Universality Classes of Spectral Learning Dynamics

Dr. Tamás Nagy

tnagyphd@gmail.com

Draft

During the preparation of this work the author used large language models in order to assist with manuscript drafting, literature search, and coding assistance. After using these tools, the author reviewed and edited the content as needed and takes full responsibility for the content of the published article.

Reader-Friendly Subtitle

One spectral language for SGD, Adam, and attention training behavior.

Technical Strapline

A finite-class characterization of optimizer dynamics via spectral invariants and perturbative equivalence.

Executive Summary (Non-Technical)

Modern training pipelines use many optimizers, but teams repeatedly observe similar large-scale behavior. This paper asks whether that similarity is accidental, or whether different optimizers follow a shared structure when viewed in the right coordinates.

The core idea is to move from optimizer-specific parameter updates to a common spectral state representation. In that representation, many algorithmic differences may become secondary effects, while the dominant dynamics follow a small number of reusable classes.

If this is correct, we get a practical win: faster model diagnostics, class-aware optimizer selection, and more predictable scaling behavior across architectures.

The paper does not claim all optimizers are identical. It claims that under explicit assumptions, they can be grouped into a finite set of spectral universality classes with measurable transition boundaries.

Abstract

Modern learning systems appear algorithmically diverse yet empirically convergent toward a small set of training regimes. We propose a universality framework in which optimization dynamics are classified by spectral invariants rather than by optimizer-specific update rules. We define a common spectral state process, identify invariant scaling signatures, and show that SGD-, Adam-, and attention-driven training trajectories fall into a finite set of universality classes under mild regularity assumptions.

1. Problem

Different optimizers look different in update equations but often similar in macroscopic outcomes. SGD, Adam, and self-attention dynamics differ at the algorithmic level but converge to a small number of observable training regimes. The central problem is: which macroscopic behaviors are optimizer-specific and which are universal features of the spectral landscape?

This matters beyond taxonomy. If universality classes exist, then optimizer selection reduces to class matching rather than trial-and-error, and scaling predictions transfer across algorithms within a class.

2. Setup

2.1 Spectral State Representation

Let $\Sigma_t = \mathbb{E}[\nabla L \nabla L^\top]$ denote the gradient covariance at step t. Diagonalize in the Hessian eigenbasis: $\Sigma_t = U \operatorname{diag}(\sigma_1(t), \dots, \sigma_d(t)) U^\top$.

Definition 1 (Spectral State). The spectral state at time t is the mode-energy profile $\mathbf{e}(t) = (e_1(t), \dots, e_K(t))$ where $e_k(t) = \sigma_k(t)/\lambda_k$ and λ_k is the k-th Hessian eigenvalue.

2.2 Spectral Invariants

Definition 2 (Class Invariants). A universality class $\mathcal{U}$ is characterized by: - Spectral exponent s: the power-law decay $\lambda_k \sim C_\lambda k^{-s}$ - Noise coupling ratio $\eta_k = \sigma_k^2/\lambda_k^2$ - Contraction signature γ_k: per-mode convergence rate

2.3 Optimizer-to-Spectral Map

For each optimizer $\mathcal{O} \in \{\text{SGD}, \text{Adam}, \text{Attention}\}$, we write the induced spectral dynamics as:

$$\frac{de_k}{dt} = -f_{\mathcal{U}}(e_k, k) + \delta_{\mathcal{O}}(e_k, k, t)$$

where $f_{\mathcal{U}}$ is the class-level drift and $\delta_{\mathcal{O}}$ is the optimizer-specific perturbation. The universality claim is that $\delta_{\mathcal{O}}$ is lower-order under explicit scaling assumptions.

Connection to existing kernel: the spectral exponent s is the same object that appears in the Scaling Laws paper (ScalingLaws/SpectralData.lean), where s governs Chinchilla-type allocation. The present paper extends s from a data property to a joint data-optimizer invariant.

3. Main Theorem

Theorem Candidate 1 (Spectral Universality). Let $\lambda_k \sim C_\lambda k^{-s}$ with $s > 1$. Suppose the optimizer $\mathcal{O}$ satisfies: - (A1) Bounded second moment: $\mathbb{E}[\|\nabla L\|^2] \le G^2$ - (A2) Hessian regularity: $\|\nabla^2 L - H\|_{\text{op}} \le \beta$ for slowly-varying H - (A3) Mode isolation: $\lambda_k/\lambda_{k+1} \ge 1 + c/k$ for some $c > 0$

Then the spectral dynamics satisfy:

$$\left\| \frac{de_k}{dt} + f_s(e_k, k) \right\| \le C \cdot k^{-(s+1)/2} \cdot \|e\|_\infty$$

where f_s depends only on s and the noise coupling structure, not on the optimizer identity.

Corollary (Class Transition). A class transition from $\mathcal{U}_s$ to $\mathcal{U}_{s'}$ occurs when the effective spectral exponent $\hat{s}_t$ crosses a critical threshold at the Marchenko-Pastur edge $k^* = n/p$.

4. Proof Sketch

1. **Common spectral form.** Write each optimizer update $\theta_{t+1} = \theta_t - \alpha_t M_t^{-1} g_t$ in the Hessian eigenbasis. The preconditioner M_t differs across optimizers but acts diagonally in the spectral basis up to off-diagonal coupling of order $O(\beta)$.
2. **Perturbation bound.** For SGD, $M_t = I$. For Adam, $M_t \approx \mathrm{diag}(\hat{v}_k^{1/2})$. The difference $\delta_{\mathcal{O}}$ satisfies $\|\delta\|_k \leq C\beta/\lambda_k$ under (A2).
3. **Scaling closure.** Under (A3), the mode-wise dynamics decouple at leading order, and the perturbation decays faster than the class drift for large k.
4. **Class identification.** The invariant (s, η, γ) is preserved along trajectories up to controlled fluctuations, yielding finite equivalence classes.

5. Empirics/Simulation

5.1 Synthetic Benchmark

- Teacher-student regression with planted spectral exponent $s \in \{1.0, 1.5, 2.0, 3.0\}$.
- Train with SGD, Adam, and linear attention. Measure $\hat{s}_t$ along trajectories.
- Prediction: trajectories with same s should cluster regardless of optimizer.

5.2 Real Benchmarks

- CIFAR-10 (ResNet-18, ViT-Tiny): measure Hessian eigenspectrum during training.
- Report class assignment stability across optimizers.

5.3 Class Transition Scan

- Sweep batch size and learning rate to locate class boundaries.
- Compare predicted vs observed transition points.

6. Limits

- **Nonstationary data:** streaming or adversarial distributions break the slowly-varying assumption (A2).
- **Heavy-tail gradients:** Lévy-stable gradient noise violates (A1) and may create additional classes.
- **Architecture dependence:** normalization layers and skip connections can alter effective spectral geometry.
- **Finite-sample invariant estimation:** practical class assignment requires sufficient training steps.

7. Related Work

- **Scaling laws:** Hoffmann et al. (2022), Kaplan et al. (2020) — empirical scaling; our Scaling Laws paper derives $L^*(C) \sim C^{-(s-1)/(s+1)}$ from spectral exponent s.
- **Optimization dynamics:** Cohen et al. (2021) on edge-of-stability; Lewkowycz et al. (2020) on catapult phase.
- **Universality in physics:** Kadanoff (1966), Wilson (1971) — renormalization group and universality classes.
- **Random matrix theory:** Marchenko-Pastur (1967), Pennington-Worah (2017) — spectral bulk/edge structure.

8. Cross-Paper Connections

This paper sits at the center of a natural network among the 9 new directions:

- **Phase Transitions (paper 5):** the spectral exponent s that defines a universality class here is also the order parameter that governs generalization phase transitions. A class transition in optimizer dynamics corresponds to a cusp in generalization error.
- **Ergodic Control (paper 7):** if the optimizer is viewed as a controller acting on mode energies, the per-mode contraction rates γ_k must respect the class structure. A controller that tries to force behavior inconsistent with the underlying universality class will exhibit instability.
- **Spectral RL (paper 10):** policy optimization is itself a learning dynamic. The universality framework should classify policy-gradient trajectories in the same way it classifies SGD/Adam.
- **Decision Functional Approximation (paper 8):** the universality-class invariants define which spectral budgets are sufficient for approximating value functionals within each class.

9. Concrete Example: SGD and Adam in the Same Class

9.1 Setup

Consider a quadratic loss $L(\theta) = \frac{1}{2}\theta^\top H\theta$ with $H = \mathrm{diag}(\lambda_1, \ldots, \lambda_d)$, $\lambda_k = k^{-s}$, and stochastic gradient noise $g_t = H\theta_t + \xi_t$, $\xi_t \sim \mathcal{N}(0, \sigma^2 I)$.

SGD dynamics (mode k):

$$\theta_k^{(t+1)} = (1 - \alpha\lambda_k)\theta_k^{(t)} - \alpha\xi_k^{(t)}$$

Spectral energy: $e_k(t) = \mathbb{E}[\theta_k^{(t)2}]$.

Drift: $de_k/dt = -2\alpha\lambda_k e_k + \alpha^2\sigma^2$.

Steady state: $e_k^* = \alpha\sigma^2/(2\lambda_k) \propto k^s$.

Adam dynamics (mode k): With bias-corrected second moment $\hat{v}_k \approx \lambda_k^2 e_k + \sigma^2$:

$$\theta_k^{(t+1)} \approx \theta_k^{(t)} - \frac{\alpha}{\sqrt{\hat{v}_k}}(H\theta)_k$$

Effective per-mode learning rate: $\alpha_k^{\mathrm{eff}} = \alpha/\sqrt{\lambda_k^2 e_k + \sigma^2}$.

At steady state, $\alpha_k^{\text{eff}} \to \alpha/\sigma$ (noise-dominated regime), so the Adam drift becomes:

$de_k/dt \approx -2(\alpha/\sigma)\lambda_k e_k + (\alpha/\sigma)^2\sigma^2$.

This has the **same functional form** as SGD with rescaled learning rate $\tilde{\alpha} = \alpha/\sigma$.

9.2 Class Identification

Both optimizers produce $e_k^* \propto k^s$ in steady state. The invariant triple is: - Spectral exponent: s (same for both, determined by data). - Noise ratio: $\eta = \sigma^2/\lambda_1^2$ (same for both, determined by data + noise). - Contraction rate: $\gamma_k = 2\tilde{\alpha}\lambda_k$ (same functional form, different constant).

The constant difference in $\tilde{\alpha}$ is a **within-class parameter**, not a class distinction. Both optimizers belong to universality class $\mathcal{U}_s$ with the same s and η.

10. Algorithm: Spectral Class Identification

Input: Training trajectory $\{\theta^{(t)}\}_{t=1}^T$, Hessian eigenbasis $\{v_k\}$.

1. Project: $c_k^{(t)} = v_k^\top \theta^{(t)}$ for each mode k.
2. Estimate mode energies: $\hat{e}_k = T^{-1}\sum_t (c_k^{(t)})^2$.
3. Fit spectral exponent: regress $\log \hat{e}_k$ on $\log k$ to obtain $\hat{s}$.
4. Estimate noise ratio: $\hat{\eta} = \hat{\sigma}^2/\hat{\lambda}_1^2$.
5. Estimate contraction: $\hat{\gamma}_k$ from autocorrelation of $c_k^{(t)}$.
6. Assign class: $\widehat{\mathcal{U}} = (\hat{s}, \hat{\eta}, \hat{\gamma})$.

Class transition detection: monitor $\hat{s}_t$ over sliding windows. If $|\hat{s}_t - \hat{s}_{t-\Delta}| > \epsilon_{\text{crit}}$, flag a class transition.

11. Outlook

- **Class-aware optimizer selection:** match optimizer to data spectral class rather than hyperparameter search.
- **Training schedule design:** transition-aware learning rate schedules that respect class boundaries.
- **Bridge to spectral RL:** extend universality classes to policy optimization dynamics (connects to paper 10, Spectral RL).
- **Lean formalization target:** the perturbation bound in Theorem 1 is a natural candidate for LeanProofs/SpectralUniversality/PerturbationBound.lean.
- **Experimental validation priority:** the quadratic-loss class-equivalence example (Section 9) is the simplest testable prediction and should be the first empirical milestone.