Grokking as Manifold Discovery: A Geometric Reinterpretation of Delayed Generalization

Author: Jin Yanyan (lmxxf@hotmail.com)

Abstract: Grokking—the phenomenon where neural networks suddenly generalize after prolonged overfitting—has accumulated multiple theoretical explanations since its discovery in 2022: Goldilocks Zone, Softmax Collapse, Lazy-Rich transition, etc. This paper reviews these theories and identifies their common blind spot: **most focus on external measurements, lacking direct characterization of representation space geometry**. Among them, the Goldilocks Zone theory touches on the "physical laws" of high-dimensional space and carries substantial theoretical value. We propose a unified framework—the **Manifold Discovery Hypothesis**: memorization is a high-dimensional jagged curve passing through all training points, generalization is discovering the low-dimensional manifold on which data is distributed, and Grokking is the phase transition from the former to the latter. In one sentence: **high-dimensional curve → low-dimensional surface**.

---

1. Introduction: Why Grokking Matters

In 2022, Power et al. at OpenAI discovered a counterintuitive phenomenon: small Transformers trained on modular arithmetic tasks would first **perfectly overfit** the training set (training loss drops to zero, test accuracy near random), then **tens of thousands or even hundreds of thousands of steps later**, test accuracy suddenly jumps from random level to nearly 100%.

They named this phenomenon **Grokking** (sudden understanding).

This discovery is important because it challenges a core assumption of deep learning:

1. **Classical assumption**: Overfitting is the enemy of generalization; once overfitting occurs, early stopping should be applied
2. **Grokking counterexample**: Overfitting can persist for a long time, then suddenly generalize

If generalization can truly occur after overfitting, the "early stopping" strategy may have killed many models that could have generalized.

The deeper question: **What happens inside the model during Grokking?**

Over the past three years, academia has accumulated multiple theoretical explanations. This paper's task is to review these theories, identify their blind spots, and propose a unified framework.

---

2. Review of Existing Theories

2.1 Goldilocks Zone Theory (Liu et al. 2022)

Core idea: Weight norm must fall within a "just right" interval for generalization.

Liu et al. at NeurIPS 2022 found that weight space contains a **hollow spherical shell**, which they called the Goldilocks Zone:

- Radius too large ($\|w\| > w_c$): Overfitting, memorizing training set
- Radius too small ($\|w\| < w_c$): Underfitting, learning nothing
- Just right on the shell ($\|w\| \approx w_c$): Generalization

Grokking mechanism: 1. Large initialization places the model outside the shell 2. Model quickly overfits first (training loss drops to zero) 3. Weight decay slowly pulls weight norm back to Goldilocks Zone 4. Once inside the shell → sudden generalization → Grokking

The true value of this paper: It suggests that high-dimensional space has its own "physical laws"—weight decay is gravity, Goldilocks Zone is the stable orbit. This is the foundation for all subsequent theories.

Limitation: Describes "where generalization happens," but doesn't explain "why it can generalize there." What is Goldilocks Zone a proxy for?

2.2 Softmax Collapse Theory (Prieto et al. 2025)

Core idea: Without weight decay, Grokking is killed by floating-point precision.

To minimize cross-entropy loss, the model aggressively amplifies the correct answer's logit (e.g., correct class = 1000, others = 1). When computing softmax, e^{1000} directly overflows, gradients become zero, training stalls.

Role of weight decay: Continuously pulls weights back, preventing infinite logit growth, keeping gradients alive.

Alternative solution: The paper proposes StableMax + orthogonal gradients (preventing gradients from going in the "amplify logit" direction), which can trigger Grokking without weight decay. However, this method may have lower convergence efficiency—weight decay is global compression with broad scope; orthogonal gradients are local constraints with narrow scope. In practice, weight decay remains the more common choice.

Contribution: Explains "what happens without weight decay."

Limitation: Only explains "why training doesn't stop," not "why it eventually generalizes."

2.3 Lazy → Rich Transition Theory (Kumar et al. 2024)

Core idea: Grokking is a phase transition from lazy training to feature learning.

Borrowing Neural Tangent Kernel (NTK) language:

- **Lazy regime**: Weights barely move, model acts like linear classifier
- **Rich regime**: Weights adjust significantly, learning true nonlinear features

Grokking occurs at the **phase transition point** from lazy → rich.

Controversy: This camp claims that under specific conditions (shallow networks + MSE loss), Grokking can be triggered without weight decay.

Assessment: The main contribution of this theory is introducing the lazy/rich conceptual framework, though there remains room for explaining "why the phase transition occurs."

2.4 Weight Efficiency Hypothesis (Varma et al. 2023)

Core idea: Weight decay favors solutions with smaller weights, and generalizing solutions are typically more weight-efficient than memorizing solutions.

- Memorization solution: Requires large weights to hard-memorize each sample
- Generalization solution: Uses concise rules to cover all samples, smaller weights
- Weight decay → penalizes large weights → favors generalization solution

Assessment: Similar to 2.3, this theory provides a useful perspective (small weights ↔ generalization), but is essentially another description of the same phenomenon, not yet revealing causal mechanisms.

2.5 Mechanistic Interpretability Perspective (Nanda et al. 2023)

Nanda et al. at ICLR 2023 (Oral) conducted solid work: **completely reverse-engineered** the algorithm the model learned.

Core finding: Modular arithmetic $(a+b) \mod p$ is essentially a **cyclic group**—0, 1, 2, ..., p-1 connected end to end, forming a discrete ring. The model decomposes this modular operation into Fourier series. Human post-hoc analysis of weight matrices found it equivalent to Fourier transform structure. The model itself just does matrix multiplication, knowing nothing about Fourier formulas.

Discussion: Modular arithmetic is naturally periodic, and expanding with Fourier series is a mathematical tool from 200 years ago. From this perspective, the model "discovering" Fourier structure is more like an inevitable response to the task's inherent periodicity, rather than an unexpected finding.

Contribution: Hard work, opened up the model to look inside.

Limitation: Doesn't explain why Weight Decay + overfitting + continued training = Fourier transform.

2.6 Common Blind Spot of Existing Theories

Theory	Question Asked	Question Not Asked
Goldilocks Zone	What weight norm interval	What's special about that interval

Theory	Question Asked	Question Not Asked
Softmax Collapse	Why training doesn't stop	Why it eventually generalizes
Lazy → Rich	How weights change	How representations change
Weight Efficiency	Which solution has smaller weights	Why small weights = generalization
Mechanistic Interp.	What circuit was learned	Why this circuit

Assessment: The Goldilocks Zone theory touches on the "physical laws" of high-dimensional space and carries substantial theoretical value. Softmax Collapse, Lazy→Rich, and Weight Efficiency each provide useful perspectives, but primarily remain at the level of external measurements (weight norm, gradient magnitude, loss curves), not yet revealing the geometric essence of representation space. Mechanistic Interpretability starts looking inside, but focuses on specific circuits, not geometric structure.

Overall, existing theories have done substantial work in "describing the phenomenon," but there remains room for improvement in "explaining the mechanism." This paper attempts to provide a complementary unified framework from the geometric perspective.

---

3. Unified Framework: The Manifold Discovery Hypothesis

We propose a unified framework: **Grokking is a phase transition from high-dimensional jagged curves to low-dimensional smooth manifolds.**

3.1 Geometric Interpretation of Memorization vs. Generalization

Memorization = Jagged Curve

When a model overfits the training set, it uses a complex jagged curve to pass through every training sample point. This curve can precisely hit all training data, but it **has no pattern**—just forcibly stringing all points together, with no structural relationship between points.

Generalization = Manifold Discovery

When the model truly "understands" the task, it discovers that training samples are actually distributed on a **low-dimensional manifold** (low-dimensional relative to the model's hidden dim).

Taking modular arithmetic $a + b \mod p$ as an example: - Input space is p^2 discrete points - But output only depends on $(a + b) \mod p$, i.e., the **congruence class** - The true structure is a one-dimensional **cyclic group** $\mathbb{Z}_p$ (only one degree of freedom: position)

Generalization means: the model discovered this cyclic group structure, rather than hard-memorizing p^2 input-output pairs.

Grokking = Phase Transition from Curve to Manifold

What happens at the Grokking moment: 1. Before: A high-dimensional jagged curve passing through all points in representation space 2. After: The curve collapses onto a low-dimensional manifold, whose topological structure corresponds to the task's true structure

This is a **topological phase transition**, not a continuous transition.

In one sentence: high-dimensional curve → low-dimensional surface.

3.2 Geometric Role of Weight Decay

In this framework, weight decay's role becomes clear:

Weight Decay = Centripetal Force that Destabilizes Jagged Curves

Without weight decay: - Gradient descent pushes the model toward "lowest loss" - For over-parameterized models, this is "perfectly memorize each training sample" - The model stabilizes on solutions formed by jagged curves

With weight decay: - There's a force opposite to loss gradient, continuously pulling weights toward "smaller" - Jagged curves need "large weights" to maintain (each inflection point needs dedicated neurons) - Smooth manifolds need "small weights" (structure sharing) - Weight decay **favors manifold encoding**

The True Meaning of Goldilocks Zone

Goldilocks Zone is not "some weight norm interval," but **activation patterns that can perceive manifold structure**.

- Weights too large: Jagged curve stable, can't see manifold
- Weights too small: Signal too weak, can't see any structure
- Just right: Jagged curve unstable + signal strong enough → manifold emerges

3.3 Geometric Interpretation of Softmax Collapse

Softmax Collapse is not just a floating-point issue, but **attention collapse**.

When a direction is over-strengthened: - That direction's logit tends toward infinity - Other directions' "voices" are drowned out - Model loses ability to explore other possibilities

This is why Softmax Collapse kills Grokking: Discovering manifolds requires simultaneously "seeing" multiple directions, while collapse shrinks the field of view to a single point.

Role of weight decay: Maintain signal balance across multiple directions, allowing the model to continue exploring.

3.4 Geometric Interpretation of Lazy → Rich

Lazy → Rich transition is not about weight dynamics, but about **representation complexity**.

- Lazy regime: Representation is linear function of input, can only encode linearly separable structures
- Rich regime: Representation is nonlinear function of input, can encode arbitrary manifolds

Grokking requires Rich regime: Because the true task structure (like cyclic groups) is nonlinear.

Why does Grokking often occur late in training? Because entering Rich regime requires sufficient weight change, and at initialization the model is in Lazy regime.

3.5 Geometric Interpretation of Weight Efficiency

The causal chain of "small weights = generalization":

1. Smooth manifolds are more "compact" than jagged curves (lower dimension)
2. Compact encoding requires fewer weights to implement
3. Weight Decay favors small weights → favors compact encoding → favors manifolds

Weight efficiency is the **result** of manifold discovery, not the **cause**.

3.6 A Set of Formalizable Mathematical Statements (Turning "Metaphors" into Provable Propositions)

This section does not attempt to "prove Grokking must happen" (that would require very strong assumptions about networks, data distributions, and optimization dynamics). Instead, we rewrite the key intuitions from this paper's framework into **mathematical propositions that are provable/verifiable under standard assumptions**:

1) **The dynamical form of weight decay** (it really is a centripetal force);
2) **The minimum norm bias of L_2 regularization** (it really favors "structure-sharing/low-complexity" interpolating solutions);
3) **Why intrinsic dimension drops when "manifold/group structure is discovered"** (at least when "representation depends only on group invariants," this can be rigorously derived).

   Notation: Parameters $w \in \mathbb{R}^m$, empirical loss $\mathcal{L}(w) = \frac{1}{n}\sum_{i=1}^{n} \ell(f_w(x_i), y_i)$, weight decay coefficient $\lambda \geq 0$, learning rate $\eta > 0$.

Lemma 3.1 (Weight decay = centripetal force in discrete dynamics) Consider the objective with L_2 regularization:

$$J(w) = \mathcal{L}(w) + \frac{\lambda}{2}\|w\|_2^2.$$

Gradient descent on J:

$$w_{t+1} = w_t - \eta \nabla J(w_t) = (1 - \eta\lambda) w_t - \eta \nabla \mathcal{L}(w_t).$$

Thus, the optimization update consists of two superimposed parts: one shrinks w_t proportionally (contraction toward origin), the other descends along the loss gradient. **Proof**: Directly expand $\nabla(\frac{\lambda}{2}\|w\|^2) = \lambda w$. □

This lemma turns Section 3.2's "centripetal force" from metaphor into an explicit dynamical term: present at all steps, independent of specific data.

Proposition 3.2 (Linear regression: regularized interpolating solution tends to minimum norm solution) Let $f_w(x) = w^\top x$, squared loss $\ell = \frac{1}{2}(f_w(x) - y)^2$, data matrix $X \in \mathbb{R}^{n \times d}$ with full row rank, and $y \in \mathbb{R}^n$. The optimal solution with L_2 regularization (ridge regression) satisfies

$$w_\lambda = \arg\min_w \frac{1}{2n}\|Xw - y\|_2^2 + \frac{\lambda}{2}\|w\|_2^2 = (X^\top X + n\lambda I)^{-1} X^\top y.$$

If an interpolating solution exists (i.e., $\exists w : Xw = y$), then as $\lambda \downarrow 0$, w_λ converges to the **minimum L_2 norm interpolating solution**

$$w_* = \arg\min_{Xw=y} \|w\|_2,$$

and w_* can be written as $w_* = X^\top (XX^\top)^{-1} y$. **Proof (key points)**: Use first-order optimality conditions to get closed-form solution. For $\lambda \to 0$, use X having full row rank to guarantee $XX^\top$ is invertible, and derive using pseudo-inverse limits or equivalent Lagrange multipliers. □

This proposition clarifies one thing in the simplest convex case: **L_2 regularization doesn't just "prevent overfitting"; when interpolation is feasible, it selects the "minimum norm" solution**. Replacing "norm" with more general function space norms (like RKHS/Sobolev) naturally yields "smoother/lower frequency" bias—consistent with this paper's "curve (high frequency) → manifold (low frequency)" intuition.

Proposition 3.3 (Separable classification: without regularization weight norm tends to infinity; with regularization optimal solution is bounded) Let binary classification linear model $f_w(x) = w^\top x$, logistic loss $\ell(w; x, y) = \log(1 + \exp(-y\, w^\top x))$, and data be linearly separable: $\exists \bar{w}$ such that for all i, $y_i\, \bar{w}^\top x_i > 0$. Then: 1) Without L_2 regularization ($\lambda = 0$), the infimum of minimum empirical loss is 0, but generally **no finite-norm optimal solution exists** (can scale w along the separable direction to make loss arbitrarily close to 0). 2) With L_2 regularization ($\lambda > 0$), the objective $J(w) = \frac{1}{n}\sum_i \ell(w; x_i, y_i) + \frac{\lambda}{2}\|w\|^2$ is **strongly convex + lower semi-continuous** on $\mathbb{R}^d$, thus there exists a unique optimal solution w_λ, with $\|w_\lambda\| < \infty$. **Proof (key points)**: 1) Use $\ell(\alpha w) \to 0$ ($\alpha \to \infty$) to show loss can be pushed arbitrarily small but not achieved. 2) The L_2 term provides strong convexity and coerciveness: $\|w\| \to \infty$ implies $J(w) \to \infty$, thus a unique minimum exists. □

This proposition corresponds to Section 2.2's Softmax/Logit explosion phenomenon: in cross-entropy/logistic loss, the shortcut to "keep improving" is often to amplify the margin to infinity; weight decay turns it into a bounded optimization problem, and gradients won't naturally vanish due to "infinite norm" escape.

Proposition 3.4 ("Discovering group structure" ⇒ representation degrees of freedom decrease: a rigorously testable sufficient condition) Using modular addition as an example, inputs are $(a, b) \in \mathbb{Z}_p^2$, let sum be $s = a+b \pmod p \in \mathbb{Z}_p$. Let some layer's representation be $h(a, b) \in \mathbb{R}^k$, and there exists a function $\phi : \mathbb{Z}_p \to \mathbb{R}^k$ such that

$$h(a, b) = \phi(s) \quad \text{depends only on } s = (a + b) \bmod p.$$

Then the representation set $\{h(a, b) : a, b \in \mathbb{Z}_p\}$ contains at most p points (exactly $\phi(\mathbb{Z}_p)$), and its effective degrees of freedom are no longer proportional to p^2, but bounded by p. Furthermore, if there exists a smooth embedding $\Phi : S^1 \to \mathbb{R}^k$ and a homomorphism $\iota : \mathbb{Z}_p \hookrightarrow S^1$ (embedding the discrete cyclic group into the circle) such that $\phi = \Phi \circ \iota$, then these representation points lie on a one-dimensional manifold (circle). **Proof**: The first part follows immediately from "function depends only on s": different (a, b) with the same congruence class map to the same h, so the number of distinct representations is bounded by the number of congruence classes p. The second part is direct substitution into the composition mapping definition: $\phi(\mathbb{Z}_p) \subseteq \Phi(S^1)$. □

This proposition provides an **operationally testable sufficient condition** for Section 5.1's "intrinsic dimension discontinuity": once some layer's representation truly "forgets (a, b)'s two degrees of freedom, keeping only s's one degree of freedom," then the degrees of freedom decrease you see using PCA/TwoNN/local dimension estimation is not metaphysics, but forced by representation factorization.

3.7 A Formulation Closer to Real LLMs (GPT-4 Class Systems): Continuous Approximation + Spectral/Low-Complexity Bias

If betting on "which mathematical path real-world GPT-4 more resembles," I would choose: **continuous approximation (S^1 / low-dimensional manifold) + spectral/low-complexity bias**, rather than doing completely rigorous algebraic derivations only in discrete finite groups. The reason is simple: real LLM training data and task distributions are closer to "samples from a continuous world," and Transformer representations typically exhibit strong "learn low frequency/simple structure first, then high frequency/exceptions" dynamical bias.

Below is a version you can directly write into papers, requiring readers to know less abstract algebra (it's compatible with Section 3.6, just replacing "discrete group $\mathbb{Z}_p$" with "continuous circle S^1 approximation").

Setup (viewing $\mathbb{Z}_p$ as evenly-spaced samples on S^1) Map $k \in \mathbb{Z}_p$ to angle $\theta_k = 2\pi k/p \in [0, 2\pi)$, and map "sum s" to circle position θ_s. Assume some layer's representation satisfies the approximate form

$$h(a, b) \approx \Phi(\theta_{(a+b) \bmod p}) \quad (\Phi : S^1 \to \mathbb{R}^d \text{ continuous/smooth}).$$

Then "discovering structure" is equivalent to: the network learns at some layer a **low-dimensional parameterization** $\theta \mapsto \Phi(\theta)$, rather than memorizing an independent representation for each (a, b).

Proposition 3.5 (Fourier expansion on the circle: low frequency = smooth structure, high frequency = jagged memorization) For each output dimension, let $\Phi_j(\theta)$ be square-integrable, then there exists a Fourier series

$$\Phi_j(\theta) = \sum_{m \in \mathbb{Z}} c_{j,m}\, e^{im\theta}.$$

If coefficients decay rapidly in frequency (e.g., $\sum_m m^2 |c_{j,m}|^2 < \infty$), then Φ is smoother in θ; conversely, if many high-frequency components are needed to fit the fine-grained differences of training points, this corresponds to "jagged/memorization" style representation. **Note**: This is not "proving networks must learn low frequency," but providing a quantifiable characterization: you can do discrete Fourier transform (DFT) on representations sampled at θ in experiments, checking whether energy spectrum concentrates from high to low frequency before and after Grokking.

Proposition 3.6 (Minimizing smoothness regularization suppresses high frequency: a clean variational conclusion) Consider the variational problem of fitting target function $g(\theta)$ on S^1:

$$\min_{\Phi} \int_{S^1} \|\Phi(\theta) - g(\theta)\|^2 \, d\theta \; + \; \alpha \int_{S^1} \|\partial_\theta \Phi(\theta)\|^2 \, d\theta, \quad \alpha > 0.$$

The optimal solution satisfies shrinkage for frequency m in Fourier domain: high-frequency components are penalized more strongly (because $\|\partial_\theta e^{im\theta}\|^2 \propto m^2$), so the solution is biased toward low-frequency/smooth structure. **Note**: In real networks you don't have explicit $\int \|\partial_\theta \Phi\|^2$ regularization, but "parameter norm/weight decay + optimization dynamics" often exhibits similar low-complexity bias; this gives you a bridge closer to engineering intuition: **weight decay → low-complexity bias → high frequency suppressed → representation smoother → more like low-dimensional manifold**.

Including this section allows you to explain Section 5's predictions using "spectral energy shifting from high to low frequency," and it aligns better with real LLM observations (learn commonalities first, then exceptions).

---

4. Reinterpreting Existing Findings

4.1 Why Weight Decay Being Too Large/Small Both Fail

Too small: Centripetal force insufficient, model stabilizes on jagged curve, never explores manifold.

Too large: Centripetal force too strong, signal suppressed, can't even draw jagged curve, let alone discover manifold.

Just right: Jagged curve unstable but signal strong enough, model forced to explore, eventually discovers manifold.

4.2 Why Data Amount Affects Grokking

Too little data: Samples on manifold too sparse, can't recover manifold structure. Model can only draw jagged curve.

Too much data: Manifold signal too strong, model directly discovers manifold, no overfitting phase, no "delayed generalization."

Just right: Manifold signal exists but not obvious, model draws jagged curve first, slowly discovers the structure behind the curve.

This explains why Grokking needs a "Goldilocks" data amount.

4.3 Why Over-parameterized Models Are More Prone to Grokking

Over-parameterization = representation space large enough

- Small model: Representation space might not fit the true manifold at all
- Large model: Representation space large enough, manifold can exist, just needs time to discover

The over-parameterization paradox: - Traditional view: Over-parameterization leads to overfitting - Grokking view: Over-parameterization is a **prerequisite** for generalization, because it provides sufficient representation space

Weight decay solves over-parameterization's "too many degrees of freedom" problem: although the space is large, centripetal force pushes the model toward low-dimensional manifolds.

4.4 Why Grokking Is Sudden

Phase transitions are not continuous

Manifold discovery is a **topological event**:

- Before: High-dimensional jagged curve
- After: Low-dimensional smooth manifold

The collapse from high to low dimension has no stable intermediate state, so Grokking is sudden.

Analogy: Iron's Curie Point

The transition of iron from ferromagnetic to paramagnetic occurs at a specific temperature (Curie point), a continuous but sudden phase transition—not gradually weakening, but suddenly losing magnetism at the critical point.

Grokking is similar: The structure of representation space undergoes qualitative change at some critical point, jagged curve "collapses" into smooth manifold.

---

5. Testable Predictions

This section is deliberately written to not "close the loop perfectly." We don't promise what some curve **must** look like, but give a set of **falsifiable observational signatures**: after you do experiments, results will push the story in some direction (support / revise / refute), rather than falling into the "can be explained either way" word game.

5.1 Intrinsic Dimension Discontinuity

Observational signature: Before and after Grokking, intrinsic dimension of intermediate layer representations may show "discontinuity/collapse," or may only show "slow drift."

Two mutually exclusive readings (giving the experiment a "choose one" verdict): - **If representation truly factorizes** (e.g., some layer starts depending approximately only on congruence class/group invariant), intrinsic dimension should drop significantly from "close to sample degrees of freedom" to "close to task degrees of freedom" (like modular addition's 1D structure). - **If generalization comes from other mechanisms** (e.g., just decision boundary becoming more stable, but representation not factorizing), you might not see obvious dimension reduction, only smooth variation in dimension estimates during training.

Experimental design: 1. Train model on modular arithmetic, record intermediate layer activations 2. Use existing algorithms (like SVD, PCA, or TwoNN) to estimate intrinsic dimension 3. Plot intrinsic dimension vs. training steps

Decision point: Whether intrinsic dimension curve shows structural inflection (discontinuity or clear turning point) near test accuracy jump.

5.2 Topological Structure of Representations

Observational signature: After Grokking, intermediate layer representations' topological structure may start "matching" task structure, or may only show weak correlation.

- Modular arithmetic $a+b \mod p$: Representation should form a one-dimensional ring
- Symmetric group tasks: Representation should form corresponding group manifold

Experimental design: 1. Extract intermediate layer representations before and after Grokking 2. Use persistent homology to compute topological invariants 3. Compare with task's true topological matching degree

Decision point: Whether Betti numbers/persistence diagrams show clear topological signatures near "grokking" (e.g., ring structure's β_1 enhancement), and stability across different random seeds.

5.3 Attention Entropy Dynamics

Observational signature: During Grokking, attention pattern entropy may show "low→high→medium" exploration-convergence process, or may completely not follow this narrative (then attention isn't the key variable).

Two distinguishable patterns: - **Exploration-convergence type**: Early low entropy (specialized patterns) → pre-critical entropy rise (pattern diversification) → afterward falls and stabilizes (shared structure). - **Monotonic/irrelevant type**: Entropy changes monotonically or is very noisy, unrelated to test jump timing—this would weaken "attention collapse"'s explanatory weight for this task.

Experimental design: 1. Record attention matrix at each training step 2. Compute attention distribution entropy 3. Plot entropy vs. training steps

Decision point: Whether entropy peaks/valleys align with test jump, and reproducible across multiple seeds.

5.4 Effect of Forced Rank Constraints

Observational signature: If "representation dimension/rank" is truly the bottleneck, forcing low-rank constraints on intermediate layers should systematically change Grokking timing; if almost no change, then the determining factor may be elsewhere (optimization/numerical stability/normalization, etc.).

- Rank constraint = task's true degrees of freedom: Accelerate Grokking (directly telling model answer's dimension)
- Rank constraint < task's true degrees of freedom: Prevent Grokking (representation space can't fit manifold)
- Rank constraint > task's true degrees of freedom: No effect or slight slowdown

Experimental design: 1. Add low-rank bottleneck to intermediate layer (like linear layer limiting output dimension) 2. Vary bottleneck dimension, record Grokking time 3. Compare with no-bottleneck baseline

Decision point: Whether Grokking time vs. bottleneck dimension curve has clear "phase transition region" (once below some threshold, completely fails), and whether threshold is same order of magnitude as task's minimum degrees of freedom.

---

6. Discussion

6.1 Methodological Limitations of Existing Research

Measurement targets of existing Grokking research: - Weight norm - Gradient magnitude - Loss curves - Test accuracy

These are all **external observables**—values measurable from outside the model.

Analogy: Studying human learning by only measuring brainwaves and pupil diameter, not asking "what does learning feel like."

This methodology can answer "under what conditions Grokking occurs," but cannot answer "what Grokking is."

6.2 Value of the Internal Perspective

The manifold discovery hypothesis proposed in this paper is an **internal perspective** explanation:

- Not asking "what weight norm interval"
- Asking "what is the structure of representation space"

The value of this perspective: 1. **Unification**: Can simultaneously explain Goldilocks Zone, Softmax Collapse, Lazy→Rich, and is compatible with Nanda et al.'s circuit findings 2. **Predictability**: Generates verifiable experimental predictions (intrinsic dimension, topological structure, etc.) 3. **Heuristic value**: Points to new research directions (directly measuring geometric properties of representation structure, not just reverse-engineering specific circuits)

6.3 Limitations and Future Directions

Limitations of this paper: 1. **Hypothesis nature**: The manifold discovery hypothesis is currently a conceptual framework, not a mathematical proof 2. **Experimental verification pending**: Section 5's predictions need experimental verification 3. **Task dependence**: Unclear if this framework applies to all Grokking tasks

Future directions: 1. Experimentally verify intrinsic dimension discontinuity hypothesis 2. Develop metrics that directly measure "degree of manifold discovery" 3. Explore whether Grokking can be controlled by manipulating representation space topology

---

7. Conclusion

Grokking is not an anomaly, but a window into deep learning—it reveals the essence of generalization.

Among existing theories, Goldilocks Zone touches on the "physical laws" of high-dimensional space and carries substantial theoretical value; the other theories each contribute, but primarily remain at the level of external measurements.

The manifold discovery hypothesis proposed in this paper provides a unified framework from the internal perspective:

- **Memorization = jagged curve**
- **Generalization = smooth manifold**
- **Grokking = topological phase transition from the former to the latter**
- **Weight decay = centripetal force that destabilizes jagged curves**

In one sentence: high-dimensional curve → low-dimensional surface.

This framework not only explains existing findings, but also generates verifiable experimental predictions.

References

1. Power, A., Burda, Y., Edwards, H., Babuschkin, I., & Misra, V. (2022). Grokking: Generalization Beyond Overfitting on Small Algorithmic Datasets. arXiv:2201.02177.
2. Liu, Z., Kitouni, O., Nolte, N., Michaud, E. J., Tegmark, M., & Williams, M. (2022). Towards Understanding Grokking: An Effective Theory of Representation Learning. NeurIPS 2022. arXiv:2205.10343.
3. Liu, Z., Michaud, E. J., & Tegmark, M. (2023). Omnigrok: Grokking Beyond Algorithmic Data. ICLR 2023. arXiv:2210.01117.
4. Prieto, L., Barsbey, M., Mediano, P. A. M., & Birdal, T. (2025). Grokking at the Edge of Numerical Stability. ICLR 2025. arXiv:2501.04697.
5. Kumar, T., Bordelon, B., Gershman, S. J., & Pehlevan, C. (2024). Grokking as the Transition from Lazy to Rich Training Dynamics. ICLR 2024. arXiv:2310.06110.
6. Varma, V., Shah, R., Kenton, Z., Kramár, J., & Kumar, R. (2023). Explaining Grokking Through Circuit Efficiency. arXiv:2309.02390.
7. Nanda, N., Chan, L., Lieberum, T., Smith, J., & Steinhardt, J. (2023). Progress Measures for Grokking via Mechanistic Interpretability. ICLR 2023 (Oral). arXiv:2301.05217.
8. Facco, E., d'Errico, M., Rodriguez, A., & Laio, A. (2017). Estimating the intrinsic dimension of datasets by a minimal neighborhood information. Scientific Reports, 7(1), 12140.
9. Carlsson, G. (2009). Topology and data. Bulletin of the American Mathematical Society, 46(2), 255-308.

"Existing research primarily asks 'under what conditions does Grokking occur'; this paper attempts to ask 'what is Grokking.' The former is an engineering question; the latter is an ontological question."