Proceedings of Machine Learning Research vol vvv:1–25, 2025

Why Neural Network Can Discover Symbolic Structures with Gradient-based Training: A Theoretic Foundation for Neurosymbolic AI with Practical Implications

Zhangyang "Atlas" Wang* ATLASWANG@UTEXAS.EDU
Peihao Wang* PEIHAOWANG@UTEXAS.EDU
VITA Group, The University of Texas at Austin

Abstract

We introduce a novel theoretical framework showing how discrete symbolic reasoning can emerge from continuous neural network training dynamics. By interpreting network optimization as a gradient flow in measure space, we prove that parameters concentrate onto low-dimensional, G-invariant manifolds. This dimension reduction, driven by symmetries and monomial-potential constraints, endows the network with approximate ring compatibility and latent symbolic structures. Crucially, the measure transitions from a high-dimensional "exploration" phase to a low-dimensional "exploitation" phase, revealing discrete-like algebraic patterns. We further derive fundamental width constraints for capturing G-invariant symbolic tasks, showing a linear dependence on $\dim(G)$ for continuous groups and a logarithmic dependence on $|G|$ for finite groups. We also unify seemingly disparate observations — such as why certain architectures or activations excel at reasoning — and yield concrete model design principles for neural reasoning tasks. Overall, our results illuminate how neural networks can internalize and exploit symbolic capabilities, charting a principled path toward robust neurosymbolic AI.

Keywords: Neurosymbolic reasoning, Mean-field analysis, Algebraic geometry

1. Introduction

The integration of neural and symbolic reasoning is a key challenge in advancing the capabilities of modern AI systems. Neural-symbolic AI (Chaudhuri et al., 2021; Garcez and Lamb, 2023) aims to combine the representational flexibility and approximation power of neural networks with the precision and compositional rigor of symbolic reasoning. Neural networks excel at learning smooth manifolds in high-dimensional parameter spaces and adapting their behavior from large-scale data. Symbolic reasoning, on the other hand, enables exact inference over discrete logical structures and algebraic constraints. Bridging these strengths promises systems that can handle both statistical and combinatorial aspects of complex tasks, leading to improved generalization, alleviated data hunger, and more transparent reasoning processes.

However, existing neural architectures often struggle to internalize true symbolic capabilities, instead relying on *statistical pattern matching* that **fails when generalizing** beyond the training distribution (Zhang et al., 2023; Valmeekam et al., 2023). This underscores the need for theoretical frameworks that explain *how* symbolic structures can emerge from continuous neural training dynamics. Understanding this emergence at a fundamental level can guide architectural choices and training strategies, ultimately shaping the design of robust neurosymbolic systems.

* Z. Wang developed the initial framework as his pet project, while P. Wang played a crucial role as the constructive reviewer and relentless critic. As a result, they contributed equally to this paper.

In this work, we propose a novel theoretical framework that reveals how discrete, symbolic reasoning constraints arise naturally from the continuous evolution of neural network parameters under gradient-based optimization. As the first framework of its kind, our analysis requires certain idealized assumptions (including *displacement convexity* and C^2*-smoothness*) that may be partially violated in practical networks. Nevertheless, these assumptions offer a tractable lens into how continuous parameter updates can lead to discrete-algebraic behaviors. By modeling neural training as a gradient flow in measure space, we show that:

- Neural networks evolving under stable displacement-convex conditions concentrate their parameters onto lower-dimensional manifolds as training progresses. This *dimension reduction* phenomenon is not a superficial artifact but a structural collapse onto submanifolds that reflect G-invariance (symmetry) and *approximate* multiplicative factorization properties encoded by Monomial Potentials (MPs) (Tian, 2024).
- The measure μ_t representing the network's parameters converges to a minimal G-invariant orbit, revealing emergent algebraic (ring-like) structures that mirror symbolic constraints. Through this process, the network transitions from high-dimensional *exploration* to low-dimensional *exploitation*, effectively discovering and internalizing the underlying symbolic patterns. While exact ring-compatibility requires a degenerate (e.g. delta) measure, we focus on an ε*-approximate* notion of factorization sufficient for typical reasoning tasks.

Critically, our theory links these geometric and algebraic insights to concrete, actionable principles. The interplay of dimension reduction and G-invariance informs minimum width requirements for networks to achieve stable ε-approximate symbolic operations. It also dictates architectural constraints—such as preserving G-equivariance, ensuring critical manifolds have bounded curvature, and maintaining ring compatibility through suitable activation functions. Although we work within a compact or effectively bounded parameter space for theoretical clarity, we discuss how this analysis can guide practical architectures. These insights provide a principled explanation for empirical observations and guide the design of architectures that excel at both continuous representation learning and discrete symbolic reasoning.

1.1. Main Contributions

Theoretical Contributions: We establish a rigorous measure-theoretic and geometric foundation for understanding the emergence of symbolic structures in neural networks, by showing that:

- **Dimension Reduction and Phase Transitions:** Neural measures evolve onto G-invariant, lower-dimensional manifolds via a sequence of critical times, each inducing a sharper, more algebraically constrained representation. In practice, this collapse is partial and incremental, but can be studied rigorously under a simplified assumption of displacement-convexity.
- **Algebraic (Ring) Structures from Continuous Dynamics:** By analyzing MP integrals, we uncover *approximate* ring-compatibility that underlies symbolic reasoning tasks, connecting continuous parameter evolution to discrete algebraic constraints. Our results focus on ε-approximate homomorphism properties relevant to the network's objective.
- **Scalable Complexity Bounds:** The theory reveals scaling laws for minimal network width as a function of group dimension or the size of finite groups, providing a formal linkage between representation capacity and symbolic complexity.

Practical Implications: After presenting our theoretical framework for dimension reduction and G-invariance in Sections 2 and 3, we translate these findings into practical principles and empirical insights for guidelines for neural-symbolic architecture design in Section 4:

- **Minimum Width Requirements:** Base this on idealized uniform-approximation arguments, we show that tasks with continuous symmetry groups (G) must scale network width linearly with $\dim(G)$, while discrete operations require widths scaling with $\log|G|/\log(1/\delta)$, offering a direct recipe for capacity planning.
- **Architectural Constraints:** Ensuring G-invariance, choosing ring-compatible backbones (transformers, GNNs) or activations (e.g., piecewise-linear or softplus), as well as using architectural components (e.g., skip connections, normalization) that preserve stable gradient flows are key to achieving robust symbolic reasoning.

We then discuss limitations and potential generalizations of our approach in Section 5.

We note that our results should be viewed as an idealized foundation rather than a fully comprehensive model. Yet, this work provides the first theoretically grounded account of how continuous neural optimization can yield discrete, symbolic reasoning structures. Our results open new avenues for building neurosymbolic systems that leverage both the flexibility of neural representations and the rigor of symbolic reasoning.

2. Geometric and Algebraic Structure of the Parameter Space

In this section, we begin by exemplifying a reasoning task following Tian (2024). We then abstract and generalize this family of problems with algebraic, probabilistic, and geometric tools.

2.1. Algebraic Structures of Reasoning Tasks

Suppose we have a finite Abelian group $(A,\cdot)$ with commutative operation $\cdot$ and cardinality $n=|A|$. We aim to train a neural network for predicting the output of $a_1\cdot a_2$ for two group elements $a_1,a_2\in A$. The network input consists of the one-hot embeddings of these two group elements, represented as $e_{a_1},e_{a_2}\in\mathbb{R}^n$. The output aims to predict $a_1\cdot a_2$, also represented in one-hot encoding.

We analyze a two-layer neural network with q hidden nodes and quadratic neurons $\sigma(x)=x^2$:

$$o(a_1,a_2)=\frac{1}{q}\sum_{j=1}^{q}w_{cj}\sigma\left(w_{aj}^\top e_{a_1}+w_{bj}^\top e_{a_2}\right), \tag{1}$$

where weight matrices $W_a,W_b,W_c\in\mathbb{R}^{n\times q}$ encode inputs e_{a_1},e_{a_2} as hidden features and decode them to predictions, respectively. Different from the conventional formulation of neural networks, we normalize the final output by $1/q$ aligned with Mei et al. (2019).

Following Tian (2024), we consider representing and learning the weight matrices in their Fourier space $w_{aj}=\sum_{k\neq0}z_{akj}\phi_k, w_{bj}=\sum_{k\neq0}z_{bkj}\phi_k, w_{cj}=\sum_{k\neq0}z_{ckj}\overline{\phi_k},\forall j\in[q]$, where $\phi_k=[\phi_k(g)]_{g\in G}\in\mathbb{C}^n$ are the scaled Fourier basis functions ($0\le k<n$), and $z_{ak},z_{bk},z_{ck}\in\mathbb{C}$ are the Fourier coefficients. We further collectively organize these coefficients as matrices $z_j=z_{akj},z_{ckj},z_{ckj}]_{0\le k<n}\in\mathbb{C}^{3\times n}$ for $j\in[q]$. We adopt L_2-loss to optimize $\{z_j\}_{j\in[q]}$ over all possible compositions of group elements:

$$H(\{z_j\}_{j\in[q]})=\sum_{a_1,a_2\in A}\left\|P^\perp\left(\frac{1}{2n}o(a_1,a_2)-e_{a_1\cdot a_2}\right)\right\|^2, \tag{2}$$

where $P^{\perp} = I - \frac{1}{n}11^{\top}$ is the zero-mean projection operator.

We conclude this section by presenting the following proposition showing that the loss function can be reformulated as a combination of a special family of polynomials:

Proposition 1 *The loss function H in Eq. 2 can be reformulated as: $H = \frac{1}{n-1}\sum_{k\neq 0}\ell_k + \frac{n-1}{n}$,*

$$\ell_k = -2r_{kkk} + \sum_{k_1,k_2}|r_{k_1k_2k}|^2 + \frac{1}{4}\left|\sum_{p\in\{a,b\}}\sum_{k'} r_{p,k',-k',k}\right|^2 + \frac{1}{4}\sum_{m\neq 0}\sum_{p\in\{a,b\}}\left|\sum_{k'} r_{p,k',m-k',k}\right|^2,$$

$$r_{k_1k_2k} = \frac{1}{q}\sum_j z_{ak_1j}z_{bk_2j}z_{ckj}, \quad r_{pk_1k_2k} = \frac{1}{q}\sum_j z_{pk_1j}z_{pk_2j}z_{ckj}.$$

The proof can be found in the Appendix A. Similar for **all proofs** hereinafter.

2.2. Lifting to Measure Space

Next, we introduce some key mathematical devices that essentially generalize the algebraic properties of the reasoning task shown in Sec. 2.1. Let $d = 3n$ and $M \subset \mathbb{C}^d$ be a C^2-smooth, finite-dimensional manifold representing the parameter space of For theoretical clarity, we assume M is compact or otherwise restrict parameters to a large but bounded region, so that standard measure-theoretic and smoothness arguments apply. Let $P(M)$ denote the space of probability measures on M endowed with the W_2-Wasserstein metric and finite second moment (Villani et al., 2009). We define a polynomial function evaluated on the parameter space:

Definition 2 (Monomial Potentials, generalized from Tian (2024)) *Let $\dim(M) = d$ and fix local coordinates $z = (z_1, \ldots, z_d)$. A* monomial potential (MP) *r is a finite linear combination of monomials $r(z) = \sum_{I\in\mathcal{I}} c_I\, z_1^{i_1}\cdots z_d^{i_d}$, for some multi-index $I = (i_1, \cdots, i_d) \in \mathbb{N}^d$ in a finite index set $\mathcal{I} \subset \mathbb{N}^d$ and coefficients $c_I \in \mathbb{C}$. The collection of all such r forms a complex algebra $\mathcal{R}$ under pointwise addition and multiplication.*

We note that terms like $z_{ak_1j}\, z_{bk_2j}\, z_{ckj}$ in Proposition 1 are all special cases of MPs. Importantly, the loss H depends on parameters $\{z_j\}_{j\in[q]}$ only through the empirical distribution $\mu^{(q)} = \frac{1}{q}\sum_{j=1}^{q}\delta_{z_j}$, and any MP can be evaluated by taking its expectation against $\mu^{(q)}$. As $q \to \infty$, $\mu^{(q)}$ converges to a limiting measure μ in distribution. This suggests generalizing H to a functional $H[\mu]$ defined for all $\mu \in P(M)$, where we track the *expected* values of MPs: $\Phi_\mu(r) = \int_M r(z)\, d\mu(z)$. Such generalization also allows for moving beyond the simple example in Sec. 2.1.

Our goal is to study how μ evolves under a gradient flow that aims to minimize the functional $H[\mu]$, and how this evolution can induce algebraic factorization properties in these MPs. Specifically, let $\{\mu_t\}_{t\geq 0}$ be the Wasserstein gradient flow for H, satisfying:

$$\partial_t\mu_t + \nabla_z \cdot \left(\mu_t\, \nabla_z\left(\frac{\delta H}{\delta \mu}[\mu_t]\right)\right) = 0. \tag{3}$$

This idealized PDE-based viewpoint assumes certain smoothness and (often) displacement-convexity conditions on H; see Ambrosio et al. (2008); Villani et al. (2009) for details. The upshot is that for each infinitesimal time interval, μ_t moves in the steepest descent direction in W_2-space: $\mu_{t+\tau} \approx \arg\min_{\mu\in P(M)}\{H(\mu) + \frac{1}{2\eta_t\tau}W_2(\mu_t, \mu)\}$ for time-variant step size η_t.

We are interested in the following property of μ_t, which recovers the ring properties of MPs in the mean-field sense.

Definition 3 (MP-Compatible (Ring-Compatible) Measures) *A measure $\mu \in P(M)$ is* MP-compatible *if the map $\Phi_\mu : \mathcal{R} \to \mathbb{C}$ defined by $\Phi_\mu(r) = \int_M r(z)\, d\mu(z)$ is a ring homomorphism, i.e. $\Phi_\mu(r_1 \cdot r_2) = \Phi_\mu(r_1)\, \Phi_\mu(r_2)$ and $\Phi_\mu(1) = 1$.*

Exact MP-compatibility (for *all* monomials) is a strong property: preserving the multiplicative structure of $\mathcal{R}$ under integration: $\int (r_1 r_2)\, d\mu = \left(\int r_1\, d\mu\right)\left(\int r_2\, d\mu\right)$. This forces μ to be a sum of delta measures (or even a single delta in most typical settings). Since we rarely want or need factorization *for all* monomials, we introduce a *mild relaxation*:

Definition 4 (ε-Approximate Ring-Compatibility) *Let $\mathcal{R}_0 \subset \mathcal{R}$ be a chosen subfamily of monomial potentials (e.g. those appearing in the training loss). Fix a norm $\|\cdot\|_{\mathcal{R}_0}$. We say μ is ε-*ring-compatible on $\mathcal{R}_0$ *if $\left|\int (r_1 r_2)\, d\mu \;-\; \left(\int r_1\, d\mu\right)\left(\int r_2\, d\mu\right)\right| \le \varepsilon \|r_1\|_{\mathcal{R}_0} \|r_2\|_{\mathcal{R}_0}$ for all $r_1, r_2 \in \mathcal{R}_0$.*

Thus, if ε is small, integrals of those *specific* polynomials behave nearly like a ring homomorphism. We will see in Section 3 that dimension reduction can yield increasingly strong ε-ring-compatibility, even when exact factorization for *all* polynomials is impossible.

We discussed several choices of $\|\cdot\|_{\mathcal{R}_0}$ in the Appendix A.2. In all cases, the key idea is that $\|r\|_{\mathcal{R}_0}$ *controls the magnitude* of r so that a difference of the form $\left|\int r_1 r_2\, d\mu - \left(\int r_1\, d\mu\right)\left(\int r_2\, d\mu\right)\right|$ can be *relatively* bounded by $\varepsilon\, \|r_1\|_{\mathcal{R}_0}\, \|r_2\|_{\mathcal{R}_0}$. Thus, Definition 4 remains flexible, as one can pick whichever norm makes sense in their setting (compact manifold, polynomial expansions, etc.).

2.3. Symmetry Groups and Actions on Measures

Many neural architectures exhibit symmetries . We can model these under our framework by letting G be a group acting smoothly on M: which might represent architectural symmetries, such as permutations of hidden units. For each $g \in G$, we have a diffeomorphism $g : M \to M$.

Definition 5 (Induced Action on $P(M)$) *For $\mu \in P(M)$, define $(g\#\mu)(S) := \mu\big(g^{-1}(S)\big)$ for any measurable set $S \subseteq M$. Thus, G acts on $P(M)$ by pushforward. If μ is MP-compatible, then $\int_M (r \circ g)(z)\, d\mu(z) = \int_M r(z)\, d(g\#\mu)(z)$, which shows that G-actions and ring structures interact naturally via integration.*

Hence, if the loss function or data distribution is G-invariant, the training dynamics often preserve these symmetries. In particular, if μ_0 is G-invariant and H is also G-invariant, the evolution $\{\mu_t\}$ may remain in the G-invariant subspace of $P(M)$. Analyzing how μ_t evolves under G-actions can then clarify whether symmetry is maintained or broken during training.

3. Phase Transitions and Emergent Structure

In this section, we investigate how measures μ_t, evolving under the gradient flow framework established in Section 2, undergo *phase transitions* that lead to dimension reduction and the emergence of algebraic structures related to MPs.

3.1. Energy-Distance Convergence and Concentration

First, we define a *time-independent* manifold $\mathcal{M}_c \subset M$ by $\mathcal{M}_c = \left\{ z \in M \;\middle|\; \frac{\delta H}{\delta \mu}\big[\mu^*\big](z) = 0 \right\}$, where $\mu^* \in P(M)$ is a fixed reference measure (e.g., an equilibrium or a measure at which we

analyze local expansions) and $\frac{\delta H}{\delta \mu}[\mu^*](z)$ denotes the first variation of H evaluated w.r.t. μ^*. Note that we define $\mathcal{M}_c$ at a chosen measure μ^* for simplicity, yielding a "constant" submanifold of M on which tangential variations vanish. If μ_t evolves in time (e.g. under a gradient flow), the actual condition $\frac{\delta H}{\delta \mu}[\mu_t](z) = 0$ might change with t. However, for local *coercivity* or *energy-distance* arguments, we fix μ^* and study expansions around $\mathcal{M}_c$ as in Chizat and Bach (2018); Rotskoff et al. (2019). We further impose the following assumptions::

Assumption 1 (Critical Manifold Regularity) *There exists $r_0 > 0$ such that $\mathcal{M}_c$ is a C^2-smooth embedded submanifold of M with uniformly bounded second fundamental form, and $\operatorname{dist}(\cdot, \mathcal{M}_c)$ is C^2-smooth on the r_0-tubular neighborhood of $\mathcal{M}_c$. Moreover, the symmetry group G preserves $\mathcal{M}_c$ and this tubular neighborhood.*

Assumption 2 (Loss Regularity) *$H : P(M) \to \mathbb{R}$ is displacement-convex and C^2-smooth.*

By standard results in Riemannian geometry (Do Carmo and Flaherty Francis, 1992; Petersen, 2006), each z with $\operatorname{dist}(z, \mathcal{M}_c) < r_0$ can be written as $z = \exp_{\Pi(z)}\big(y(z)\big)$ for a unique $\Pi(z) \in \mathcal{M}_c$ and $y(z)$ in the normal bundle of $\mathcal{M}_c$. This decomposition is G-equivariant if G acts isometrically.

Now we analyze how deviations from $\mathcal{M}_c$ affect $H[\mu]$, obtaining a classical *coercivity* bound:

Theorem 6 (Energy-Distance Relation) *Suppose $\mu \in P(M)$ is supported in the r_0-tubular neighborhood of $\mathcal{M}_c$ (Assumption 1) and functional H satisfies Assumption 2. Then there exists $c_0 > 0$ such that $H[\mu] - H[\Pi\mu] \;\geq\; c_0 \int_M \operatorname{dist}^2\big(z, \mathcal{M}_c\big)\, d\mu(z)$, where where $(\Pi\mu)(S) = \mu(\Pi^{-1}(S))$ is the pushforward of μ by the projection $z \mapsto \Pi(z)$.*

As training evolves, the parameter distribution μ_t follows a gradient flow minimizing the functional H. In other words, μ_t is guided by the steepest descent direction in the Wasserstein space, continually adjusting itself to reduce $H[\mu_t]$. The energy-distance relation (Theorem 6) plays a crucial role here: it establishes that straying from the critical manifold $\mathcal{M}_c$ incurs a quadratic cost in terms of the functional H. Specifically, any mass of the measure μ_t that remains at a positive distance from $\mathcal{M}_c$ is penalized, thereby creating a strong incentive for μ_t to "push" its support closer to $\mathcal{M}_c$.

Next, we show that the empirical measures $\mu_t^{(q)}$ of q i.i.d. samples of μ_t approximate μ_t at the optimal statistical rate $q^{-1/2}$ and thus exhibit a similar convergence in the W_2 metric:

Theorem 7 (Concentration and Empirical Rates) *Under Assumption 1 and 2, assume also that M is compact so that μ_t has finite moments of order $p > 2$ in dimension $d < \infty$. Then there exist constants $C_1, C_2 > 0$ such that for any $\varepsilon > 0$, $W_2\big(\mu_t^{(q)}, \mu_t\big) < \varepsilon$ with probability at least $1 - C_1\, e^{-C_2\, q\, \varepsilon^2}$. Moreover, $q^{-1/2}$ is a minimax optimal rate for estimating μ_t in the W_2 metric.*

3.2. Phase Transitions and Dimension Reduction Reveal Algebraic Factorization

We now address a core emergent phenomenon: *phase transitions* at critical times t_k, leading to dimension reduction and the incremental emergence of algebraic constraints among monomial potentials (MPs). As μ_t collapses onto lower-dimensional submanifolds, its integrals of MPs exhibit increasingly factorizable behavior, revealing near ring-compatibility under certain conditions.

Flat Directions in the Hessian. Recall that we denote the Hessian operator of H at μ_t by $L(t) = -\frac{\delta^2 H}{\delta \mu^2}[\mu_t]$. Since $L(t)$ is self-adjoint with a discrete spectrum (under elliptic regularity and compactness assumptions), its eigenvalues $\{\lambda_j(t)\}$ vary continuously with t. A *flat direction* at time t is any nonzero vector $v \in T_{\mu_t} P(M)$ such that $L(t)v = 0$, i.e. $v \in \ker(L(t))$. An *approximately flat direction* refers to v whose associated eigenvalue is close to zero but not strictly zero.

Theorem 8 (Main Theorem - Existence of Critical Times and Dimension Reduction.) *Suppose M is compact and Assumption 2 holds. Then there are finitely many $0 < t_1 < t_2 < \cdots < t_N$ where an eigenvalue of $L(t)$ crosses zero. At each t_k, $\dim(\ker(L(t)))$ increases, forcing μ_t to concentrate on a strictly lower-dimensional submanifold S_k. Between t_k and t_{k+1}, the support dimension reduces further, and MP integrals exhibit* increasingly factorizable (ring-like) *structures.*

In essence, each time an eigenvalue of $L(t)$ crosses zero, it creates a newly available flat direction in the second variation of H. Along this direction, the measure μ_t can re-distribute mass without incurring second-order penalties, inducing a collapse onto a submanifold $S_k \subset M$ of lower dimension. Consequently, integrals of certain monomials become more closely factorized, approximating a ring-homomorphism property.

Dimension reduction effectively "removes" degrees of freedom in which monomials r_1, r_2 can vary jointly, so the integral $\int r_1 r_2 \, d\mu_t$ gets closer to $\left(\int r_1 \, d\mu_t\right)\left(\int r_2 \, d\mu_t\right)$ for $r_1, r_2 \in \mathcal{R}_0$. Hence, after each critical time t_k, we see *improved* ε-approximate factorization, culminating in the final minimal submanifold where no further kernel directions remain.

Exploration-Exploitation Transition and Algebraic Growth. We can have a fine-grained understanding of how μ_t transitions from *exploration* in higher dimensions to *exploitation* of kernel directions and lower-dimensional manifolds, ultimately uncovering richer polynomial constraints. Initially, μ_t "explores" a higher-dimensional region of M where no kernel directions are available. After t_1, it "exploits" the newly available *flat direction* (defined in §3.2), collapsing onto a lower-dimensional orbit. Each subsequent t_k further reduces dimension, introducing additional algebraic constraints detectable via MPs. Ultimately, μ_t will settle on a manifold S_N where no further dimension reduction can occur. In the presence of group symmetries G, the final manifold S_N will be a G-invariance orbit. We formalize this process in Appendix A.6 (stated as Theorem 11).

3.3. Interpreting Dimension Reduction Through the Lens of Factorization

In turn, dimension reduction can be also understood through the lens of MP factorization, elucidating why specific polynomial identities become "active" precisely at those times.

Polynomial Constraints and the Role of r-Coordinates. Recall that each MP $r \in \mathcal{R}$ is a polynomial-like function capturing relevant algebraic behavior of the network parameters. As μ_t localizes onto lower-dimensional manifolds, certain polynomial identities in $r_1, r_2, \ldots$ become effectively enforced. Formally, these identities manifest as $\nabla_r H = 0$ or $\int (r_1 r_2) d\mu_t \approx (\int r_1 \, d\mu_t)(\int r_2 \, d\mu_t)$ for the subset of r_i that dominate the loss H. While initially (before the critical time), $\nabla_r H$ may indicate that *small* polynomial adjustments can still reduce H, after the critical time, the newly "activated" identity forbids any further decrease in H via that polynomial direction.

Dimension Reduction as Satisfying Polynomial Identities. Each zero-eigenvalue crossing effectively "locks in" an algebraic constraint among the MPs $r_1, r_2, \ldots$. Concretely, once $\lambda_j(t_k) = 0$

introduces a kernel direction, any variation in r-space along that direction no longer decreases H. Geometrically, no second-order penalty arises, so μ_t reorients its mass to a manifold $S_k \subset M$ where the relevant polynomial identity (or identities) is *fully satisfied* (cf. Theorem 8, Step 3). As a result, the *effective dimension* of the support shrinks each time such an identity is enforced, removing the degrees of freedom that previously allowed H to vary.

In short, *dimension reduction emerges as the natural consequence of consecutively satisfying a growing set of polynomial constraints* in r-space. These constraints progressively reduce the dimensional subspace on which H can still decrease, thus aligning **geometry** (support dimension) and **algebra** (factorization).

Minimal Orbits. Eventually, μ_t settles on a *minimal G-invariant orbit*, i.e. a manifold $S_N \subset M$ to which no further dimension-lowering or complexity simplification applies. Interpreted through polynomial factorizations, this final state corresponds to having $\nabla_r H(r_i) = 0$ for all relevant monomial coordinates r_i, implying no polynomial direction remains that can further reduce H. Mapping these monomial directions *back* through $\partial r/\partial z$ yields no direction in parameter space $z \in M$ that lowers H. The minimal orbit thus represents a terminal *algebraic equilibrium* where all essential polynomial identities are activated, ensuring maximal *ring compatibility* (Def. 3) and G-invariance.

Hence, the dimension reduction and factorization perspectives coalesce: *once enough polynomial constraints lock in, the measure μ_t can no longer drift off that low-dimensional manifold without raising H*. This final configuration manifests *both* geometric minimality (S_N has the smallest dimension consistent with the constraints) *and* maximal algebraic factorization in the MPs.

4. Practical Implications for Neuralsymbolic AI System Design

Our measure-theoretic and dimension-reduction framework (Sections 2–3) provides a stable theoretical foundation for understanding neural network training dynamics, G-invariance, and approximate symbolic operations. In this section, we connect these insights to practical considerations in designing *neural-symbolic* AI systems. Two major themes emerge:

1. **Capacity Requirements:** Realizing ε-approximate symbolic operations (invariant under G) imposes fundamental lower bounds on network width.

2. **Architectural Constraints:** Preserving G-equivariance, ensuring well-structured manifolds at critical times, and maintaining ring-compatibility through appropriate activations or module designs shape how we choose layers, activation functions, and initialization procedures.

4.1. Minimum Width for Symbolic Reasoning

Dimension-reduction results from Section 3 suggest that achieving stable ε-approximate symbolic operations—those remaining consistent with G-invariance—requires capturing the underlying group symmetries in the network's representational capacity. Let G act on $\mathcal{X} \subset \mathbb{R}^d$ (as in Section 2.3), and consider tasks requiring a neural network to emulate G-invariant algebraic structures (e.g., transformations or polynomial identities) identified by the ring-factorization arguments of Section 3.2.

Theorem 9 (Informal, Minimal Width for Symbolic Reasoning) *Let $\mathcal{X} \subset \mathbb{R}^d$ be compact, and let $F_h : \mathcal{X} \to \mathcal{X}$ be a feedforward neural network of hidden width h (with sufficient smoothness to implement a G-invariant function), under mild assumption, any feedforward neural network F_h*

satisfying ε-approximate G-invariant operations with probability at least $1-\delta$ must have hidden width h satisfying

$$h \quad \geq \quad \begin{cases} \dfrac{\log\big(|G|\big)}{\log\big(1/\delta\big)}, & \text{if } G \text{ is finite}, \\ \dim(G), & \text{if } G \text{ is infinite or continuous}. \end{cases}$$

Discussion. We interpret Theorem 9 as follows:

- **Discrete G:** If G is finite, $\dim(G)$ is effectively zero. Then the statistical bound $\log|G|/\log(1/\delta)$ controls h. Thus, networks can learn discrete symbolic tasks (like group multiplication or boolean logic) with widths scaling only logarithmically in $|G|$, *provided* enough samples are available ($q \geq C_1 \log|G|/\log(1/\delta)$).
- **Continuous G:** If G is infinite or continuous, $\dim(G)$ is positive. The dimension-lowering arguments in Section 3.2 show that stabilizing a G-invariant structure requires representational capacity scaling at least linearly in $\dim(G)$. Hence tasks with high-dimensional continuous symmetries (e.g., large Lie groups) demand significantly larger network widths.

These scaling behaviors align with observed practice in neural-symbolic AI: discrete logic tasks rarely need large widths, whereas continuous/analog tasks (arithmetic, geometry) can require wide architectures. Grounding these requirements in a measure-theoretic, dimension-reduction argument clarifies why certain domains inherently demand more capacity.

4.2. Geometric Constraints on Architecture Design

We now derive concrete architectural guidelines for neural reasoning tasks. In particular, stable gradient flows, dimension reduction, G-invariance, and MP-based factorization collectively shape how one should choose layers, activations, and initialization to build robust *neural-symbolic* systems:

1. **Preserving G-Equivariance.** Many tasks require F_h to remain (approximately) G-equivariant, as discussed in Theorem 9 and Theorem 11. Architectures should incorporate structural symmetries if the group G is finite (e.g. permutations) or continuous (e.g. Lie group of dimension k). Failing to do so impedes dimension reduction in the measure flow, as no kernel direction can effectively map to G-invariant submanifolds.
2. **Maintaining a Well-Behaved Critical Manifold Structure.** Dimension reduction (Theorem 8) presupposes that the Hessian $\nabla^2 H$ does not produce pathological curvature or ill-defined tubular neighborhoods around critical manifolds. Architectural decisions — like ensuring Lipschitz continuity, bounding layer gradients, or mildly smooth activation functions — help ensure that $\mathcal{M}_c = \{\delta H/\delta r = 0\}$ is a smooth submanifold with finite curvature.
3. **Ensuring Ring Compatibility for Monomial Potentials.** As Section 3.2 shows, achieving factorization of MP integrals up to an ε-ring-compatibility standard (Definition 3) requires that the network *not* destroy polynomial interactions crucial to the symbolic task. Hence choosing activation functions or module designs that respect or approximate multiplicative structures is essential for stable symbolic reasoning.

We now illustrate these geometric principles with *representative architectures*:

a. Transformer Architectures Transformers can realize G-equivariance through their self-attention mechanism, provided attention patterns respect input symmetries. This design inherently suits the invariance preservation requirement from Section 3. Moreover, Theorem 9 prescribes how the network's embedding dimension d must scale in discrete or continuous G settings:

- For *finite* G, $d \gtrsim \log |G|$, matching the mild requirement from dimension reduction plus finite-sample uniform approximation arguments.
- For *continuous* G, $d \gtrsim \dim(G)$, reflecting the linear capacity growth needed to capture a continuous symmetry (McLeish et al., 2024; Wang et al., 2024).

Hence, carefully configured transformers can satisfy both invariance and approximate ring compatibility (with polynomial-friendly attention or feedforward layers). Their capacity matches theoretical predictions if the embedding dimension is scaled commensurately.

b. Graph Neural Networks (GNNs) GNNs naturally respect structural invariances, such as node permutations or isomorphisms, making them ideal for tasks with graph-based G actions (e.g. adjacency symmetries). By design, message-passing layers preserve G-equivariance. Additionally, local hierarchical structures can keep the second fundamental form bounded, easing stable dimension reduction. As a result, GNNs often satisfy the measure-based geometric requirements for tasks mapping onto graph symmetries or combinatorial reasoning.

c. RNNs and MLPs Standard recurrent networks (RNNs) and basic MLPs lack inherent G-equivariance or ring compatibility. Absent modifications, they may fail to exploit kernel directions discovered by dimension reduction. Nevertheless, for simpler tasks with small or trivial G, a minimal dimension is enough. In more complex settings, G-equivariant layers, linear constraints, or specialized activations might be introduced to preserve the group's structure.

We defer discussion on nonlinearity and normalization in Appendix B. We also discuss in Appendix B on **example architectures that fail to satisfy our theoretical findings.** Overall, the geometric perspective guides *why* specific architectural choices (e.g. group-convolution layers, polynomial activations, skip connections) are crucial for robust, G-invariant symbolic reasoning. By respecting the constraints enumerated above, one can design neural architectures that fully exploit the dimension-reduction phenomenon to achieve stable, scalable symbolic operations.

5. Conclusion and More Discussions

We have developed a measure-theoretic and geometric framework that explains how discrete symbolic capabilities can emerge in neural networks through continuous gradient-flow training. By rigorously analyzing how network parameters concentrate on low-dimensional, G-invariant manifolds, we demonstrated a natural pathway to approximate ring compatibility and the formation of symbolic-like algebraic structures. This viewpoint offers a unified explanation for the transition from high-dimensional "exploration" to low-dimensional "exploitation," clarifies why certain architecture choices facilitate stable dimension reduction, and provides principled width constraints tied to group symmetries These results establish concrete guidelines that enable robust symbolic reasoning in neural architectures. In short, the measure-based gradient flow perspective reveals that symbolic operations emerge naturally from the interplay of symmetries, algebraic constraints, and dimension reduction, culminating in a deeper theoretical foundation for neural-symbolic AI.

We have further outlined our **future research opportunities** in Appendix C, highlighting areas we aim to explore immediately should we receive a **DARPA Disruptive Idea Award**.

Acknowledgments

Z. Wang is supported by DAPRA ANSR (RTX CW2231110), DARPA TIAMAT (HR0011-24-9-0431), and ARL StAmant (W911NF-23-S-0001). We sincerely appreciate the insightful discussions from Elisenda Grigsby, Yuandong Tian, Kathryn Lindsey, Boris Hanin, and Alvaro Velasquez.

References

Luigi Ambrosio, Nicola Gigli, and Giuseppe Savaré. *Gradient flows: in metric spaces and in the space of probability measures*. Springer Science & Business Media, 2008.

Swarat Chaudhuri, Kevin Ellis, Oleksandr Polozov, Rishabh Singh, Armando Solar-Lezama, Yisong Yue, et al. Neurosymbolic programming. *Foundations and Trends® in Programming Languages*, 7(3):158–243, 2021.

Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

Manfredo Perdigao Do Carmo and J Flaherty Francis. *Riemannian geometry*. Springer, 1992.

Nicolas Fournier and Arnaud Guillin. On the rate of convergence in wasserstein distance of the empirical measure. *Probability theory and related fields*, 162(3):707–738, 2015.

Artur d'Avila Garcez and Luis C Lamb. Neurosymbolic ai: The 3 rd wave. *Artificial Intelligence Review*, 56(11):12387–12406, 2023.

Tosio Kato. *Perturbation theory for linear operators*. Springer Science & Business Media, 2013.

Sean McLeish, Arpit Bansal, Alex Stein, Neel Jain, John Kirchenbauer, Brian R Bartoldson, Bhavya Kailkhura, Abhinav Bhatele, Jonas Geiping, Avi Schwarzschild, et al. Transformers can do arithmetic with the right embeddings. *arXiv preprint arXiv:2405.17399*, 2024.

Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on learning theory*, pages 2388–2464. PMLR, 2019.

Amnon Pazy. *Semigroups of linear operators and applications to partial differential equations*, volume 44. Springer Science & Business Media, 2012.

P Petersen. *Riemannian geometry*. Springer-Verlarg, 2006.

Grant Rotskoff, Samy Jelassi, Joan Bruna, and Eric Vanden-Eijnden. Global convergence of neuron birth-death dynamics. *arXiv preprint arXiv:1902.01843*, 2019.

Yuandong Tian. Composing global optimizers to reasoning tasks via algebraic objects in neural nets. *arXiv preprint arXiv:2410.01779*, 2024.

Leslie G Valiant. A theory of the learnable. *Communications of the ACM*, 27(11):1134–1142, 1984.

Karthik Valmeekam, Matthew Marquez, Sarath Sreedharan, and Subbarao Kambhampati. On the planning abilities of large language models-a critical investigation. *Advances in Neural Information Processing Systems*, 36:75993–76005, 2023.

Cédric Villani et al. *Optimal transport: old and new*, volume 338. Springer, 2009.

Zhiwei Wang, Yunji Wang, Zhongwang Zhang, Zhangchen Zhou, Hui Jin, Tianyang Hu, Jiacheng Sun, Zhenguo Li, Yaoyu Zhang, and Zhi-Qin John Xu. Towards understanding how transformer perform multi-step reasoning with matching operation. *arXiv preprint arXiv:2405.15302*, 2024.

Jonathan Weed and Francis Bach. Sharp asymptotic and finite-sample rates of convergence of empirical measures in wasserstein distance. *arXiv preprint arXiv:1707.00087*, 2017.

Honghua Zhang, Liunian Harold Li, Tao Meng, Kai-Wei Chang, and Guy Van Den Broeck. On the paradox of learning to reason from data. In *Proceedings of the Thirty-Second International Joint Conference on Artificial Intelligence*, pages 3365–3373, 2023.

Appendix A. Proofs

A.1. Proof of Proposition 1

Proof Let $\tilde{\boldsymbol{w}}_{cj} = \frac{1}{q}\boldsymbol{w}_{cj}$, and define $\tilde{z}_{cjk}$ accordingly. Then by Theorem 1 of Tian (2024), the ℓ_k in the loss decomposition can be written as:

$$\ell_k = -2\tilde{r}_{kkk} + \sum_{k_1,k_2} |\tilde{r}_{k_1k_2k}|^2 + \frac{1}{4}\left|\sum_{p\in\{a,b\}}\sum_{k'}\tilde{r}_{p,k',-k',k}\right|^2 + \frac{1}{4}\sum_{m\neq 0}\sum_{p\in\{a,b\}}\left|\sum_{k'}\tilde{r}_{p,k',m-k',k}\right|^2$$

where

$$\tilde{r}_{k_1k_2k} = \sum_j z_{ak_1j}z_{bk_2j}\tilde{z}_{ckj}, \quad \tilde{r}_{pk_1k_2k} = \sum_j z_{pk_1j}z_{pk_2j}\tilde{z}_{ckj}.$$

We can conclude the proof by noticing that $\tilde{z}_{ckj} = \frac{1}{q}z_{ckj}$ by the linearity of Fourier transform. ■

A.2. Choice of $\|\cdot\|_{\mathcal{R}_0}$ in Definition 4

The norm $\|\cdot\|_{\mathcal{R}_0}$ is a device for measuring the "size" of monomial potentials in $\mathcal{R}_0$. Its exact specification depends on the setting:

- **Supremum norm on a compact manifold.** If M is compact and each $r \in \mathcal{R}_0$ is continuous on M, one natural choice is
$$\|r\|_{\mathcal{R}_0} = \sup_{z\in M}\big|r(z)\big|.$$
This allows us to bound integrals of r by $\|r\|_\infty$ directly, and is particularly convenient when proving uniform convergence statements.

- **Coefficient-based norm.** Alternatively, if each $r \in \mathcal{R}_0$ is a polynomial (or finite linear combination of monomials) with coordinate representation

$$r(z) \;=\; \sum_{\alpha} c_\alpha \, z^\alpha,$$

one may define

$$\|r\|_{\mathcal{R}_0} \;=\; \sum_{\alpha} \big|c_\alpha\big|, \quad \text{or equivalently} \quad \Big(\sum_{\alpha} |c_\alpha|^p\Big)^{1/p},$$

for some fixed p, ensuring that bounding coefficients suffices for bounding the polynomial's variation on restricted domains.

- **L^p-based norms.** In some measure-theoretic contexts, one might define

$$\|r\|_{\mathcal{R}_0} \;=\; \Big(\int_M |r(z)|^p \, d\nu(z)\Big)^{1/p}$$

for a reference measure ν, though this is less common for "exact vs. approximate ring-compatibility" arguments unless one expects certain integrability conditions.

A.3. Proof of Theorem 6

Proof Fix a reference measure $\mu^* \in P(M)$, and define

$$\mathcal{M}_c \;=\; \Big\{ z \in M : \tfrac{\delta H}{\delta \mu}[\mu^*](z) = 0 \Big\},$$

as a C^2-smooth submanifold by Assumption 1. We show that for any probability measure $\mu \in P(M)$ whose support lies in the r_0-tubular neighborhood of $\mathcal{M}_c$, there is a constant $c_0 > 0$ such that

$$H[\mu] \;-\; H[\Pi\mu] \;\geq\; c_0 \int_M \operatorname{dist}^2\big(z, \mathcal{M}_c\big) \, \mathrm{d}\mu(z),$$

where $\Pi(z)$ is the unique normal projection of z onto $\mathcal{M}_c$ within that tubular neighborhood, and $\Pi\mu$ is the pushforward measure defined by $S \mapsto \mu(\Pi^{-1}(S))$.

Local Expansions in z. While H is a functional $H : P(M) \to \mathbb{R}$, we may write $H[\delta_z]$ in a small neighborhood of $\mathcal{M}_c$ to indicate the local cost or integrand that H induces around each point $z \in M$. Such expansions are standard in PDE-based mean-field analyses (Chizat and Bach, 2018; Rotskoff et al., 2019), by expressing $H[\mu]$ as $\int(\int h(z) + \int g(z)g(z')d\mu(z'))d\mu(z)$.

Step 1: Second-Order Expansion Near $\mathcal{M}_c$. By definition of $\mathcal{M}_c$, every $z \in \mathcal{M}_c$ satisfies

$$\frac{\delta H}{\delta \mu}\big[\mu^*\big](z) = 0 \quad \text{(tangential derivative vanishes on } \mathcal{M}_c).$$

Since H is C^2-smooth and displacement-convex, its second variation in normal directions is strictly positive near $\mathcal{M}_c$. Concretely, if $z \in \mathcal{M}_c$ and $v \in N_{\delta_z} P(M)$ is a normal vector at δ_z, then the second derivative in that direction is strictly positive. By the positive semi-definiteness property, there is a uniform $\alpha > 0$ such that

$$\Big\langle v, \; \frac{\delta^2 H}{\delta \mu^2}[\delta_z] v \Big\rangle \;\geq\; \alpha \, \|v\|^2,$$

for all normal v at $z \in \mathcal{M}_c$. This implies a quadratic penalty for deviations from $\mathcal{M}_c$.

Step 2: Tubular Neighborhood and Normal Coordinates. By Assumption 1, for each z with $\mathrm{dist}(z, \mathcal{M}_c) < r_0$, we can write

$$z = \exp_{\Pi(z)}\big(y(z)\big),$$

where $\Pi(z) \in \mathcal{M}_c$ is the closest point (normal exponential), and $y(z)$ lies in the normal bundle. Hence $\|y(z)\| = \mathrm{dist}\big(z, \mathcal{M}_c\big)$.

Using these coordinates, we do a second-order expansion of $H[\delta_z]$ around $\Pi(z)$. Because $\frac{\delta H}{\delta \mu}\big[\mu^*\big](\Pi(z)) = 0$ in tangential directions, and H has strict convexity in normal directions, there is $c_0 > 0$ (depending on α and curvature) such that

$$H[\delta_z] \geq H[\delta_{\Pi(z)}] + c_0 \, \|y(z)\|^2.$$

The derivation follows from: (1) first-order tangential derivative at $\delta_{\Pi(z)}$ is zero, because $\Pi(z) \in \mathcal{M}_c$; (2) second-order convexity ensures a strictly positive quadratic form in normal directions; and (3) geodesic or normal coordinate argument $|\delta_z - \delta_{\Pi(z)}| \sim |z - \Pi(z)|$ ensures $|y(z)|^2$ arises.

Step 3: Summing Over the Measure μ. Integrating over z in $\mathrm{supp}(\mu)$,

$$\int_M H[\delta_z] \, \mathrm{d}\mu(z) \geq \int_M \big(H[\delta_{\Pi(z)}] + c_0 \|y(z)\|^2\big) \, d\mu(z).$$

Define $(\Pi\mu)(S) = \mu(\Pi^{-1}(S))$. By a change of variables,

$$\int_M H[\delta_z] d\mu(z) - \int_M H[\delta_z] d\Pi\mu(z) \geq c_0 \int_M \|y(z)\|^2 d\mu(z).$$

If we identify $H[\Pi\mu]$ with $\int_M H[z] \, \mathrm{d}\Pi\mu(z)$ consistently at $\mathcal{M}_c$, we conclude

$$H[\mu] - H[\Pi\mu] \geq c_0 \int_M \mathrm{dist}^2\,(z, \mathcal{M}_c) \, d\mu(z).$$

Step 4: Conclusion. Since this inequality holds for all μ supported in that r_0-tubular neighborhood, Theorem 6 follows immediately.

■

Remark 10 *The submanifold $\mathcal{M}_c$ is defined w.r.t. a fixed μ^*, making it time-independent. Even if μ_t evolves in time, expansions around $\mathcal{M}_c$ still yield a coercivity-like bound: any measure μ that leaves $\mathcal{M}_c$ in the normal direction must pay a quadratic cost. This idea parallels standard Euclidean "coercivity" lemmas and underpins dimension-reduction arguments in mean-field PDE analyses (Chizat and Bach, 2018; Rotskoff et al., 2019).*

A.4. Proof of Theorem 7

Proof The goal is twofold:

1. Show the empirical measure $\mu_t^{(q)} = \frac{1}{q} \sum_{j=1}^{q} \delta_{X_j}$, where X_j are i.i.d. samples from μ_t, converges to μ_t in W_2 with high probability:

$$\Pr\big(W_2(\mu_t^{(q)}, \mu_t) > \varepsilon\big) \;\leq\; C_1 \, \exp\big(-C_2 \, q \, \varepsilon^2\big).$$

2. Prove that no estimator can do better than $q^{-1/2}$ in W_2 (minimax lower bound).

The result itself — that the empirical measure obtains a sub-Gaussian tail in W_2, and $q^{-1/2}$ is the minimax rate — is not new but a known fundamental statement from modern measure concentration and optimal transport theory. However, the application to a mean-field neural network measure is interesting and essential to our paper's theoretical framework.

Step 1: Finite-Sample Deviation for the Empirical Measure. Because M is compact, μ_t has finite moments of all orders, and $d = \dim(M)$ is finite. Then by classical non-asymptotic results on empirical measures in Wasserstein distance (Fournier and Guillin, 2015, Theorem 1), there exist $C_1, C_2 > 0$ such that for all $\varepsilon > 0$,

$$\Pr\big(W_2(\mu_t^{(q)}, \mu_t) > \varepsilon\big) \;\le\; C_1 \exp\big(-C_2\, q\, \varepsilon^2\big).$$

The compactness plus bounded dimension ensure the required finite-moment conditions for $p > 2$.

Step 2: Minimax Optimality (Weed–Bach). Additionally, Weed and Bach (2017) show that on a compact (or suitably bounded) metric space, no estimator can uniformly achieve a W_2 convergence rate faster than $q^{-1/2}$. Formally, for any estimator $\widehat{\mu}$,

$$\inf_{\widehat{\mu}} \sup_{\mu_t \in \mathcal{P}} \mathbb{E}\big[W_2(\widehat{\mu}, \mu_t)\big] \;\ge\; C\, q^{-1/2},$$

for some constant $C > 0$ and some problem class $\mathcal{P}$. Since the empirical measure $\mu_t^{(q)}$ already achieves an upper bound of order $q^{-1/2}$, we conclude $q^{-1/2}$ is the minimax-optimal rate.

Conclusion. Combining Step 1 and Step 2, Theorem 7 follows:

- We have an exponential tail bound $\Pr[W_2(\mu_t^{(q)}, \mu_t) > \varepsilon] \le C_1 \exp(-C_2\, q\, \varepsilon^2)$.
- No estimator can beat $q^{-1/2}$ in a minimax sense, so $q^{-1/2}$ is optimal.

■

A.5. Proof of Theorem 8

Proof We prove the following claims:

1. Only finitely many times t_k exist where an eigenvalue of $L(t)$ crosses zero within any finite interval $[0, T]$.

2. At each such time t_k, the measure μ_t localizes onto a strictly lower-dimensional submanifold $S_k \subset M$, thereby enhancing the factorization properties of integrals of monomial potentials (MPs).

Step 1: Spectral Properties of $L(t)$. *(a) Self-adjointness and Compact Resolvent.* Since H is C^2-smooth, displacement-convex, and defined on a compact manifold M, its Hessian operator at μ_t, under standard elliptic and geometric assumptions,

$$L(t) \;=\; -\frac{\delta^2 H}{\delta r^2}\big[\mu_t\big],$$

is self-adjoint on an appropriate Hilbert space, e.g. $L^2(\mu_t)$. By classical arguments in semigroup theory and spectral analysis (Pazy, 2012), $L(t)$ then admits a compact resolvent. Consequently, the eigenvalues $\{\lambda_j(t)\}_{j=1}^{\infty}$ of $L(t)$ form a discrete real spectrum tending to $\pm\infty$.

(b) Continuous Dependence on t. The measure μ_t evolves via a metric gradient flow in the W_2-Wasserstein space (Ambrosio et al., 2008), ensuring $t \mapsto \mu_t$ is continuous in t. As $L(t)$ depends continuously on μ_t, Kato's perturbation theory (Kato, 2013) implies each eigenvalue $\lambda_j(t)$ is itself a continuous function of t.

Hence, for each fixed index j, the map $t \mapsto \lambda_j(t)$ is continuous, and the spectrum of $L(t)$ is discrete for each t.

Step 2: Finite Number of Zero-Crossings on $[0, T]$. Define a *critical time* t_k if some eigenvalue $\lambda_j(t)$ crosses zero at t_k, i.e. $\lambda_j(t_k) = 0$ and $\lambda_j(t)$ changes sign in a neighborhood of t_k. Suppose, for contradiction, that infinitely many zero-crossings occurred within a finite interval $[0, T]$.

Since $\lambda_j(t)$ is continuous[1] in t by Step 1(b), any eigenvalue cannot "cross zero" an unbounded number of times without either (i) remaining identically zero on a sub-interval of $[0, T]$ (implying no strict sign change) or (ii) oscillating with infinitely many roots that accumulate at a finite point in time, contradicting the properties of a non-constant continuous function. Thus, for each j, only finitely many zero-crossings can occur on $[0, T]$. Moreover, displacement-convexity and smoothness of H rule out degenerate behaviors such as repeated instantaneous vanishings of the same eigenvalue. Hence each eigenvalue $\lambda_j(t)$ can cross zero at most finitely many times on $[0, T]$.

By covering $[0, \infty)$ by disjoint intervals $[nT, (n+1)T]$, we conclude that there are finitely many crossing times in each finite sub-interval. Collecting those times (if any) in ascending order gives $0 < t_1 < t_2 < \cdots < t_N < \cdots$ with $N < \infty$ on each finite interval.

Step 3: Dimension Reduction via Energy-Distance. We now show how a zero eigenvalue at time t_k forces μ_t to concentrate on a strictly lower-dimensional submanifold $S_k \subset M$, while noting that no crossing may occur at all.

(a) Zero-Crossing & Kernel Direction in Measure Space. Suppose an eigenvalue $\lambda_j(t)$ indeed crosses zero at time t_k, giving $\lambda_j(t_k) = 0$ with $v \in \ker(L(t_k))$ a nonzero kernel vector in $T_{\mu_{t_k}} P(M)$. By definition, $L(t_k)v = 0$ indicates no second-order cost for variations of μ_{t_k} along v. If *no* eigenvalue crosses zero on $[0, T]$, no dimension reduction occurs, yet Theorem 8 still holds in that sign changes cannot accumulate infinitely.

(b) Defining S_k in M via a Parameterization Map. To interpret v as a "flat direction" in the parameter manifold M (rather than producing delta measures), we note that in a mean-field or

1. If one allows arbitrary continuous (non-analytic) dependence on t, we cannot automatically exclude infinitely many zero-crossings. In PDE/spectral settings, real-analytic or at least sufficiently regular dependence typically holds (Kato, 2013), ensuring each eigenvalue is real-analytic and cannot exhibit infinitely many sign changes unless identically zero. We adopt or reference such regularity assumptions to exclude pathological oscillations.

infinite-width setting, *each measure* $\mu \in P(M)$ typically emerges from a higher-level parameter $\theta \in \Theta$ (or from an entire family of local coordinates). Formally, there is a map

$$\Phi : \Theta \ \to \ P(M), \quad \text{with } \mu_\theta = \Phi(\theta),$$

describing how each θ (often representing an entire distribution of neurons or sub-parameters) induces a measure $\mu_\theta \in P(M)$, which is a full distribution over M. Thus, if at $\mu_{t_k} = \mu_{\theta_{t_k}}$ we have $v = d\Phi(\theta_{t_k})(w)$ for some $w \in T_{\theta_{t_k}}\Theta$, we see that second variations vanish along w. Translating w back to local coordinates in M yields a submanifold $S_k \subset M$ of strictly lower dimension, on which these second-order variations remain zero (the "flat direction"). Formally, we define S_k as:

$$S_k = \bigcup_{\mu \in \mathcal{U}} \operatorname{supp}(\mu), \quad \mathcal{U} = \left\{ \mu \in P(M) : \frac{\delta H}{\delta \mu}[\mu] \in \ker(L(t_k)) \right\},$$

$|\mathcal{U}| = 1$ if H is strongly convex.

(c) Strict Dimension Drop: $\dim(S_k) < \dim(S_{k-1})$**.** Because v must be linearly independent of previously discovered kernel directions, S_k is contained in but strictly lower-dimensional than S_{k-1}. Concretely, each new zero-crossing reduces the rank of the Hessian, forcing the measure to confine itself to fewer degrees of freedom. Thus

$$\dim(S_k) \ < \ \dim(S_{k-1}),$$

and if multiple eigenvalues cross simultaneously, an even larger dimension drop can occur.

(d) Localization of μ_t **onto** S_k**.** By the energy-distance relation (Theorem 6), any mass of μ_t lying away from S_k incurs a positive second-order penalty. Since μ_t follows a steepest descent flow, it continuously "moves" mass toward S_k for $t > t_k$. We do *not* claim an instantaneous jump at t_k; instead, one typically shows

$$\int_M \operatorname{dist}\big(z, S_k\big)^2 \, d\mu_t(z) \ \to \ 0 \quad \text{as } t \downarrow t_k + \epsilon,$$

meaning μ_t localizes near S_k after a short time. Hence from $t > t_k + \delta$ onward, $\operatorname{supp}(\mu_t)$ is effectively in S_k, enforcing dimension reduction in μ_t's support.

(e) If No Zero-Crossing, No Dimension Reduction. If no $\lambda_j(t)$ crosses zero, no S_k is defined, and no dimension-lowering arises. The theorem's statement that infinitely many sign changes cannot appear remains true, but no crossing is mandated.

Step 3 thus confirms: (1) zero-crossings are *finite* in number, and (2) any actual crossing yields a strict dimension drop by localizing μ_t on a lower-dimensional S_k. We next show (Step 4) how factorization of monomial potentials is enhanced at each dimensional reduction.

Step 4: Approximate Factorization of Monomial Potentials. Once μ_t is predominantly (or entirely) supported on a submanifold $S_k \subset M$ of dimension $d_k < \dim(M)$, we claim that for $r_1, r_2 \in \mathcal{R}_0$, the difference

$$\int (r_1 r_2) \, d\mu_t \ - \ \Big(\int r_1 \, d\mu_t \Big) \Big(\int r_2 \, d\mu_t \Big)$$

can be made arbitrarily small (up to a factor $\|r_1\|_{\mathcal{R}_0}\,\|r_2\|_{\mathcal{R}_0}$), thus implying ε-ring-compatibility. Specifically, bounding such products on a manifold of dimension d_k ensures near factorization.

The concrete bounding arguments are standard for polynomial factorization. To be self-contained, we outline a short bounding approach step by step below.

(a) Reduction to a Covariance-Like Quantity. Define the function

$$F_{r_1,r_2}(z) \;=\; r_1(z)\,r_2(z) \;-\; \Big(\int r_1\,d\mu_t\Big)\,r_2(z) \;-\; \Big(\int r_2\,d\mu_t\Big)\,r_1(z) \;+\; \Big(\int r_1\,d\mu_t\Big)\Big(\int r_2\,d\mu_t\Big).$$

One easily checks that

$$\int_M F_{r_1,r_2}\,d\mu_t \;=\; \int (r_1 r_2)\,d\mu_t \;-\; \Big(\int r_1\,d\mu_t\Big)\Big(\int r_2\,d\mu_t\Big).$$

Hence controlling $\big|\int F_{r_1,r_2} d\mu_t\big|$ is precisely controlling how closely $\int (r_1 r_2)$ factorizes as $(\int r_1)(\int r_2)$.

(b) Bounding F_{r_1,r_2} on a Lower-Dimensional Manifold. Since $r_1, r_2 \in \mathcal{R}_0$ are (finite) linear combinations of monomials of bounded degree, their product $r_1 r_2$ is also a polynomial of bounded total degree. Consequently, $F_{r_1,r_2}(z)$ is again a polynomial in z (albeit with some constant shifts).

Now, $S_k \subset M$ is of dimension $d_k < d$. There are two main ways to show that $F_{r_1,r_2}(z)$ cannot vary too much on S_k:

1. **Uniform Bound on a Compact Manifold.** If S_k is a *compact* (or at least closed and bounded) submanifold of dimension d_k, then F_{r_1,r_2} is a continuous function on S_k. Hence

   $$\sup_{z\in S_k} \big|F_{r_1,r_2}(z)\big| \;\le\; C(\mathcal{R}_0, S_k)\;\|r_1\|_{\mathcal{R}_0}\,\|r_2\|_{\mathcal{R}_0}$$

   for some constant $C(\mathcal{R}_0, S_k)$ that depends on (i) the polynomial degrees in $\mathcal{R}_0$ and (ii) the geometry of S_k. Thus

   $$\Big|\int_M F_{r_1,r_2}(z)\,d\mu_t(z)\Big| \;\le\; \sup_{z\in S_k}\big|F_{r_1,r_2}(z)\big| \;\approx\; \varepsilon\,\|r_1\|_{\mathcal{R}_0}\,\|r_2\|_{\mathcal{R}_0},$$

   provided that μ_t is predominantly supported on S_k and $C(\mathcal{R}_0, S_k)$ can be made small or is finite while $\varepsilon > 0$ captures the desired approximation level.

2. **Variance / Covariance Argument.** One can interpret $\int F_{r_1,r_2} d\mu_t$ as something akin to $\mathrm{Cov}(r_1, r_2; \mu_t)$ plus a constant shift. If μ_t is restricted to S_k, the dimension d_k can limit the "joint variation" of $(r_1(z), r_2(z))$. Formally, bounding second moments on S_k or employing dimension-based constraints on polynomials can show that $\mathrm{Cov}(r_1, r_2; \mu_t)$ is forced below any positive $\varepsilon \cdot \|r_1\|\|r_2\|$. Hence again

   $$\Big|\int (r_1 r_2)\,d\mu_t \;-\; \int r_1\,d\mu_t \int r_2\,d\mu_t\Big| \;\le\; \varepsilon\,\|r_1\|_{\mathcal{R}_0}\,\|r_2\|_{\mathcal{R}_0}.$$

Thus in both approaches, the key fact is that restricting μ_t to a manifold of dimension d_k below d precludes large "cross-terms," ensuring factorization up to a small $\varepsilon > 0$.

Since we achieve an upper bound of the form

$$\Big|\int (r_1 r_2)\,d\mu_t \;-\; \big(\int r_1\,d\mu_t\big)\,\big(\int r_2\,d\mu_t\big)\Big| \;\;\le\;\; \varepsilon\,\|r_1\|_{\mathcal{R}_0}\,\|r_2\|_{\mathcal{R}_0}$$

for all $r_1, r_2 \in \mathcal{R}_0$, it follows that μ_t is ε-ring-compatible on $\mathcal{R}_0$. Hence dimension reduction at time t_k improves the factorization properties of μ_t, completing the argument for Step 4.

Summary. Combining these steps, we conclude:

- Zero-crossings of $L(t)$'s eigenvalues can only occur finitely many times in any finite interval $[0, T]$.
- Each zero-crossing lowers dimension by producing an additional kernel direction, forcing μ_t onto a strictly lower-dimensional manifold $S_k \subset M$.
- Finally, dimension reduction entails enhanced approximate factorization of MP integrals, completing the proof of Theorem 8.

■

A.6. Formal Statement and Proof of Theorem 11

Theorem 11 (Exploration–Exploitation Characterization) *Let M be a compact Riemannian manifold and H satisfies Assumption 2. Assume also that a group G acts smoothly on M (Definition 5), and H, μ_t remain G-invariant (no spontaneous symmetry-breaking).*

Then, there exists a finite set of "critical times" $0 < t_1 < t_2 < \cdots < t_N$, each corresponding to a zero eigenvalue crossing of the Hessian operator $L(t)$, such that:

1. *For $t < t_1$, μ_t has no nontrivial kernel direction in $L(t)$ and thus stays in a higher-dimensional region of M, "exploring" without collapsing.*
2. *At t_1, a zero eigenvalue emerges, creating a kernel direction in $L(t_1)$ which forces μ_t to localize on a lower-dimensional submanifold $S_1 \subset M$. This reduces the "effective dimension" of its support and increases algebraic factorization for monomial potentials.*
3. *Repeating at each t_k ($k = 2, \ldots, N$), μ_t aligns with an even lower-dimensional manifold $S_k \subset S_{k-1}$ after a new zero-crossing appears.*
4. *Ultimately, no further dimension reduction occurs; μ_t settles on a final G-invariant manifold (or orbit) S_N that cannot be simplified further, exhibiting maximal algebraic (ring-like) factorization among relevant monomial potentials.*

Proof We prove the four statements in turn, highlighting how dimension reduction ensues from zero-eigenvalue crossings and why G-invariance endures throughout.

Step 1: Initial Phase ($t < t_1$): Full-Dimensional Exploration. By Theorem 8, an eigenvalue of the Hessian operator $L(t) = -\delta^2 H/\delta r^2[\mu_t]$ crosses zero only finitely many times, and none occurs before t_1. Hence for $0 \leq t < t_1$, no eigenvalue $\lambda_j(t)$ is zero, implying $L(t)$ is strictly non-degenerate in all directions.

Thus, there is no kernel direction ("flat direction") in the second variation of H at μ_t. Any attempt to confine μ_t to a strictly lower-dimensional manifold in M would incur a positive second-order penalty, preventing measure collapse. Therefore, μ_t remains relatively extended in a higher-dimensional region of M during $t < t_1$. Moreover, if H and μ_0 are G-invariant, displacement-convexity ensures uniqueness of the gradient flow solution, implying μ_t stays G-invariant for $t < t_1$. No external impetus exists to break the group symmetry.

Step 2: Emergence of a Zero Eigenvalue at t_1: First Exploitation. At time t_1, Theorem 8 indicates an eigenvalue $\lambda_j(t)$ crosses zero for the first time, creating a nontrivial kernel direction $v \in \ker\big(L(t_1)\big)$. By definition, $L(t_1)\, v = 0$ implies no second-order growth of H in that direction, enabling μ_t to "rearrange" its mass onto a lower-dimensional submanifold $S_1 \subset M$ associated with v (see the dimension-reduction argument of Theorem 8, Step 3). The measure thus reduces $H[\mu_t]$ by confining its support to S_1, lowering the effective dimensionality of its support.

Since G-invariance is assumed, we may choose S_1 to lie within a G-invariant orbit or to be itself G-invariant, thus preserving the group action throughout. Hence the time t_1 marks a transition from a fully high-dimensional "exploratory" regime to an "exploitive" regime, where μ_t localizes on S_1. MP integrals now exhibit partial factorization due to dimension reduction (see Step 4 in Theorem 8 on approximate ring-compatibility).

Step 3: Iterative Crossings and Cumulative Dimension Reduction. Each subsequent time t_k (for $k = 2, \ldots, N$) similarly corresponds to another zero-crossing of $\lambda_j(t)$, providing a fresh kernel direction. Repeating the dimension-reduction argument yields a chain of submanifolds

$$S_1 \supset S_2 \supset \cdots \supset S_k \supset \cdots,$$

each strictly lower-dimensional than its predecessor. Because H remains G-invariant and the gradient flow is unique, μ_t must likewise remain G-invariant. No spurious oscillations or jumps occur, as displacement-convex gradient flows do not spontaneously revert dimension. Meanwhile, MP integrals factorize more closely at each step, reflecting the diminished degrees of freedom in M on which μ_t has nonnegligible mass.

Step 4: Stabilization onto a Minimal G-Invariant Orbit. Since Theorem 8 asserts only finitely many zero-crossings can occur, at time t_N the last such crossing takes place. No further kernel directions are introduced afterward; thus μ_t stabilizes onto a final submanifold $S_N \subset M$ of dimension at most $d' < d$. By G-invariance, S_N is (or can be chosen to be) G-invariant. No additional dimension-lowering is possible without increasing H, so S_N is a "minimal G-invariant orbit."

In this final configuration, MP integrals

$$\int_M r_1(z) r_2(z)\, d\mu_t(z)$$

factorize up to an ε-ring-compatibility degree (Definition 4), reflecting a "maximal algebraic complexity" that emerges once no further dimension-lowering is available.

Conclusion. Hence we obtain the *exploration–exploitation* characterization:

- **Exploration** ($t < t_1$): no kernel directions, thus no dimension-lowering.
- **Exploitation** ($t > t_1$): each zero eigenvalue crossing reduces dimension, leading to more "algebraic structure" (factorization) in MP integrals.
- **Final Stabilization**: a minimal G-invariant submanifold (or orbit) S_N of M with no further dimension-lowering possible, attaining *maximal* factorization properties under G-symmetry.

This completes the proof of Theorem 11. ■

A.7. Formal Statement and Proof of Theorem 9

Theorem 12 (Minimal Width for Symbolic Reasoning) *Let $\mathcal{X} \subset \mathbb{R}^d$ be compact, and let $F_h : \mathcal{X} \to \mathcal{X}$ be a feedforward neural network of hidden width h (with sufficient smoothness to implement a G-invariant function). Suppose:*

- *μ_t emerges from a measure-based gradient flow that is G-invariant under displacement-convexity, and localizes onto a lower-dimensional submanifold as per Theorem 11.*
- *For each $x \in \mathcal{X}$ and $g \in G$, F_h is required to be ε-approximate G-equivariant, i.e.$\| F_h(g \cdot x) - g \cdot F_h(x)\| \leq \varepsilon$, so that F_h captures the symbolic operations stable under group action.*
- *When G is finite of cardinality $|G|$, we also assume an available sample size $q \geq C_1 \log\big(|G|/\delta\big)$ ensures uniform approximation across the $|G|$ transformations with probability at least $1-\delta$.*
- *When G is infinite (or a continuous Lie group), let $\dim(G)$ be its dimension. The dimension-reduction arguments (Theorem 8, 11) imply that capturing G-invariant orbits requires at least $\dim(G)$ degrees of freedom in the final submanifold.*

Then any feedforward neural network F_h satisfying these ε-approximate G-invariant operations with probability at least $1-\delta$ must have hidden width h satisfying

$$h \quad \geq \quad \begin{cases} \dfrac{\log\big(|G|\big)}{\log\big(1/\delta\big)}, & \text{if } G \text{ is finite}, \\ \dim(G), & \text{if } G \text{ is infinite or continuous}. \end{cases}$$

Proof Step 1 (Continuous G): Representation Constraint from Dimension Reduction. Assume G is an infinite (continuous) group with $\dim(G) > 0$. By Theorem 11, once μ_t localizes on a minimal G-invariant orbit $\mathcal{O} \subseteq M$, that orbit has dimension at most $\dim(G)$. Any neural network F_h that expresses a stable G-invariant symbolic operation must, at minimum, be capable of resolving all degrees of freedom inherent in $\mathcal{O}$.

Concretely, capturing a $\dim(G)$-dimensional continuous symmetry calls for at least $\dim(G)$ "directions" in function space for implementing the group action. In feedforward architectures, the effective dimension of the hypothesis class is typically bounded by a polynomial in h, but to *exactly* encode a group of dimension $\dim(G)$, a linear scaling $h \geq \dim(G)$ is *necessary* in typical universal-approximation arguments (cf. McLeish et al., 2024; Wang et al., 2024).

If $h < \dim(G)$, the network lacks the capacity to represent a $\dim(G)$-dimensional family of G-equivariant transformations stably. Hence for continuous G, we obtain

$$h \quad \geq \quad \dim(G).$$

Step 2 (Finite G): Statistical Complexity and Uniform Approximation. Now assume G is a finite group of cardinality $|G| < \infty$. To remain ε-approximate G-invariant on *all* $|G|$ transformations with probability $\geq 1-\delta$, standard PAC or uniform-convergence theory states we need at least $\log(|G|/\delta)$ bits (or "units of capacity") in the hypothesis class (Valiant, 1984). Specifically, with $q \geq C_1 \log(|G|/\delta)$ i.i.d. training samples, to maintain uniform error $\leq \varepsilon$ across $|G|$ transformations w.p. $\geq 1-\delta$, the network's expressive power must be at least $\Theta(\log|G|/\log(1/\delta))$.

In a feedforward neural network, the width h provides a principal bottleneck on capacity. Typically, if $h < (\log |G|)/(\log(1/\delta))$, no architecture can guarantee ε-approximate invariance across *all* $|G|$ transformations. Thus we get

$$h \geq \frac{\log |G|}{\log(1/\delta)}.$$

Step 3: Combining Results and Concluding the Lower Bound.
For a discrete group G, $\dim(G)$ is effectively zero, so the representational constraint from Step 1 is trivial. The statistical bound from Step 2 then dominates:

$$h \geq \frac{\log |G|}{\log(1/\delta)}.$$

For an infinite or continuous group G, Step 2 becomes irrelevant, whereas Step 1's dimension-limiting argument imposes the stronger requirement $h \geq \dim(G)$.

Consequently, we combine these two scenarios into the piecewise expression

$$h \geq \begin{cases} \dfrac{\log |G|}{\log(1/\delta)} & \text{if } |G| < \infty, \\ \dim(G) & \text{if } |G| = \infty. \end{cases}$$

This completes the proof. ■

Remark 13 (Remark on Necessity vs. Sufficiency.) *Our argument shows a* necessary *condition on h: even if one had an ideal optimization procedure or infinite training time, a network with width h below these thresholds* cannot *systematically realize the required G-invariant symbolic mappings. Of course, reaching such a solution in practice may require additional inductive biases, but the established bound clarifies that no optimization method can circumvent this fundamental representational limit.*

Appendix B. Further Discussion on Architectural Design

Activation Functions and Ring Compatibility Maintaining ring compatibility is critical: polynomial factorization arguments (Section 3.2) show how multiplicative structures can degrade if activations over-mix terms. *Discrete tasks* often tolerate piecewise-linear activations like ReLU, which do not overly distort polynomial relations. *Continuous tasks* benefit from smoother or polynomial-like activations (e.g. softplus) that help retain essential monomial interactions for dimension-reduction arguments on continuous G. Strong nonlinearities such as sigmoid, tanh, Swish, or GeLU can hinder ring factorization by breaking multiplicative consistency. These theoretical considerations clarify empirical findings that certain activations better suit symbolic tasks despite similar performance on conventional benchmarks.

Stabilization Components A bounded second fundamental form around critical manifolds (Theorem 6) requires that gradient flows remain stable. Techniques like skip connections (residual blocks) and normalization layers (BN, LN) help regulate Hessian curvature in parameter space. From a dimension-reduction perspective, these are not mere engineering heuristics but *theoretical enablers* ensuring a well-defined tubular neighborhood around critical submanifolds and facilitating measure localization onto lower-dimensional sets.

Anti-Patterns in Architectural Design Certain design decisions directly clash with our theoretical lens:

- *Breaking G-equivariance:* Inserting operations that fundamentally violate the group's symmetry (without compensations) obstructs the dimension-reduction synergy from Section 3.
- *Deep or brittle flows:* Excessively deep networks lacking proper stabilization risk undefined or highly curved Hessians, impeding the "flat" directions that drive dimension-lowering.
- *Violations of ring compatibility:* Combining neural modules (e.g. attention) with symbolic modules (e.g. exact solvers) in ways that break multiplicative composition fosters discontinuities, undermining stable G-invariant reasoning.

Case Study: Mixed Architecture Incompatibility. Consider a hybrid system mixing a neural encoder (e.g. transformer) with a symbolic solver:

- The neural encoder applies nonlinear attention and extraction functions: $\sigma_{\text{attn}}, \sigma_{\text{extract}}$.
- The symbolic solver applies exact algebraic rules: r_{exact}.

While conceptually appealing, if $\sigma_{\text{attn}}, \sigma_{\text{extract}}$ do not compose multiplicatively with r_{exact}, we obtain:

$$\sigma_{\text{extract}} \circ \sigma_{\text{attn}}\big(r(z_1 * z_2)\big) \quad \neq \quad r_{\text{exact}}\big(\sigma_{\text{extract}}(\sigma_{\text{attn}}(z_1))\big) \ * \ r_{\text{exact}}\big(\sigma_{\text{extract}}(\sigma_{\text{attn}}(z_2))\big),$$

thus violating the ring-compatibility needed for stable symbolic reasoning. Empirically, one observes instability or discontinuous logic when the neural and symbolic components are not aligned with the measure-based geometry or the group symmetry. A more principled design either (i) uses differentiable approximations of r_{exact}, or (ii) ensures each neural transformation $\sigma_{\text{attn}}, \sigma_{\text{extract}}$ respects ring multiplicative laws *to the degree necessary* for stable G-invariant synergy.

Appendix C. Future Directions

Should our team receive a DARPA Disruptive Idea Award, we will immediately launch an intensive research program targeting several crucial expansions of the present framework, both theoretically (Sec .C.1 and C.2) and practically (Sec .C.3 and C.4).

First, we seek to handle multiple symmetry groups and more general (non-polynomial) function spaces, capturing the broader range of algebraic constraints found in real-world tasks. Second, we aim to formulate and empirically validate neural scaling laws specific to neurosymbolic models, clarifying how group structure and ring compatibility reduce required model capacity. Finally, we plan to translate our theoretical findings into concrete design principles for practical architectures, to tackle the growing complexity of modern neurosymbolic applications.

C.1. Beyond MP-Based and Single-Group Settings

Our reliance on monomial potentials (MPs) and a single symmetry group G provided a clean algebraic framework but constrained the scope to tasks that fit neatly into polynomial or monomial potential formulations. This limitation becomes apparent in *practical* scenarios where multiple groups or non-polynomial transformations arise. For instance, consider a speech-recognition system that must simultaneously respect time-translation symmetry (a continuous group) *and* discrete

symmetries (like phoneme permutations). Extending our approach to accommodate *multiple* interacting groups would capture far more complex tasks, such as multi-modal integration (vision plus language), each contributing distinct G-invariant structures.

Moreover, real-world data often exhibits *piecewise-defined* operations or partial symmetries not purely polynomial in nature. Handling these requires moving beyond strictly MP-based expansions, perhaps introducing piecewise-polynomial or spline-like potentials. While technically challenging—especially regarding ring factorizations that do not hold globally—these generalizations could yield a deeper understanding of how neural networks encode diverse symbolic transformations.

C.2. Beyond One Fixed-Width Parameter Space

Although our theory emphasized a single, fixed-width setting, many practical architectures adapt capacity on-the-fly—via layer expansion, architecture search, or model compression—to handle varying task complexities. For instance, consider an evolving LLM that adds layers whenever it encounters insufficient capacity for capturing novel symbolic patterns in code or reasoning tasks. Another example is one same architecture family of LLMs, yet at different parameter counts. Building on Tian (2024), we could treat the overall parameter space as a *union* of subspaces $\{M_w\}_{w\in\mathcal{W}}$ for widths w, letting the measure μ_t move *across* these subspaces rather than remain in a single M.

Analyzing dimension reduction under such dynamic expansions is significantly more involved but promises valuable insights. If G is large or if multiple groups $G_1, \ldots, G_r$ are in play, the capacity must scale in ways partially reflecting $\sum \dim(G_i)$. By permitting μ_t to *migrate* to a higher-capacity subspace $M_{w'}$ when beneficial, we might uncover a more natural synergy between data complexity, symbolic constraints, and network sizes.

C.3. Exploring Neural Scaling Laws for Neurosymbolic Models

A central insight of our framework is that minimal network width h for implementing G-invariant symbolic tasks must scale *at least* linearly in $\dim(G)$ or logarithmically in $|G|$ - as a necessary condition stated in Sec. 4.1. This motivates exploring *formal scaling laws* for neurosymbolic models, akin to how large-scale language models exhibit emergent behaviors upon crossing certain width or depth thresholds. Concretely, one could hypothesize a "hybrid" scaling law such as

$$h \;\gtrsim\; \dim(G) \;+\; f(\text{problem complexity}),$$

where $\dim(G)$ captures the algebraic constraints needed for G-invariance, and $f(\text{problem complexity})$ encodes the data-driven or combinatorial aspects (akin to existing LLM scaling laws).

Such a perspective would not merely restate that h must be large but predict how *combined* symbolic structure (group symmetry) and data complexity (like input length, vocabulary size) determine architecture growth. We conjecture that, if a neurosymbolic model *already* encodes the relevant symmetry or ring-compatibility prior, it could exhibit *provably better* scaling than a conventional model that attempts to learn these structures from scratch. Empirical exploration could involve systematically varying $\dim(G)$ or the difficulty of a ring-based reasoning task, testing whether our predicted dimension-limiting threshold indeed triggers emergent symbolic performance.

C.4. Guidelines for Designing More Practical Neurosymbolic Architectures

Our analysis of smoothness, G-equivariance, and ring-compatible activations focuses on idealized conditions (e.g., C^2 activations). Many real networks rely on ReLU or piecewise-linear units—often

non-smooth but still amenable to a *modified* displacement-convexity and stable measure evolution. Extending the theory to these practical activations would align well with widely deployed architectures in computer vision or language.

Beyond activations, the dimension-reduction principle suggests several design heuristics:

- *Enforcing G-invariance or approximate group symmetries*: e.g., group-convolution layers or transformer self-attention that respect known symmetries (permutations, rotations, etc.).
- *Structuring layers to preserve multiplicative relationships*: e.g., factored or diagonal weight matrices to better maintain ring compatibility for tasks involving polynomials or factorable transformations.
- *Stabilizing gradient flows*: employing skip connections, residual blocks, or normalization layers that regulate the Hessian's curvature to keep critical manifolds well-defined.

In practice, a system that must, for instance, parse algebraic expressions and reason about geometric transformations might *(i)* adopt specialized group-convolution for local transformations, *(ii)* incorporate ring-friendly activation functions in deeper layers, and *(iii)* employ skip connections to ensure smooth dimension-lowering.

By systematically applying these heuristics across varied tasks—from discrete logic puzzles to continuous transformations in robotics—one can concretely test the measure-based dimension reduction hypothesis in real neurosymbolic models. The hope is to show that explicitly embedding G-invariance and ring-compatibility, along with the appropriate scaling laws for network size, enables robust and efficient neural reasoning, bridging symbolic and sub-symbolic paradigms.