(\lambda_v - \lambda)] g_u g_v \omega_{v \rightarrow u},$$

hence

$$\partial_\lambda \mathcal{F}_{\text{coup}} = - \sum_{u,v} g_u g_v \omega_{v \rightarrow u} \delta(\lambda_u - \lambda).$$

This exactly cancels the coupling term appearing in $\partial_t \varepsilon$.

E.5 Exact conservation law

Combining the two pieces,

$$\mathcal{F}(\lambda, t) = \mathcal{F}_{\text{drift}}(\lambda, t) + \mathcal{F}_{\text{coup}}(\lambda, t),$$

that is,

$$\mathcal{F}(\lambda, t) = \frac{1}{2} \sum_u g_u^2(t) \dot{\lambda}_u(t) \delta(\lambda_u(t) - \lambda) + \sum_{u,v} h(\lambda_u(t) - \lambda) [1 - h(\lambda_v(t) - \lambda)] g_u(t) g_v(t) \omega_{v \rightarrow u}(t),$$

we obtain

$$\partial_\lambda \mathcal{F}(\lambda, t) = -\frac{1}{2} \sum_u g_u^2 \dot{\lambda}_u \delta'(\lambda_u - \lambda) - \sum_{u,v} g_u g_v \omega_{v \rightarrow u} \delta(\lambda_u - \lambda).$$

Substituting this into the expression for $\partial_t \varepsilon$ proves that

$$\partial_t \varepsilon(\lambda, t) + \partial_\lambda \mathcal{F}(\lambda, t) = -2\lambda \varepsilon(\lambda, t)$$

in the sense of distributions.

E.6 Recovery of the shell balance law

We now verify that the microscopic conservation law reproduces the exact shell bookkeeping derived in Appendix D.

Let $I_\alpha = [b_{\alpha-1}, b_\alpha]$ be a shell interval. Integrating the conservation law

$$\partial_t \varepsilon(\lambda, t) + \partial_\lambda \mathcal{F}(\lambda, t) = -2\lambda \varepsilon(\lambda, t)$$

over I_α yields

$$\frac{d}{dt} \int_{I_\alpha} \varepsilon(\lambda, t) d\lambda + \mathcal{F}(b_\alpha, t) - \mathcal{F}(b_{\alpha-1}, t) = -2 \int_{I_\alpha} \lambda \varepsilon(\lambda, t) d\lambda.$$

Using the definition of ε ,

$$\varepsilon(\lambda, t) = \frac{1}{2} \sum_u g_u^2(t) \delta(\lambda_u(t) - \lambda),$$

the leftmost term becomes

$$\frac{d}{dt} \int_{I_\alpha} \varepsilon(\lambda, t) d\lambda = \frac{d}{dt} E_\alpha(t).$$

We now examine the two contributions to the flux

$$\mathcal{F} = \mathcal{F}_{\text{drift}} + \mathcal{F}_{\text{coup}}.$$

Drift contribution. Recall

$$\mathcal{F}_{\text{drift}}(\lambda, t) = \frac{1}{2} \sum_u g_u^2(t) \dot{\lambda}_u(t) \delta(\lambda_u(t) - \lambda).$$

Evaluating the boundary term gives

$$\mathcal{F}_{\text{drift}}(b_\alpha, t) - \mathcal{F}_{\text{drift}}(b_{\alpha-1}, t) = \frac{1}{2} \sum_u g_u^2(t) \dot{\lambda}_u(t) [\delta(\lambda_u(t) - b_\alpha) - \delta(\lambda_u(t) - b_{\alpha-1})] = -B_\alpha(t).$$

This is exactly minus the boundary-crossing term $B_\alpha(t)$ derived in Appendix D. Hence the drift part of the microscopic flux accounts for changes in shell membership due to eigenvalue drift.

Coupling contribution. The coupling flux is

$$\mathcal{F}_{\text{coup}}(\lambda, t) = \sum_{u,v} h(\lambda_u(t) - \lambda) [1 - h(\lambda_v(t) - \lambda)] g_u(t) g_v(t) \omega_{v \rightarrow u}(t).$$

Evaluating its boundary difference yields

$$\mathcal{F}_{\text{coup}}(b_\alpha, t) - \mathcal{F}_{\text{coup}}(b_{\alpha-1}, t) = - \sum_{u \in S_\alpha(t)} \sum_{v \notin S_\alpha(t)} g_u(t) g_v(t) \omega_{v \rightarrow u}(t) = - \sum_{\beta \neq \alpha} \mathcal{F}_{\beta \rightarrow \alpha}(t),$$

which is precisely minus the inter-shell exchange term appearing in the shell energy balance.

Shell balance. Substituting these identities into the integrated conservation law gives

$$\frac{d}{dt} E_\alpha(t) = - \sum_{u \in S_\alpha(t)} \lambda_u(t) g_u^2(t) + \sum_{\beta \neq \alpha} F_{\beta \rightarrow \alpha}(t) + B_\alpha(t),$$

which coincides with the exact shell energy equation established in Appendix D.

Therefore the distribution-valued flux $\mathcal{F}(\lambda, t)$ provides an exact microscopic realization of the shell-level bookkeeping.

F Spectral estimation via (stochastic) Lanczos and shell binning

This appendix describes the Lanczos-based estimator used to obtain log-shell energies without forming the operator $M = J_e J_e^*$ explicitly.

F.1 Motivation

Directly forming the operator

$$M(\theta) = J_{e(\theta)} J_{e(\theta)}^*$$

is infeasible for modern neural networks due to its dimensionality. However, our analysis only requires access to spectral information and low-rank projections of M and its directional derivative.

The Lanczos method provides an efficient way to approximate the action of M on a low-dimensional Krylov subspace using only matrix–vector products. This makes it particularly well-suited for probing spectral structure in high-dimensional learning systems.

F.2 Automatic separation of learnable and irreducible components.

The model has N parameters, so $M = J_e J_e^*$ has at most N nonzero eigenvalues. The error decomposes as $e = e_{\parallel} + e_{\perp}$, where $e_{\parallel} \in \text{Range}(M)$ and $e_{\perp} \in \text{Null}(M)$. The Lanczos procedure, which builds a Krylov subspace via repeated application of M , probes only $e_{\parallel}$: from the second Krylov vector onward, the component $e_{\perp}$ is annihilated. Shell energies at $\lambda > 0$ therefore reflect only the learnable component. Under the binary KL formulation, the use of a reference model further eliminates the irreducible entropy floor: $D_{\text{bin}} \rightarrow 0$ as $q_{\theta} \rightarrow q_{\text{ref}}$, so the spectral probe measures a quantity that vanishes at convergence.

F.3 Goal: shell energies from the spectral measure of M

Fix a checkpoint θ and let $M := J_e J_e^*$ be the PSD operator governing error dynamics in function space. Let e denote the error observable ($e(x) = \sqrt{2 D_{\text{bin}}(x)}$ as defined in Section 4). Formally diagonalizing $M\phi_u = \lambda_u \phi_u$ and expanding $e = \sum_u g_u \phi_u$, the shell-integrated energy is

$$E_{\alpha} := \sum_{u: \lambda_u \in [\lambda_{\alpha}, \lambda_{\alpha+1})} g_u^2.$$

Equivalently, define the spectral measure μ_e of M induced by e so that for any test function f ,

$$\langle e, f(M)e \rangle = \int f(\lambda) d\mu_e(\lambda), \quad d\mu_e(\lambda) = \sum_u g_u^2 \delta(\lambda - \lambda_u) d\lambda.$$

Then E_{α} is exactly the mass of μ_e in the bin $[\lambda_{\alpha}, \lambda_{\alpha+1})$.

F.4 Lanczos tridiagonalization and quadrature

Lanczos constructs, from repeated applications of M to vectors, an orthonormal basis of the Krylov subspace $\mathcal{K}_m(M, v_0) = \text{span}\{v_0, Mv_0, \dots, M^{m-1}v_0\}$ and a symmetric tridiagonal matrix $T_m \in \mathbb{R}^{m \times m}$ such that

$$MV_m \approx V_m T_m, \quad V_m = [v_0, \dots, v_{m-1}], \quad V_m^* V_m = I.$$

We choose the starting vector to be the normalized error direction

$$v_0 := \frac{e}{\|e\|}.$$

Let $T_m = U \text{diag}(\tilde{\lambda}_i) U^*$ be the eigendecomposition of the tridiagonal matrix. Then Gauss quadrature yields an approximation of the spectral measure μ_e by the discrete measure

$$\mu_e(d\lambda) \approx \|e\|^2 \sum_{i=1}^m w_i \delta(\lambda - \tilde{\lambda}_i) d\lambda, \quad w_i := (U_{1i})^2,$$

where U_{1i} is the first component of the i -th eigenvector of T_m .

Algorithm 1 Lanczos shell-energy estimation at a fixed checkpoint

Require: Operator-vector product routine $u \mapsto Mu$; error vector e ; Lanczos steps m ; log-shell edges $\{\lambda_\alpha\}_{\alpha=0}^B$.

Ensure: Estimated shell energies $\hat{E}_\alpha$.

```

1:  $v_0 \leftarrow e/\|e\|$ ;  $\beta_0 \leftarrow 0$ ;  $v_{-1} \leftarrow 0$ 
2: for  $k = 0, \dots, m-1$  do
3:    $w \leftarrow Mv_k - \beta_k v_{k-1}$ 
4:    $\alpha_k \leftarrow \langle v_k, w \rangle$ 
5:    $\tilde{w} \leftarrow w - \alpha_k v_k$ 
6:    $\beta_{k+1} \leftarrow \|\tilde{w}\|$ 
7:   if  $\beta_{k+1} = 0$  then break
8:   end if
9:    $v_{k+1} \leftarrow \tilde{w}/\beta_{k+1}$ 
10: end for
11: Form tridiagonal  $T$  with diagonal entries  $\alpha_k$  and off-diagonals  $\beta_{k+1}$ 
12: Compute eigendecomposition  $T = U \text{diag}(\tilde{\lambda}_i) U^*$ 
13: Set weights  $w_i \leftarrow (U_{1i})^2$ 
14: Initialize  $\hat{E}_\alpha \leftarrow 0$  for all  $\alpha = 0, \dots, B-1$ 
15: for  $i = 1, \dots, m$  do
16:   Find  $\alpha$  such that  $\tilde{\lambda}_i \in [\lambda_\alpha, \lambda_{\alpha+1})$ 
17:   if such  $\alpha$  exists then
18:      $\hat{E}_\alpha \leftarrow \hat{E}_\alpha + \|e\|^2 w_i$ 
19:   end if
20: end for
21: return  $\{\hat{E}_\alpha\}_{\alpha=0}^{B-1}$ 

```

F.5 Log-shell binning

Given logarithmic bin edges $\{\lambda_\alpha\}_{\alpha=0}^B$, define the estimated shell energies by accumulating quadrature masses:

$$\hat{E}_\alpha := \|e\|^2 \sum_{i: \tilde{\lambda}_i \in [\lambda_\alpha, \lambda_{\alpha+1})} w_i.$$

We report $\log \hat{E}_\alpha$ as a function of $\log \lambda_\alpha^{\text{ctr}}$, where $\lambda_\alpha^{\text{ctr}}$ is the geometric bin center.

F.6 Data-induced averaging (no extra randomness)

We do not rely on a global random seed. Variance reduction is achieved by averaging the resulting shell estimates across many independent data batches: each batch provides an empirical error vector e (computed from a large set of tokens), and the above Lanczos+binning pipeline produces $\hat{E}_\alpha$ for that batch; we then average $\hat{E}_\alpha$ over batches.

F.7 Pseudo-code

Implementation note. The only required primitive is the product $u \mapsto Mu = J_e(J_e^* u)$. In modern autodiff frameworks this can be implemented via a VJP followed by a JVP, where the Jacobian is that of the error observable $e(x) = \sqrt{2 D_{\text{bin}}(x)}$. The chain rule through the nonlinear transformation is handled automatically.

G Power-Law Decay under Power-Law Initial Data

When $\mathcal{F} = 0$, Equation (20) reduces to $\partial_t \varepsilon(\lambda, t) = -2\lambda \varepsilon(\lambda, t)$, with solution $\varepsilon(\lambda, t) = \varepsilon(\lambda, 0) e^{-2\lambda t}$. The integrated loss is

$$L(t) = \int_0^\infty \varepsilon(\lambda, t) d\lambda = \int_0^\infty \varepsilon(\lambda, 0) e^{-2\lambda t} d\lambda.$$

Suppose the initial spectral density follows a power law: $\varepsilon(\lambda, 0) = C \lambda^{-\beta}$ for λ in the relevant spectral range. Then by the substitution $u = 2\lambda t$,

$$L(t) = C \int_0^\infty \lambda^{-\beta} e^{-2\lambda t} d\lambda = C (2t)^{\beta-1} \int_0^\infty u^{-\beta} e^{-u} du = C \Gamma(1-\beta) (2t)^{\beta-1}.$$

Thus $L(t) \propto t^{\beta-1}$, which is a power law in t with exponent $\beta - 1 < 0$. The exponent is determined by the spectral slope β of the initial error distribution.

H Proof of Theorem 5.1

We establish the corrected loss formula (24) by a first-order perturbation analysis of the transport–dissipation PDE, then derive each claim of Theorem 5.1 from the resulting expression.

Setup

Write the net spectral flux as $\mathcal{F}(\lambda, t) = v(\lambda, t) \varepsilon(\lambda, t)$, where v is a signed velocity field encoding the direction and magnitude of spectral transport. The transport–dissipation equation (20) becomes

$$\partial_t \varepsilon(\lambda, t) = -2\lambda \varepsilon(\lambda, t) - \partial_\lambda [v(\lambda, t) \varepsilon(\lambda, t)]. \quad (25)$$

We treat v as a small perturbation and expand $\varepsilon = \varepsilon^{(0)} + \varepsilon^{(1)} + O(v^2)$, where $\varepsilon^{(0)}$ is the unperturbed solution and $\varepsilon^{(1)}$ is the first-order correction.

Unperturbed solution

Setting $v = 0$ in (25) gives the pure dissipation equation $\partial_t \varepsilon^{(0)} = -2\lambda \varepsilon^{(0)}$, with unique solution

$$\varepsilon^{(0)}(\lambda, t) = \varepsilon(\lambda, 0) e^{-2\lambda t}. \quad (26)$$

The corresponding unperturbed loss is

$$L_{\text{PL}}(t) = \int_0^\infty \varepsilon^{(0)}(\lambda, t) d\lambda = \int_0^\infty \varepsilon(\lambda, 0) e^{-2\lambda t} d\lambda, \quad (27)$$

which decays as a power law whenever $\varepsilon(\lambda, 0)$ is power-law distributed (see Appendix G).

First-order correction

Substituting $\varepsilon = \varepsilon^{(0)} + \varepsilon^{(1)}$ into (25) and collecting $O(v)$ terms gives

$$\partial_t \varepsilon^{(1)} = -2\lambda \varepsilon^{(1)} - \partial_\lambda [v \varepsilon^{(0)}]. \quad (28)$$

This is a linear inhomogeneous equation for $\varepsilon^{(1)}$ with the same dissipation operator as the unperturbed problem. Using the integrating factor $e^{2\lambda t}$, the unique solution with $\varepsilon^{(1)}(\lambda, 0) = 0$ is

$$\varepsilon^{(1)}(\lambda, t) = -e^{-2\lambda t} \int_0^t e^{2\lambda s} \partial_\lambda [v(\lambda, s) \varepsilon^{(0)}(\lambda, s)] ds. \quad (29)$$

Integrating over λ to obtain the first-order correction to the loss,

$$L^{(1)}(t) = \int_0^\infty \varepsilon^{(1)}(\lambda, t) d\lambda = - \int_0^\infty e^{-2\lambda t} \int_0^t e^{2\lambda s} \partial_\lambda [v \varepsilon^{(0)}] ds d\lambda. \quad (30)$$

Integrating by parts in λ (boundary terms vanish since ε decays as $\lambda \rightarrow 0^+$ and $\lambda \rightarrow \infty$),

$$L^{(1)}(t) = - \int_0^\infty \int_0^t 2(t-s) e^{-2\lambda(t-s)} v(\lambda, s) \varepsilon^{(0)}(\lambda, s) ds d\lambda. \quad (31)$$

Substituting $\varepsilon^{(0)}(\lambda, s) = \varepsilon(\lambda, 0) e^{-2\lambda s}$,

$$L^{(1)}(t) = -2 \int_0^t (t-s) \int_0^\infty v(\lambda, s) \varepsilon(\lambda, 0) e^{-2\lambda t} d\lambda ds. \quad (32)$$

Reduction to the scalar correction

Define, at each observation time t , the normalized spectral weight $w_t(\lambda) := \varepsilon(\lambda, 0)e^{-2\lambda t}$, whose total mass is exactly the unperturbed loss, $\int_0^\infty w_t(\lambda) d\lambda = L_{\text{PL}}(t)$, and let

$$\bar{v}(s) := \frac{\int_0^\infty v(\lambda, s) w_t(\lambda) d\lambda}{\int_0^\infty w_t(\lambda) d\lambda} \quad (33)$$

be the spectrally weighted average of v at time s . Note that the weight w_t , and hence $\bar{v}$, depends on the observation time t ; we suppress this in the notation since only the bound $\bar{v} \geq v_{\min}$ is used below. Equation (32) then reads

$$L^{(1)}(t) = -2 L_{\text{PL}}(t) \int_0^t (t-s) \bar{v}(s) ds = -2 \bar{V}(t) L_{\text{PL}}(t), \quad (34)$$

where the accumulated drift is defined with the memory kernel $(t-s)$,

$$\bar{V}(t) := \int_0^t (t-s) \bar{v}(s) ds. \quad (35)$$

The total loss to first order is therefore

$$L(t) = L_{\text{PL}}(t) + L^{(1)}(t) + O(\bar{V}^2) = L_{\text{PL}}(t) [1 - 2 \bar{V}(t)] + O(\bar{V}^2), \quad (36)$$

which is Equation (24). $\square$

Proof of claims (i)–(iii)

Claim (i): Local power-law. When $2\bar{V}(t) \ll 1$, the correction term is negligible and $L(t) = L_{\text{PL}}(t) [1 + O(2\bar{V})] \approx L_{\text{PL}}(t)$. Since L_{PL} is a power law (Appendix G), so is $L(t)$ to leading order. $\square$

Claim (ii): Systematic undershoot. Since $v \geq v_{\min} > 0$, the weighted average obeys $\bar{v}(s) \geq v_{\min}$ for every s , so $\bar{V}(t) = \int_0^t (t-s) \bar{v}(s) ds > 0$ and the correction factor $1 - 2\bar{V}(t) < 1$, giving $L(t) < L_{\text{PL}}(t)$. Moreover $\bar{V}(t) \geq v_{\min} t^2/2$, so the relative undershoot $(L_{\text{PL}} - L)/L_{\text{PL}} = 2\bar{V}(t)$ grows without bound. $\square$

Claim (iii): Irreversibility. Suppose one fits a power law $\tilde{L}_{\text{PL}}$ to the true trajectory $L(t)$ over a later window $[t_1, t_2]$ with t_1 large enough that $2\bar{V}(t_1)$ is non-negligible. Over this window, $L(t) = L_{\text{PL}}(t) [1 - 2\bar{V}(t)]$. Since $2\bar{V}(t)$ is itself varying in t , the true trajectory is not a pure power law: fitting a power law to $L(t)$ over $[t_1, t_2]$ yields an effective exponent $\tilde{\alpha}$ that differs from both the early-window exponent and the true instantaneous exponent. Since $\bar{V}(t)$ continues to grow for $t > t_2$, the relative discrepancy between $L(t)$ and any fixed power law fitted on $[t_1, t_2]$ keeps increasing throughout the regime in which the first-order expansion remains valid. No finite re-fit removes the ongoing accumulation. $\square$

I Fitting Procedure for Power-Law Validation

Following the standard three-parameter form of Hoffmann et al. (2022), we fit

$$L(t) = C_\infty + A \cdot t^\alpha \quad (37)$$

to all checkpoints with $t \leq 32,000$ steps, a window that, by the spectral analysis of Section 4, lies well within the quasi-static phase where $\bar{V}(t)$ remains small and power-law decay holds locally. The floor C_∞ , amplitude A , and exponent α are estimated jointly: we perform a grid search over $C_\infty \in (-\infty, \min_t L(t))$, with closed-form ordinary least squares in log-space for A and α at each fixed C_∞ , followed by scalar refinement via bounded minimization.⁴ The fit quality within the training window is uniformly high (RMSE < 0.025 nats across all three models). The resulting curves are extrapolated to the remainder of training without further adjustment.

⁴We fit the raw cross-entropy $\mathcal{L}_{\text{CE}}$ rather than the KL divergence, following the convention of Hoffmann et al. (2022). The irreducible entropy $H[p]$ appears as a constant offset absorbed into C_∞ and does not affect the predicted undershoot.

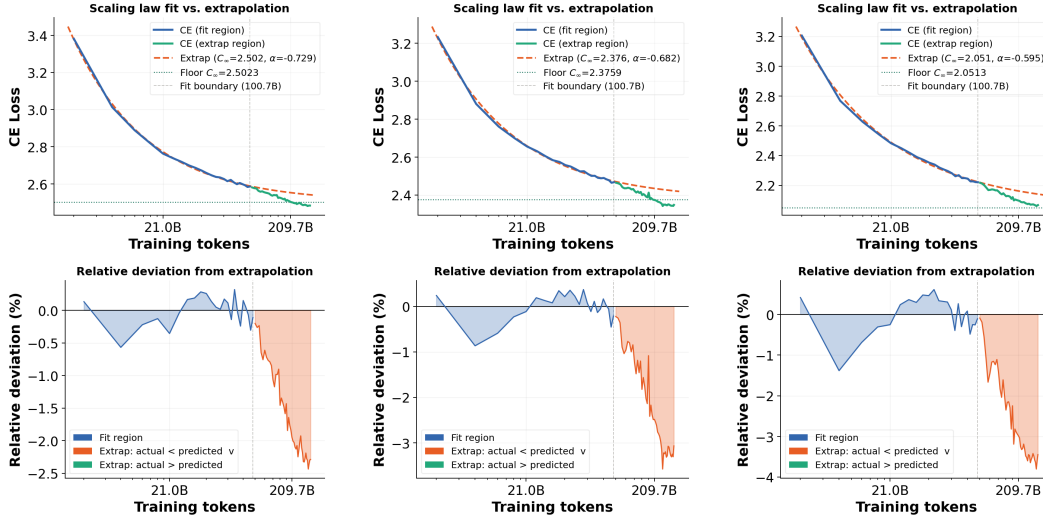


Figure 3: Robustness check of Theorem 5.1 on Pythia-410M, 1B, and 7B (left to right) with fit boundary at 48k steps (100.7B tokens). Layout follows Figure 2. **Top row:** CE loss trajectories (solid) against the power-law fit extrapolated from steps 1k–48k (dashed); dotted line shows fitted floor C_∞ . **Bottom row:** relative deviation $(L_{\text{actual}} - L_{\text{pred}})/L_{\text{pred}}$. The systematic undershoot persists across all three model scales, confirming that the phenomenon is not an artifact of the 32k-step fit boundary used in the main text.

J Robustness of the Undershoot to the Fit Window

The main text (Section 5) fits the power-law model $L(t) = C_\infty + A \cdot t^\alpha$ to checkpoints with $t \leq 32,000$ steps (corresponding to 67.1B tokens) and extrapolates to the remainder of training. A natural concern is whether the systematic undershoot reported in Figure 2 is an artifact of this particular choice of fit window. To address this, we repeat the identical fitting and extrapolation procedure using two alternative fit boundaries: $t \leq 48,000$ steps (100.7B tokens) and $t \leq 64,000$ steps (134.2B tokens).

Figures 3 and 4 report the results. In both cases, and across all three model scales (Pythia-410M, 1B, 7B), the qualitative pattern is unchanged: almost every held-out checkpoint beyond the fit boundary falls below the extrapolated power law, and the undershoot grows monotonically. The fitted parameters (C_∞, A, α) shift slightly as the fit window expands, absorbing a small portion of the accumulated drift $\bar{V}(t)$ into the effective exponent, but the residual undershoot remains unambiguously one-sided. This is precisely what Theorem 5.1 predicts: since the correction $2\bar{V}(t)$ grows monotonically, no finite fit window can fully absorb the drift, and extrapolation from any window will eventually underestimate the true loss reduction.

These results confirm that the undershoot is not sensitive to the choice of fit boundary and reflects a genuine dynamical effect rather than a fitting artifact.

K Spectral and Loss Evidence under a WSD Schedule

We trained a ~ 109 M Llama-style model (10 layers, 8 heads, $d_{\text{model}} = 768$, sequence length 512, GPT-2 tokenizer) on SlimPajama (Shen et al., 2023) for 12,500 steps under a Warmup-Stable-Decay schedule (Hu et al., 2024) using the public framework of Hägele et al. (2024): 300 warmup steps, a constant learning rate of 10^{-3} , and a linear cooldown over the final 20% of training (steps 10,000–12,500). Checkpoints were saved every 500 steps; the spectral probe uses the final checkpoint as the binary-KL reference and is otherwise identical to the procedure of Section 4.

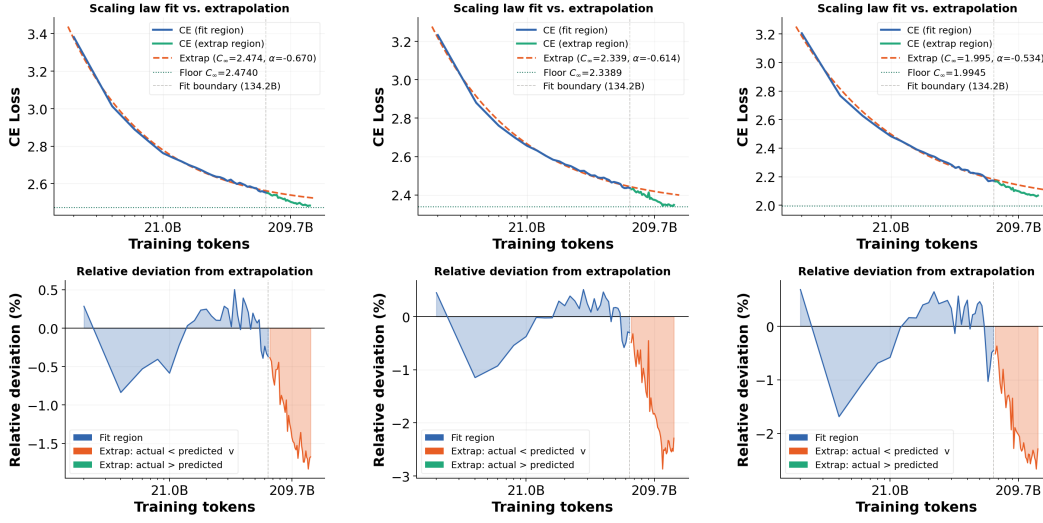


Figure 4: Robustness check of Theorem 5.1 on Pythia-410M, 1B, and 7B (left to right) with fit boundary at 64k steps (134.2B tokens). Layout follows Figure 2. **Top row:** CE loss trajectories (solid) against the power-law fit extrapolated from steps 1k–64k (dashed); dotted line shows fitted floor C_∞ . **Bottom row:** relative deviation $(L_{\text{actual}} - L_{\text{pred}})/L_{\text{pred}}$. Even with the fit window extended to cover roughly half the total training budget, the undershoot beyond the fit boundary remains strictly one-sided and monotonically growing, consistent with the prediction of Theorem 5.1.

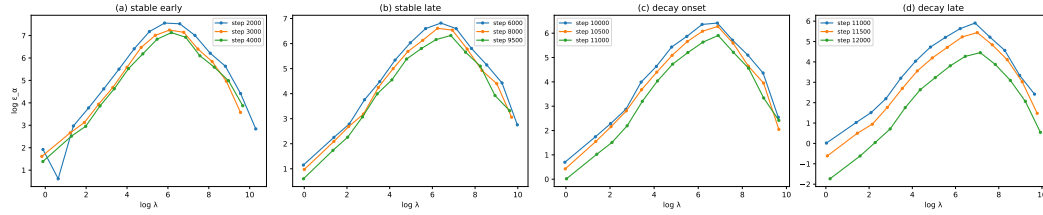


Figure 5: We applied the stochastic-Lanczos shell diagnostic on the ~ 109 M model trained under a WSD schedule. The left shoulders are approximately aligned during the constant-learning-rate (stable) phase (the $\mathcal{F} \approx 0$ signature) and migrate systematically toward larger λ during the cooldown.

Figure 5 shows the shell-resolved spectral profiles. During the stable phase the left shoulders of successive checkpoints are aligned ($\mathcal{F} \approx 0$); during the cooldown the shoulder migrates systematically toward larger λ , indicating a nonzero net drift. We note that the earliest checkpoints, shortly after the short warmup, have not yet settled into the quasi-static regime, so the stable-phase shoulders are read from checkpoints past this initial transient, the same situation as in Pythia, where the quasi-static regime also emerges after early training rather than from step zero.

Figure 6 shows the loss analysis on the same run. A power law fit to the stable phase tracks the loss with relative deviation near zero throughout that phase; the deviation turns negative at the onset of the cooldown and grows monotonically to about -3 to -4% by the end of the run, comparable to the 3–5% observed on Pythia. The same stable-then-cooldown pattern is visible in the public WSD logs of Hägele et al. (2024) and Schaipp et al. (2025).

L Validation on OLMo-2 7B

To test whether the systematic undershoot extends beyond the Pythia family, we repeat the fit-and-extrapolate procedure on OLMo-2 7B, which differs from Pythia in architecture,

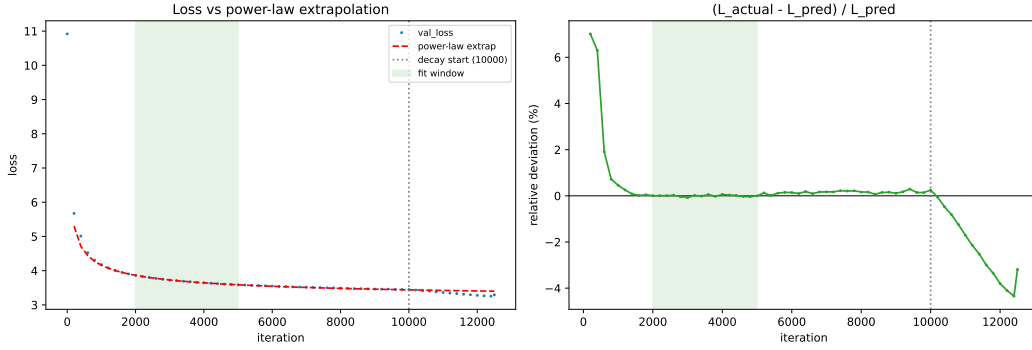


Figure 6: A power-law fit to the stable phase of the WSD run tracks the loss with relative deviation near zero throughout that phase; the deviation turns negative at the onset of the cooldown and grows to a few percent by the end.

optimizer, and training data, and is trained on roughly an order of magnitude more tokens. Rather than the coarse checkpoint subsampling of an earlier version, we fit directly to the full public training-loss trajectory (from the released WandB logs), which gives dense coverage of the loss curve.

Figure 7 shows the result for two choices of fit boundary. In both cases the loss tracks the power-law fit closely within the fit region and then falls one-sidedly and monotonically below the extrapolation, reaching about -3% by the end of training—comparable to the 3–5% observed on Pythia, and one-sided throughout. That the undershoot is essentially unchanged when the fit boundary is moved from $\sim 220k$ to $\sim 440k$ steps confirms it is not an artifact of a particular fit window, mirroring the robustness checks on Pythia (Appendix J).

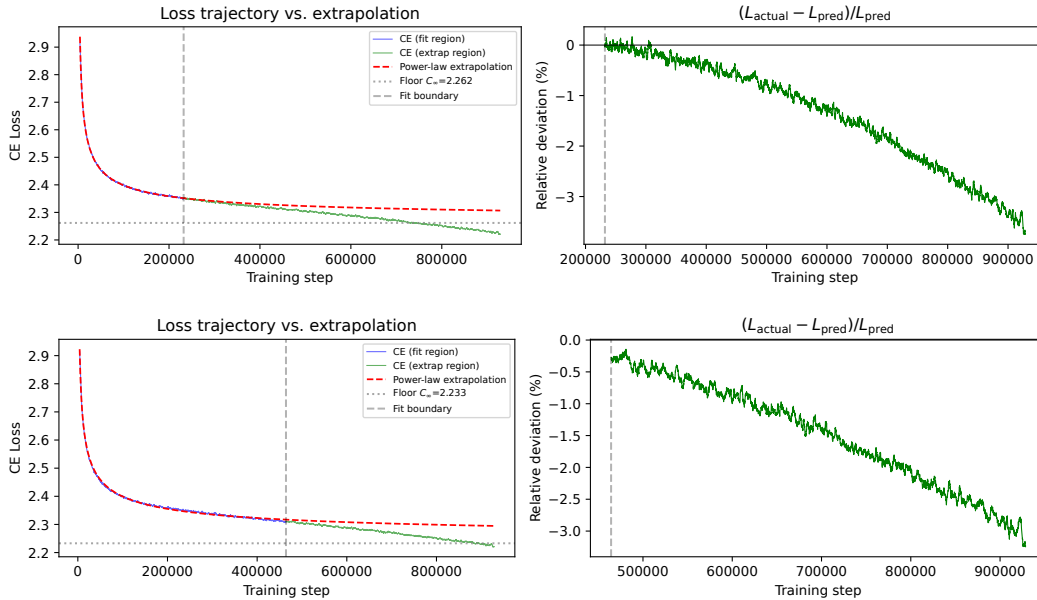


Figure 7: Power-law fit and undershoot for OLMo-2 7B, fit to the full public training-loss trajectory. **Top:** fit boundary at $\sim 220k$ steps ($C_\infty = 2.262$). **Bottom:** fit boundary at $\sim 440k$ steps ($C_\infty = 2.233$). **Left:** CE loss (solid) vs. power-law extrapolation (dashed); dotted line is the fitted floor C_∞ . **Right:** relative deviation $(L_{\text{actual}} - L_{\text{pred}}) / L_{\text{pred}}$. Under both fit boundaries the loss undershoots the extrapolation one-sidedly and monotonically, reaching about -3% by the end of training.

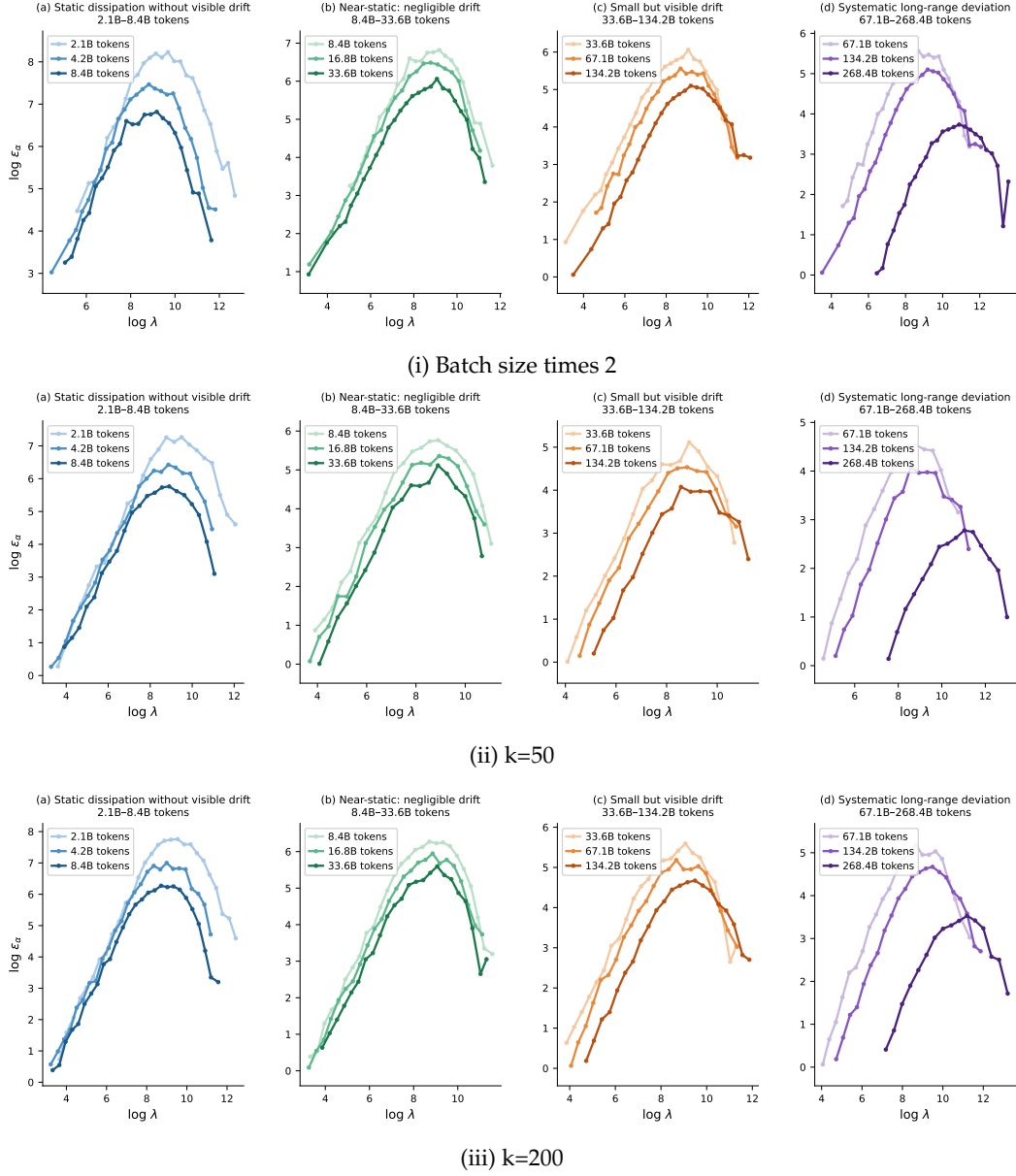


Figure 8: The shoulder diagnostic is stable to the Lanczos truncation and calibration batch size: over $k = 50/200$ and a doubled calibration batch, the shoulder positions shift by less than about half a shell ($\Delta \log \lambda \lesssim 0.5$), an order of magnitude below the drift signal ($\Delta \log \lambda \approx 3$ across panels).

M Robustness of the Spectral Diagnostic

In Figure 8, we verify that the shoulder diagnostic does not depend on the Lanczos resolution or the calibration batch size. Repeating the analysis at $k = 50, 200$ Lanczos steps and at a doubled calibration batch, the left-shoulder positions shift by less than about half a shell ($\Delta \log \lambda \lesssim 0.5$), an order of magnitude below the drift signal ($\Delta \log \lambda \approx 3$ across panels).