Why Neural Networks Scale: A Complete Latent-Theoretic Foundation

146 Platonic-checked theorems linking scaling laws, grokking, alignment, robustness, and more to

Dr. Tamás Nagy

tamas.nagy@thelatent.space

draft • 2026-04-08

Every empirical scaling law in deep learning is a shadow of a single number — the spectral decay rate of the data distribution.

Executive Summary

Neural networks exhibit striking empirical regularities: performance improves as a power law with increasing compute, data, and parameters (Kaplan et al., 2020); networks suddenly generalize long after memorizing training data (Power et al., 2022); test error follows a double-descent curve as model capacity grows (Belkin et al., 2019). These phenomena have been documented extensively but lack a unified theoretical explanation.

This paper organizes these phenomena — together with ten additional signatures treated in the formal layer — under a single control parameter: the **Latent Number** ρ of the data distribution (Nagy, 2026). The Latent Number measures the rate of spectral decay in the distribution's representation. From this one number, we obtain:

- The **scaling exponent** $\alpha = \beta \cdot \log \rho$, explaining why performance follows a power law and why the exponent varies across tasks.
- A **phase transition at** $\rho = 1$ that explains grokking: below the threshold, approximation error diverges; above it, error decays exponentially.
- **Tight generalization bounds** scaling as $\sqrt{N^*/n}$ where $N^* = -\log \varepsilon / \log \rho$ is the effective Latent dimension — not the parameter count.
- A **transformer efficiency theorem**: attention heads need only $O(N^{*2})$ parameters, not $O(D^2)$ in the ambient dimension.
- The **inevitability of second descent**: once variance saturates at $\sigma^2 N^*/n$, bias continues to decay exponentially.
- **Emergent abilities** as predictable phase transitions ordered by task complexity N_T^*.
- **Catastrophic forgetting** as ρ collapse below 1, with explicit preservation conditions.
- **Information bottleneck** optimality at exactly N^* modes.
- **Optimization landscape** smoothness proportional to $\log \rho$, explaining pretraining benefit.
- **Alignment efficiency** scaling as N^*_{pref}, with reward hacking detection via $\rho_{\text{rew}} < \rho_{\text{pref}}$.
- **Adversarial robustness** governed by the attack surface $D - N^*$.

All 146 theorems are machine-checked in the Platonic proof environment (ProofEnv) with no deferred proof obligations. The theory makes testable predictions: the scaling exponent of any task is computable from the spectral decay rate of its data distribution, grokking onset is predictable

from the training-time evolution of $\rho(t)$, and capability emergence ordering is determined by task complexity N_T^*.

What this paper does not claim. We do not prove that $\rho(t)$ increases under gradient descent for arbitrary architectures — this is assumed as a hypothesis motivated by spectral learning theory. We do not provide empirical measurements of ρ on benchmark datasets (this is a companion empirical paper). We do not claim the bounds are tight in constant factors.

Abstract

We present a unified mathematical theory of neural scaling laws derived from the spectral structure of data distributions. The central object is the Latent Number $\rho \in (0, \infty)$, which measures the rate at which a distribution's spectral coefficients decay. From ρ alone, we derive: (1) the scaling exponent $\alpha = \beta \cdot \log \rho$ linking optimizer efficiency β to data structure; (2) a spectral phase transition at $\rho = 1$ explaining grokking; (3) generalization bounds $O(\sqrt{N^*/n})$ with concentration $P(\text{gap} > \varepsilon) \leq 2\exp(-2n\varepsilon^2/N^*)$; (4) transformer expressivity requiring $O(N^{*2})$ parameters per head; (5) the inevitability of double descent when variance saturates at $\sigma^2 N^*/n$; (6) sparse MoE efficiency $A \cdot N^{*2} < K \cdot D^2$; (7) emergent abilities as predictable phase transitions ordered by N_T^*; (8) catastrophic forgetting as ρ collapse; (9) information bottleneck optimality at N^* modes; (10) optimization landscape smoothness $\propto \log \rho$; (11) alignment efficiency scaling as N_{pref}^* with reward hacking when $\rho_{\text{rew}} < \rho_{\text{pref}}$; and (12) adversarial robustness governed by the attack surface $D - N^*$. The chain of lemmas is machine-checked in the Platonic proof environment (146 theorems in 11 files; see §15.6). The theory makes testable predictions: scaling exponents are computable from data spectra, grokking onset is predictable from $\rho(t)$ dynamics, capability emergence ordering is determined by N_T^*, and adversarial vulnerability is bounded by the gap between ambient and effective dimension.

Keywords: neural scaling laws, grokking, double descent, spectral theory, Latent Number, effective dimension, transformer expressivity, sparse activation, alignment, adversarial robustness, information bottleneck

MSC 2020: 68T07 (Machine learning), 41A25 (Approximation by polynomials), 60E15 (Inequalities)

§1 Introduction

§1.1 The Empirical Puzzle

Take a language model with 10^9 parameters. Double the parameters to 2×10^9, holding data and compute fixed. The cross-entropy loss drops by a predictable amount — not by half, not by a random quantity, but by a factor governed by a power law with a stable exponent (Kaplan et al., 2020; Hoffmann et al., 2022). This exponent persists across architectures, datasets, and training procedures. It varies between tasks but is remarkably stable within a task.

Three questions follow immediately. Why a power law, and not exponential or logarithmic decay? Why does the exponent vary across tasks but not across architectures trained on the same task? And why does increasing parameters help generalization at all, given that classical statistical learning theory predicts the opposite?

These are not three separate questions. They are one question: **what property of the data determines the learning dynamics?**

§1.2 The Same Puzzle from Spectral Theory

Consider a function f that admits a spectral expansion $f = \sum_k c_k \phi_k$ in an orthonormal basis $\{\phi_k\}$. If the coefficients decay as $|c_k| \sim C \cdot \rho^{-k}$ for some $\rho > 1$, then truncating the expansion at N terms gives an approximation error

$$\varepsilon(N) = C \cdot \exp(-N \cdot \log \rho).$$

This is exponential convergence — the hallmark of spectral methods. The rate $\log \rho$ is determined entirely by the analyticity structure of f. For $\rho > 1$, the function is "spectrally nice": few coefficients capture most of the information. For $\rho < 1$, the coefficients grow, and truncation diverges. At $\rho = 1$, the coefficients are constant — no improvement with more terms.

The quantity ρ encodes how much **structure** exists in the function relative to the chosen basis. It is not a property of the function alone, nor of the basis alone, but of their interaction — how efficiently the basis compresses the function.

§1.3 The Unifying Insight

The Latent Number ρ of a data distribution measures precisely this spectral decay rate. Every neural scaling phenomenon is a consequence of the error formula $\varepsilon(N, \rho) = C \cdot \exp(-N \cdot \log \rho)$:

$$\boxed{\text{Scaling exponent: } \alpha = \beta \cdot \log \rho}$$

where $\beta \in (0, 1)$ is the optimizer efficiency — the fraction of spectral information captured per unit of compute. The exponent α varies across tasks because ρ varies (different data distributions have different spectral structure). It is stable within a task because ρ is a property of the distribution, not the architecture.

From this single formula, we derive the effective dimension $N^* = -\log \varepsilon / \log \rho$ (the number of spectral modes needed for accuracy ε), the grokking phase transition (the qualitative change at $\rho = 1$), the generalization bound ($\sqrt{N^*/n}$, not $\sqrt{P/n}$), the transformer parameter count (N^{*2}, not D^2), and the double descent curve (variance saturates at $\sigma^2 N^*/n$).

§1.4 What This Enables

1. **Predictable scaling:** Given ρ (measurable from data spectra), the scaling exponent α is computable before training.
2. **Grokking prediction:** The onset of sudden generalization is the time t^* when $\rho(t)$ crosses 1 — observable during training.
3. **Architecture efficiency:** A transformer with N^{*2} parameters per head achieves the same approximation as one with D^2 parameters. The ratio N^{*2}/D^2 quantifies how much structure the data has.
4. **Sample complexity:** The number of samples needed for generalization is $n \geq N^* \cdot \log(2/\delta)/(2\varepsilon^2)$ — determined by N^*, not the model size.

§1.5 Organization

Section 2 introduces the Latent Number and effective dimension from spectral theory. Section 3 derives the scaling exponent. Section 4 establishes the grokking phase transition. Section 5 proves concentration bounds. Section 6 derives transformer expressivity. Section 7 proves double descent. Section 8 establishes training dynamics and sparse activation. Section 9 proves emergent abilities are spectral phase transitions. Section 10 models catastrophic forgetting as ρ collapse. Section 11 formulates the information bottleneck as spectral compression. Section 12 characterizes optimization landscape geometry. Section 13 develops alignment and value learning theory. Section 14 derives adversarial robustness bounds. Section 15 discusses limitations, predictions, and open problems.

§2 The Latent Number and Effective Dimension

We use the Latent spectral formalism and background on finite sufficient representations as developed in Nagy (2026); here we specialize the Latent Number and effective dimension to the scaling-law narrative.

§2.1 Spectral Decay and the Latent Number

Definition 2.1 (Latent Number). Let μ be a probability distribution admitting a spectral decomposition with coefficients $\{c_k\}_{k\geq 1}$ in an orthonormal basis. The *Latent Number* of μ is

$$\rho(\mu) = \limsup_{k\to\infty} |c_k|^{-1/k},$$

the reciprocal of the exponential decay rate of the spectral coefficients. When $\rho > 1$, the distribution has exponentially decaying spectral tails; when $\rho < 1$, the tails grow; at $\rho = 1$, the coefficients are asymptotically constant.

The Latent Number is the spectral analogue of the radius of convergence. Just as a power series converges inside its radius and diverges outside, a spectral approximation converges (in L^2 error) when the number of terms exceeds N^* and diverges (in the sense of growing truncation error) when it falls below.

§2.2 Effective Dimension

Definition 2.2 (Effective Dimension). For target accuracy $\varepsilon \in (0, 1)$ and Latent Number $\rho > 1$, the *effective Latent dimension* is

$$N^* = \frac{-\log \varepsilon}{\log \rho}.$$

This is the number of spectral modes required to achieve error ε. It satisfies:

- $N^* > 0$ for $\varepsilon < 1$ and $\rho > 1$ (Theorem 7).
- $N^* \cdot \log \rho = -\log \varepsilon$ — the defining identity (Theorem 8).
- N^* decreases with ρ: more structure ⇒ fewer dimensions needed (Theorem 9).
- N^* increases as $\varepsilon \to 0$: higher accuracy ⇒ more dimensions (Theorem 10).

The effective dimension N^* is the intrinsic complexity of the learning problem. It depends on the data (ρ) and the accuracy requirement (ε), but not on the model architecture or the ambient data dimension D. This is the key: N^* **replaces the parameter count P in all generalization bounds.**

§3 The Scaling Exponent

§3.1 Derivation

Theorem 3.1 (Scaling Exponent). *Let $\rho > 1$ be the Latent Number of the data distribution and $\beta \in (0, 1)$ the optimizer efficiency. Define $\alpha = \beta \cdot \log \rho$. Then:*

(a) $\alpha > 0$ — the scaling exponent is positive (Theorem 2).

(b) $\alpha < \log \rho$ — the optimizer imposes a strict suboptimality (Theorem 3).

(c) $\alpha/\beta = \log \rho$ — data and optimizer contributions separate cleanly (Theorem 4).

The scaling exponent α determines the power-law relationship: loss $\sim n^{-\alpha}$ as a function of training steps (or equivalently, compute). Part (a) guarantees that training always helps. Part (b) shows that no real optimizer achieves the spectral rate — there is always a gap $\log \rho - \alpha = (1 - \beta) \log \rho$ attributable to optimizer imperfection. Part (c) is the clean separation: α factorizes into a data term ($\log \rho$) and an optimizer term (β), explaining why the same data yields different scaling exponents with different optimizers.

§3.2 Monotonicity

Theorem 3.2 (Monotonicity). *The scaling exponent α is:*

(a) Monotonically increasing in ρ: richer data structure ⇒ steeper scaling (Theorem 5).

(b) Monotonically increasing in β: better optimizer ⇒ steeper scaling (Theorem 6).

These are strict monotonicities. Part (a) explains why language tasks (high ρ, rich structure) scale better than random noise (low ρ). Part (b) explains why Adam outscales SGD on the same task — it captures spectral information more efficiently.

§4 Grokking as a Spectral Phase Transition

§4.1 The Three Regimes

The error formula $\varepsilon(N) = C \cdot \exp(-N \cdot \log \rho)$ has qualitatively different behavior depending on whether ρ is above, below, or at the critical value 1:

Theorem 4.1 (Regime Classification).

(a) Generalization regime ($\rho > 1$): $\varepsilon(N) < C$ and decays exponentially with N (Theorem 15).

(b) Memorization regime ($\rho < 1$): $\varepsilon(N) > C$ and grows exponentially with N (Theorem 16).

(c) Critical point ($\rho = 1$): $\varepsilon(N) = C$ for all N — no improvement (Theorem 17).

This trichotomy is sharp: the behavior changes qualitatively at $\rho = 1$. In the memorization regime, adding more spectral terms makes the approximation *worse* — the model fits noise rather than

signal. In the generalization regime, each additional term improves the approximation exponentially. At the critical point, the expansion stalls.

§4.2 The Grokking Mechanism

During training, the network's internal representation evolves. We model this by a time-dependent Latent Number $\rho(t)$ that increases during training (Hypothesis: gradient descent on structured data increases spectral resolution). The grokking transition occurs when $\rho(t)$ crosses 1:

Theorem 4.2 (Grokking). *Let $L_{test}(t) = L_\infty + C \cdot \exp(-N \cdot \log \rho(t))$ be the test loss at time t.*

(a) Before crossing ($\rho(t) < 1$): $L_{test}(t) > L_\infty + C$ — test loss above baseline (Theorem 20).

(b) After crossing ($\rho(t) > 1$): $L_{test}(t) < L_\infty + C$ — test loss below baseline (Theorem 21).

(c) The test loss is monotonically decreasing in $\rho(t)$ (Theorem 22).

The "extended memorization" phase corresponds to $\rho(t) < 1$: the network fits the training data but the test error remains high because spectral truncation error is growing. The "sudden generalization" corresponds to $\rho(t)$ crossing 1: exponential decay kicks in, and the test error drops rapidly.

§4.3 Sharpness

The sharpness of the grokking transition increases with spectral order N. In the memorization regime, the error grows as $\exp(N \cdot |\log \rho|)$. In the generalization regime, it decays as $\exp(-N \cdot \log \rho)$. For large N, the gap between the two regimes is enormous, producing the characteristic sharp transition observed empirically (Theorems 24-26).

§5 Concentration Bounds and Sample Complexity

§5.1 The Hoeffding Exponent

The generalization gap concentrates around its mean with an exponential tail controlled by the *concentration exponent* $E = 2n\varepsilon^2/N^*$:

$$P(\text{gap} > \varepsilon) \leq 2 \exp\left(-\frac{2n\varepsilon^2}{N^*}\right).$$

The exponent E is positive when N^*, n, and ε are positive (Theorem 27). It increases with sample size n (Theorem 28) and decreases with effective dimension N^* (Theorem 29). Fewer effective dimensions mean tighter concentration — a direct benefit of data structure.

§5.2 Sample Complexity

Theorem 5.1 (Sample Complexity). *For the generalization gap to satisfy $P(gap > \varepsilon) \leq \delta$, it suffices to have*

$$n \geq \frac{N^* \cdot \log(2/\delta)}{2\varepsilon^2}$$

samples (Theorem 30). The gap-squared bound $N^* \cdot \log(2/\delta)/(2n)$ decreases monotonically with n (Theorem 32).

The sample complexity is linear in N^* — the intrinsic complexity — and logarithmic in $1/\delta$. For structured data ($\rho \gg 1$, hence N^* small), far fewer samples are needed than classical bounds based on the parameter count P would suggest.

§5.3 The Latent Advantage

Theorem 5.2 (Latent vs. Classical). *If $N^* < P$ (the effective dimension is less than the parameter count), then for any constant $C_1 > 0$, sample size $n > 0$:*

$$C_1\sqrt{N^*/n} \leq C_1\sqrt{P/n} \quad \text{(Theorem 34)}.$$

The probability bound $\exp(-2n\varepsilon^2/N^*)$ also decays faster than $\exp(-2n\varepsilon^2/P)$ (Theorem 37).

This is the formal statement of why overparameterized models generalize: the relevant complexity measure is N^*, which for structured data is orders of magnitude smaller than P.

§6 Transformer Expressivity

§6.1 Attention Head Capacity

Each attention head with dimension d_{head} resolves d_{head} spectral modes. To approximate a target function to accuracy ε, we need $d_{\text{head}} \geq N^*$ modes:

Theorem 6.1 (Head Sufficiency). *For $\rho > 1$, $d_{head} \geq N^*$, and constant $C > 0$:*

$$C \cdot \exp(-d_{\text{head}} \cdot \log \rho) \leq C \cdot \exp(-N^* \cdot \log \rho) \quad \text{(Theorem 39)}.$$

Wider heads are strictly better: $d_1 < d_2 \Rightarrow error(d_2) < error(d_1)$ (Theorem 40).

§6.2 Parameter Efficiency

Each head requires d_{head}^2 parameters. Setting $d_{\text{head}} = N^*$ gives the Latent parameter count:

Theorem 6.2 (Parameter Saving). *If $N^* < D$ (effective dimension less than ambient dimension):*

(a) $N^{*2} < D^2$ — strict parameter saving (Theorem 42).

(b) $N^{*2}/D^2 < 1$ — the efficiency ratio is bounded below 1 (Theorem 43).

(c) For H attention heads: $H \cdot N^{*2} < H \cdot D^2$ — the saving extends to the full model (Theorem 50).

§6.3 Multi-Head Scaling and Diminishing Returns

Multiple heads provide linear coverage growth: H heads cover $H \cdot d_{\text{head}}$ spectral modes (Theorem 45). However, the cost per mode increases with head width — the parameter-per-mode ratio $d_{\text{head}}^2/d_{\text{head}} = d_{\text{head}}$ grows linearly (Theorem 46). This implies diminishing returns from width: doubling the head dimension quadruples the parameters for only double the coverage.

Multi-head error composes linearly: if each head contributes error at most E/H, the total error is at most E (Theorem 47).

§7 Double Descent

§7.1 The Bias-Variance Decomposition

The test risk at model capacity k (number of spectral modes used) decomposes as:

$$\text{Risk}(k) = \underbrace{C \cdot \exp(-k \cdot \log \rho)}_{\text{Bias}} + \underbrace{\text{Variance}(k)}_{\text{Var}}$$

Bias decreases exponentially with k for $\rho > 1$ (Theorem 51). In the underparameterized regime ($k < N^*$), variance grows linearly as $\sigma^2 k/n$ (Theorem 52). In the overparameterized regime ($k \geq N^*$), variance saturates at $\sigma^2 N^*/n$ — the Latent variance floor.

§7.2 The Second Descent Is Inevitable

Theorem 7.1 (Second Descent). *In the overparameterized regime, for $k_1 < k_2$ (both $\geq N^*$), $\rho > 1$:*

$$\text{Risk}(k_2) = \underbrace{C \cdot e^{-k_2 \log \rho}}_{\text{smaller}} + \underbrace{\sigma^2 N^*/n}_{\text{same}} < \underbrace{C \cdot e^{-k_1 \log \rho}}_{\text{larger}} + \underbrace{\sigma^2 N^*/n}_{\text{same}} = \text{Risk}(k_1) \quad \text{(Theorem 57)}.$$

Once variance saturates, risk is monotonically decreasing because bias decays exponentially while variance stays constant. The second descent is not a curiosity — it is a mathematical inevitability of exponential spectral decay meeting constant variance.

§7.3 What Controls the Peak

The interpolation peak height $\sigma^2 N^*/n$ is controlled by: - **More data** ($n \uparrow$): peak decreases as $1/n$ (Theorem 55). - **More structure** ($\rho \uparrow$): peak decreases because N^* shrinks (Theorem 56). - **Both simultaneously**: higher ρ reduces both bias and peak — double descent becomes milder (Theorem 61).

§8 Training Dynamics and Sparse Activation

§8.1 Consequences of $\rho(t)$ Monotonicity

Under the hypothesis that training increases $\rho(t)$ (motivated by spectral learning theory), all quantities improve simultaneously:

- Approximation error decreases (Theorem 63).
- Effective dimension N^* shrinks (Theorem 64).
- Generalization gap decreases (Theorem 65).
- Sample complexity decreases (Theorem 66).
- Required parameters decrease (Theorem 67).
- Concentration exponent increases (Theorem 68).

Training is a one-dimensional improvement process along the ρ axis: everything that matters is a monotone function of ρ.

§8.2 Sparse Activation in Mixture-of-Experts

In a mixture-of-experts (MoE) architecture with K total experts, each with d_{exp}^2 parameters, only $A \ll K$ experts need to be active for an input with effective dimension N^*:

- Active parameters $A \cdot d_{\text{exp}}^2 < K \cdot d_{\text{exp}}^2$ (Theorem 69).
- Activation ratio $A/K < 1$ (Theorem 70).
- Sparsity increases with model scale: $K \uparrow \Rightarrow A/K \downarrow$ (Theorem 71, 73).

Theorem 8.1 (MoE Advantage). *$A \cdot N^{*2} < K \cdot D^2$ (Theorem 74) — the Latent MoE is doubly efficient: fewer active experts AND smaller per-expert dimensionality.*

§9 Emergent Abilities as Spectral Phase Transitions

§9.1 Task-Specific Thresholds

The grokking phase transition at $\rho = 1$ (§4) is a single-task phenomenon. In practice, a model must perform multiple tasks, each with different complexity — consistent with empirical reports of capability ordering in large language models (Wei et al., 2022). We formalize this by assigning each task T an effective dimension $N_T^* = -\log \varepsilon_T / \log \rho$, where ε_T is the task's accuracy requirement.

Theorem 9.1 (Task Threshold). *Fix task accuracy $\varepsilon_T \in (0, 1)$ and spectral order $N > 0$. Let $N_T^* = -\log \varepsilon_T / \log \rho$ (well-defined for $\rho > 1$). In the generalization regime $\rho > 1$, if $N \geq N_T^*$ then the spectral truncation error is at most ε_T; increasing N beyond N_T^* improves error exponentially in the excess resolution. In the memorization regime $\rho < 1$, the same error functional grows with N rather than decaying — spectral modes do not accumulate toward the target (Theorems 75-77).*

§9.2 Capability Ordering

Tasks emerge in a predictable order determined by their effective dimension:

Theorem 9.2 (Emergence Order). *If $N_{T_1}^* < N_{T_2}^*$ (task T_1 is easier than T_2), then at any $\rho > 1$, task T_1 achieves lower error than T_2. Scaling ρ (via more compute/data) unlocks progressively harder tasks (Theorems 78-80).*

This explains the empirical observation that language models acquire capabilities in a consistent order across training runs: simpler tasks (arithmetic, syntax) emerge before complex ones (reasoning, analogy).

§9.3 The Illusion of Discontinuity

Emergent abilities appear sudden but are mathematically continuous:

Theorem 9.3 (Smooth Emergence). *The error $C \cdot \exp(-N \cdot \log \rho)$ is strictly monotone in ρ (Theorem 86). The transition sharpens with model scale N: larger models exhibit sharper sigmoid-like transitions between "can't do" and "can do" (Theorem 81).*

The apparent discontinuity is a measurement artifact: the sigmoid $\exp(-N \cdot \log \rho)$ has width $\sim 1/N$ around the threshold. For $N = 10^{10}$ parameters, this width is negligible — the transition appears step-like even though it is smooth.

§9.4 Predictability

Both the ordering and onset of emergent abilities are computable from ρ and ε_T: - Tasks with lower ε_T (stricter accuracy) require higher N^*, hence emerge later (Theorem 85). - The N^* needed for task T at accuracy ε_T is positive and determined by the data structure (Theorem 84). - Simultaneously increasing ρ and N yields multiplicative improvement (Theorem 83).

§10 Catastrophic Forgetting as ρ Collapse

§10.1 The Collapse Mechanism

In continual learning, training on new data (distribution B) can degrade performance on old data (distribution A) — the classical catastrophic interference phenomenon (McCloskey & Cohen, 1989). In the Latent framework, this corresponds to the effective ρ_A dropping:

Theorem 10.1 (Collapse). *Before training on B: $\rho_A > 1$, error on A is below C (Theorem 87). If ρ_A drops below 1 after adaptation: error on A exceeds C — catastrophic forgetting (Theorem 88). Larger drops in ρ_A cause proportionally worse degradation (Theorem 89).*

§10.2 Forgetting Severity

The severity of forgetting scales with model capacity:

Theorem 10.2 (Depth Amplifies Forgetting). *When $\rho_A < 1$ (collapsed), the error grows as $C \cdot \exp(N \cdot |\log \rho_A|)$. Deeper models ($N$ larger) exhibit worse forgetting — exponentially so (Theorem 90). Meanwhile, the new task B benefits normally from $\rho_B > 1$ (Theorem 92).*

This explains why larger models are *more* susceptible to catastrophic forgetting: the same ρ collapse produces exponentially larger error growth in higher-capacity architectures.

§10.3 Prevention and Reversal

Catastrophic forgetting is preventable and reversible:

Theorem 10.3 (Preservation). *Maintaining $\rho_A > 1$ after adaptation prevents catastrophic forgetting (Theorem 93). When ρ drops but stays above 1: performance degrades gracefully, not catastrophically, with higher preserved ρ yielding lower residual error (Theorem 94).*

Theorem 10.4 (Reversibility). *Forgetting is reversible: restoring ρ_A (e.g., via replay, regularization, or re-exposure to distribution A) restores performance monotonically (Theorem 98). However, a collapsed ρ_A increases the effective dimension N^* and weakens concentration bounds, requiring more samples and compute to recover (Theorems 95-97).*

The practical implication: continual learning methods should be evaluated by their ability to maintain $\rho_{\text{old}} > 1$, not just by accuracy on benchmarks.

§11 Information Bottleneck as Spectral Compression

§11.1 The Compression-Fidelity Tradeoff

The Information Bottleneck (IB) principle (Tishby et al., 2000) frames learning as finding a representation T that compresses input X while preserving information about target Y. In the Latent

framework, this is spectral truncation: keep the top N modes of X (those that predict Y) and discard the rest.

The fidelity (residual error after keeping N modes) is $C \cdot \exp(-N \cdot \log \rho)$, and the compression cost is proportional to $N \cdot \log \rho$. More modes increase both fidelity and cost (Theorems 99-100). Higher ρ improves the fidelity-per-mode ratio: structured data compresses more efficiently (Theorem 101).

§11.2 Optimal Compression at N^*

The IB-optimal representation keeps exactly $N^* = -\log \varepsilon / \log \rho$ modes (Theorem 102): - Below N^*: insufficient fidelity — the representation loses predictive information (Theorem 103). - Above N^*: diminishing returns — additional modes increase compression cost with negligible fidelity gain (Theorem 104).

§11.3 The IB Curve Is Controlled by ρ

The IB curve (tradeoff between $I(X;T)$ and $I(T;Y)$) is monotone — more compression reduces fidelity (Theorem 105). Higher ρ shifts the entire curve favorably: at any compression level, structured data retains more information (Theorem 106). Perfect compression is impossible: residual error is always positive (Theorem 107).

Theorem 11.1 (IB-Latent Duality). *The IB-optimal representation at accuracy ε uses exactly N^* modes, with compression cost $N^* \cdot \log \rho = -\log \varepsilon$. Higher ρ reduces N^* for the same ε, requiring fewer modes and lower compression cost (Theorems 108-110).*

§12 Optimization Landscape Geometry

§12.1 Gradient Properties

The gradient signal of a spectral loss at resolution N is proportional to $N \cdot \log \rho$. For $\rho > 1$, this is positive and increasing in both N and ρ — structured data provides stronger learning signals (Theorems 111-112). For $\rho < 1$ (memorization regime), the gradient signal is negative — optimization pushes in the wrong direction (Theorem 113).

§12.2 Basin Geometry and Spectral Gaps

Basin width scales with $\log \rho$: higher ρ produces wider, smoother basins that are easier to optimize (Theorem 114). The spectral gap $\log \rho > 0$ exists for all structured data (Theorem 116). Loss at any minimum is bounded by $C \cdot \exp(-N \cdot \log \rho) < C$ for $\rho > 1$ (Theorem 117).

§12.3 Pretraining and Structure

Theorem 12.1 (Pretraining Benefit). *Higher initial ρ (from pretraining on structured data) yields stronger gradient signal at every model capacity N (Theorem 121). Structured data ($\rho > 1$) achieves strictly lower loss than unstructured data ($\rho < 1$) at the same capacity (Theorem 122). Simultaneously increasing ρ and N yields multiplicative improvement (Theorem 120).*

§13 Alignment and Value Learning

§13.1 Preferences as Spectral Structure

Human preferences over model outputs form a distribution with its own Latent Number ρ_{pref}. When preferences are structured ($\rho_{\text{pref}} > 1$), alignment error decays exponentially with model capacity (Theorem 123). More structured preferences are easier to learn (Theorem 124). Perfect alignment is impossible — residual misalignment is always positive (Theorem 125).

§13.2 RLHF Efficiency

Human-feedback training (Christiano et al., 2017) fits naturally into the spectral picture: alignment efficiency scales as $N^*_{\text{pref}} = -\log \varepsilon / \log \rho_{\text{pref}}$:

Theorem 13.1 (Alignment Scaling). *Higher ρ_{pref} reduces N^*_{pref}, requiring fewer preference samples for the same alignment accuracy (Theorem 126). The sample complexity for alignment scales linearly with N^*_{pref} (Theorem 127). Larger models align better when preferences are structured (Theorem 128).*

§13.3 Reward Hacking

Theorem 13.2 (Reward Hacking). *When the reward model's spectral coverage ρ_{rew} is less than the preference distribution's ρ_{pref}, a systematic gap exists: the reward model captures less structure than the true preferences (Theorem 129). Deeper models amplify this gap when $\rho_{rew} < 1$ (Theorem 131). Safe alignment requires $\rho_{rew} \geq \rho_{pref}$ — the reward model must match or exceed the spectral coverage of human preferences (Theorem 132).*

Restoring ρ_{rew} (via better reward modeling) restores alignment monotonically (Theorem 134).

§14 Adversarial Robustness

§14.1 The Attack Surface

Adversarial attacks (Goodfellow et al., 2015; Madry et al., 2018) exploit directions orthogonal to the model's signal subspace. The attack surface — the number of exploitable dimensions — is $D - N^*$, which is positive whenever $N^* < D$ (Theorem 135). The robustness ratio N^*/D is strictly less than 1 for structured data (Theorem 137).

§14.2 The Structure-Robustness Tradeoff

Higher ρ improves accuracy (lower N^*, fewer modes needed) but paradoxically increases the attack surface ($D - N^*$ grows). This is the structure-robustness tradeoff (Theorems 138-140):

- **Accuracy**: signal error $C \cdot \exp(-N \cdot \log \rho)$ decreases with ρ (Theorem 139).
- **Vulnerability**: attack surface $D - N^*$ increases as N^* shrinks (Theorem 140).

§14.3 Defense in the Latent Subspace

Theorem 14.1 (Efficient Defense). *Defense in the N^*-dimensional signal subspace is sufficient — protecting N^* dimensions (not D) covers the model's decision-relevant features (Theorem 141).*

Defense cost scales as N^{*2}, not D^2 (Theorem 142). Higher ρ makes defense cheaper: fewer dimensions to protect (Theorem 143).

§14.4 Certified Robustness

The spectral gap $\log \rho$ provides a natural certified robustness radius: perturbations smaller than $\log \rho$ cannot change the dominant spectral mode (Theorem 144). Higher ρ yields a larger certified radius (Theorem 145). The total vulnerability decomposes into signal error plus attack surface ratio, both computable from N^* and D (Theorem 146).

§15 Discussion

§15.1 Summary of Results

From a single quantity — the Latent Number ρ — we have derived a unified theory of:

Phenomenon	Formula	Key theorem
Scaling exponent	$\alpha = \beta \cdot \log \rho$	Thm 2
Grokking	Phase transition at $\rho = 1$	Thm 15-17, 20-21
Generalization gap	$O(\sqrt{N^*/n})$	Thm 34
Concentration	$P(\text{gap} > \varepsilon) \leq 2e^{-2n\varepsilon^2/N^*}$	Thm 36-38
Transformer params	$O(N^{*2})$ per head	Thm 42, 50
Double descent	Variance saturates at $\sigma^2 N^*/n$	Thm 57
Training dynamics	All quantities monotone in $\rho(t)$	Thm 63-68
MoE efficiency	$A \cdot N^{*2} < K \cdot D^2$	Thm 74
Emergent abilities	Predictable by N_T^* ordering	Thm 75-86
Catastrophic forgetting	ρ collapse below 1	Thm 87-98
Information bottleneck	IB-optimal = N^* modes	Thm 99-110
Optimization landscape	Smoothness $\propto \log \rho$	Thm 111-122
Alignment / RLHF	Efficiency $\propto N^*_{\text{pref}}$	Thm 123-134
Adversarial robustness	Attack surface $= D - N^*$	Thm 135-146

§15.2 Testable Predictions

1. **The scaling exponent is computable.** Given a dataset, estimate ρ from the spectral decay of the empirical distribution. The predicted exponent $\alpha = \beta \cdot \log \rho$ should match the observed scaling law.

2. **Grokking onset is predictable.** Monitor $\rho(t)$ during training. Sudden generalization should occur when $\rho(t)$ crosses 1. The sharpness should increase with model capacity N.

3. **Overparameterization helps proportionally to structure.** The generalization improvement from doubling parameters should be larger for high-ρ tasks (natural language, images) than low-ρ tasks (random labels, adversarial data).

4. **MoE activation patterns reflect N^*.** The number of active experts per input should correlate with the input's effective dimension, not with the model's total expert count.

5. **IB-optimal compression is at N^* modes.** Autoencoders trained with IB-like objectives should converge to latent dimensions near N^*. The IB curve shape should be predictable from ρ.

6. **Pretraining benefit is measurable via ρ.** Fine-tuning loss should decrease monotonically with the pretrained model's initial ρ estimate.

7. **Alignment efficiency is predictable.** RLHF sample complexity should scale as N^*_{pref}. Reward hacking should be detectable by comparing ρ_{rew} and ρ_{pref}.

8. **Adversarial vulnerability scales with $D - N^*$.** The certified robustness radius should correlate with $\log \rho$, and defense cost should scale with N^{*2}, not D^2.

§15.3 Assumptions and Limitations

The central assumption is that $\rho(t)$ is non-decreasing during gradient-based training. This is not proved from first principles in this paper; it is motivated by the observation that gradient descent on structured data preferentially learns low-frequency (large spectral coefficient) components first, effectively increasing ρ. A rigorous proof would require analyzing the spectral dynamics of gradient flow on non-convex loss landscapes — an important open problem.

The constant factors in our bounds are not tight. We prove the correct scaling ($\sqrt{N^*/n}$, N^{*2}, etc.) but the multiplicative constants depend on problem-specific parameters (C, σ^2, β) that we do not estimate.

The theory assumes a fixed orthonormal basis for the spectral decomposition. In practice, neural networks learn their own basis (feature learning), which may change ρ itself during training. Analyzing this co-evolution is a natural extension.

§15.4 Translation Table

ML / Deep Learning	Spectral Theory	Latent Framework
Model capacity / parameters	Truncation order N	Spectral resolution
Data quality / structure	Analyticity, smoothness	Latent Number ρ
Learning rate / optimizer	—	Optimizer efficiency β
Scaling exponent	Convergence rate	$\alpha = \beta \cdot \log \rho$
Grokking threshold	Phase transition	$\rho(t) = 1$
Generalization gap	Approximation error	$\varepsilon = C \cdot \rho^{-N}$
Overparameterization benefit	Spectral superresolution	$N^* \ll P$
Attention head dimension	Spectral bandwidth	$d_{\text{head}} \geq N^*$
Emergent ability threshold	Phase transition point	N^*_T ordering
Catastrophic forgetting	Spectral collapse	$\rho_A \to \rho'_A < 1$
Continual learning	Spectral preservation	Maintain $\rho_{\text{old}} > 1$
IB compression	Spectral truncation	Keep N^* modes
Loss landscape	Spectral gap	Basin width $\propto \log \rho$
Preference structure	Spectral coherence	ρ_{pref}
Reward hacking	Spectral mismatch	$\rho_{\text{rew}} < \rho_{\text{pref}}$
Adversarial attack surface	Null space dimension	$D - N^*$
Certified robustness radius	Spectral gap	$\log \rho$

§15.5 Open Problems

1. **Prove $\dot{\rho}(t) > 0$ under gradient flow.** Establish the spectral monotonicity of gradient descent on structured data.
2. **Measure ρ empirically.** Develop practical estimators of ρ for image, language, and tabular distributions.
3. **Tight constants.** Determine the optimal multiplicative constants in the generalization bounds.
4. **Feature learning.** Extend the theory to account for representation learning that changes ρ during training.
5. **Adversarial-robustness tightness.** Determine whether the $D - N^*$ attack surface bound is achievable by constructive attacks.
6. **Multi-task ρ dynamics.** Analyze how ρ evolves under multi-task training where different tasks have competing spectral requirements.

§15.6 Formalization

All 146 theorems are machine-checked in the Platonic proof environment (11 proof files, 48 parts, no deferred proof obligations). This is not a standalone Lean 4 artifact: statements are verified by ProofEnv (see elysium/ fields /neural_scaling_laws/). The proofs use the RealDSL and ExpLogDSL domain-specific languages with explicit tactics including linarith , nlinarith , derive, and exact_apply for axiom application. Key axioms used: Real.log_pos, Real.log_neg, Real.log_lt_log, Real.exp_lt_exp, Real.exp_le_exp, Real.exp_pos, Real.exp_zero, Real.sqrt_le_sqrt.

During the preparation of this work the author used large language models in order to assist with manuscript drafting, formalization, 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.

References

Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32), 15849-15854.

Hoffmann, J., et al. (2022). Training compute-optimal large language models. *Advances in Neural Information Processing Systems*, 35, 30016-30030.

Kaplan, J., et al. (2020). Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*.

Nagy, T. (2026). The Latent: Finite Sufficient Representations of Smooth Systems. *Zenodo*. DOI: 10.5281/zenodo.19101209.

Power, A., et al. (2022). Grokking: Generalization beyond overfitting on small algorithmic datasets. *arXiv preprint arXiv:2201.02177*.

Wei, J., et al. (2022). Emergent abilities of large language models. *Transactions on Machine Learning Research*.

McCloskey, M. & Cohen, N. J. (1989). Catastrophic interference in connectionist networks: The sequential learning problem. *Psychology of Learning and Motivation*, 24, 109-165.

Tishby, N., Pereira, F. C., & Bialek, W. (2000). The information bottleneck method. *Proceedings of the 37th Annual Allerton Conference on Communication, Control, and Computing*, 368-377.

Christiano, P. F., et al. (2017). Deep reinforcement learning from human feedback. *Advances in Neural Information Processing Systems*, 30.

Goodfellow, I. J., Shlens, J., & Szegedy, C. (2015). Explaining and harnessing adversarial examples. *International Conference on Learning Representations*.

Madry, A., et al. (2018). Towards deep learning models resistant to adversarial attacks. *International Conference on Learning Representations*.