Spectral Phase Transitions in Generalization

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

Why generalization can change abruptly instead of gradually.

Technical Strapline

Sharp threshold theorems from mode occupancy, noise coupling, and effective-rank geometry.

Executive Summary (Non-Technical)

Generalization is often described as a smooth trade-off, but in practice performance can shift suddenly. This paper treats those shifts as genuine phase transitions, not noise.

The key idea is to define spectral order parameters that govern regime changes. When these cross critical boundaries, the dominant error mechanism changes and test behavior can move abruptly.

A successful theory here would provide practical transition diagnostics: where training is still in a safe regime and where small changes can trigger large performance shifts.

The paper does not claim all tasks have sharp transitions. It characterizes conditions under which sharp threshold behavior should appear.

Abstract

We develop a phase-transition theory of generalization based on spectral order parameters. Instead of smooth monotonic scaling, we show that model performance exhibits threshold behavior when mode occupancy, noise-floor coupling, and effective rank cross critical boundaries. We derive sharp transition conditions and finite-sample scaling laws, with implications for architecture and training-budget design.

1. Problem

Generalization in modern ML shows abrupt regime changes: double descent, grokking, and sudden capability emergence. Existing theory explains these as smooth interpolation phenomena, but the sharpness of observed transitions suggests a phase-transition mechanism.

We propose that spectral order parameters — effective rank, noise-floor coupling, and mode occupancy — govern sharp generalization thresholds in the same way thermodynamic order parameters govern phase transitions in physics.

2. Setup

2.1 Spectral Profile of Data and Model

Let Σ_{data} have eigenvalues $\lambda_1 \geq \lambda_2 \geq \cdots$ with power-law decay $\lambda_k \sim Ck^{-s}$.

The model captures K modes with N parameters and D data points.

Definition 1 (Effective Rank).

$$r_{\text{eff}} = \frac{(\sum_k \lambda_k)^2}{\sum_k \lambda_k^2}$$

Definition 2 (Mode Occupancy).

$$\phi_k = \min(1, D\lambda_k/\sigma_\xi^2)$$

2.2 Order Parameters and Critical Surface

Definition 3 (Spectral Order Parameter).

$$\gamma = r_{\text{eff}}/D$$

The Marchenko-Pastur edge predicts a critical surface at $\gamma^* = 1$, where the bulk eigenvalue distribution transitions from signal-dominated to noise-dominated.

2.3 Generalization Error Decomposition

$$\mathcal{E}(K, D) = \underbrace{\sum_{k>K} \lambda_k}_{\text{bias (unlearned modes)}} + \underbrace{\sum_{k\leq K} \frac{\sigma_\xi^2}{D\lambda_k}}_{\text{variance (noisy estimation)}}$$

Connection to existing kernel: the bias-variance decomposition is structurally identical to ScalingLaws/BiasVariance.lean. The phase transition arises from the dominance swap between these two terms, which the Scaling Laws paper treats smoothly but which can be sharp near criticality.

3. Main Theorem

Theorem Candidate 1 (Spectral Generalization Phase Transition). Let $\lambda_k = Ck^{-s}$ with $s > 1$. Define:

$$K^*(D) = \arg\min_K \mathcal{E}(K, D)$$

Then: 1. For $\gamma < \gamma^*$ (under-parameterized): $\mathcal{E}$ is dominated by bias, $\mathcal{E} \sim C_1 D^{-(s-1)/(s+1)}$, and reducing model size helps.

2. For $\gamma > \gamma^*$ (over-parameterized): $\mathcal{E}$ is dominated by variance, $\mathcal{E} \sim C_2 K/D$, and adding data helps.

3. At $\gamma = \gamma^*$: the derivative $\partial\mathcal{E}/\partial\gamma$ is discontinuous. The generalization error has a cusp with exponent depending on s:
$$\mathcal{E}(\gamma^* + \delta) - \mathcal{E}(\gamma^*) \sim |\delta|^{(s-1)/s}$$

Theorem Candidate 2 (Grokking as Spectral Timescale Separation). If mode k has learning timescale $t_k \sim k^s/D$, then grokking occurs when $t_K/t_1 \gg 1$: early modes converge quickly (training accuracy high), while late modes converge much later (test accuracy delayed).

4. Proof Sketch

1. **Mode-wise error analysis.** Each mode contributes independently to generalization error. Mode k contributes bias $\lambda_k(1-\phi_k)$ and variance $\sigma_\xi^2\phi_k/D$.

2. **Dominance swap.** The bias term decreases in K, the variance term increases. The crossover happens at K^* where $\lambda_{K^*} = \sigma_\xi^2/D$.

3. **Sharpness.** Near the crossover, the sum transitions from bias-dominated to variance-dominated. The power-law structure of λ_k makes this transition sharp (cusp) rather than smooth.

4. **Finite-sample correction.** For finite D, the transition is smoothed by $O(D^{-1/2})$, but becomes arbitrarily sharp as $D \to \infty$.

5. Empirics/Simulation

5.1 Synthetic Data

- Teacher-student with planted $s \in \{1.0, 1.5, 2.0, 3.0\}$.
- Sweep D and K to locate transition surfaces.
- Validate cusp exponents against predicted $(s-1)/s$.

5.2 Grokking Experiments

- Modular arithmetic and algorithmic tasks.
- Measure per-mode convergence timescales.
- Validate timescale separation prediction.

5.3 Real Data

- CIFAR-10 with controlled model size.
- Track effective rank and mode occupancy during training.
- Report correlation between predicted and observed transition points.

6. Limits

- **Regularization:** explicit regularization can smooth the cusp and eliminate the sharp transition.

- **Non-power-law spectra:** spectra with bumps or plateaus produce more complex transition geometry.
- **Finite-sample boundaries:** practical transition detection requires sufficient scale separation.
- **Nonstationary data:** distribution drift can move the critical surface during training.

7. Related Work

- **Double descent:** Belkin et al. (2019), Nakkiran et al. (2021).
- **Scaling laws:** Kaplan et al. (2020), our Scaling Laws paper.
- **Random matrix theory:** Marchenko-Pastur (1967), Baik-Ben Arous-Péché (2005) on phase transitions in eigenvalue distributions.
- **Grokking:** Power et al. (2022), Nanda et al. (2023).
- **Statistical physics of learning:** Engel-Van den Broeck (2001), Advani-Saxe (2017).

8. Cross-Paper Connections

- **Universality (paper 1):** the spectral exponent s that defines universality classes is the same parameter that controls the cusp exponent $(s-1)/s$ at the phase transition. A universality class transition in paper 1 is simultaneously a generalization phase transition here.
- **Minimal Sufficient State (paper 4):** the spectral gap ρ used in paper 4 can undergo a phase transition at the Marchenko-Pastur edge. When this happens, $K^* \to \infty$, meaning the minimal sufficient state diverges at criticality.
- **Causal Identifiability (paper 3):** the spectral gap condition $\mathrm{SG}(\delta)$ degrades at the phase boundary. Causal identifiability fails precisely when the generalization regime changes.
- **Ergodic Control (paper 7):** if the system operates near a phase transition, the contraction rates γ_k in paper 7 become nearly zero for critical modes, and the regret bound diverges.

9. Multi-Parameter Phase Diagram

9.1 Order Parameters

The generalization phase is governed by three parameters simultaneously: 1. **Data complexity** $\gamma = r_{\text{eff}}/D$ (effective rank vs sample size). 2. **Model capacity** $\kappa = N/K$ (parameters per captured mode). 3. **Noise level** $\nu = \sigma_\xi^2/\lambda_1$ (noise-to-signal ratio for the top mode).

9.2 Phase Diagram Structure

The three-parameter space (γ, κ, ν) admits the following phases:

Phase	γ	κ	ν	Regime
I: Underfitting	Low	Low	Any	Bias-dominated, more parameters help

Phase	γ	κ	ν	Regime
II: Classical	Low	High	Low	Variance-dominated, more data helps
III: Double descent peak	~ 1	High	Moderate	Near-interpolation, error spikes
IV: Benign overfitting	High	High	Low	Many parameters, low noise, good generalization
V: Tempered overfitting	High	High	High	Overfitting but with bounded excess risk

Phase boundaries are codimension-1 surfaces in (γ, κ, ν) space, corresponding to cusps in the generalization error surface.

9.3 Double Descent as Two-Step Transition

Theorem Candidate 3 (Double Descent Decomposition). Classical double descent is the composition of two phase transitions:

1. **First transition** at $\gamma = \gamma_1^*$: bias $\to$ variance dominance swap (classical U-curve minimum).
2. **Second transition** at $\gamma = \gamma_2^* \approx 1$: interpolation threshold where the variance peak occurs and then declines as implicit regularization takes over.

The two transitions have different cusp exponents: - First: $\mathcal{E} \sim |\gamma - \gamma_1^*|^{(s-1)/s}$ (same as Theorem 1). - Second: $\mathcal{E} \sim |\gamma - 1|^{-1}$ (pole, not cusp) in the interpolation regime.

The overall double-descent curve is the envelope of these two effects.

9.4 Grokking and Timescale Separation

Theorem 2 can be sharpened: grokking occurs when the phase diagram path crosses from Phase I to Phase IV along the time axis, and the crossing time for late modes is exponentially longer than for early modes.

Prediction: grokking is more likely for tasks with large spectral exponent s (steep mode hierarchy), because $t_K/t_1 \sim K^s$ grows faster.

10. Outlook

- **Transition-aware training:** schedules that detect approach to criticality and adjust accordingly. The multi-parameter phase diagram provides explicit monitoring targets.

- **Early-warning diagnostics:** track $\widehat{\gamma}(t)$, $\widehat{\kappa}(t)$, $\widehat{\nu}(t)$ during training and alert when the trajectory approaches a phase boundary.
- **Architecture design:** use the phase diagram to size models relative to data spectral structure. Given estimated s and ν, choose N to stay in a desired phase.
- **Grokking prediction:** use the timescale separation criterion to predict which tasks and architectures will exhibit grokking before training begins.
- **Lean formalization:** the cusp exponent theorem is a natural target for LeanProofs/SpectralPhaseTransition The double-descent decomposition (Theorem 3) is a second formalization target.