THE GEOMETRY OF ROBUST SYMBOLIC COMPUTATION IN NEURAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

When can neural modules preserve symbolic semantics through composition? Correct decoding on canonical inputs does not ensure correctness on intermediate representations. For fixed encoders and deterministic decoders, we study correctness through reachable-set containment. Our main result pairs an exact closed-ball ℓ_∞ capacity bound under a common inner-reachability radius with a finite-ReLU realization of arbitrary finite deterministic transition families at a prescribed horizon. A realization with an identity transition attains the bound without a network-size constraint. A complementary endpoint certificate derives reachable inner balls from joint disturbance propagation, even when each disturbance is rank deficient. Polyhedral module certificates reduce compositional correctness to regional linear programs. A separation construction shows that identical canonical tables and exact global Lipschitz constants can conceal different robust behavior. Outer-radius and quantization-aware bounds provide additional sufficient conditions. Learned arithmetic modules distinguish canonical accuracy from representation robustness and terminal accuracy from pathwise correctness. Regional certificates hold for trained four-dimensional ReLU modules, and a fixed-gate model with learned translations attains the integer capacity bound. High-dimensional endpoint tests yield no positive target-centered inner-radius estimate.

1 INTRODUCTION

A long-standing line of neural-symbolic research uses neural representations to encode symbolic objects or compositional symbolic structures, and neural transformations to realize symbolic operators, inference steps, or program-state transitions (Smolensky, 1990; Pollack, 1990). Representative interface-level approaches include (Manhaeve et al., 2018; Koh et al., 2020; Mao et al., 2019; Andreas et al., 2016), while representative multi-step reasoning and execution approaches include (Avila Garcez & Zaverucha, 1999; Evans & Grefenstette, 2018; Rocktäschel & Riedel, 2017; Dong et al., 2019; Reed & de Freitas, 2016). Canonical-input accuracy alone does not establish that non-canonical intermediate states remain valid inputs to subsequent modules.

Stable neural realizations of finite automata already address preservation of discrete state through arbitrarily long inputs (Giles et al., 1992; Omlin & Giles, 1996). Noise-dependent limits on analog computation also relate computational capacity to representation and disturbance parameters (Maass & Orponen, 1996). Approximate simulation and bisimulation provide a general framework for relating continuous and symbolic transition systems (Girard & Pappas, 2007a;b; Tabuada, 2008; Pola et al., 2008), while regional neural verification uses abstraction–refinement, SMT solving, and abstract interpretation (Pulina & Tacchella, 2010; Katz et al., 2017; Gehr et al., 2018; Singh et al., 2019). Contraction and recurrent-network verification provide further stability and robustness guarantees (Revay & Manchester, 2020; Ko et al., 2019). We build on these ideas for a given symbolic system, encoder, and decoder, keeping sufficient correctness conditions separate from necessary representation constraints.

Our main result pairs an exact ℓ_∞ capacity bound with a matching finite-ReLU realization (Theorem 4 and Proposition 1). The construction uses one encoder and decoder for every transition in a prescribed finite deterministic family and permits every full-ball disturbance sequence through a fixed horizon. This is a robust capacity result, not a bound on noiseless representation or efficient learning. Two complementary certificates address its application and compositional correctness. The

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

endpoint certificate obtains inner balls from disturbances propagated across several steps. The polyhedral certificate checks the representations passed between modules, and a separation construction identifies information absent from canonical tables and exact global Lipschitz constants. The endpoint argument uses quantitative openness, while the polyhedral test specializes regional ReLU verification. Experiments contrast empirical accuracy with regional guarantees and examine capacity attainment with learned translations and fixed geometry. Complete baseline statements and proofs appear in the appendix.

2 Formal Setting

2.1 Quotient Symbolic Transition Systems

For each type $\tau \in \mathsf{Ty}$, let S_τ be a symbolic state set with an equivalence relation $\sim_\tau$, and write $\bar{S}_\tau = S_\tau / \sim_\tau$. States carry their types, so $\bar{S}_\tau \cap \bar{S}_{\tau'} = \varnothing$ for $\tau \neq \tau'$. A typed symbolic operator is a map $\bar{\alpha} : \bar{S}_{\tau_1} \times \cdots \times \bar{S}_{\tau_k} \to \bar{S}_{\tau_0}$. An operator on representatives induces such a map precisely when it respects the equivalence relations. A transition label r from τ to τ' has a choice set C_r and a partial map $\bar{r} : \bar{S}_\tau \times C_r \rightharpoonup \bar{S}_{\tau'}$. The choice resolves all symbolic nondeterminism. A valid trace π satisfies $\bar{r}_t(\bar{s}_{t-1}, c_t) = \bar{s}_t$ for $1 \le t \le m$; undefined pairs are excluded. Full definitions appear in Appendix A.1.

2.2 Neural Execution Semantics

For each type, let (F_τ, d_τ) be a metric space, with closed balls $\mathbb{B}_\tau(f, a)$. The encoder $\eta_\tau : \bar{S}_\tau \to F_\tau$ and deterministic decoder $\delta_\tau : F_\tau \to \bar{S}_\tau$ satisfy $\delta_\tau \circ \eta_\tau = \mathrm{id}$. Define the decoder cell $\mathcal{C}_\tau(\bar{s}) = \delta_\tau^{-1}(\{\bar{s}\})$ and a certified decoding radius by $\mathbb{B}_\tau(\eta_\tau(\bar{s}), \rho_\tau(\bar{s})) \subseteq \mathcal{C}_\tau(\bar{s})$.

A nominal transition is $T_r : F_\tau \times G_r \to F_{\tau'}$, with a choice encoder $\chi_r : C_r \to G_r$. Nonempty initial sets $I_\tau(\bar{s})$ and nonempty successor sets $\mathcal{D}_r(f, \chi_r(c))$ specify admissible execution. Along π, abbreviate $T_t(f) = T_{r_t}(f, \chi_{r_t}(c_t))$ and $\mathcal{D}_t(f) = \mathcal{D}_{r_t}(f, \chi_{r_t}(c_t))$. For sets, $\mathcal{D}_t(A) = \bigcup_{f \in A} \mathcal{D}_t(f)$. The reachable sets are $\mathsf{Tube}_0^\pi = I_{\tau_0}(\bar{s}_0)$ and $\mathsf{Tube}_t^\pi = \mathcal{D}_t(\mathsf{Tube}_{t-1}^\pi)$. An admissible trajectory starts in $I_{\tau_0}(\bar{s}_0)$ and satisfies $f_t \in \mathcal{D}_t(f_{t-1})$. Execution is decoded-correct when every such trajectory satisfies $\delta_{\tau_t}(f_t) = \bar{s}_t$ at every step. Nonempty successor sets exclude deadlocks and explicit execution failures.

Running Example: A Cyclic Counter Encode $n \ge 3$ cyclic states by the standard basis vectors $e_i \in \mathbb{R}^n$, use nearest-code decoding with fixed tie-breaking, and let Π_{cyc} cyclically permute coordinates. Any radius $\rho < 1/2$ is certified. With $T(f) = \Pi_{\mathrm{cyc}} f$ and disturbances in $\mathbb{B}_\infty(0, \sigma)$, canonical initialization gives $\mathsf{Tube}_t = \mathbb{B}_\infty(e_{i+t \bmod n}, t\sigma)$. A disturbance set merely contained in that ball gives only the corresponding outer inclusion. The complete example and a finite-ReLU counterexample to robustness inferred from canonical agreement are given in the appendix.

3 Exact Finite-Horizon Neural Simulation

In this section, we characterize when every neural state reachable along a fixed symbolic trace decodes to the intended symbolic state at every time. We then extend this characterization to bounded-horizon trace families and introduce an exact outer radius for decoder-ball certification.

Theorem 1 (Exact finite-horizon simulation)**.** *Let $\pi : \bar{s}_0 \xrightarrow{(r_1,c_1)} \cdots \xrightarrow{(r_m,c_m)} \bar{s}_m$ be a valid symbolic trace, with $\bar{s}_t \in \bar{S}_{\tau_t}$ for every $t \in \{0, \dots, m\}$. Let $\mathcal{D}_t$ and Tube^π be the step-specific successor maps and generated tube defined above. The following statements are equivalent:*

(i) The neural execution is decoded-correct along π.

(ii) For every $t \in \{0, \dots, m\}$, $\mathsf{Tube}_t^\pi \subseteq \mathcal{C}_{\tau_t}(\bar{s}_t)$.

(iii) There exist sets $R_t^\pi \subseteq F_{\tau_t}$, $t \in \{0, \dots, m\}$, such that $\begin{array}{c} I_{\tau_0}(\bar{s}_0) \subseteq R_0^\pi \subseteq \mathcal{C}_{\tau_0}(\bar{s}_0), \\ \mathcal{D}_t(R_{t-1}^\pi) \subseteq R_t^\pi \subseteq \mathcal{C}_{\tau_t}(\bar{s}_t) \end{array}$ for every $t \in \{1, \dots, m\}$.

Moreover, if a family $R_t^\pi \subseteq F_{\tau_t}$ satisfies $I_{\tau_0}(\bar{s}_0) \subseteq R_0^\pi$ and $\mathcal{D}_t(R_{t-1}^\pi) \subseteq R_t^\pi$ for every $t \in \{1, \dots, m\}$, then $\mathsf{Tube}_t^\pi \subseteq R_t^\pi$ for every $t \in \{0, \dots, m\}$. Thus Tube^π is the least forward-closed family containing the admissible initial set.

Corollary 1 (Reachable failure witness)**.** *If $\mathsf{Tube}_t^\pi \not\subseteq \mathcal{C}_{\tau_t}(\bar{s}_t)$ for some t, then there exists an admissible trajectory prefix $f_0, \dots, f_t$ such that $\delta_{\tau_t}(f_t) \neq \bar{s}_t$.*

An outer approximation that extends beyond the intended decoder cell does not by itself establish failure, because the excess states may be unreachable. By contrast, an exact tube violation gives a reachable failure witness. Since every forward-closed certificate contains the generated tube, repairing such a failure requires changing the execution semantics or the verified trace family, not merely choosing different certificate sets.

Definition 1 (Bounded-horizon trace family and aggregated reachable sets)**.** *Fix $T \in \mathbb{Z}_{\geq 0}$. A bounded-horizon trace family $\mathcal{K}_T$ is a set of valid symbolic traces such that every $\pi \in \mathcal{K}_T$ contains at most T transitions. The family $\mathcal{K}_T$ need not be finite. For $\bar{s} \in \bar{S}_\tau$ and $0 \leq t \leq T$, define $U_t(\bar{s}) \subseteq F_\tau$ by $U_t(\bar{s}) = \bigcup \left\{ \mathsf{Tube}_t^\pi : \begin{array}{l} \pi \in \mathcal{K}_T,\ \pi \text{ has at least } t \text{ transitions}, \\ \text{and the time-}t \text{ state of } \pi \text{ is } \bar{s} \end{array} \right\}$.*

Corollary 2 (Bounded-horizon trace-family characterization)**.** *Let $\mathcal{K}_T$ be a bounded-horizon trace family. The neural execution is decoded-correct on $\mathcal{K}_T$ iff $U_t(\bar{s}) \subseteq \mathcal{C}_\tau(\bar{s})$ for every $0 \leq t \leq T$, every $\tau \in \mathsf{Ty}$, and every $\bar{s} \in \bar{S}_\tau$ for which $U_t(\bar{s})$ is nonempty.*

Remark 1 (Trace-indexed and state-indexed certificates)**.** *A trace-indexed certificate keeps traces separate and checks only the neural states and transitions that actually occur along each trace. A state-indexed simulation relation instead combines neural states reached through different histories whenever they represent the same symbolic state, and requires every permitted transition to be correct from the entire combined set. This introduces state–transition combinations that may occur on no trace in the family, so the state-indexed condition is generally stronger.*

For a nonempty tube, define its exact outer radius around the canonical target by $\mathrm{rad}_t^\pi = \sup_{f \in \mathsf{Tube}_t^\pi} d_{\tau_t}(f, \eta_{\tau_t}(\bar{s}_t)) \in [0, \infty]$. The decoder-ball certificate at time t holds if and only if $\mathrm{rad}_t^\pi \leq \rho_{\tau_t}(\bar{s}_t)$. Failure of this inequality does not imply $\mathsf{Tube}_t^\pi \not\subseteq \mathcal{C}_{\tau_t}(\bar{s}_t)$, because the decoder cell need not be a ball centered at $\eta_{\tau_t}(\bar{s}_t)$.

4 Operator Execution and Recanonicalization

A finite typed derivation tree is evaluated by applying each module T_α to its children's neural outputs, starting from encoded leaves. Correct decoding on encoded input tuples requires $\delta(T_\alpha(\eta(\bar{s}_1), \dots, \eta(\bar{s}_k))) = \bar{\alpha}(\bar{s}_1, \dots, \bar{s}_k)$; exact realization on canonical inputs requires the output itself to equal the target encoding. The former condition does not imply direct compositional correctness (Appendix Example 1). Closure of certified input neighborhoods into the certified target neighborhood does imply it (Theorem 6).

Alternatively, inserting $P_\tau = \eta_\tau \circ \delta_\tau$ between modules restores canonical inputs and makes correct canonical decoding compose (Appendix Theorem 7). This external operation changes the neural execution. Complete tables with exact canonical outputs already compose without it in noiseless execution, provided they cover every input tuple used by the tree. The distinction is between exact encoding and correct decoding, as well as between canonical execution and a prescribed perturbation model.

5 Outer-Radius Robustness Certificates

We derive a ball outer approximation to the exact tube and certify correctness when this ball remains inside the intended decoder cell.

Fix a symbolic trace π with m transitions. For each $t \in \{1, \dots, m\}$, let $T_t(f) = T_{r_t}(f, \chi_{r_t}(c_t))$ be the step-specific nominal transition map, and define the encoded transition error by $\epsilon_t = d_{\tau_t}\left(T_t\left(\eta_{\tau_{t-1}}(\bar{s}_{t-1})\right), \eta_{\tau_t}(\bar{s}_t)\right)$.

Theorem 2 (Outer-radius propagation)**.** *Fix a symbolic trace π with m transitions. Let $b_0 \geq 0$ be an upper bound on the distance between the admissible initial states and the canonical encoding of $\bar{s}_0$.*

For each $t \in \{1, \dots, m\}$, let $q_{t-1} \geq 0$ be the radius of the neighborhood centered at $\eta_{\tau_{t-1}}(\bar{s}_{t-1})$ on which the Lipschitz bound is imposed, let $L_t \geq 0$ be the corresponding Lipschitz constant of T_t, and let $u_t \geq 0$ bound the distance between each admissible successor and the nominal output $T_t(f)$. Assume the following conditions:

(i) The initial set satisfies $\mathsf{Tube}_0^\pi \subseteq \mathbb{B}_{\tau_0}\left(\eta_{\tau_0}(\bar{s}_0), b_0\right)$.

(ii) For each $t \in \{1, \dots, m\}$, the map T_t is L_t-Lipschitz on $\mathbb{B}_{\tau_{t-1}}\left(\eta_{\tau_{t-1}}(\bar{s}_{t-1}), q_{t-1}\right)$.

(iii) For each $t \in \{1, \dots, m\}$ and every $f \in \mathbb{B}_{\tau_{t-1}}\left(\eta_{\tau_{t-1}}(\bar{s}_{t-1}), q_{t-1}\right)$, one has $\mathcal{D}_t(f) \subseteq \mathbb{B}_{\tau_t}\left(T_t(f), u_t\right)$.

Define $B_0 = b_0$ and, for each $t \in \{1, \dots, m\}$, $B_t = \epsilon_t + L_t B_{t-1} + u_t$. If $B_{t-1} \leq q_{t-1}$ for every $t \in \{1, \dots, m\}$, then $\mathsf{Tube}_t^\pi \subseteq \mathbb{B}_{\tau_t}\left(\eta_{\tau_t}(\bar{s}_t), B_t\right)$ for every $t \in \{0, \dots, m\}$.

If, in addition, $B_t \leq \rho_{\tau_t}(\bar{s}_t)$ for every $t \in \{0, \dots, m\}$, then the trace is decoded-correct.

Although u_t enters the radius recurrence additively, the proof uses only the triangle inequality and distance bounds; no vector-space operations on neural states are required.

Corollary 3 (Uniform outer-radius and contraction bounds)**.** *Under the hypotheses of Theorem 2, suppose that $\epsilon_t \leq \epsilon$, $u_t \leq u$, and $L_t \leq L$ at every step, and set $\nu = \epsilon + u$. Then $B_t \leq L^t b_0 + \nu \sum_{j=0}^{t-1} L^j$. In particular, if $0 \leq L < 1$, then $B_t \leq \max\left\{b_0, \frac{\nu}{1-L}\right\}$.*

Now let $\mathcal{K}_{\mathrm{fin}}$ be any family of valid finite traces, with no common upper bound on trace length, for which the preceding assumptions hold uniformly with common bounds b_0, ϵ, u, and $0 \leq L < 1$. Let $q > 0$ and $\rho > 0$ be common lower bounds on the trusted-neighborhood radii and certified decoding radii, respectively. If $\max\left\{b_0, \frac{\nu}{1-L}\right\} \leq \min\{q, \rho\}$, then the neural execution is decoded-correct on $\mathcal{K}_{\mathrm{fin}}$.

The finite-horizon estimate and the uniform trace-family guarantee are proved in Appendix Sections A.10 and A.11. For $L \geq 1$, even a fixed positive per-step error can eventually leave any bounded decoder cell, as shown by Proposition 9 in the Appendix.

6 Inner Reachability

An inner approximation reverses the inclusion. In real normed state spaces, suppose $T_t(\mathbb{B}(x, a)) \supseteq \mathbb{B}(T_t(x), \mu_t a)$ throughout the relevant radii and $\mathbb{B}(T_t(f), v_t) \subseteq \mathcal{D}_t(f)$ on the predecessor ball. If the initial radius-a_0 ball is reachable and $\mu_t a_{t-1} + v_t \geq \epsilon_t$, then $a_t = \mu_t a_{t-1} + v_t - \epsilon_t$ gives a reachable ball centered at the intended encoding (Appendix Theorem 8). A negative expression cannot be replaced by zero, since the target encoding may itself be unreachable. The next result uses the disturbance sequence jointly and does not require positive full-dimensional noise at each step. The one-step certificate can miss disturbances confined to proper subspaces whose propagated effects span the terminal state space. We therefore also consider the map from an entire disturbance sequence to its terminal state. The following result applies a standard right-inverse and fixed-point argument for quantitative openness (Graves, 1950).

Theorem 3 (Multistep inner reachability)**.** *Fix a valid trace π with $m \geq 1$ transitions, and assume that F_{τ_m} is a finite-dimensional real normed space of positive dimension, with $d_{\tau_m}(x, y) = \|x - y\|_{\tau_m}$. Let $\mathcal{W}$ be a finite-dimensional real normed space, let $h > 0$, and fix $f^\circ \in \mathsf{Tube}_0^\pi$. Suppose that $\Psi : \mathbb{B}_{\mathcal{W}}(0, h) \to F_{\tau_m}$ is continuous and that, for every $w \in \mathbb{B}_{\mathcal{W}}(0, h)$, $\Psi(w)$ is the endpoint of an admissible trajectory along π starting at f°.*

Let $J : \mathcal{W} \to F_{\tau_m}$ be a surjective linear map, and let a linear map $\Lambda : F_{\tau_m} \to \mathcal{W}$ satisfy $J\Lambda = \mathrm{id}_{F_{\tau_m}}$. Set $\gamma_{\mathrm{inv}} = \|\Lambda\|$. Suppose that $\zeta \geq 0$ satisfies $\|\Psi(w) - \Psi(0) - Jw\|_{\tau_m} \leq \zeta$ for every $w \in \mathbb{B}_{\mathcal{W}}(0, h)$. Define $\epsilon_{\mathrm{end}}^\pi = \|\Psi(0) - \eta_{\tau_m}(\bar{s}_m)\|_{\tau_m}$ and $a_{\mathrm{end}}^\pi = h/\gamma_{\mathrm{inv}} - \zeta - \epsilon_{\mathrm{end}}^\pi$. If $a_{\mathrm{end}}^\pi \geq 0$, then $\mathbb{B}_{F_{\tau_m}}\left(\eta_{\tau_m}(\bar{s}_m), a_{\mathrm{end}}^\pi\right) \subseteq \mathsf{Tube}_m^\pi$. Before recentering, whenever $h/\gamma_{\mathrm{inv}} \geq \zeta$, one has $\mathbb{B}_{F_{\tau_m}}\left(\Psi(0), h/\gamma_{\mathrm{inv}} - \zeta\right) \subseteq \Psi\left(\mathbb{B}_{\mathcal{W}}(0, h)\right)$.

The admissibility assumption applies to every parameter sequence in the ball. The certificate supplies a terminal inner radius for Section 7 even when intermediate reachable sets have empty interior. If

$a^{\pi}_{\mathrm{end}} < 0$, it cannot be replaced by zero without separately establishing reachability of the target encoding.

Appendix A.14 proves the result and gives two remainder certificates. If $J = D\Psi(0)$ and the derivative is Lipschitz with constant ℓ_{der} on the parameter ball, then $\zeta = \ell_{\mathrm{der}} h^2/2$ is valid. For affine disturbances and a ReLU execution with a common activation pattern throughout that ball, the endpoint map is affine and $\zeta = 0$. Neither case requires global smoothness.

For example, let $x_0 = 0$, $x_t = Ax_{t-1} + e_1 w_t$ for $t \in \{1, 2\}$, and $A(x_1, x_2) = (0, x_1)$. Each update has rank-one nominal dynamics and rank-one disturbance injection, yet $\Psi(w_1, w_2) = (w_2, w_1)$. With the ℓ_∞ norm and $|w_t| \le h$, the two-step reachable set is exactly $\mathbb{B}_\infty(0, h)$.

7 Capacity from Inner Reachability

Fix a nonempty family $\mathcal{K}_T$ of valid traces of length T, with a common terminal type τ, and let S_T be its finite set of terminal states, with $N_T = |S_T|$. Write $A_T(\bar{s}) = \bigcup_{\pi \in \mathcal{K}_T : \bar{s}^{\pi}_T = \bar{s}} \mathsf{Tube}^{\pi}_T$. Terminal robustness means $A_T(\bar{s}) \subseteq \mathcal{C}_\tau(\bar{s})$, so these sets are disjoint for distinct states. If each contains $\mathbb{B}_\tau(\eta_\tau(\bar{s}), a_T)$ with $a_T > 0$, their number is bounded by the number of disjoint ambient balls centered in the encoding set (Appendix Theorem 9). Metric-entropy and doubling-dimension versions are given in the appendix.

The general packing bound admits an exact evaluation for ℓ_∞ encodings in $[-B, B]^d$, together with a matching finite-horizon realization.

Theorem 4 (Sharp ℓ_∞ horizon-capacity bound)**.** *Let $F_\tau = \mathbb{R}^d$ carry the ℓ_∞ metric, with $d \ge 1$, $B \ge 0$, and $a_T > 0$. Suppose that S_T consists of $N_T \ge 1$ distinct terminal quotient states, that their encodings $\eta_\tau(\bar{s})$ lie in $[-B, B]^d$, and that a_T is a common inner reachability radius. If the deterministic decoder is terminally robust, then $B = 0$ implies $N_T \le 1$. If $B > 0$, set $n_{\mathrm{cap}} = \lceil B/a_T \rceil$. Then $N_T \le n^d_{\mathrm{cap}} \le (1 + B/a_T)^d$.*

The integer bound is the exact maximum number of pairwise disjoint closed radius-a_T balls with centers in $[-B, B]^d$. In particular, $a_T \ge B$ implies $N_T \le 1$. If $N_T > 1$, then $n_{\mathrm{cap}} \ge 2$ and $d \ge \log N_T / \log n_{\mathrm{cap}}$. The dimension bound $d \ge \log N_T / \log(1 + B/a_T)$ obtained from volume remains valid whenever $B > 0$.

The integer bound concerns states actually reached at horizon T. The following realization chooses a common encoder and decoder for a prescribed horizon and allows arbitrary finite deterministic transition tables.

Proposition 1 (Finite-horizon realization)**.** *Fix positive integers T, d, let $B \ge 0$, and let $\sigma > 0$. Set $a_T = T\sigma$, and set $n_{\mathrm{cap}} = 1$ when $B = 0$ and $n_{\mathrm{cap}} = \lceil B/a_T \rceil$ otherwise. Let $\bar{S}$ be nonempty with $|\bar{S}| \le n^d_{\mathrm{cap}}$, and let $\mathcal{R}_{\mathrm{cap}}$ index finitely many maps $\bar{r} : \bar{S} \to \bar{S}$. There exist a fixed encoder $\eta : \bar{S} \to [-B, B]^d$, a fixed deterministic decoder δ, and finite feed-forward ReLU maps $T_r : \mathbb{R}^d \to \mathbb{R}^d$ such that $T_r(f) = f + \eta(\bar{r}(\bar{s})) - \eta(\bar{s})$ for $f \in \mathbb{B}_\infty(\eta(\bar{s}), a_T)$.*

Under $I(\bar{s}) = \{\eta(\bar{s})\}$ and $\mathcal{D}_r(f) = \mathbb{B}_\infty(T_r(f), \sigma)$, every symbolic trace π of length at most T satisfies, at every prefix, $\mathsf{Tube}^{\pi}_t = \mathbb{B}_\infty(\eta(\bar{s}_t), t\sigma) \subseteq \delta^{-1}(\{\bar{s}_t\})$. Thus all transitions are exact on canonical encodings and locally open with coefficient 1 up to radius a_T. If $\mathcal{R}_{\mathrm{cap}}$ includes the identity and $|\bar{S}| = n^d_{\mathrm{cap}}$, the horizon-$T$ capacity bound is attained.

The construction permits many-to-one transitions and all admissible disturbance sequences. Its network size is unrestricted and may grow with the transition tables; it does not assert that prescribed trained modules attain the bound. The experiments report the volume expression as U_T. Corollary 9 in the appendix gives its specialization to $a_T = T\sigma$, while Proposition 12 provides a separate static grid construction.

7.1 Digital Addressing and State-Quantized Execution

A deterministic quantizer $\mathsf{Q}_\tau : F_\tau \to \mathcal{Q}_\tau$ changes the initial set to $\mathsf{Q}_{\tau_0}(I_{\tau_0}(\bar{s}_0))$ and each successor set to $\mathsf{Q}_{\tau_t}(\mathcal{D}_t(f))$. If its errors are bounded by κ_0 on the initial set and by κ_t on all pre-quantization successors from the trusted predecessor neighborhood, Appendix Theorem 10 uses $B^{\mathsf{Q}}_0 = b_0 + \kappa_0$

and $B_t^{\mathsf{Q}} = \epsilon_t + L_t B_{t-1}^{\mathsf{Q}} + u_t + \kappa_t$, subject to the same neighborhood and decoding checks. Saturation must be included in κ_t.

Separately, p bits in each of d coordinates provide at most 2^{pd} addresses, so robust representation of N_T terminal states requires $pd \geq \lceil \log_2 N_T \rceil$ (Appendix Proposition 5). This cardinality condition does not ensure correct dynamics. In a positive-dimensional real normed space, a finite quantized reachable set contains no positive-radius ambient ball, so the continuous inner-ball argument cannot be applied to it directly.

8 Neural-Network Certificates

For feed-forward networks with 1-Lipschitz activations, the product of induced layer norms is an upper bound on the network Lipschitz constant (Appendix Proposition 6). Within a single affine activation region, an invertible linear part on equal input and output spaces gives a local openness certificate (Appendix Proposition 3); a norm upper bound alone does not. Finite ReLU interpolation realizes any prescribed finite table exactly (Appendix Proposition 7), but supplies no specified neighborhood robustness. We next certify the module's image directly.

8.1 Polytope certificates for module composition

Layerwise norms can obscure the behavior of a module on the states supplied by other modules. A complementary certificate specializes standard regional ReLU verification to these images (Katz et al., 2017; Singh et al., 2019). Fix finite collections of symbolic states and permitted operator tuples, with $F_\tau = \mathbb{R}^{n_\tau}$. For each state $\bar{s}$, choose a finite union $K_{\bar{s}} = \bigcup_j P_{\bar{s},j}$ of nonempty compact polytopes contained in its decoder cell. Require admissible leaf and initial states to belong to the corresponding sets. For nearest-code decoding in ℓ_∞, decoder containment follows if every component lies within a code-centered box whose radius is strictly less than half the minimum code separation.

For a permitted tuple $\bar{\boldsymbol{s}} = (\bar{s}_1, \ldots, \bar{s}_k)$ with output $\bar{s}_0 = \bar{\alpha}(\bar{s}_1, \ldots, \bar{s}_k)$, subdivide every product of input components into finitely many polytopes Q_ι. Let T_α be a finite feed-forward ReLU module. Let $D_{\alpha,\bar{\boldsymbol{s}}}$ be a nonempty compact disturbance polytope, and assume admissible module outputs are contained in $T_\alpha(x) + D_{\alpha,\bar{\boldsymbol{s}}}$. Assign each source polytope a target component $P_{\bar{s}_0,j_\iota} = \{y : H_\iota y \leq h_\iota\}$. Write $T_\alpha(x) = A_\beta x + b_\beta$ on each closed activation region $\mathcal{V}_\beta$.

Theorem 5 (Finite polytope certificate)**.** *For every source polytope, target inequality, and activation region with $Q_\iota \cap \mathcal{V}_\beta \neq \varnothing$, suppose $\max_{\substack{x \in Q_\iota \cap \mathcal{V}_\beta \\ \xi \in D_{\alpha,\bar{\boldsymbol{s}}}}} H_{\iota,i}(A_\beta x + b_\beta + \xi) \leq h_{\iota,i}$. These finitely many linear-programming inequalities are equivalent to the prescribed inclusions $T_\alpha(Q_\iota) + D_{\alpha,\bar{\boldsymbol{s}}} \subseteq P_{\bar{s}_0,j_\iota}$. If they hold for every permitted tuple, every admissible finite derivation tree using only checked operator tuples is correct at every node. The same construction for the unary maps $f \mapsto T_r(f, \chi_r(c))$ certifies every finite serial execution using the checked transitions, without a common horizon bound.*

The certificate retains affine cancellations and target-facet directions. It requires neither global contraction nor multiplication of layerwise norms. Its cost can be exponential in the number of ReLU units; the theorem is a finite verification reduction, not a polynomial-time algorithm. Appendix A.31 proves the result, specifies exact arithmetic checks, and discusses the stronger requirements introduced by combining histories within one symbolic state.

Proposition 2 (Equal global Lipschitz constants, different robust behavior)**.** *For every $0 < r < 1/2$ and $L_{\mathrm{sep}} \geq 4/r$, two finite scalar ReLU networks implement the same two-state canonical transition table and both have global Lipschitz constant exactly L_{sep}. Under additive disturbances of radius $r/2$, one has an interval certificate for every finite horizon; the other has a two-step decoding failure from a canonical initial state. Even the exact global Lipschitz constant fails to certify the second step of the first network by the scalar radius recurrence.*

This separation concerns global bounds, not all local Lipschitz methods. Section 9.2 contrasts inconclusive high-dimensional enclosures with successful regional verification on four-dimensional MLPs. The 128-dimensional diagnostics use interval and zonotope enclosures rather than exact activation-region linear programs and do not establish closure of the tested neighborhoods.

9 EXPERIMENTS

The arithmetic modules implement addition, subtraction, and multiplication modulo 64 with codes in $\{-0.5, 0.5\}^{128}$, nearest-code ℓ_∞ decoding, and $\rho = 0.49$. Each has four width-1024 ReLU hidden layers. Canonical-input and neighborhood supervision use ten paired seeds; evaluations reuse frozen checkpoints. Four-dimensional experiments examine regional verification and capacity attainment. Neural evaluations use FP64 on an RTX 4090; uncertainty is the sample standard deviation across seeds. The guarantees and limitations of each verification procedure are specified in Appendix A.33.

9.1 CANONICAL COMPOSITION AND REPRESENTATION STRESS

All 60 arithmetic modules decode the $4{,}096$-entry table correctly, holding canonical correctness fixed rather than testing symbolic generalization. All $15{,}360$ unperturbed tree evaluations are correct at every node, with or without canonicalization. Exp1b replaces each nonroot internal-node output by $\widetilde{f} = \eta(\delta(f)) + \xi$, with $\|\xi\|_\infty \le r < \rho$. This preserves the preceding prediction but changes the execution; even $r = 0$ transmits a canonical code. The comparison applies $P = \eta \circ \delta$ after this intervention.

Table 1: Representation stress in 512-operator trees (Exp1b). Percentages are mean $\pm$ sample standard deviation across ten seeds, with 64 trees per seed and condition. "Stressed" passes $\eta(\delta(f)) + \xi$ to the parent, whereas "Canonicalized" applies $\eta \circ \delta$ before evaluation. Whole-tree accuracy requires correct decoding at every internal node.

Method	Root accuracy	Whole-tree accuracy
Balanced, corners; $r = 0.20$		
Point, stressed	39.38 ± 8.62	1.41 ± 2.49
Neighborhood, stressed	99.69 ± 0.66	98.28 ± 1.72
Point, canonicalized	100.00 ± 0.00	100.00 ± 0.00
Balanced, uniform; $r = 0.48$		
Point, stressed	13.75 ± 6.02	0.00 ± 0.00
Neighborhood, stressed	79.38 ± 10.07	52.97 ± 13.27
Point, canonicalized	100.00 ± 0.00	100.00 ± 0.00
Chain, uniform; $r = 0.48$		
Point, stressed	87.66 ± 4.74	0.00 ± 0.00
Neighborhood, stressed	100.00 ± 0.00	97.34 ± 2.22
Point, canonicalized	100.00 ± 0.00	100.00 ± 0.00
Balanced, corners; $r = 0.48$		
Point, stressed	6.88 ± 2.23	0.00 ± 0.00
Neighborhood, stressed	7.81 ± 3.90	0.00 ± 0.00
Point, canonicalized	100.00 ± 0.00	100.00 ± 0.00

Neighborhood supervision improves whole-tree accuracy under radius-0.20 corner perturbations, but both fail on the largest balanced trees under radius-0.48 corner perturbations. Root accuracy can conceal intermediate errors. The following regional tests distinguish an inconclusive bound from a witnessed failure of the assigned inclusion.

9.2 OUTER-RADIUS CERTIFICATES AND THEIR CONSERVATISM

Exp2 composes the learned operators serially, holds the encoded right operand fixed at each step, and adds bounded disturbances after each update. We evaluate the recurrence in Theorem 2 using either the layerwise induced-ℓ_∞ norm bound or a local bound obtained by interval propagation and absolute-weight sensitivity propagation. The local construction removes ReLUs certified inactive over the entire predecessor box. Its trusted radius is $q = 0.25$. Local propagation is stopped permanently when the preceding radius exceeds q, and path certification requires all decoding-radius tests through the current step.

For the 128-dimensional modules, all 2,560 tested serial executions remain pathwise correct through 512 steps, but global and local bounds certify only step 1 (Appendix Table 4). Even without disturbances, mean second-step local radii exceed 10^4 while mean observed errors stay below 0.02. The step-2 local bounds remain valid, but exceed both the trusted and decoding radii, so local propagation stops at step 3. The analytic separation shows that conservative bounds can coexist with correct execution; it does not explain the behavior of these particular learned weights.

An analytic ReLU counter provides a complementary test with prescribed local slope L and canonical error $1/1024$. Positive-corner errors agree with the local recurrence within 2.3×10^{-15} across 17,160 valid records. The $L = 0.5$ controls satisfy the uniform contraction condition and remain certified through 512 steps. With $L = 1$ and no disturbances, certification ends at step 65, although no decoding failure is observed through step 512. With $L = 1$ and disturbance radius $1/1024$, certification ends at step 33 and the first sampled failure occurs at step 40. Thus certificate loss and execution failure need not coincide.

Regional verification of learned modules. The 128-dimensional addition networks pass all 12,800 distinct additive-disturbance trajectories, but none of 11,520 enclosure checks proves closure (Appendix A.33.6). We also train 4–64–4 ReLU modules for increment, decrement, and doubling modulo 16, with all weights and biases optimized from random initialization; verification outcomes do not enter checkpoint selection. Both supervision regimes attain complete canonical accuracy. Across all eight radius pairs, activation-region verification certifies $2432/3840$ neighborhood-trained transition checks, with 1050 witnessed assigned-inclusion violations and 358 inconclusive checks, whereas all 3840 point-trained checks have witnessed violations. At source and target radius $1/8$ and output-disturbance radius $1/256$, the neighborhood-trained certified fraction is $91.25 \pm 4.03\%$. Every neighborhood-trained doubling module has all 16 transitions certified, but no complete three-operator family is certified. Global-norm and interval bounds certify none (Appendix A.33.7).

9.3 Inner reachability, collisions, and capacity

Exp3 compares recursive and exact inner radii for affine controls. No inner-inclusion violation occurs in 3,876 valid records. For coordinate-permutation controls, full-ball disturbances produce radius-$T\sigma$ terminal balls. Shared-scalar disturbances yield the same outer radius but confine reachable sets to affine lines. The multistep rank-one example supplies a separate case in which propagating rank-one disturbances through the dynamics produces full-dimensional terminal reachability.

Table 2: Full-ball reachability and fixed-grid packing (Exp3), with $d = 2$, nine levels per coordinate, $B = 1$, $\Delta = 1/4$, and $\sigma = 1/1024$. Here $a_T = T\sigma$, $U_T = (1 + B/a_T)^d$, $n_{\mathrm{cap}}^d = \lceil B/a_T \rceil^d$, and M_T is the maximum fixed-grid packing size.

T	a_T	Full-grid N_T	U_T	n_{cap}^d	M_T	Full grid robust
64	$1/16$	81	289	256	81	Yes
128	$1/8$	81	81	64	25	No
256	$1/4$	81	25	16	9	No
512	$1/2$	81	9	4	4	No

The integer-bound column follows directly from Theorem 4. It excludes the first-contact configuration at $T = 128$, which the volume bound does not exclude. Among 420 full-ball settings, the complete grid is robust in 200 and has intersecting reachable balls in 220. The 60 collision pairs have different terminal symbols and identical neural endpoints. Proposition 1 establishes attainability with a chosen encoding and ReLU realization; it does not change the fixed-grid decoder used here.

Multistep endpoint diagnostics. In the 128-dimensional addition networks, some low-rank disturbance sequences yield full numerical endpoint rank, but no tested setting gives a positive target-centered inner-radius estimate. Right-inverse residuals and target offsets limit these estimates (Appendix A.33.6).

Capacity attainment with learned translations. The 16-state increment experiment fixes ReLU gates and a local linear coefficient $\alpha = 0.82$, learning only 64 translation parameters. On radius-0.40 code neighborhoods, the module is affine with local openness coefficient α. With full-ball disturbances of radius 0.06, all 160 state–seed inclusion checks pass, with minimum margin approximately 0.012, and all 5120 initial-state–step inner-reachability records remain valid through step 32. The common inner radius first exceeds 0.25 at step 7; at the reported horizons it grows from $a_8 \approx 0.265195$ to $a_{32} \approx 0.332751$. Bijectivity preserves $N_T = 16$; with $B = 0.5$ and $d = 4$, Theorem 4 gives $\lceil B/a_T \rceil^d = 16$ at every reported horizon from 8 to 32 in all ten seeds. An exact scalar recheck with a conservative error budget preserves these conclusions (Appendix A.33.8).

9.4 State quantization versus address capacity

Exp4 compares unquantized execution, terminal-only quantization, and quantization of the initial state and every update. The uniform quantizer has 2^p levels per coordinate in $[-1, 1]$; weights and nominal computations remain FP64. Quantization-error bounds include saturation.

Table 3: State quantization (Exp4), with $T = 512$ and $u = 0.005$. Accuracy (%) is mean $\pm$ sample standard deviation across ten seeds. Bit ranges group identical results at tested precisions.

Training	Quantization	Bits per coordinate	Terminal accuracy	Path accuracy
Point	None	–	100.00 ± 0.00	100.00 ± 0.00
Point	Terminal only	2–16	100.00 ± 0.00	100.00 ± 0.00
Point	Every step	2	74.69 ± 10.25	0.00 ± 0.00
Point	Every step	3	95.31 ± 5.56	62.81 ± 17.83
Point	Every step	4–16	100.00 ± 0.00	100.00 ± 0.00
Neighborhood	None	–	100.00 ± 0.00	100.00 ± 0.00
Neighborhood	Terminal only	2–16	100.00 ± 0.00	100.00 ± 0.00
Neighborhood	Every step	2	96.88 ± 2.95	0.00 ± 0.00
Neighborhood	Every step	3	100.00 ± 0.00	100.00 ± 0.00
Neighborhood	Every step	4–16	100.00 ± 0.00	100.00 ± 0.00

With two-bit every-step quantization, every tested 512-step path has an intermediate error, even though the neighborhood-trained modules retain $96.88 \pm 2.95\%$ terminal accuracy. Four bits suffice empirically for all tested learned paths. These are thresholds for the tested models and quantizer, not universal precision bounds. The every-step protocol also quantizes initialization, so its effect is not isolated from later quantization. The learned norm bounds provide no first-step certificate under this protocol. Analytic contracting controls provide the nonvacuous quantization-aware certificates; omitting quantization terms leaves 63.44% of their sampled final errors outside the resulting radius. Address enumeration separately confirms the cardinality condition, which cannot establish dynamic correctness.

10 Discussion, Limitations, and Conclusion

The capacity theorem requires a specified disturbance model and a common reachable inner ball; its matching construction has no size or training guarantee. The structured experiment learns translations but fixes the geometry and local openness. The endpoint certificate requires admissibility, a right inverse, and a uniform remainder bound; numerical rank alone is insufficient. Regional verification can be exponential, and the unstructured 128-dimensional modules remain uncertified. External canonicalization changes execution. Probabilistic decoding, distribution-dependent guarantees, and explicit failure states are outside the framework.

The exact ℓ_∞ bound and matching ReLU dynamics characterize robust capacity under these assumptions. Endpoint and polyhedral certificates address distinct questions: which representations are reachable, and whether subsequent modules preserve their symbolic meaning. The experiments separate these universal conditions from canonical, terminal, and sampled pathwise accuracy.

AI USE STATEMENT

Generative AI tools were used to aid and polish the writing of this paper, to support the retrieval and discovery of related work, and to assist in checking and revising portions of the mathematical arguments and proofs, and to support the verification of portions of the experimental code. All AI-assisted writing was independently verified by the authors against the stated assumptions and derivations, and all references identified with AI assistance were checked against the original publications. The authors take full responsibility for the final content of this work.

REPRODUCIBILITY STATEMENT

Complete baseline statements appear in Appendix A.1, with proofs in the subsequent appendix sections. Section 9 reports the results; Appendix A.33 specifies protocols, statistical units, and numerical acceptance rules. The supplementary source code reproduces the reported experiments and numerical evaluations.

REFERENCES

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. Neural module networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 39–48, 2016.

Artur S Avila Garcez and Gerson Zaverucha. The connectionist inductive learning and logic programming system. *Applied Intelligence*, 11(1):59–77, 1999.

Dimitri Bertsekas. *Dynamic programming and optimal control: Volume I*, volume 4. Athena scientific, 2012.

George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of Control, Signals and Systems*, 2(4):303–314, 1989.

Honghua Dong, Jiayuan Mao, Tian Lin, Chong Wang, Lihong Li, and Denny Zhou. Neural logic machines. In *7th International Conference on Learning Representations (ICLR 2019)*. OpenReview.net, 2019. URL https://openreview.net/forum?id=B1xY-hRctX.

Richard Evans and Edward Grefenstette. Learning explanatory rules from noisy data. *Journal of Artificial Intelligence Research*, 61:1–64, 2018.

Timon Gehr, Matthew Mirman, Dana Drachsler-Cohen, Petar Tsankov, Swarat Chaudhuri, and Martin Vechev. AI2: Safety and robustness certification of neural networks with abstract interpretation. In *2018 IEEE Symposium on Security and Privacy*, pp. 3–18. IEEE, 2018.

C Lee Giles, Clifford B Miller, Dong Chen, Hsing-Hen Chen, Guo-Zheng Sun, and Yee-Chun Lee. Learning and extracting finite state automata with second-order recurrent neural networks. *Neural Computation*, 4(3):393–405, 1992.

Antoine Girard and George J. Pappas. Approximation metrics for discrete and continuous systems. *IEEE Transactions on Automatic Control*, 52(5):782–798, 2007a. doi: 10.1109/TAC.2007.895849.

Antoine Girard and George J. Pappas. Approximation metrics for discrete and continuous systems. In *Hybrid Systems: Computation and Control*, volume 4416 of *Lecture Notes in Computer Science*, pp. 244–258. Springer, 2007b.

Lawrence M. Graves. Some mapping theorems. *Duke Mathematical Journal*, 17:111–114, 1950. doi: 10.1215/S0012-7094-50-01713-3.

Kurt Hornik. Approximation capabilities of multilayer feedforward networks. *Neural Networks*, 4 (2):251–257, 1991.

Guy Katz, Clark Barrett, David L. Dill, Kyle Julian, and Mykel J. Kochenderfer. Reluplex: An efficient smt solver for verifying deep neural networks. In Rupak Majumdar and Viktor Kunčak (eds.), *Computer Aided Verification*, pp. 97–117, Cham, 2017. Springer International Publishing.

Ching-Yun Ko, Zhaoyang Lyu, Lily Weng, Luca Daniel, Ngai Wong, and Dahua Lin. POPQORN: Quantifying robustness of recurrent neural networks. In Kamalika Chaudhuri and Ruslan Salakhutdinov (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3468–3477. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/ko19a.html.

Pang Wei Koh, Thao Nguyen, Yew Siang Tang, Stephen Mussmann, Emma Pierson, Been Kim, and Percy Liang. Concept bottleneck models. In *International conference on machine learning*, pp. 5338–5348. PMLR, 2020.

Andrei Nikolaevich Kolmogorov and Vladimir Mikhailovich Tikhomirov. ε-entropy and ε-capacity of sets in function spaces. *Uspekhi Matematicheskikh Nauk*, 14(2):3–86, 1959.

Wolfgang Maass and Pekka Orponen. On the effect of analog noise in discrete-time analog computations. In M.C. Mozer, M. Jordan, and T. Petsche (eds.), *Advances in Neural Information Processing Systems*, volume 9. MIT Press, 1996. URL https://proceedings.neurips.cc/paper_files/paper/1996/file/33ceb07bf4eeb3da587e268d663aba1a-Paper.pdf.

Robin Manhaeve, Sebastijan Dumancic, Angelika Kimmig, Thomas Demeester, and Luc De Raedt. DeepProbLog: Neural probabilistic logic programming. In *Advances in Neural Information Processing Systems*, 2018.

Jiayuan Mao, Chuang Gan, Pushmeet Kohli, Joshua B. Tenenbaum, and Jiajun Wu. The neuro-symbolic concept learner: Interpreting scenes, words, and sentences from natural supervision. In *7th International Conference on Learning Representations (ICLR 2019)*. OpenReview.net, 2019. URL https://openreview.net/forum?id=rJgMlhRctm.

Christian W Omlin and C Lee Giles. Constructing deterministic finite-state automata in recurrent neural networks. *Journal of the ACM (JACM)*, 43(6):937–972, 1996.

Giordano Pola, Antoine Girard, and Paulo Tabuada. Approximately bisimilar symbolic models for nonlinear control systems. *Automatica*, 44(10):2508–2516, 2008.

Jordan B Pollack. Recursive distributed representations. *Artificial Intelligence*, 46(1-2):77–105, 1990.

Luca Pulina and Armando Tacchella. An abstraction-refinement approach to verification of artificial neural networks. In *International Conference on Computer Aided Verification*, pp. 243–257. Springer, 2010.

Scott Reed and Nando de Freitas. Neural programmer-interpreters. In *Proceedings of the 4th International Conference on Learning Representations*, San Juan, Puerto Rico, 2016.

Max Revay and Ian Manchester. Contracting implicit recurrent neural networks: Stable models with improved trainability. In *Learning for Dynamics and Control*, pp. 393–403. PMLR, 2020.

Tim Rocktäschel and Sebastian Riedel. End-to-end differentiable proving. *Advances in neural information processing systems*, 30, 2017.

Gagandeep Singh, Timon Gehr, Markus Püschel, and Martin Vechev. An abstract domain for certifying neural networks. *Proceedings of the ACM on Programming Languages*, 3(POPL):1–30, 2019.

Paul Smolensky. Tensor product variable binding and the representation of symbolic structures in connectionist systems. *Artificial Intelligence*, 46(1–2):159–216, 1990.

Paulo Tabuada. An approximate simulation approach to symbolic control. *IEEE Transactions on Automatic Control*, 53(6):1406–1418, 2008.

A.1 FULL DEFINITIONS AND BASELINE CERTIFICATES

A.1.1 FULL STATEMENT: QUOTIENT SYMBOLIC TRANSITION SYSTEMS

This section defines the symbolic states, operators, and transitions that neural execution is meant to realize. Semantically equivalent states are treated as a single quotient state. A symbolic state may still have several valid successors, so an explicit choice selects one transition branch and makes each valid step determine a unique successor.

Definition 2 (Typed quotient state space). *Let* Ty *be a set of types. For each* $\tau \in \mathsf{Ty}$*, let* S_τ *be a symbolic state set, and let* $\sim_\tau$ *be an equivalence relation on* S_τ*. Write* $\bar{S}_\tau = S_\tau/\sim_\tau$*. We call the indexed family* $(\bar{S}_\tau)_{\tau\in\mathsf{Ty}}$ *a typed quotient state space, and we call each element* $\bar{s} \in \bar{S}_\tau$ *a quotient state of type* τ*. We regard quotient states as type-tagged elements, so state sets of different types are disjoint.*

Definition 3 (Typed symbolic operator). *Let* $\tau_0, \dots, \tau_k \in \mathsf{Ty}$ *be types, and let* $\bar{\alpha} : \bar{S}_{\tau_1} \times \cdots \times \bar{S}_{\tau_k} \longrightarrow \bar{S}_{\tau_0}$ *be a map. We call* $\bar{\alpha}$ *a typed symbolic operator of type* $(\tau_1, \dots, \tau_k) \to \tau_0$*.*

A map $\alpha : S_{\tau_1} \times \cdots \times S_{\tau_k} \longrightarrow S_{\tau_0}$ induces a well-defined typed symbolic operator on the quotient state spaces if and only if, for all $x_i, y_i \in S_{\tau_i}$, $i \in \{1, \dots, k\}$, whenever $x_i \sim_{\tau_i} y_i$ for every $i \in \{1, \dots, k\}$, we have $\alpha(x_1, \dots, x_k) \sim_{\tau_0} \alpha(y_1, \dots, y_k)$.

Definition 4 (Choice-determined transition). *Let* r *be a transition label from type* τ *to* τ'*, let* C_r *be a symbolic choice space, and let* $\bar{r} : \bar{S}_\tau \times C_r \rightharpoonup \bar{S}_{\tau'}$ *be a partial map. We call* $\bar{r}$ *a choice-determined transition. We call a pair* $(\bar{s}, c)$ *valid when* $\bar{r}(\bar{s}, c)$ *is defined.*

A nondeterministic symbolic transition can be represented in this form when each transition branch is identified by a choice. We assume that c resolves all symbolic nondeterminism and contains every branch, rule instance, witness, selected premise, input symbol, or other datum needed to determine the successor quotient state. If multiple symbolic successors remain possible for a fixed quotient state and choice, the symbolic transition must be represented as a set-valued map, and correctness must relate neural successors to the set of admissible symbolic successors.

Definition 5 (Symbolic trace). *Let* $\pi : \bar{s}_0 \xrightarrow{(r_1,c_1)} \bar{s}_1 \xrightarrow{(r_2,c_2)} \cdots \xrightarrow{(r_m,c_m)} \bar{s}_m$ *be a finite sequence. We call* π *a symbolic trace if* $\bar{r}_t(\bar{s}_{t-1}, c_t) = \bar{s}_t$ *for every* $1 \le t \le m$*.*

If $\bar{r}(\bar{s}, c)$ is undefined, then $(\bar{s}, c)$ is not a symbolic transition step and cannot occur in a valid symbolic trace. All results below are stated for valid symbolic traces.

A.1.2 FULL STATEMENT: NEURAL EXECUTION SEMANTICS

Definition 6 (Typed neural state space). *Let* $\mathfrak{F} = \{(F_\tau, d_\tau) : \tau \in \mathsf{Ty}\}$ *be a family of metric spaces. We call* $\mathfrak{F}$ *a typed neural state space. In the finite-dimensional setting,* F_τ *is a subset of a normed vector space and* d_τ *is the induced metric.*

For $f \in F_\tau$ and $r \ge 0$, write $\mathbb{B}_\tau(f, r) = \{g \in F_\tau : d_\tau(f, g) \le r\}$. When $F_\tau = \mathbb{R}^d$ is equipped with the ℓ_∞ metric, $d_\tau(x, y) = \|x - y\|_\infty$ and $\mathbb{B}_\tau(x, r) = \mathbb{B}_\infty(x, r) = \{y \in \mathbb{R}^d : \|y - x\|_\infty \le r\}$.

Definition 7 (Encoder, decoder, and decoder cell). *Let* $\tau \in \mathsf{Ty}$ *be a type, and let* $\eta_\tau : \bar{S}_\tau \to F_\tau, \delta_\tau : F_\tau \to \bar{S}_\tau$ *be maps satisfying* $\delta_\tau(\eta_\tau(\bar{s})) = \bar{s}$ *for every* $\bar{s} \in \bar{S}_\tau$*. We call* η_τ *an encoder and* δ_τ *a decoder, and we call this condition code consistency. Let* $\mathcal{C}_\tau(\bar{s}) = \delta_\tau^{-1}(\{\bar{s}\}) = \{f \in F_\tau : \delta_\tau(f) = \bar{s}\}$*. We call* $\mathcal{C}_\tau(\bar{s})$ *the decoder cell of* $\bar{s}$*. Let* $\rho_\tau : \bar{S}_\tau \to \mathbb{R}_{\ge 0}$ *be a function satisfying* $\mathbb{B}_\tau(\eta_\tau(\bar{s}), \rho_\tau(\bar{s})) \subseteq \mathcal{C}_\tau(\bar{s})$ *for every* $\bar{s} \in \bar{S}_\tau$*. We call* ρ_τ *a state-dependent certified decoding radius.*

Code consistency implies that each encoder η_τ is injective.

Definition 8 (Neural module and nominal neural transition map). *Let* $\bar{\alpha} : \bar{S}_{\tau_1} \times \cdots \times \bar{S}_{\tau_k} \to \bar{S}_{\tau_0}$ *be a symbolic operator, and let* $T_\alpha : F_{\tau_1} \times \cdots \times F_{\tau_k} \to F_{\tau_0}$ *be a map. We call* T_α *a neural module for* $\bar{\alpha}$*. Let* r *be a transition label from type* τ *to type* τ'*, let* G_r *be a space, let* $\chi_r : C_r \to G_r$ *be a map, and let* $T_r : F_\tau \times G_r \to F_{\tau'}$ *be a map. We call* G_r *a neural choice space,* χ_r *a choice encoder, and* T_r *a nominal neural transition map.*

To represent nonempty sets of possible neural successors, for any set X, let $\mathcal{P}_+(X) = \{A \subseteq X : A \ne \varnothing\}$ denote the collection of all nonempty subsets of X. For a neural state $f \in F_\tau$ and a

symbolic choice $c \in C_r$, the value $T_r(f, \chi_r(c))$ gives the deterministic nominal successor, while the admissible successor map introduced below is evaluated at the same encoded choice $\chi_r(c)$ and collects all neural successors allowed by perturbations, approximation errors, or stochastic updates.

Definition 9 (Serial neural execution semantics). *Let $I_\tau : \bar{S}_\tau \to \mathcal{P}_+(F_\tau)$ be a function. We call I_τ an admissible initial-set map. Let r be a transition label from τ to τ', and let $\mathcal{D}_r : F_\tau \times G_r \to \mathcal{P}_+(F_{\tau'})$ be a function. We call $\mathcal{D}_r$ an admissible neural successor map. For $A \subseteq F_\tau$ and $g \in G_r$, define $\mathcal{D}_r(A, g) = \bigcup_{f \in A} \mathcal{D}_r(f, g)$. We call the successor semantics deterministic and noiseless when $\mathcal{D}_r(f, g) = \{T_r(f, g)\}$ for every $f \in F_\tau$ and $g \in G_r$.*

Because $\mathcal{D}_r$ takes values in $\mathcal{P}_+(F_{\tau'})$, every pair $(f, g) \in F_\tau \times G_r$ has at least one admissible successor. Thus the neural execution semantics is serial. Deadlocks, rejected updates, and numerical failures are not modeled in the present framework; incorporating them would require an explicit failure state and a totalized successor semantics.

Definition 10 (Admissible trajectory and generated tube). *Let $\pi : \bar{s}_0 \xrightarrow{(r_1,c_1)} \cdots \xrightarrow{(r_m,c_m)} \bar{s}_m$ be a symbolic trace. For each $1 \le t \le m$, define the step-specific successor map on individual neural states by $\mathcal{D}_t(f) = \mathcal{D}_{r_t}(f, \chi_{r_t}(c_t))$ for every $f \in F_{\tau_{t-1}}$, and define its action on sets of neural states by $\mathcal{D}_t(A) = \mathcal{D}_{r_t}(A, \chi_{r_t}(c_t))$ for every $A \subseteq F_{\tau_{t-1}}$. Let $f_0, \dots, f_m$ be a sequence satisfying $f_0 \in I_{\tau_0}(\bar{s}_0)$ and $f_t \in \mathcal{D}_t(f_{t-1})$ for every $1 \le t \le m$. We call $f_0, \dots, f_m$ an admissible neural trajectory along π. Let $\mathsf{Tube}_0^\pi = I_{\tau_0}(\bar{s}_0), \mathsf{Tube}_t^\pi = \mathcal{D}_t(\mathsf{Tube}_{t-1}^\pi)$ for every $1 \le t \le m$. We call $\mathsf{Tube}^\pi = (\mathsf{Tube}_0^\pi, \dots, \mathsf{Tube}_m^\pi)$ the generated neural tube along π.*

Definition 11 (Pathwise decoded correctness). *Let $\pi : \bar{s}_0 \xrightarrow{(r_1,c_1)} \cdots \xrightarrow{(r_m,c_m)} \bar{s}_m$ be a valid symbolic trace, with $\bar{s}_t \in \bar{S}_{\tau_t}$ for every $t \in \{0, \dots, m\}$. We say that the neural execution is* decoded-correct *along π if every admissible neural trajectory $f_0, \dots, f_m$ along π satisfies $\delta_{\tau_t}(f_t) = \bar{s}_t$ for every $t \in \{0, \dots, m\}$. For a family $\mathcal{K}$ of valid symbolic traces, we say that the neural execution is* decoded-correct on $\mathcal{K}$ *if it is decoded-correct along every trace $\pi \in \mathcal{K}$.*

A.1.3 Full statement: Running Example: A Cyclic Counter

Fix an integer $n \ge 3$ and a type τ. Let $\bar{S}_\tau = \{q_0, \dots, q_{n-1}\}$ be a quotient state set of type τ with $|\bar{S}_\tau| = n$. Let r be a transition label from τ to τ, let $C_r = \{c_*\}$, and define $\bar{r}(q_i, c_*) = q_{i+1 \bmod n}$ for every $i \in \{0, \dots, n-1\}$. For simplicity, we use the standard basis vectors as canonical encodings of the quotient states only in this example, not as an assumption of the general framework. Let $F_\tau = \mathbb{R}^n$ be equipped with the ℓ_∞ norm and its induced metric $d_\tau(x, y) = \|x - y\|_\infty$. Let $e_0, \dots, e_{n-1}$ denote the standard basis vectors of $\mathbb{R}^n$ under zero-based indexing, and define the encoder by $\eta_\tau(q_i) = e_i$ for every $i \in \{0, \dots, n-1\}$. Define $\delta_\tau(f) = q_j$ whenever e_j is a nearest basis vector to f, and we use a fixed rule to resolve ties. Then $\delta_\tau(\eta_\tau(q_i)) = q_i$ for every i. For any fixed $\rho \in [0, 1/2)$, define $\rho_\tau(q_i) = \rho$ for every $i \in \{0, \dots, n-1\}$. Then $\mathbb{B}_\tau(e_i, \rho) = \mathbb{B}_\infty(e_i, \rho) \subseteq \mathcal{C}_\tau(q_i)$ for every $i \in \{0, \dots, n-1\}$. Let $G_r = \{g_*\}$, define $\chi_r(c_*) = g_*$, and let Π_{cyc} be the cyclic permutation matrix satisfying $\Pi_{\mathrm{cyc}} e_i = e_{i+1 \bmod n}$. Define the nominal neural transition map by $T_r(f, g_*) = \Pi_{\mathrm{cyc}} f$.

The map $f \mapsto \Pi_{\mathrm{cyc}} f$ is the chosen nominal realization of the cyclic transition and is exact on every canonical encoding. Agreement on the canonical encodings, however, does not control behavior on noncanonical states. For every $i \in \{0, \dots, n-1\}$ and every $\varepsilon > 0$, there exists a finite feed-forward ReLU map $T : \mathbb{R}^n \to \mathbb{R}^n$ such that $T(e_j) = \Pi_{\mathrm{cyc}} e_j$ for every j, but for some f satisfying $0 < \|f - e_i\|_\infty < \varepsilon$ and $\delta_\tau(f) = q_i$, one has $\delta_\tau(T(f)) \ne q_{i+1 \bmod n}$. The explicit construction and its parameter dependence are given in Proposition 8.

The cyclic counter also separates outer and inner reachable-set guarantees. Suppose that the exact map Π_{cyc} is used and $\mathcal{D}(f) = \{\Pi_{\mathrm{cyc}} f + \xi : \xi \in \Xi\}$. If $\Xi \subseteq \mathbb{B}_\infty(0, \sigma)$, then, from a canonical initial state e_i, $\mathsf{Tube}_t \subseteq \mathbb{B}_\infty(e_{i+t \bmod n}, t\sigma)$. This outer inclusion bounds the maximum deviation but does not imply that every point of the ball is reachable. By contrast, if $\Xi = \mathbb{B}_\infty(0, \sigma)$, then $\mathsf{Tube}_t = \mathbb{B}_\infty(e_{i+t \bmod n}, t\sigma)$. Hence $t\sigma$ is an inner reachability radius in the full-ball disturbance case.

A.1.4 Full statement: Operator Execution and Recanonicalization

Definition 12 (Correctness and exactness on canonical inputs). *Let $\bar{\alpha} : \bar{S}_{\tau_1} \times \cdots \times \bar{S}_{\tau_k} \to \bar{S}_{\tau_0}$. A neural module $T_\alpha : F_{\tau_1} \times \cdots \times F_{\tau_k} \to F_{\tau_0}$ is correct on canonical inputs for $\bar{\alpha}$ if, for every*

$(\bar{s}_1, \ldots, \bar{s}_k) \in \bar{S}_{\tau_1} \times \cdots \times \bar{S}_{\tau_k}$, $\delta_{\tau_0}\left(T_\alpha\left(\eta_{\tau_1}(\bar{s}_1), \ldots, \eta_{\tau_k}(\bar{s}_k)\right)\right) = \bar{\alpha}(\bar{s}_1, \ldots, \bar{s}_k)$. *It realizes $\bar{\alpha}$ exactly on canonical inputs if, for every such tuple, $T_\alpha\left(\eta_{\tau_1}(\bar{s}_1), \ldots, \eta_{\tau_k}(\bar{s}_k)\right) = \eta_{\tau_0}\left(\bar{\alpha}(\bar{s}_1, \ldots, \bar{s}_k)\right)$.*

Definition 13 (Symbolic derivation tree and neural execution)**.** *A finite symbolic derivation tree is a finite rooted tree in which each node v has a type $\tau_v \in \mathsf{Ty}$. Each leaf v is labeled by a quotient state $\bar{s}_v \in \bar{S}_{\tau_v}$. Each internal node v with ordered children $v_1, \ldots, v_k$ is labeled by a symbolic operator $\bar{\alpha}_v : \bar{S}_{\tau_{v_1}} \times \cdots \times \bar{S}_{\tau_{v_k}} \to \bar{S}_{\tau_v}$.*

The exact symbolic value at each internal node is defined recursively by $\bar{s}_v = \bar{\alpha}_v\left(\bar{s}_{v_1}, \ldots, \bar{s}_{v_k}\right)$.

The direct execution of the tree assigns $f_v = \eta_{\tau_v}(\bar{s}_v)$ at each leaf and $f_v = T_{\alpha_v}\left(f_{v_1}, \ldots, f_{v_k}\right)$ at each internal node.

Definition 14 (Closure of certified neighborhoods)**.** *A neural module T_α satisfies closure of certified neighborhoods if, for every input tuple $\bar{s}$ and output $\bar{s}_0 = \bar{\alpha}(\bar{s})$, $T_\alpha\left(\prod_{j=1}^k \mathbb{B}_{\tau_j}(\eta_{\tau_j}(\bar{s}_j), \rho_{\tau_j}(\bar{s}_j))\right) \subseteq \mathbb{B}_{\tau_0}(\eta_{\tau_0}(\bar{s}_0), \rho_{\tau_0}(\bar{s}_0))$.*

Theorem 6 (Neighborhood closure gives compositional correctness)**.** *Assume every operator module in a finite symbolic derivation tree satisfies closure of certified neighborhoods. Initialize each leaf at its canonical encoding and evaluate the tree by direct execution. Then, for every node v, $f_v \in \mathbb{B}_{\tau_v}\left(\eta_{\tau_v}(\bar{s}_v), \rho_{\tau_v}(\bar{s}_v)\right)$. Consequently, $\delta_{\tau_v}(f_v) = \bar{s}_v$ for every node v.*

A complementary numerical certificate is stated in Appendix A.7. It assigns each internal node v a point error ϵ_v and childwise Lipschitz constants $L_{v,j}$, and recursively defines $B_v = \epsilon_v + \sum_{j=1}^k L_{v,j} B_{v_j}$, with $B_v = 0$ at leaves. Theorem 11 shows that $B_v \le \rho_{\tau_v}(\bar{s}_v)$ at every node is sufficient for decoded correctness of the direct derivation-tree execution. A module output may decode correctly without matching the canonical encoding of the corresponding symbolic state, and may be unsuitable as a direct input to the next module. We use exact external canonicalization to address this mismatch by decoding each intermediate output and re-encoding the resulting symbolic state before it is passed to the next module.

Definition 15 (Exact external canonicalizer)**.** *For each type τ, let $P_\tau = \eta_\tau \circ \delta_\tau : F_\tau \to F_\tau$ be the map defined by $P_\tau(f) = \eta_\tau\left(\delta_\tau(f)\right)$. We call P_τ the exact external canonicalizer of type τ. By code consistency, $P_\tau\left(P_\tau(f)\right) = P_\tau(f)$ and $\delta_\tau\left(P_\tau(f)\right) = \delta_\tau(f)$ for every $f \in F_\tau$.*

Definition 16 (Recanonicalized derivation-tree execution)**.** *Let a finite symbolic derivation tree be given. At each leaf v, set $f_v = \eta_{\tau_v}(\bar{s}_v)$. After the state f_u has been computed at a non-root node u, set $\widehat{f}_u = P_{\tau_u}(f_u)$. If v is an internal node with ordered children $v_1, \ldots, v_k$, define $f_v = T_{\alpha_v}\left(\widehat{f}_{v_1}, \ldots, \widehat{f}_{v_k}\right)$. We call the resulting assignment the recanonicalized execution of the tree.*

Recanonicalization leaves the decoded quotient state unchanged, but applies the external map $P_\tau = \eta_\tau \circ \delta_\tau$ between modules without assuming a neural implementation.

Theorem 7 (Canonical-input correctness composes under exact external canonicalization)**.** *Assume that, for every internal node v, the neural module T_{α_v} is correct on canonical inputs for $\bar{\alpha}_v$. Under the recanonicalized execution of Definition 16, $\delta_{\tau_v}(f_v) = \bar{s}_v$ for every node v.*

The following example shows that ensuring correct decoding for each module on canonical inputs is not sufficient to guarantee correctness under direct composition.

Example 1 (Canonical-input correctness with an incorrect composition)**.** *Let $\bar{S} = \{0, 1\}$, $\bar{\alpha} = \bar{\beta} = \mathrm{id}_{\bar{S}}$, and $F = \mathbb{R}$. Define $\eta(0) = 0$, $\eta(1) = 1$, and $\delta(x) = \begin{cases} 0, & x < 1/2, \\ 1, & x \ge 1/2. \end{cases}$ Let $T_\alpha(x) = 0.4 + 0.6x$ and $T_\beta(x) = 2.5x$. Both modules decode correctly on the canonical inputs for the identity: $\delta(T_\alpha(\eta(s))) = \delta(T_\beta(\eta(s))) = s$ for $s \in \{0, 1\}$. However, $\delta\left(T_\beta(T_\alpha(\eta(0)))\right) = \delta(1) = 1 \neq 0 = \bar{\beta}(\bar{\alpha}(0))$. Thus correctness on canonical inputs is not preserved by direct composition.*

A.1.5 Full statement: Inner Reachability

The previous section derived outer-radius certificates that enclose the reachable tube. This section studies the reverse inclusion. We give conditions under which a full norm ball around the target encoding is contained in the reachable tube and derive a recursive lower bound on its radius.

Definition 17 (Closed norm ball)**.** *Let $(E, \|\cdot\|_E)$ be a real normed linear space. For $x \in E$ and $r \geq 0$, define $\mathbb{B}_E(x, r) = \{y \in E : \|y - x\|_E \leq r\}$.*

Definition 18 (Local openness coefficient)**.** *Let $(E, \|\cdot\|_E)$ and $(E', \|\cdot\|_{E'})$ be real normed linear spaces, and let $T : E \to E'$ be a map. We say that T is locally open at $x \in E$ with coefficient $\mu \geq 0$ up to radius $a \geq 0$ if $T\left(\mathbb{B}_E(x, r)\right) \supseteq \mathbb{B}_{E'}\left(T(x), \mu r\right)$ for every $r \in [0, a]$.*

Definition 19 (Full-ball disturbance inclusion)**.** *Let $(E, \|\cdot\|_E)$ and $(E', \|\cdot\|_{E'})$ be real normed linear spaces. Let $T : E \to E'$ be a nominal map, let $\mathcal{D} : E \to \mathcal{P}_+(E')$ be a successor map, let $A \subseteq E$, and let $v \geq 0$. We say that $\mathcal{D}$ satisfies full-ball disturbance inclusion relative to T on A with radius v if $\mathbb{B}_{E'}\left(T(f), v\right) \subseteq \mathcal{D}(f)$ for every $f \in A$.*

Theorem 8 (Inner-radius propagation)**.** *Fix a symbolic trace $\pi : \bar{s}_0 \xrightarrow{(r_1,c_1)} \cdots \xrightarrow{(r_m,c_m)} \bar{s}_m$. For every $t \in \{0, \ldots, m\}$, assume that F_{τ_t} is a real normed linear space with norm $\|\cdot\|_{\tau_t}$ and that the metric used by the neural execution semantics is the induced metric, $d_{\tau_t}(x, y) = \|x - y\|_{\tau_t}$ for all $x, y \in F_{\tau_t}$. Thus the typed metric ball $\mathbb{B}_{\tau_t}(x, r)$ coincides with the norm ball $\mathbb{B}_{F_{\tau_t}}(x, r)$.*

Let $a_0 \geq 0$ and assume $\mathbb{B}_{F_{\tau_0}}\left(\eta_{\tau_0}(\bar{s}_0), a_0\right) \subseteq \mathsf{Tube}_0^\pi$. For each $t \in \{1, \ldots, m\}$, suppose that the following conditions hold with the already defined radius a_{t-1}.

(i) The map $T_t : F_{\tau_{t-1}} \to F_{\tau_t}$ is locally open at $\eta_{\tau_{t-1}}(\bar{s}_{t-1})$ with coefficient $\mu_t \geq 0$ up to radius a_{t-1}.

(ii) The successor map $\mathcal{D}_t$ satisfies full-ball disturbance inclusion relative to T_t on $\mathbb{B}_{F_{\tau_{t-1}}}\left(\eta_{\tau_{t-1}}(\bar{s}_{t-1}), a_{t-1}\right)$ with radius $v_t \geq 0$.

(iii) With $\epsilon_t = \left\| T_t\left(\eta_{\tau_{t-1}}(\bar{s}_{t-1})\right) - \eta_{\tau_t}(\bar{s}_t) \right\|_{\tau_t}$, one has $\mu_t a_{t-1} + v_t \geq \epsilon_t$.

Define recursively $a_t = \mu_t a_{t-1} + v_t - \epsilon_t$ for $t \in \{1, \ldots, m\}$. Then $a_t \geq 0$ and $\mathbb{B}_{F_{\tau_t}}\left(\eta_{\tau_t}(\bar{s}_t), a_t\right) \subseteq \mathsf{Tube}_t^\pi$ for every $t \in \{0, \ldots, m\}$.

When $\mu_t a_{t-1} + v_t < \epsilon_t$, the preceding argument does not produce an inner ball centered at the target encoding. In particular, the negative quantity $\mu_t a_{t-1} + v_t - \epsilon_t$ cannot be truncated to zero, because the assumptions do not guarantee that the target encoding itself is reachable.

Corollary 4 (Exact transitions with unit openness and full-ball disturbances)**.** *Fix a symbolic trace π with m transitions. For every $t \in \{0, \ldots, m\}$, assume that F_{τ_t} is a real normed linear space with norm $\|\cdot\|_{\tau_t}$ and $d_{\tau_t}(x, y) = \|x - y\|_{\tau_t}$. Assume $\mathsf{Tube}_0^\pi = \{\eta_{\tau_0}(\bar{s}_0)\}$. Let $\sigma_t \geq 0$ for $t \in \{1, \ldots, m\}$, and define $a_0 = 0,\ a_t = a_{t-1} + \sigma_t$. For every $t \in \{1, \ldots, m\}$, assume:*

(i) $T_t(\eta_{\tau_{t-1}}(\bar{s}_{t-1})) = \eta_{\tau_t}(\bar{s}_t)$.

(ii) T_t is locally open at $\eta_{\tau_{t-1}}(\bar{s}_{t-1})$ with coefficient 1 up to radius a_{t-1}.

(iii) For every $f \in \mathbb{B}_{F_{\tau_{t-1}}}\left(\eta_{\tau_{t-1}}(\bar{s}_{t-1}), a_{t-1}\right)$, $\mathbb{B}_{F_{\tau_t}}\left(T_t(f), \sigma_t\right) \subseteq \mathcal{D}_t(f)$.

Then $\mathbb{B}_{F_{\tau_t}}\left(\eta_{\tau_t}(\bar{s}_t), a_t\right) \subseteq \mathsf{Tube}_t^\pi$ for every $t \in \{0, \ldots, m\}$. Moreover, $a_t = \sum_{j=1}^t \sigma_j$. If $\sigma_t = \sigma$ at every step, then $a_t = t\sigma$.

Remark 2 (Surjective isometries)**.** *A surjective isometry between normed linear spaces maps every closed norm ball onto the closed norm ball of the same radius. Its local openness coefficient is therefore 1. Surjectivity is essential: a non-surjective isometric embedding need not contain any positive-radius ball in the ambient output space.*

Proposition 3 (Affine openness certificate)**.** *Let $(F, \|\cdot\|_F)$ be a finite-dimensional normed linear space with $\dim F \geq 1$. Let $A : F \to F$ be an invertible linear map, let $b \in F$, and define $T(f) = Af + b$. Then, for every $x \in F$ and $r \geq 0$, $T\left(\mathbb{B}_F(x, r)\right) \supseteq \mathbb{B}_F\left(T(x), \mu r\right)$, where $\mu = \frac{1}{\|A^{-1}\|}$. If A is a surjective linear isometry, one may take $\mu = 1$.*

A rank-deficient affine map can satisfy an outer Lipschitz bound while having no positive local-openness coefficient in the ambient output space. Thus an outer Lipschitz bound alone does not provide an inner-reachability guarantee.

A.1.6 FULL STATEMENT: CAPACITY FROM INNER REACHABILITY

Under terminal robustness, inner reachable balls for distinct terminal states are pairwise disjoint, yielding packing bounds on the number of robustly representable states.

Let $T \in \mathbb{Z}_{\geq 0}$, and let $\mathcal{K}_T \neq \varnothing$ be a possibly infinite family of valid symbolic traces, each with exactly T transitions. Assume that all states reached at horizon T have the same type τ, and write $\bar{s}_T^{\pi}$ for the terminal state of $\pi \in \mathcal{K}_T$. Define $S_T = \{\bar{s}_T^{\pi} : \pi \in \mathcal{K}_T\}$. Assume that S_T is finite, and let $N_T = |S_T| \geq 1$.

Definition 20 (Terminal robustness at a horizon)**.** *Let $\delta_\tau : F_\tau \to \bar{S}_\tau$ be a deterministic decoder such that, for every $\pi \in \mathcal{K}_T$ and every admissible trajectory $f_0, \ldots, f_T$ along π, one has $\delta_\tau(f_T) = \bar{s}_T^{\pi}$. We call δ_τ terminally robust for $\mathcal{K}_T$ at horizon T.*

We use the term horizon (Bertsekas, 2012) to denote the number of transition steps covered by the analysis or execution.

Definition 21 (Terminal reachable set)**.** *Let $\bar{s} \in \bar{S}_\tau$ be a quotient state occurring at horizon T. Let $A_T(\bar{s}) = \bigcup \{\mathsf{Tube}_T^{\pi} : \pi \in \mathcal{K}_T \text{ and } \bar{s}_T^{\pi} = \bar{s}\}$. We call $A_T(\bar{s})$ the terminal reachable set of $\bar{s}$ at horizon T.*

Proposition 4 (Terminal decoder-cell characterization)**.** *A deterministic decoder is terminally robust for $\mathcal{K}_T$ at horizon T if and only if $A_T(\bar{s}) \subseteq \mathcal{C}_\tau(\bar{s})$ for every quotient state $\bar{s}$ occurring at horizon T. Whenever these equivalent conditions hold, $A_T(\bar{s}) \cap A_T(\bar{s}') = \varnothing$ for any two distinct horizon-T quotient states $\bar{s}$ and $\bar{s}'$.*

Terminal robustness guarantees correctness only at the final time T, whereas pathwise decoded correctness guarantees preservation of the intended symbolic semantics at every step of the execution.

Definition 22 (Common inner reachability radius)**.** *Let $\varnothing \neq S_T \subseteq \bar{S}_\tau$ be a set of quotient states occurring at horizon T, and let $a_T > 0$. We call a_T a common inner reachability radius for S_T if $\mathbb{B}_\tau(\eta_\tau(\bar{s}), a_T) \subseteq A_T(\bar{s})$ for every $\bar{s} \in S_T$.*

Definition 23 (Packing numbers)**.** *Let (F, d) be a metric space, let $\varnothing \neq K \subseteq F$, and let $r > 0$. For $x \in F$ and $s \geq 0$, write $\mathbb{B}_F(x, s) = \{y \in F : d(x, y) \leq s\}$. Here, X denotes a nonempty finite subset of K, and $|X|$ denotes its cardinality.*

Define the packing number by $\mathrm{Pack}(K, r) = \sup\Big\{|X| : \varnothing \neq X \subseteq K,\ X \text{ is finite}, d(x, y) \geq r \text{ for all } x, y \in X \text{ with } x \neq y\Big\} \in \mathbb{N} \cup \{\infty\}$.

Define the ambient ball-packing number by $\mathrm{BPack}_F(K, r) = \sup\Big\{|X| : \varnothing \neq X \subseteq K,\ X \text{ is finite}, \mathbb{B}_F(x, r) \cap \mathbb{B}_F(y, r) = \varnothing \text{ for all } x, y \in X \text{ with } x \neq y\Big\} \in \mathbb{N} \cup \{\infty\}$.

In either definition, the value is ∞ exactly when packings of arbitrarily large finite cardinality exist.

The packing number (Kolmogorov & Tikhomirov, 1959) $\mathrm{Pack}(K, r)$ quantifies the supremum of the cardinalities of finite r-separated subsets of K, whereas the ambient ball-packing number $\mathrm{BPack}_F(K, r)$ quantifies the supremum of the numbers of pairwise disjoint ambient balls of radius r with centers in K. The ambient ball-packing number gives the general upper bound, while the packing number does so when F has the midpoint property.

Theorem 9 (General horizon packing bound)**.** *Let $\varnothing \neq S_T \subseteq \bar{S}_\tau$ be a finite set of quotient states occurring at horizon T, and let $N_T = |S_T| \geq 1$. Let $K \subseteq F_\tau$ contain $\eta_\tau(\bar{s})$ for every $\bar{s} \in S_T$. Assume that $a_T > 0$ is a common inner reachability radius for S_T and that the decoder is terminally robust at horizon T. Then $N_T \leq \mathrm{BPack}_{F_\tau}(K, a_T)$.*

In addition, if F_τ has the midpoint property, meaning that for every $x, y \in F_\tau$ there exists $z \in F_\tau$ such that $d(x, z) = d(y, z) = \frac{d(x,y)}{2}$, then $N_T \leq \mathrm{Pack}(K, 2a_T)$.

The general packing theorem also yields metric-entropy and doubling-dimension obstructions. Precise definitions and quantitative specializations are given in Appendix Corollaries 7 and 8.

A.1.6.1 DIGITAL ADDRESSING AND STATE-QUANTIZED EXECUTION

Digital address capacity and robustness under quantization are different questions. Address capacity is a cardinality constraint. Quantization robustness requires a quantizer to be placed explicitly in the execution semantics and requires its approximation error to be propagated through the reachable tube.

Definition 24 (State-quantized execution). *For each type $\tau \in \mathsf{Ty}$, let $\mathcal{Q}_\tau \subseteq F_\tau$ be a set of quantized neural states, and let $\mathsf{Q}_\tau : F_\tau \to \mathcal{Q}_\tau$ be a deterministic quantizer. For $A \subseteq F_\tau$, write $\mathsf{Q}_\tau(A) = \{\mathsf{Q}_\tau(f) : f \in A\}$. Along a symbolic trace π, define the quantized initial set and quantized successor maps by $I^{\mathsf{Q}}_{\tau_0}(\bar{s}_0) = \mathsf{Q}_{\tau_0}\left(I_{\tau_0}(\bar{s}_0)\right)$ and $\mathcal{D}^{\mathsf{Q}}_t(f) = \mathsf{Q}_{\tau_t}\left(\mathcal{D}_t(f)\right) = \{\mathsf{Q}_{\tau_t}(y) : y \in \mathcal{D}_t(f)\}$. The corresponding quantized tube is defined by $\mathsf{Tube}^{\pi,\mathsf{Q}}_0 = I^{\mathsf{Q}}_{\tau_0}(\bar{s}_0),\ \mathsf{Tube}^{\pi,\mathsf{Q}}_t = \mathcal{D}^{\mathsf{Q}}_t\left(\mathsf{Tube}^{\pi,\mathsf{Q}}_{t-1}\right)$. A sequence $f_0, \dots, f_m$ is an admissible quantized trajectory along π if $f_0 \in I^{\mathsf{Q}}_{\tau_0}(\bar{s}_0)$ and $f_t \in \mathcal{D}^{\mathsf{Q}}_t(f_{t-1})$ for every $t \in \{1, \dots, m\}$. We say that the state-quantized execution is decoded-correct along π if every admissible quantized trajectory satisfies $\delta_{\tau_t}(f_t) = \bar{s}_t$ for every $t \in \{0, \dots, m\}$. Thus the initial state is quantized once, and every admissible successor is quantized after the corresponding neural update.*

Theorem 10 (Quantization-aware outer-radius propagation). *Fix a symbolic trace π with m transitions and the state-quantized execution of Definition 24. Let $b_0 \geq 0$, and for every $t \in \{1, \dots, m\}$ let $q_{t-1}, L_t, u_t \geq 0$. Let $\kappa_t \geq 0$ for every $t \in \{0, \dots, m\}$. Assume the following conditions.*

(i) The unquantized initial set satisfies $I_{\tau_0}(\bar{s}_0) \subseteq \mathbb{B}_{\tau_0}\left(\eta_{\tau_0}(\bar{s}_0), b_0\right)$.

(ii) The initial quantization error satisfies $d_{\tau_0}\left(\mathsf{Q}_{\tau_0}(y), y\right) \leq \kappa_0$ for every $y \in I_{\tau_0}(\bar{s}_0)$.

(iii) For every $t \in \{1, \dots, m\}$, the nominal map T_t is L_t-Lipschitz on $\mathbb{B}_{\tau_{t-1}}\left(\eta_{\tau_{t-1}}(\bar{s}_{t-1}), q_{t-1}\right)$.

(iv) For every $t \in \{1, \dots, m\}$ and every $f \in \mathbb{B}_{\tau_{t-1}}\left(\eta_{\tau_{t-1}}(\bar{s}_{t-1}), q_{t-1}\right)$, one has $\mathcal{D}_t(f) \subseteq \mathbb{B}_{\tau_t}\left(T_t(f), u_t\right)$.

(v) For every $t \in \{1, \dots, m\}$, every such f, and every $y \in \mathcal{D}_t(f)$, $d_{\tau_t}\left(\mathsf{Q}_{\tau_t}(y), y\right) \leq \kappa_t$.

Define $\epsilon_t = d_{\tau_t}\left(T_t\left(\eta_{\tau_{t-1}}(\bar{s}_{t-1})\right), \eta_{\tau_t}(\bar{s}_t)\right)$, and set $B^{\mathsf{Q}}_0 = b_0 + \kappa_0,\ B^{\mathsf{Q}}_t = \epsilon_t + L_t B^{\mathsf{Q}}_{t-1} + u_t + \kappa_t$. If $B^{\mathsf{Q}}_{t-1} \leq q_{t-1}$ for every $t \in \{1, \dots, m\}$, then $\mathsf{Tube}^{\pi,\mathsf{Q}}_t \subseteq \mathbb{B}_{\tau_t}\left(\eta_{\tau_t}(\bar{s}_t), B^{\mathsf{Q}}_t\right)$ for every $t \in \{0, \dots, m\}$. If, in addition, $B^{\mathsf{Q}}_t \leq \rho_{\tau_t}(\bar{s}_t)$ for every t, then the state-quantized execution is decoded-correct along π.

Proposition 5 (Digital address counting bound). *Let $\tau \in \mathsf{Ty}$, and let $p, d, N_T \in \mathbb{N}$ satisfy $p \geq 1$, $d \geq 1$, and $N_T \geq 1$. Let $\mathcal{A}_{p,d} = (\{0,1\}^p)^d$ be the set of digital addresses, let $\psi : \mathcal{A}_{p,d} \to F_\tau$ interpret each address as a neural state, and define $\mathcal{Q}_{p,d} = \psi(\mathcal{A}_{p,d})$. Let $\delta_\tau : F_\tau \to \bar{S}_\tau$ be a deterministic decoder. If there exists a set $S_T \subseteq \bar{S}_\tau$ such that $|S_T| = N_T$ and $S_T \subseteq \delta_\tau(\mathcal{Q}_{p,d})$, then $pd \geq \lceil \log_2 N_T \rceil$.*

Corollary 5 (Address requirement for quantized terminal execution). *Let $\mathcal{K}_T$ be a nonempty family of symbolic traces with exactly T transitions and a common terminal type τ. Let $S_T = \{\bar{s}^\pi_T : \pi \in \mathcal{K}_T\}$ be finite, with $|S_T| = N_T$. Assume the state-quantized execution of Definition 24 satisfies $\mathcal{Q}_\tau \subseteq \mathcal{Q}_{p,d}$ for a digital representation set $\mathcal{Q}_{p,d}$ as in Proposition 5. Assume also that every admissible quantized trajectory along every $\pi \in \mathcal{K}_T$ satisfies $\delta_\tau(f_T) = \bar{s}^\pi_T$. Then $pd \geq \lceil \log_2 N_T \rceil$.*

Theorem 10 provides a sufficient robustness certificate for explicitly quantized execution. Proposition 5 and Corollary 5 give necessary bounds on digital address capacity, while Proposition 13 gives the corresponding geometric constraint before quantization. The inner-reachability argument for the unquantized system does not apply directly to a finite quantized tube, since, in a positive-dimensional real normed space, such a tube contains no positive-radius norm ball.

A.1.7 FULL STATEMENT: NEURAL-NETWORK CERTIFICATES

This section specializes those conditions to neural networks. We derive outer and inner certificates from standard architectural quantities and show that exact interpolation of a finite operator table is not sufficient for robust multi-step execution.

Proposition 6 (Layerwise operator-norm Lipschitz certificate). *Let $m \in \mathbb{N}$ with $m \geq 1$, and let $(E_j, \|\cdot\|_j)$ be normed spaces for $j \in \{0, \dots, m\}$. For every $j \in \{1, \dots, m\}$, let $W_j : E_{j-1} \to E_j$ be a bounded linear map and let $b_j \in E_j$. For every $j \in \{1, \dots, m-1\}$, let $\phi_j : E_j \to E_j$ be 1-Lipschitz with respect to $\|\cdot\|_j$.*

Set $z_0 = x$, define $z_j = \phi_j(W_j z_{j-1} + b_j)$ for $j \in \{1, \dots, m-1\}$, and define $T(x) = W_m z_{m-1} + b_m$. Then $\mathrm{Lip}(T) \leq \prod_{j=1}^{m} \|W_j\|$, where $\|W_j\|$ is the operator norm induced by $\|\cdot\|_{j-1}$ and $\|\cdot\|_j$.

Suppose additionally that $k \in \mathbb{N}$, $k \geq 1$, each $(E_{0,i}, \|\cdot\|_{0,i})$ is a normed space, and $E_0 = E_{0,1} \times \cdots \times E_{0,k}$. Assume that there exist bounded linear maps $W_{1,i} : E_{0,i} \to E_1$, $i \in \{1, \dots, k\}$, such that $W_1(x_1, \dots, x_k) = \sum_{i=1}^{k} W_{1,i} x_i$. Then $L_i = \left(\prod_{j=2}^{m} \|W_j\|\right) \|W_{1,i}\|$ is a childwise Lipschitz constant for input block i. When $m = 1$, the empty product is understood to be 1.

The resulting layerwise bounds instantiate the Lipschitz constants in Theorem 2 and Theorem 11.

On each activation region R, a ReLU network agrees with an affine map $x \mapsto Ax + b$. Suppose that $x \in R$ and that there exists $a > 0$ such that $\mathbb{B}(x, a) \subseteq R$. If the input and output spaces coincide and A is invertible, then Proposition 3 implies that the network is locally open at x with coefficient $1/\|A^{-1}\|$ up to radius a. If A is rank deficient, no positive local-openness coefficient in the ambient output space follows from that affine representation. At a boundary point of an activation region, invertibility of the linear part on a single adjacent region is not sufficient; local openness must be verified using all activation regions intersecting the neighborhood. Thus a layerwise operator-norm bound alone does not provide the inner-reachability condition required by the capacity bounds above.

Proposition 7 (Exact ReLU interpolation of a finite operator table). *Let $X = \{x_1, \dots, x_M\} \subset \mathbb{R}^p$ be a finite set of distinct encoded input tuples, where $M \geq 1$, and let $y_1, \dots, y_M \in \mathbb{R}^q$ be prescribed encoded outputs. If $M \geq 2$, choose $0 < r < \frac{1}{2} \min_{i \neq j} \|x_i - x_j\|_\infty$. If $M = 1$, choose any $r > 0$.*

For each $i \in \{1, \dots, M\}$, define $\phi_i(x) = \max\left\{0, 1 - \frac{\|x - x_i\|_\infty}{r}\right\}$, and let $G(x) = \sum_{i=1}^{M} \phi_i(x) y_i$. Then $G(x_i) = y_i$ for every $i \in \{1, \dots, M\}$. Moreover, G is exactly realizable by a finite feed-forward ReLU network with an affine output layer.

Finite-table interpolation proves exact agreement on the selected encoded points. Complete tables with exact canonical outputs compose correctly from canonical inputs in noiseless execution when they cover every tuple used by the computation. Interpolation alone does not provide a prescribed neighborhood-closure or perturbation guarantee, uniform robustness over unbounded horizons, or systematic behavior on tuples not covered by the table. Universal approximation theorems show that neural networks can approximate continuous target functions (Cybenko, 1989; Hornik, 1991), but they do not provide the reachability or decoder-cell guarantees required here.

A.2 FINITE-RELU FAILURE ON NONCANONICAL STATES

The map $f \mapsto \Pi_{\mathrm{cyc}} f$ is the chosen nominal neural realization of the cyclic transition and satisfies $\Pi_{\mathrm{cyc}} \eta_\tau(q_i) = \eta_\tau(q_{i+1 \bmod n})$ for every i. The following proposition shows that a finite ReLU network can agree with this transition at every canonical encoding $\eta_\tau(q_i) = e_i$ while behaving differently on noncanonical states.

Proposition 8 (Finite-ReLU failure on noncanonical states). *Let $e_0, \dots, e_{n-1}$ be the standard basis vectors of $\mathbb{R}^n$, let Π_{cyc} be the cyclic permutation matrix satisfying $\Pi_{\mathrm{cyc}} e_j = e_{j+1 \bmod n}$, and let δ_τ be the nearest-basis decoder with a fixed rule for resolving ties. Fix $i \in \{0, \dots, n-1\}$ and $0 < \rho < 1/2$. For every $a \in (0, \rho/2)$, there exist $\lambda > 0$ and a map $T_\lambda : \mathbb{R}^n \to \mathbb{R}^n$ such that $T_\lambda(e_j) = \Pi_{\mathrm{cyc}} e_j$ for every $j \in \{0, \dots, n-1\}$, while, for $f_a = (1 + a)e_i$, one has $\delta_\tau(f_a) = q_i$ and $\delta_\tau(T_\lambda(f_a)) = q_i \neq q_{i+1 \bmod n}$. Consequently, for every $\varepsilon > 0$, there exist $\lambda > 0$ and $f \in \mathbb{R}^n$ such that $0 < \|f - e_i\|_\infty < \varepsilon$, $\delta_\tau(f) = q_i$, and $\delta_\tau(T_\lambda(f)) \neq q_{i+1 \bmod n}$. For every fixed $\lambda > 0$, the constructed map T_λ is exactly realizable by a finite feed-forward ReLU network with an affine output layer.*

Proof. Fix $i \in \{0, \dots, n-1\}$ and $0 < \rho < 1/2$. For $f \in \mathbb{R}^n$, define $r_i(f) = \|f - e_i\|_\infty$ and $g_i(f) = \min\{r_i(f), \max\{0, \rho - r_i(f)\}\}$. Equivalently, $g_i(f) = \begin{cases} r_i(f), & 0 \le r_i(f) \le \rho/2, \\ \rho - r_i(f), & \rho/2 < r_i(f) < \rho, \\ 0, & r_i(f) \ge \rho. \end{cases}$

For $j = i$, one has $r_i(e_i) = 0$, and therefore $g_i(e_i) = 0$. For $j \neq i$, one has $r_i(e_j) = 1 > \rho$, and hence $g_i(e_j) = 0$. Thus $g_i(e_j) = 0$ for every $j \in \{0, \dots, n-1\}$.

For $\lambda > 0$, define $T_\lambda(f) = \Pi_{\mathrm{cyc}} f + \lambda g_i(f)(e_i - e_{i+1 \bmod n})$. Since $g_i(e_j) = 0$, we obtain $T_\lambda(e_j) = \Pi_{\mathrm{cyc}} e_j$ for every canonical encoding e_j.

Now fix $a \in (0, \rho/2)$. Since $a < 1$, the interval $\left(\frac{1+a}{2}, 1\right)$ is nonempty. Choose $\theta \in \left(\frac{1+a}{2}, 1\right)$, set $\lambda = \frac{\theta}{a}$, and let $f_a = (1+a)e_i$. Then $\|f_a - e_i\|_\infty = a$, whereas $\|f_a - e_j\|_\infty = 1 + a$ for every $j \neq i$. Therefore, $\delta_\tau(f_a) = q_i$.

Because $r_i(f_a) = a < \rho/2$, we have $g_i(f_a) = a$. Using $\Pi_{\mathrm{cyc}} e_i = e_{i+1 \bmod n}$ and $\lambda a = \theta$, we obtain $T_\lambda(f_a) = \theta e_i + (1 + a - \theta) e_{i+1 \bmod n}$. Let $y = T_\lambda(f_a)$. Since $\theta > a$, $\|y - e_i\|_\infty = 1 + a - \theta$. Moreover, $\|y - e_{i+1 \bmod n}\|_\infty = \theta$. The choice $\theta > (1+a)/2$ implies $1 + a - \theta < \theta$. For every $j \notin \{i, i+1 \bmod n\}$, one has $\|y - e_j\|_\infty = 1$, while $1 + a - \theta < 1$. Thus e_i is the unique nearest basis vector to y, and therefore $\delta_\tau(T_\lambda(f_a)) = \delta_\tau(y) = q_i \neq q_{i+1 \bmod n}$.

Given $\varepsilon > 0$, choose $a \in (0, \min\{\rho/2, \varepsilon\})$. The preceding construction then gives $0 < \|f_a - e_i\|_\infty < \varepsilon$, together with the required decoding failure.

It remains to prove exact ReLU realizability. Let $\sigma_{\mathrm{sc}}(z) = \max\{z, 0\}$. The identities $|z| = \sigma_{\mathrm{sc}}(z) + \sigma_{\mathrm{sc}}(-z)$, $\max\{x, y\} = \sigma_{\mathrm{sc}}(x - y) + y$, and $\min\{x, y\} = x - \sigma_{\mathrm{sc}}(x - y)$ show that scalar absolute value, binary maximum, and binary minimum are exactly realizable using affine maps and scalar ReLU units.

Repeated application of these identities gives an exact finite ReLU realization of $r_i(f) = \max_{0 \le j \le n-1} |f_j - (e_i)_j|$ and therefore of $g_i(f) = \min\{r_i(f), \max\{0, \rho - r_i(f)\}\}$. Each coordinate can also be propagated through ReLU units using $f_j = \sigma_{\mathrm{sc}}(f_j) - \sigma_{\mathrm{sc}}(-f_j)$. Hence the affine map $f \mapsto \Pi_{\mathrm{cyc}} f$ and the scalar map $f \mapsto g_i(f)$ can be computed in parallel and combined by an affine output layer. Consequently, $T_\lambda(f) = \Pi_{\mathrm{cyc}} f + \lambda g_i(f)(e_i - e_{i+1 \bmod n})$ is exactly realizable by a finite feed-forward ReLU network with an affine output layer. □

Remark 3. *The parameter λ in Proposition 8 is allowed to depend on the offset a. The proposition does not assert that a single fixed map T_λ fails at points arbitrarily close to e_i.*

A.3 Exact finite-horizon simulation

Proof. For each $t \in \{0, \dots, m\}$, let

$$A_t = \left\{ f \in F_{\tau_t} : \begin{array}{l} \text{there exist } f_0, \dots, f_t \text{ with } f_t = f, \\ f_0 \in I_{\tau_0}(\bar{s}_0), \\ f_j \in \mathcal{D}_j(f_{j-1}) \text{ for every } j \in \{1, \dots, t\} \end{array} \right\}.$$

We first show by induction that $A_t = \mathsf{Tube}_t^\pi$ for every $t \in \{0, \dots, m\}$. At $t = 0$, both sets equal $I_{\tau_0}(\bar{s}_0)$. If $A_{t-1} = \mathsf{Tube}_{t-1}^\pi$, then $A_t = \bigcup_{f \in A_{t-1}} \mathcal{D}_t(f) = \mathcal{D}_t(A_{t-1}) = \mathcal{D}_t(\mathsf{Tube}_{t-1}^\pi) = \mathsf{Tube}_t^\pi$.

Assume (i). Fix $t \in \{0, \dots, m\}$ and $f \in \mathsf{Tube}_t^\pi = A_t$. By the definition of A_t, the state f is the endpoint of an admissible trajectory prefix of length t. Since each set $\mathcal{D}_j(g)$ is nonempty, this prefix extends to an admissible trajectory of length m. Statement (i) therefore gives $\delta_{\tau_t}(f) = \bar{s}_t$. Hence $\mathsf{Tube}_t^\pi \subseteq \mathcal{C}_{\tau_t}(\bar{s}_t)$, which proves (ii).

Conversely, assume (ii) and let $f_0, \dots, f_m$ be an admissible trajectory along π. For each $t \in \{0, \dots, m\}$, the prefix $f_0, \dots, f_t$ shows that $f_t \in A_t = \mathsf{Tube}_t^\pi$. Thus $f_t \in \mathcal{C}_{\tau_t}(\bar{s}_t)$, and consequently $\delta_{\tau_t}(f_t) = \bar{s}_t$. This proves (i).

Assume (ii). Taking $R_t^\pi = \mathsf{Tube}_t^\pi$ for every t gives (iii).

Now assume (iii). At time zero, $\mathsf{Tube}_0^\pi = I_{\tau_0}(\bar{s}_0) \subseteq R_0^\pi$. If $\mathsf{Tube}_{t-1}^\pi \subseteq R_{t-1}^\pi$, then monotonicity of $\mathcal{D}_t$ gives $\mathsf{Tube}_t^\pi = \mathcal{D}_t(\mathsf{Tube}_{t-1}^\pi) \subseteq \mathcal{D}_t(R_{t-1}^\pi) \subseteq R_t^\pi$. Therefore $\mathsf{Tube}_t^\pi \subseteq R_t^\pi$ for every t. Since $R_t^\pi \subseteq \mathcal{C}_{\tau_t}(\bar{s}_t)$, statement (ii) follows.

The same induction applies to any family satisfying only the initial and successor inclusions. Hence every such family contains Tube^π componentwise. □

A.4 Reachable failure witness

Proof. Choose $f_t \in \mathsf{Tube}_t^\pi \setminus \mathcal{C}_{\tau_t}(\bar{s}_t)$. By the reachable-set identity $A_t = \mathsf{Tube}_t^\pi$ established in the proof of Theorem 1, Tube_t^π is exactly the set of endpoints of admissible trajectory prefixes. Hence there exist states $f_0, \dots, f_{t-1}$ such that $f_0 \in I_{\tau_0}(\bar{s}_0)$ and $f_j \in \mathcal{D}_j(f_{j-1})$ for every $1 \le j \le t$. Thus $f_0, \dots, f_t$ is an admissible trajectory prefix. Moreover, since $\mathcal{C}_{\tau_t}(\bar{s}_t) = \delta_{\tau_t}^{-1}(\{\bar{s}_t\})$, the condition $f_t \notin \mathcal{C}_{\tau_t}(\bar{s}_t)$ implies $\delta_{\tau_t}(f_t) \neq \bar{s}_t$. □

A.5 Bounded-horizon trace-family characterization

Proof. For each trajectory $\pi \in \mathcal{K}_T$, write $\bar{s}_t^\pi \in \bar{S}_{\tau_t^\pi}$ for its symbolic state at time t. By Theorem 1, decoded correctness along π is equivalent to $\mathsf{Tube}_t^\pi \subseteq \mathcal{C}_{\tau_t^\pi}(\bar{s}_t^\pi)$ at every t occurring in π.

Fix $0 \le t \le T$, $\tau \in \mathsf{Ty}$, and $\bar{s} \in \bar{S}_\tau$ such that $U_t(\bar{s})$ is nonempty. By definition, $U_t(\bar{s})$ is the union of the time-t tubes of all traces in $\mathcal{K}_T$ whose time-t symbolic state is $\bar{s}$. Therefore, $U_t(\bar{s}) \subseteq \mathcal{C}_\tau(\bar{s})$ iff every such tube is contained in $\mathcal{C}_\tau(\bar{s})$. Applying this equivalence at every t, τ, and quotient state proves the result. □

A.6 Neighborhood closure gives compositional correctness

Proof. We proceed by structural induction on the tree. If v is a leaf, then $f_v = \eta_{\tau_v}(\bar{s}_v)$, so $f_v \in \mathbb{B}_{\tau_v}\left(\eta_{\tau_v}(\bar{s}_v), \rho_{\tau_v}(\bar{s}_v)\right)$.

Now let v be an internal node with children $v_1, \dots, v_k$. By the induction hypothesis, each child output f_{v_j} lies in the corresponding trusted neighborhood. Closure of certified neighborhoods therefore gives $f_v = T_{\alpha_v}(f_{v_1}, \dots, f_{v_k}) \in \mathbb{B}_{\tau_v}\left(\eta_{\tau_v}(\bar{s}_v), \rho_{\tau_v}(\bar{s}_v)\right)$. Since $f_v \in \mathbb{B}_{\tau_v}\left(\eta_{\tau_v}(\bar{s}_v), \rho_{\tau_v}(\bar{s}_v)\right)$ and $\mathbb{B}_{\tau_v}\left(\eta_{\tau_v}(\bar{s}_v), \rho_{\tau_v}(\bar{s}_v)\right) \subseteq \mathcal{C}_{\tau_v}(\bar{s}_v)$, we have $f_v \in \mathcal{C}_{\tau_v}(\bar{s}_v)$. By the definition of the decoder cell, $\mathcal{C}_{\tau_v}(\bar{s}_v) = \delta_{\tau_v}^{-1}(\{\bar{s}_v\}) = \{f \in F_{\tau_v} : \delta_{\tau_v}(f) = \bar{s}_v\}$. Therefore, $\delta_{\tau_v}(f_v) = \bar{s}_v$. □

A.7 Derivation-tree error certificate

Definition 25 (Childwise Lipschitz constant)**.** *Fix a symbolic input tuple $\bar{\boldsymbol{s}} = (\bar{s}_1, \dots, \bar{s}_k)$. Let $L_j \ge 0$ be a constant such that, whenever two input tuples $f = (f_1, \dots, f_k)$ and $g = (g_1, \dots, g_k)$ belong to the trusted product neighborhood and satisfy $f_\ell = g_\ell$ for every $\ell \neq j$, we have $d_{\tau_0}\left(T_\alpha(f), T_\alpha(g)\right) \le L_j d_{\tau_j}(f_j, g_j)$. We call L_j a childwise Lipschitz constant for input j on the trusted product neighborhood.*

Definition 26 (Point error)**.** *For $\bar{\boldsymbol{s}} = (\bar{s}_1, \dots, \bar{s}_k) \in \bar{S}_{\tau_1} \times \cdots \times \bar{S}_{\tau_k}$, define $\epsilon_\alpha(\bar{\boldsymbol{s}}) = d_{\tau_0}\left(T_\alpha\left(\eta_{\tau_1}(\bar{s}_1), \dots, \eta_{\tau_k}(\bar{s}_k)\right), \eta_{\tau_0}\left(\bar{\alpha}(\bar{\boldsymbol{s}})\right)\right)$.*

If $\epsilon_\alpha(\bar{\boldsymbol{s}}) \le \rho_{\tau_0}\left(\bar{\alpha}(\bar{\boldsymbol{s}})\right)$ for every symbolic input tuple $\bar{\boldsymbol{s}}$, then T_α is correct on canonical inputs for $\bar{\alpha}$.

Definition 27 (Nodewise error bound)**.** *Let V be the set of nodes of a finite symbolic derivation tree. For each internal node $v \in V$ with ordered children $v_1, \dots, v_k$, labeled by the symbolic operator $\bar{\alpha}_v$ and implemented by the neural module T_{α_v}, let $L_{v,j}$ be a childwise Lipschitz constant for the jth input on the trusted product neighborhood $\prod_{\ell=1}^k \mathbb{B}_{\tau_{v_\ell}}\left(\eta_{\tau_{v_\ell}}(\bar{s}_{v_\ell}), \rho_{\tau_{v_\ell}}(\bar{s}_{v_\ell})\right)$ for each $j \in \{1, \dots, k\}$. Let $\epsilon_v = \epsilon_{\alpha_v}\left(\bar{s}_{v_1}, \dots, \bar{s}_{v_k}\right)$. Let $(B_v)_{v \in V}$ be the family of nonnegative real numbers defined recur-*

sively by $B_v = \begin{cases} 0, & \text{if } v \text{ is a leaf,} \\ \epsilon_v + \sum_{j=1}^{k} L_{v,j} B_{v_j}, & \text{if } v \text{ is an internal node.} \end{cases}$ *We call* B_v *the nodewise error bound at node* v.

Theorem 11 (Derivation-tree error certificate)**.** *Assume that each* $L_{v,j}$ *is a valid childwise Lipschitz constant on the trusted product neighborhood associated with node* v*. If* $B_v \le \rho_{\tau_v}(\bar{s}_v)$ *for every node* v*, then* $d_{\tau_v}(f_v, \eta_{\tau_v}(\bar{s}_v)) \le B_v$ *for every node* v*. Consequently,* $\delta_{\tau_v}(f_v) = \bar{s}_v$ *at every node.*

Proof. If v is a leaf, then $f_v = \eta_{\tau_v}(\bar{s}_v)$ and $B_v = 0$. Therefore, $d_{\tau_v}(f_v, \eta_{\tau_v}(\bar{s}_v)) = 0 = B_v$. Now let v be an internal node with ordered children $v_1, \dots, v_k$. By the induction hypothesis, for every $j \in \{1, \dots, k\}$, $d_{\tau_{v_j}}\left(f_{v_j}, \eta_{\tau_{v_j}}(\bar{s}_{v_j})\right) \le B_{v_j}$. The hypothesis of the theorem, applied to the child node v_j, gives $B_{v_j} \le \rho_{\tau_{v_j}}(\bar{s}_{v_j})$. Therefore, $d_{\tau_{v_j}}\left(f_{v_j}, \eta_{\tau_{v_j}}(\bar{s}_{v_j})\right) \le \rho_{\tau_{v_j}}(\bar{s}_{v_j})$ for every $j \in \{1, \dots, k\}$. Hence $f_{v_j} \in \mathbb{B}_{\tau_{v_j}}\left(\eta_{\tau_{v_j}}(\bar{s}_{v_j}), \rho_{\tau_{v_j}}(\bar{s}_{v_j})\right)$ for every $j \in \{1, \dots, k\}$, and thus $(f_{v_1}, \dots, f_{v_k}) \in \prod_{j=1}^{k} \mathbb{B}_{\tau_{v_j}}\left(\eta_{\tau_{v_j}}(\bar{s}_{v_j}), \rho_{\tau_{v_j}}(\bar{s}_{v_j})\right)$.

Replacing the child inputs by their canonical encodings one at a time, the triangle inequality and the childwise Lipschitz bounds give $d_{\tau_v}\left(T_{\alpha_v}(f_{v_1}, \dots, f_{v_k}), T_{\alpha_v}\left(\eta_{\tau_{v_1}}(\bar{s}_{v_1}), \dots, \eta_{\tau_{v_k}}(\bar{s}_{v_k})\right)\right)$ $\le \sum_{j=1}^{k} L_{v,j} d_{\tau_{v_j}}\left(f_{v_j}, \eta_{\tau_{v_j}}(\bar{s}_{v_j})\right) \le \sum_{j=1}^{k} L_{v,j} B_{v_j}$.

By the definition of ϵ_v and the triangle inequality, $d_{\tau_v}(f_v, \eta_{\tau_v}(\bar{s}_v)) \le d_{\tau_v}\left(T_{\alpha_v}(f_{v_1}, \dots, f_{v_k}), T_{\alpha_v}\left(\eta_{\tau_{v_1}}(\bar{s}_{v_1}), \dots, \eta_{\tau_{v_k}}(\bar{s}_{v_k})\right)\right) + \epsilon_v \le \sum_{j=1}^{k} L_{v,j} B_{v_j} + \epsilon_v = B_v$.

Since $B_v \le \rho_{\tau_v}(\bar{s}_v)$, we have $f_v \in \mathbb{B}_{\tau_v}(\eta_{\tau_v}(\bar{s}_v), \rho_{\tau_v}(\bar{s}_v)) \subseteq \mathcal{C}_{\tau_v}(\bar{s}_v)$. By the definition of the decoder cell, $\mathcal{C}_{\tau_v}(\bar{s}_v) = \delta_{\tau_v}^{-1}(\{\bar{s}_v\}) = \{f \in F_{\tau_v} : \delta_{\tau_v}(f) = \bar{s}_v\}$, we have $\delta_{\tau_v}(f_v) = \bar{s}_v$. □

A.8 Canonical-input correctness composes under exact external canonicalization

Proof. We proceed by structural induction. If v is a leaf, then $f_v = \eta_{\tau_v}(\bar{s}_v)$, and code consistency gives $\delta_{\tau_v}(f_v) = \bar{s}_v$.

Now let v be an internal node with ordered children $v_1, \dots, v_k$. By the induction hypothesis, $\delta_{\tau_{v_j}}(f_{v_j}) = \bar{s}_{v_j}$ for every $j \in \{1, \dots, k\}$. Therefore, $\widehat{f}_{v_j} = P_{\tau_{v_j}}(f_{v_j}) = \eta_{\tau_{v_j}}\left(\delta_{\tau_{v_j}}(f_{v_j})\right) = \eta_{\tau_{v_j}}(\bar{s}_{v_j})$ for every $j \in \{1, \dots, k\}$. Hence $f_v = T_{\alpha_v}\left(\eta_{\tau_{v_1}}(\bar{s}_{v_1}), \dots, \eta_{\tau_{v_k}}(\bar{s}_{v_k})\right)$. Since T_{α_v} is correct on canonical inputs for $\bar{\alpha}_v$, $\delta_{\tau_v}(f_v) = \bar{\alpha}_v(\bar{s}_{v_1}, \dots, \bar{s}_{v_k}) = \bar{s}_v$. □

A.9 Outer-radius propagation

Proof. For $t = 0$, condition *(i)* and $B_0 = b_0$ give $\mathsf{Tube}_0^{\pi} \subseteq \mathbb{B}_{\tau_0}(\eta_{\tau_0}(\bar{s}_0), B_0)$.

Now fix $t \in \{1, \dots, m\}$ and assume that $\mathsf{Tube}_{t-1}^{\pi} \subseteq \mathbb{B}_{\tau_{t-1}}\left(\eta_{\tau_{t-1}}(\bar{s}_{t-1}), B_{t-1}\right)$. Let $y \in \mathsf{Tube}_t^{\pi}$. By the tube recursion, $\exists f \in \mathsf{Tube}_{t-1}^{\pi}$ such that $y \in \mathcal{D}_t(f)$. Since $f \in \mathsf{Tube}_{t-1}^{\pi}$ and, by the induction hypothesis, $\mathsf{Tube}_{t-1}^{\pi} \subseteq \mathbb{B}_{\tau_{t-1}}\left(\eta_{\tau_{t-1}}(\bar{s}_{t-1}), B_{t-1}\right)$, we have $d_{\tau_{t-1}}\left(f, \eta_{\tau_{t-1}}(\bar{s}_{t-1})\right) \le B_{t-1}$. Since $B_{t-1} \le q_{t-1}$, it follows that $d_{\tau_{t-1}}\left(f, \eta_{\tau_{t-1}}(\bar{s}_{t-1})\right) \le q_{t-1}$, and hence $f \in \mathbb{B}_{\tau_{t-1}}\left(\eta_{\tau_{t-1}}(\bar{s}_{t-1}), q_{t-1}\right)$. Thus conditions *(ii)* and *(iii)* apply at f. By the triangle inequality, $d_{\tau_t}(y, \eta_{\tau_t}(\bar{s}_t)) \le d_{\tau_t}(y, T_t(f)) + d_{\tau_t}\left(T_t(f), T_t\left(\eta_{\tau_{t-1}}(\bar{s}_{t-1})\right)\right) + d_{\tau_t}\left(T_t\left(\eta_{\tau_{t-1}}(\bar{s}_{t-1})\right), \eta_{\tau_t}(\bar{s}_t)\right) \le u_t + L_t d_{\tau_{t-1}}\left(f, \eta_{\tau_{t-1}}(\bar{s}_{t-1})\right) + \epsilon_t \le u_t + L_t B_{t-1} + \epsilon_t = B_t$. Since y was arbitrary, $\mathsf{Tube}_t^{\pi} \subseteq \mathbb{B}_{\tau_t}(\eta_{\tau_t}(\bar{s}_t), B_t)$.

Finally, if $B_t \le \rho_{\tau_t}(\bar{s}_t)$, then $\mathsf{Tube}_t^{\pi} \subseteq \mathbb{B}_{\tau_t}(\eta_{\tau_t}(\bar{s}_t), B_t) \subseteq \mathbb{B}_{\tau_t}(\eta_{\tau_t}(\bar{s}_t), \rho_{\tau_t}(\bar{s}_t)) \subseteq \mathcal{C}_{\tau_t}(\bar{s}_t)$. Theorem 1 therefore implies decoded correctness along π. □

A.10 UNIFORM FINITE-HORIZON BOUND

Proof. The recurrence in Theorem 2 gives $B_t \le LB_{t-1} + \nu$ for every $t \in \{1, \dots, m\}$. At $t = 1$, $B_1 \le Lb_0 + \nu = L^1 b_0 + \nu \sum_{j=0}^{0} L^j$. Now let $t \in \{2, \dots, m\}$ and assume that $B_{t-1} \le L^{t-1} b_0 + \nu \sum_{j=0}^{t-2} L^j$. Then $B_t \le LB_{t-1} + \nu \le L^t b_0 + \nu \sum_{j=1}^{t-1} L^j + \nu = L^t b_0 + \nu \sum_{j=0}^{t-1} L^j$. This proves the first bound.

If $0 \le L < 1$, then $B_t \le L^t b_0 + (1 - L^t) \frac{\nu}{1-L}$. Since $0 \le L^t \le 1$, the right-hand side is a convex combination of b_0 and $\nu/(1 - L)$. Hence $B_t \le \max\left\{b_0, \frac{\nu}{1-L}\right\}$, which proves the second bound. □

A.11 UNIFORM CONTRACTION CERTIFICATE OVER ARBITRARY FINITE HORIZONS

Corollary 6 (Detailed trace-indexed contraction certificate)**.** *Let $\mathcal{K}_{\mathrm{fin}}$ be a nonempty family of valid finite symbolic traces, each containing at least one transition, with no common upper bound on trace length. For every $\pi \in \mathcal{K}_{\mathrm{fin}}$, let m_π be the number of transitions in π, and use a superscript π for the trace-dependent quantities in Theorem 2. In particular, define $B_0^\pi = b_0^\pi$ and $B_t^\pi = \epsilon_t^\pi + L_t^\pi B_{t-1}^\pi + u_t^\pi$ for every $t \in \{1, \dots, m_\pi\}$.*

Assume that the hypotheses of Theorem 2 hold along every $\pi \in \mathcal{K}_{\mathrm{fin}}$. Suppose there exist constants $b_0, \epsilon, u \ge 0$ and $0 \le L < 1$ such that $b_0^\pi \le b_0$, $\epsilon_t^\pi \le \epsilon$, $u_t^\pi \le u$, and $L_t^\pi \le L$ for every $\pi \in \mathcal{K}_{\mathrm{fin}}$ and every $t \in \{1, \dots, m_\pi\}$.

Define $q = \inf\left\{q_{t-1}^\pi : \pi \in \mathcal{K}_{\mathrm{fin}}, 1 \le t \le m_\pi\right\}$ and $\rho = \inf\left\{\rho_{\tau_t^\pi}(\bar{s}_t^\pi) : \pi \in \mathcal{K}_{\mathrm{fin}}, 0 \le t \le m_\pi\right\}$. Assume $q > 0$ and $\rho > 0$, and let $\nu = \epsilon + u$. If $\max\left\{b_0, \frac{\nu}{1-L}\right\} \le \min\{q, \rho\}$, then, for every $\pi \in \mathcal{K}_{\mathrm{fin}}$, $B_0^\pi \le b_0 \le \max\left\{b_0, \frac{\nu}{1-L}\right\}$, and, for every $t \in \{1, \dots, m_\pi\}$, $B_t^\pi \le L^t b_0 + \nu \frac{1-L^t}{1-L} \le \max\left\{b_0, \frac{\nu}{1-L}\right\}$. Consequently, $\mathsf{Tube}_t^\pi \subseteq \mathbb{B}_{\tau_t^\pi}\left(\eta_{\tau_t^\pi}(\bar{s}_t^\pi), B_t^\pi\right) \subseteq \mathcal{C}_{\tau_t^\pi}(\bar{s}_t^\pi)$ for every $\pi \in \mathcal{K}_{\mathrm{fin}}$ and every $t \in \{0, \dots, m_\pi\}$. Thus the neural execution is decoded-correct on $\mathcal{K}_{\mathrm{fin}}$.

Proof. Fix $\pi \in \mathcal{K}_{\mathrm{fin}}$. Set $\widehat{B}_0 = b_0$. For each $t \in \{1, \dots, m_\pi\}$, define $\widehat{B}_t = L^t b_0 + \nu \frac{1-L^t}{1-L}$. A direct calculation gives $\widehat{B}_t = L\widehat{B}_{t-1} + \nu$ for every $t \in \{1, \dots, m_\pi\}$.

At $t = 0$, $\widehat{B}_0 = b_0 \le \max\left\{b_0, \frac{\nu}{1-L}\right\}$. For every $t \in \{1, \dots, m_\pi\}$, $\widehat{B}_t = L^t b_0 + (1 - L^t) \frac{\nu}{1-L}$. Since $0 \le L^t \le 1$, this is a convex combination of b_0 and $\frac{\nu}{1-L}$. Therefore, $\widehat{B}_t \le \max\left\{b_0, \frac{\nu}{1-L}\right\} \le \min\{\rho, q\}$ for every $t \in \{0, \dots, m_\pi\}$.

At $t = 0$, $B_0^\pi = b_0^\pi \le b_0 = \widehat{B}_0$. Now let $1 \le t \le m_\pi$ and assume that $B_{t-1}^\pi \le \widehat{B}_{t-1}$. The uniform bound established above gives $\widehat{B}_{t-1} \le q$, and the definition of q gives $q \le q_{t-1}^\pi$. Therefore, $B_{t-1}^\pi \le \widehat{B}_{t-1} \le q \le q_{t-1}^\pi$. Hence the local assumptions of Theorem 2 are valid at step t. Using $\epsilon_t^\pi \le \epsilon$, $u_t^\pi \le u$, and $L_t^\pi \le L$, we obtain $B_t^\pi = \epsilon_t^\pi + L_t^\pi B_{t-1}^\pi + u_t^\pi \le \epsilon + L\widehat{B}_{t-1} + u = L\widehat{B}_{t-1} + \nu = \widehat{B}_t$. Therefore, $B_t^\pi \le \widehat{B}_t$ for every $0 \le t \le m_\pi$.

Theorem 2 now gives $\mathsf{Tube}_t^\pi \subseteq \mathbb{B}_{\tau_t^\pi}\left(\eta_{\tau_t^\pi}(\bar{s}_t^\pi), B_t^\pi\right)$ for every $0 \le t \le m_\pi$. Moreover, $B_t^\pi \le \widehat{B}_t \le \rho \le \rho_{\tau_t^\pi}(\bar{s}_t^\pi)$. It follows that $\mathsf{Tube}_t^\pi \subseteq \mathbb{B}_{\tau_t^\pi}\left(\eta_{\tau_t^\pi}(\bar{s}_t^\pi), B_t^\pi\right) \subseteq \mathbb{B}_{\tau_t^\pi}\left(\eta_{\tau_t^\pi}(\bar{s}_t^\pi), \rho_{\tau_t^\pi}(\bar{s}_t^\pi)\right) \subseteq \mathcal{C}_{\tau_t^\pi}(\bar{s}_t^\pi)$. By Theorem 1, the neural execution is decoded-correct along π. Since π was arbitrary, the neural execution is decoded-correct on $\mathcal{K}_{\mathrm{fin}}$. □

Remark 4 (Finite-prefix quantification)**.** *Corollary 3 certifies every finite trace satisfying the common bounds. Consequently, if every positive-length finite prefix of an infinite valid symbolic trace belongs to the certified family, every admissible infinite neural trajectory along that trace decodes correctly at every finite time. This follows by applying the finite-prefix guarantee at each time. The*

corollary does not assert the existence of such an infinite symbolic trace or provide a state-indexed invariant set closed under transitions absent from the checked trace family.

A.12 UNBOUNDED DRIFT WITHOUT CONTRACTION

Proposition 9 (Unbounded drift without contraction)**.** *Let $\bar{s}$ be a symbolic state with a self-loop, and assume its decoder cell in $F = \mathbb{R}$ satisfies $\mathcal{C}(\bar{s}) \subseteq [-\rho, \rho]$ for some finite $\rho > 0$. Assume there exists an admissible neural trajectory along repeated applications of the self-loop satisfying $f_{t+1} = Lf_t + \epsilon$ with $f_0 = 0$, $L \geq 1$, and $\epsilon > 0$. Then decoded correctness fails along the repeated self-loop at some finite horizon.*

Proof. If $L = 1$, then $f_t = t\epsilon$. If $L > 1$, then $f_t = \epsilon \frac{L^t - 1}{L - 1}$. In either case, there exists a finite time T such that $f_T > \rho$. Since $\mathcal{C}(\bar{s}) \subseteq [-\rho, \rho]$, it follows that $f_T \notin \mathcal{C}(\bar{s})$. By the definition of the decoder cell, $\delta(f_T) \neq \bar{s}$. Hence decoded correctness fails at horizon T. □

A.13 INNER-RADIUS PROPAGATION

Proof. The assertion at $t = 0$ is the assumed initial inclusion. Fix $t \in \{1, \ldots, m\}$ and assume $\mathbb{B}_{F_{\tau_{t-1}}}\left(\eta_{\tau_{t-1}}(\bar{s}_{t-1}), a_{t-1}\right) \subseteq \mathsf{Tube}^{\pi}_{t-1}$. Set $x_{t-1} = \eta_{\tau_{t-1}}(\bar{s}_{t-1})$ and $x_t = \eta_{\tau_t}(\bar{s}_t)$. By the tube recursion, monotonicity of the set extension of $\mathcal{D}_t$, and full-ball disturbance inclusion, $\mathsf{Tube}^{\pi}_t = \mathcal{D}_t\left(\mathsf{Tube}^{\pi}_{t-1}\right) \supseteq \bigcup_{f \in \mathbb{B}_{F_{\tau_{t-1}}}(x_{t-1}, a_{t-1})} \mathcal{D}_t(f) \supseteq \bigcup_{f \in \mathbb{B}_{F_{\tau_{t-1}}}(x_{t-1}, a_{t-1})} \mathbb{B}_{F_{\tau_t}}\left(T_t(f), v_t\right) = T_t\left(\mathbb{B}_{F_{\tau_{t-1}}}(x_{t-1}, a_{t-1})\right) + \mathbb{B}_{F_{\tau_t}}(0, v_t)$, where the last expression is a Minkowski sum in F_{τ_t}.

Local openness gives $T_t\left(\mathbb{B}_{F_{\tau_{t-1}}}(x_{t-1}, a_{t-1})\right) \supseteq \mathbb{B}_{F_{\tau_t}}\left(T_t(x_{t-1}), \mu_t a_{t-1}\right)$. For every normed linear space E, every $x \in E$, and all $r, s \geq 0$, $\mathbb{B}_E(x, r) + \mathbb{B}_E(0, s) = \mathbb{B}_E(x, r + s)$. The inclusion from left to right follows from the triangle inequality. For the reverse inclusion, the case $r + s = 0$ is immediate. If $r + s > 0$ and $z \in \mathbb{B}_E(x, r + s)$, define $u = x + \frac{r}{r+s}(z - x)$ and $w = \frac{s}{r+s}(z - x)$. Then $u \in \mathbb{B}_E(x, r)$, $w \in \mathbb{B}_E(0, s)$, and $z = u + w$. Consequently, $\mathsf{Tube}^{\pi}_t \supseteq \mathbb{B}_{F_{\tau_t}}\left(T_t(x_{t-1}), \mu_t a_{t-1} + v_t\right)$.

Let $z \in \mathbb{B}_{F_{\tau_t}}(x_t, a_t)$. By the definition of a_t, the triangle inequality, and the definition of ϵ_t, $\|z - T_t(x_{t-1})\|_{\tau_t} \leq \|z - x_t\|_{\tau_t} + \|x_t - T_t(x_{t-1})\|_{\tau_t} \leq a_t + \epsilon_t = \mu_t a_{t-1} + v_t$. Thus z belongs to the ball already shown to lie in Tube^{π}_t. Since z was arbitrary, $\mathbb{B}_{F_{\tau_t}}(x_t, a_t) \subseteq \mathsf{Tube}^{\pi}_t$. Induction completes the proof. □

A.14 MULTISTEP INNER REACHABILITY

The parameter ball in Theorem 3 describes a set of entire disturbance sequences that are permitted by the execution semantics. It need not describe every permitted sequence. Restricting to such a subset is sufficient for an inner approximation, provided that every included sequence is admissible. The norm on the parameter space may encode different disturbance bounds at different times. The endpoint map must use a fixed valid symbolic trace, including its specified choices; varying a symbolic choice is not a disturbance unless the resulting execution still follows that trace.

Proof of Theorem 3. Because F_{τ_m} has positive dimension and $J\Lambda = \mathrm{id}_{F_{\tau_m}}$, the linear map Λ is nonzero. Its operator norm γ_{inv} is therefore positive. Define the continuous remainder map $e(w) = \Psi(w) - \Psi(0) - Jw$ on $\mathbb{B}_{\mathcal{W}}(0, h)$. By assumption, $\|e(w)\|_{\tau_m} \leq \zeta$ on this ball.

First assume $h/\gamma_{\mathrm{inv}} \geq \zeta$, and fix any $y \in \mathbb{B}_{F_{\tau_m}}\big(\Psi(0), h/\gamma_{\mathrm{inv}} - \zeta\big)$. Define a continuous map on the parameter ball by $\mathcal{H}_y(w) = \Lambda\big(y - \Psi(0) - e(w)\big)$. For every $w \in \mathbb{B}_{\mathcal{W}}(0, h)$, the operator-norm bound gives

$$\begin{aligned}\|\mathcal{H}_y(w)\|_{\mathcal{W}} &\leq \gamma_{\mathrm{inv}}\big(\|y - \Psi(0)\|_{\tau_m} + \|e(w)\|_{\tau_m}\big) \\ &\leq \gamma_{\mathrm{inv}}\big(h/\gamma_{\mathrm{inv}} - \zeta + \zeta\big) = h.\end{aligned}$$

Thus $\mathcal{H}_y$ maps $\mathbb{B}_{\mathcal{W}}(0,h)$ into itself. This ball is nonempty, compact, and convex because $\mathcal{W}$ is finite dimensional. Brouwer's fixed-point theorem supplies $w_y \in \mathbb{B}_{\mathcal{W}}(0,h)$ satisfying $w_y = \mathcal{H}_y(w_y)$. Applying J and using the right-inverse identity yields $Jw_y = y - \Psi(0) - e(w_y)$. Consequently, $\Psi(w_y) = \Psi(0) + Jw_y + e(w_y) = y$. Since y was arbitrary, this proves $\mathbb{B}_{F_{\tau_m}}\big(\Psi(0), h/\gamma_{\mathrm{inv}} - \zeta\big) \subseteq \Psi\big(\mathbb{B}_{\mathcal{W}}(0,h)\big)$.

Now assume $a^{\pi}_{\mathrm{end}} \geq 0$. This implies $h/\gamma_{\mathrm{inv}} \geq \zeta$. For any $z \in \mathbb{B}_{F_{\tau_m}}\big(\eta_{\tau_m}(\bar{s}_m), a^{\pi}_{\mathrm{end}}\big)$, the triangle inequality gives

$$\begin{aligned}\|z - \Psi(0)\|_{\tau_m} &\leq \|z - \eta_{\tau_m}(\bar{s}_m)\|_{\tau_m} + \epsilon^{\pi}_{\mathrm{end}} \\ &\leq a^{\pi}_{\mathrm{end}} + \epsilon^{\pi}_{\mathrm{end}} = h/\gamma_{\mathrm{inv}} - \zeta.\end{aligned}$$

The preceding inclusion therefore implies

$$\mathbb{B}_{F_{\tau_m}}\big(\eta_{\tau_m}(\bar{s}_m), a^{\pi}_{\mathrm{end}}\big) \subseteq \Psi\big(\mathbb{B}_{\mathcal{W}}(0,h)\big).$$

Every point in this image is the endpoint of an admissible trajectory along π from $f^{\circ} \in \mathsf{Tube}^{\pi}_0$. Hence the image is contained in Tube^{π}_m, proving the asserted inner approximation. □

Remark 5 (Centering and the initial state)**.** *The zero parameter need not produce a trajectory consisting of the intended canonical encodings. All initial-state deviation and nominal approximation error that affect its terminal state are included in $\epsilon^{\pi}_{\mathrm{end}}$. The only initial-state requirement is that the fixed state f° is actually admissible. If the certificate is applied to a suffix of a longer trace, its starting state must belong to the reachable set at that suffix's starting time; concatenating a prefix reaching that state with each certified suffix gives the required terminal reachability.*

A negative value of a^{π}_{end} cannot be truncated to zero. For example, with $\mathcal{W} = F_{\tau_m} = \mathbb{R}$, $h = 1$, $\Psi(w) = 2 + w$, $J = \Lambda = \mathrm{id}_{\mathbb{R}}$, and target encoding zero, one has $\gamma_{\mathrm{inv}} = 1$, $\zeta = 0$, and $a^{\pi}_{\mathrm{end}} = -1$. The endpoint image is $[1, 3]$, so the target encoding is not reachable.

Proposition 10 (Derivative remainder certificate)**.** *Use the endpoint map and parameter ball of Theorem 3. Suppose that Ψ extends to a continuously differentiable map on an open neighborhood of $\mathbb{B}_{\mathcal{W}}(0,h)$, and let $J = D\Psi(0)$. Assume that a constant $\ell_{\mathrm{der}} \geq 0$ satisfies $\|D\Psi(w) - D\Psi(0)\| \leq \ell_{\mathrm{der}}\|w\|_{\mathcal{W}}$ for every $w \in \mathbb{B}_{\mathcal{W}}(0,h)$, where the derivative norm is induced by the norms on $\mathcal{W}$ and F_{τ_m}. Then $\|\Psi(w) - \Psi(0) - Jw\|_{\tau_m} \leq \frac{1}{2}\ell_{\mathrm{der}}\|w\|^2_{\mathcal{W}}$ throughout the parameter ball. In particular, $\zeta = \ell_{\mathrm{der}}h^2/2$ is valid. The derivative condition holds whenever $D\Psi$ is Lipschitz with constant ℓ_{der} on the ball.*

Proof. Fix $w \in \mathbb{B}_{\mathcal{W}}(0,h)$. The segment $\{\theta w : 0 \leq \theta \leq 1\}$ lies in the ball. Applying the fundamental theorem of calculus to the continuously differentiable function $\theta \mapsto \Psi(\theta w)$ gives $\Psi(w) - \Psi(0) - Jw = \int_0^1 \big(D\Psi(\theta w) - D\Psi(0)\big)w\, d\theta$. Consequently,

$$\begin{aligned}\|\Psi(w) - \Psi(0) - Jw\|_{\tau_m} &\leq \int_0^1 \ell_{\mathrm{der}}\theta\|w\|^2_{\mathcal{W}}\, d\theta \\ &= \tfrac{1}{2}\ell_{\mathrm{der}}\|w\|^2_{\mathcal{W}} \leq \tfrac{1}{2}\ell_{\mathrm{der}}h^2.\end{aligned}$$

If $D\Psi$ is Lipschitz with the stated constant, its defining inequality applied to w and zero yields the assumed radial derivative bound. □

Proposition 11 (Affine endpoint certificate)**.** *Fix the setting of Theorem 3. Assume also that every intermediate F_{τ_t} is a real normed linear space with its induced metric, and write the parameter space as $\mathcal{W} = \mathcal{W}_1 \times \cdots \times \mathcal{W}_m$. Suppose every $w = (w_1, \ldots, w_m) \in \mathbb{B}_{\mathcal{W}}(0,h)$ produces an admissible trajectory from the same state f°, with updates $f_t(w) = A_t f_{t-1}(w) + b_t + M_t w_t$, where $A_t : F_{\tau_{t-1}} \to F_{\tau_t}$ and $M_t : \mathcal{W}_t \to F_{\tau_t}$ are linear and $b_t \in F_{\tau_t}$. These update identities need hold only on the states generated by the parameter ball. Then the endpoint map $\Psi(w) = f_m(w)$ is affine on that ball, with linear part $Jw = \sum_{j=1}^{m} A_m A_{m-1} \cdots A_{j+1} M_j w_j$, where the product for $j = m$ is the identity on F_{τ_m}. If J is surjective, any right inverse Λ gives the certificate of Theorem 3 with $\zeta = 0$.*

Proof. Let $f_t(0)$ denote the state produced by the zero parameter. Since the initial state is fixed, $f_0(w) - f_0(0) = 0$. Subtracting the two update equations gives $f_t(w) - f_t(0) = A_t\big(f_{t-1}(w) - f_{t-1}(0)\big) + M_t w_t$. Induction on t therefore yields $f_t(w) - f_t(0) = \sum_{j=1}^{t} A_t A_{t-1} \cdots A_{j+1} M_j w_j$. At $t = m$ this is $\Psi(w) - \Psi(0) = Jw$, so the remainder is identically zero. The endpoint map is continuous because it is affine on the ball. Surjectivity of a linear map between finite-dimensional spaces guarantees a linear right inverse: choose a basis of the output space, choose a preimage of each basis vector, and extend the resulting assignment linearly. Theorem 3 now applies. □

For finite ReLU transition networks with affine disturbance injection, a sufficient condition for Proposition 11 is that the unrolled execution has a common activation pattern for every parameter in $\mathbb{B}_{\mathcal{W}}(0, h)$. Each transition then has an affine expression valid on all predecessor states generated by that ball. This is a regional condition, not a global differentiability assumption. It can be certified by verifying that every relevant preactivation retains its prescribed sign throughout the parameter ball. A pattern chosen only at the zero parameter is insufficient. If an activation boundary is crossed, Theorem 3 remains applicable whenever a continuous endpoint map and a valid uniform remainder bound are available; the derivative certificate is then optional.

The operator J in Proposition 11 is the usual finite-horizon disturbance-to-state map of linear systems. Its surjectivity can hold even when each disturbance injection, and each individual state-transition matrix, has deficient rank. No positive one-step openness coefficient is required for the terminal certificate.

Example 2 (Two rank-one updates with a full-dimensional endpoint image)**.** *Equip the state space $\mathbb{R}^2$ and the parameter space $\mathcal{W} = \mathbb{R}^2$ with the ℓ_∞ norm. Let $h > 0$, let $e_1 = (1, 0)$, and define $A(x_1, x_2) = (0, x_1)$. Consider a symbolic self-loop whose canonical encoding is zero, with fixed neural initial state $x_0 = 0$ and successor map $\mathcal{D}(x) = \{Ax + e_1 v : |v| \le h\}$. Every $(w_1, w_2) \in \mathbb{B}_\infty(0, h)$ specifies an admissible two-step trajectory. Direct substitution gives $x_1 = (w_1, 0),\ x_2 = (w_2, w_1)$. Thus the endpoint map is $\Psi(w_1, w_2) = (w_2, w_1)$. Its linear part is the coordinate permutation J, which is its own inverse and is an ℓ_∞ isometry. Taking $\Lambda = J$ gives $\gamma_{\mathrm{inv}} = 1$, $\zeta = 0$, and $\epsilon^{\pi}_{\mathrm{end}} = 0$. The theorem therefore yields $\mathbb{B}_\infty(0, h) \subseteq \mathsf{Tube}^{\pi}_2$. The explicit expression for x_2 proves equality.*

Both A and the disturbance injection $v \mapsto e_1 v$ have rank one. At each fixed predecessor state, $\mathcal{D}(x)$ is a line segment and contains no positive-radius ball in the ambient state space. The one-step inner-radius recurrence therefore cannot initiate a positive ambient radius from $x_0 = 0$. The two independent disturbance coordinates nevertheless span both terminal coordinates after propagation.

Remark 6 (What the certificate establishes)**.** *The theorem provides an inner approximation to the terminal reachable set. It does not establish decoded correctness at intermediate times or at the terminal time. Terminal robustness must still be proved or assumed before the resulting radius can be used in the packing bounds. The conclusion is conditional on an admissible parameter domain, surjectivity, a right-inverse norm bound, and a remainder bound; it does not assert that these quantities are easy to compute for an arbitrary network.*

A state-quantized endpoint map with finite range cannot have a positive-radius norm ball in its image. In particular, the continuous endpoint-map assumption and the positive-radius conclusion cannot be transferred unchanged to a finite quantized execution. The quantization-aware outer-radius certificate remains the appropriate sufficient condition in that setting.

A.15 Exact transitions with unit openness and full-ball disturbances

Proof. Condition *(i)* gives $\epsilon_t = 0$. Conditions *(ii)* and *(iii)* are the local-openness and full-ball disturbance assumptions of Theorem 8 with $\mu_t = 1$ and $v_t = \sigma_t$. Hence the theorem gives the claimed inclusions and the recurrence $a_t = a_{t-1} + \sigma_t$. Starting from $a_0 = 0$, induction yields $a_t = \sum_{j=1}^{t} \sigma_j$. The constant-disturbance formula follows immediately. □

A.16 AFFINE OPENNESS CERTIFICATE

Proof. Fix $x \in F$ and $r \geq 0$, and let $y \in \mathbb{B}_F(T(x), \mu r)$. Define $h = A^{-1}(y - T(x))$. Then $\|h\|_F \leq \|A^{-1}\|\|y - T(x)\|_F \leq \|A^{-1}\|\mu r = r$. Hence $x + h \in \mathbb{B}_F(x, r)$, and $T(x + h) = A(x + h) + b = T(x) + Ah = y$. Therefore $y \in T(\mathbb{B}_F(x, r))$. Since y was arbitrary, the claimed inclusion follows. If A is a surjective linear isometry, then A^{-1} is also an isometry, so $\|A^{-1}\| = 1$. □

A.17 TERMINAL DECODER-CELL CHARACTERIZATION

Proof. Assume first that the decoder is terminally robust for $\mathcal{K}_T$ at horizon T. Fix a quotient state $\bar{s}$ occurring at horizon T, and let $f \in A_T(\bar{s})$. By the definition of $A_T(\bar{s})$, $\exists \pi \in \mathcal{K}_T$ such that $\bar{s}_T^\pi = \bar{s}$ and $f \in \mathsf{Tube}_T^\pi$. By Theorem 1, Tube_T^π is the set of states reachable at time T along admissible trajectories. Hence there exists an admissible trajectory $f_0, \ldots, f_T$ along π with $f_T = f$. Terminal robustness gives $\delta_\tau(f) = \delta_\tau(f_T) = \bar{s}_T^\pi = \bar{s}$. Therefore, $f \in \mathcal{C}_\tau(\bar{s})$. Since f was arbitrary, $A_T(\bar{s}) \subseteq \mathcal{C}_\tau(\bar{s})$.

Conversely, assume that $A_T(\bar{s}) \subseteq \mathcal{C}_\tau(\bar{s})$ for every quotient state $\bar{s}$ occurring at horizon T. Fix $\pi \in \mathcal{K}_T$ and an admissible trajectory $f_0, \ldots, f_T$ along π. Then $f_T \in \mathsf{Tube}_T^\pi \subseteq A_T(\bar{s}_T^\pi) \subseteq \mathcal{C}_\tau(\bar{s}_T^\pi)$. By the definition of the decoder cell, $\delta_\tau(f_T) = \bar{s}_T^\pi$. Thus the decoder is terminally robust for $\mathcal{K}_T$ at horizon T.

Assume now that the equivalent conditions hold, and let $\bar{s}$ and $\bar{s}'$ be distinct quotient states occurring at horizon T. If $f \in A_T(\bar{s}) \cap A_T(\bar{s}')$, then $f \in \mathcal{C}_\tau(\bar{s}) \cap \mathcal{C}_\tau(\bar{s}')$. Therefore, $\delta_\tau(f) = \bar{s}$ and $\delta_\tau(f) = \bar{s}'$, which contradicts $\bar{s} \neq \bar{s}'$. Hence $A_T(\bar{s}) \cap A_T(\bar{s}') = \varnothing$. □

A.18 GENERAL HORIZON PACKING BOUND

Proof. For every $\bar{s} \in S_T$, the common inner reachability assumption gives $\mathbb{B}_\tau(\eta_\tau(\bar{s}), a_T) \subseteq A_T(\bar{s})$. Since the decoder is terminally robust, Proposition 4 implies that the sets $A_T(\bar{s})$, $\bar{s} \in S_T$, are pairwise disjoint. Therefore, for distinct $\bar{s}, \bar{s}' \in S_T$, $\mathbb{B}_\tau(\eta_\tau(\bar{s}), a_T) \cap \mathbb{B}_\tau(\eta_\tau(\bar{s}'), a_T) \subseteq A_T(\bar{s}) \cap A_T(\bar{s}') = \varnothing$. Hence the radius-$a_T$ balls centered at the encodings of the states in S_T are pairwise disjoint. Because their centers belong to K, they form an admissible ambient ball packing of cardinality N_T. Hence $N_T \leq \mathrm{BPack}_{F_\tau}(K, a_T)$.

Now assume that F_τ has the midpoint property. Let $\bar{s}, \bar{s}' \in S_T$ be distinct, and set $x = \eta_\tau(\bar{s})$ and $y = \eta_\tau(\bar{s}')$. Suppose that $d(x, y) \leq 2a_T$. By the midpoint property, there exists $z \in F_\tau$ such that $d(x, z) = d(y, z) = \frac{d(x,y)}{2} \leq a_T$. It follows that $z \in \mathbb{B}_\tau(x, a_T) \cap \mathbb{B}_\tau(y, a_T)$, contradicting the pairwise disjointness established above. Therefore, $d(x, y) > 2a_T$ for every pair of distinct states $\bar{s}, \bar{s}' \in S_T$. In particular, the set $\{\eta_\tau(\bar{s}) : \bar{s} \in S_T\}$ is $2a_T$ -separated and has cardinality N_T. Consequently, $N_T \leq \mathrm{Pack}(K, 2a_T)$. □

A.19 METRIC-ENTROPY OBSTRUCTION

Definition 28 (Diameter covering number and metric entropy)**.** *Let (K, d) be a nonempty metric space, let $r > 0$, and let $\mathcal{A}$ be a finite family of subsets of K. We call $\mathcal{A}$ an r -diameter cover of K if $K \subseteq \bigcup_{A \in \mathcal{A}} A$ and $\mathrm{diam}(A) < r$ for every $A \in \mathcal{A}$. We call $\mathrm{Cov}(K, r) = \inf\{|\mathcal{A}| : \mathcal{A}$ is an r-diameter cover of $K\}$ the diameter covering number of K at scale r. If no finite r -diameter cover of K exists, we set $\mathrm{Cov}(K, r) = \infty$. Thus, $\mathrm{Cov}(K, r) \in \mathbb{N} \cup \{\infty\}$.*

We call $\mathrm{Ent}(K, r) = \log \mathrm{Cov}(K, r) \in [0, \infty]$ the metric entropy of K at scale r. Here, $\log$ denotes the natural logarithm on finite positive numbers, and we set $\log \infty = \infty$. Whenever $\mathrm{Pack}(K, r)$ or $\mathrm{BPack}_F(K, r)$ is infinite, we set its logarithm equal to ∞ likewise.

We use the diameter-cover convention, under which the scale parameter r bounds the diameter of each covering set rather than the radius of a covering ball.

Corollary 7 (Metric-entropy obstruction)**.** *Assume the hypotheses of Theorem 9, and suppose that F_τ has the midpoint property. Then $\log N_T \leq \log \mathrm{Pack}(K, 2a_T) \leq \mathrm{Ent}(K, 2a_T)$.*

Proof. By Theorem 9, $N_T \leq \mathrm{Pack}(K, 2a_T)$. It therefore remains to compare the packing and covering numbers at scale $2a_T$.

Let $x_1, \ldots, x_n \in K$ be any finite $2a_T$ -separated family. A subset of K with diameter strictly smaller than $2a_T$ cannot contain two distinct points from this family. Consequently, every $2a_T$ -diameter cover of K must contain at least n sets. Hence $n \leq \mathrm{Cov}(K, 2a_T)$.

Since this inequality holds for every finite $2a_T$ -separated family in K, taking the supremum over their cardinalities yields $\mathrm{Pack}(K, 2a_T) \leq \mathrm{Cov}(K, 2a_T)$. This inequality also holds when either quantity is infinite, under the extended-value conventions introduced above. Applying the logarithm gives $\log N_T \leq \log \mathrm{Pack}(K, 2a_T) \leq \log \mathrm{Cov}(K, 2a_T) = \mathrm{Ent}(K, 2a_T)$. □

A.20 DOUBLING-DIMENSION SPECIALIZATION

Definition 29 (Doubling constant and doubling dimension)**.** *Let (K, d) be a nonempty metric space. For $x \in K$ and $r > 0$, let $\mathbb{B}_K(x, r) = \{y \in K : d(x, y) \leq r\}$ be the closed ball in K centered at x with radius r.*

Let λ_K be the least $\lambda \in \mathbb{N}$ such that every closed ball $\mathbb{B}_K(x, r)$ can be covered by at most λ closed balls of radius $r/2$ with centers in K. If no such finite λ exists, let $\lambda_K = \infty$. We call $\lambda_K \in \mathbb{N} \cup \{\infty\}$ the doubling constant of K.

Let $\mathsf{ddim}(K) = \log_2 \lambda_K$, with the convention $\log_2 \infty = \infty$. We call $\mathsf{ddim}(K)$ the doubling dimension of K. We say that K has doubling dimension at most m if $\mathsf{ddim}(K) \leq m$, or equivalently, $\lambda_K \leq 2^m$.

Corollary 8 (Doubling-dimension specialization)**.** *Assume the hypotheses of Theorem 9, and assume that F_τ has the midpoint property stated there. Assume additionally that $\mathrm{diam}(K) \leq D$ for some $D \geq 0$, and that $\mathsf{ddim}(K) \leq m < \infty$. If $D \leq 2a_T$, then $N_T \leq 1$. If $D > 2a_T$, then $N_T \leq \left(\frac{2D}{a_T}\right)^m$. If $N_T > 1$, then $D > 2a_T$ and $m \geq \frac{\log N_T}{\log(2D/a_T)}$.*

Proof. Let $X_T = \{\eta_\tau(\bar{s}) : \bar{s} \in S_T\}$. By code consistency, the encoder η_τ is injective. Hence $|X_T| = |S_T| = N_T$. Moreover, by the argument in Theorem 9, the midpoint property and the pairwise disjointness of the radius- a_T balls imply that, for any distinct $x, y \in X_T$, $d(x, y) > 2a_T$. Thus X_T is $2a_T$ -separated.

Assume first that $D \leq 2a_T$. For any $x, y \in K$, we have $d(x, y) \leq \mathrm{diam}(K) \leq D \leq 2a_T$. On the other hand, any two distinct points of X_T have distance strictly greater than $2a_T$. Hence K cannot contain two distinct points of X_T, and therefore $N_T \leq 1$.

Now assume that $D > 2a_T$. Choose any $x_0 \in K$. Since $\mathrm{diam}(K) \leq D$, we have $K \subseteq \mathbb{B}_K(x_0, D)$. Let λ_K be the doubling constant of K. By definition, $\mathsf{ddim}(K) = \log_2 \lambda_K$. Since $\mathsf{ddim}(K) \leq m < \infty$, the doubling constant λ_K is finite, and $\log_2 \lambda_K \leq m$. Because the map $x \mapsto 2^x$ is increasing, it follows that $\lambda_K \leq 2^m$.

Set $k = \left\lfloor \log_2 \left(\frac{D}{a_T}\right) \right\rfloor + 1$. Iterating the doubling property k times shows that $\mathbb{B}_K(x_0, D)$ can be covered by at most λ_K^k closed balls in K of radius $\frac{D}{2^k}$. By the choice of k, we have $\frac{D}{2^k} < a_T$. Hence each covering ball has diameter strictly smaller than $2a_T$. Since X_T is $2a_T$ -separated, each covering ball contains at most one point of X_T. It follows that $N_T \leq \lambda_K^k$.

By the definition of k, we have $k = \left\lfloor \log_2 \left(\frac{D}{a_T}\right) \right\rfloor + 1 \leq \log_2 \left(\frac{D}{a_T}\right) + 1 = \log_2 \left(\frac{2D}{a_T}\right)$. Since $m \geq 0$, it follows that $mk \leq m \log_2 \left(\frac{2D}{a_T}\right)$. Therefore, $N_T \leq \lambda_K^k \leq 2^{mk} \leq 2^{m \log_2(2D/a_T)} = \left(\frac{2D}{a_T}\right)^m$.

Finally, assume that $N_T > 1$. Then X_T contains distinct points x and y. As established above, $d(x, y) > 2a_T$. Therefore, $\mathrm{diam}(K) > 2a_T$, and the assumption $\mathrm{diam}(K) \leq D$ gives $D > 2a_T$. The preceding bound yields $\log N_T \leq m \log \left(\frac{2D}{a_T}\right)$. Since $D > 2a_T$ makes the logarithm on the right positive, division gives $m \geq \frac{\log N_T}{\log(2D/a_T)}$. □

A.21 SHARP ℓ_∞ HORIZON-CAPACITY BOUND

Proof of Theorem 4. By Proposition 4, terminal robustness implies that $A_T(\bar{s})$ and $A_T(\bar{s}')$ are disjoint whenever $\bar{s} \neq \bar{s}'$. The common inner radius therefore makes the closed balls $\mathbb{B}_\infty(\eta_\tau(\bar{s}), a_T)$ pairwise disjoint. For two different centers x and y, the inequality $\|x - y\|_\infty \leq 2a_T$ would put their midpoint in both balls. Hence all different centers have distance strictly greater than $2a_T$.

If $B = 0$, every center is the origin, so there can be at most one state. Assume henceforth that $B > 0$ and write $n_{\mathrm{cap}} = \lceil B/a_T \rceil \geq 1$. Cover $[-B, B]$ by the n_{cap} closed intervals

$$\left[-B + \frac{2Bj}{n_{\mathrm{cap}}}, -B + \frac{2B(j+1)}{n_{\mathrm{cap}}}\right], \ 0 \leq j < n_{\mathrm{cap}}.$$

Each interval has length $2B/n_{\mathrm{cap}} \leq 2a_T$. Their Cartesian products cover $[-B, B]^d$ with n_{cap}^d sets of ℓ_∞ diameter at most $2a_T$. Each such set contains at most one center, by strict separation. The covers may overlap at their boundaries; this causes no difficulty, since the number of centers in their union is no larger than the sum of the numbers in the covering sets. Therefore $N_T \leq n_{\mathrm{cap}}^d$. Since $n_{\mathrm{cap}} \leq 1 + B/a_T$, the stated volume bound follows as well.

To prove that the integer count is exact for closed-ball packing, define the scalar grid

$$\Gamma_{\mathrm{cap}} = \begin{cases} \{0\}, & n_{\mathrm{cap}} = 1, \\ \{-B + 2Bj/(n_{\mathrm{cap}} - 1) : j = 0, \ldots, n_{\mathrm{cap}} - 1\}, & n_{\mathrm{cap}} \geq 2. \end{cases}$$

This grid lies in $[-B, B]$ and has n_{cap} elements. If $n_{\mathrm{cap}} \geq 2$, then $n_{\mathrm{cap}} - 1 < B/a_T$, so its spacing $2B/(n_{\mathrm{cap}} - 1)$ is strictly greater than $2a_T$. Consequently any two distinct points of Γ_{cap}^d differ by more than $2a_T$ in at least one coordinate. The closed radius-a_T balls centered at these n_{cap}^d points are pairwise disjoint. For $n_{\mathrm{cap}} = 1$, a single ball attains the bound. This proves the exact count, including the strict-separation requirement at integer values of B/a_T.

If $a_T \geq B > 0$, then $n_{\mathrm{cap}} = 1$, so $N_T \leq 1$; the case $B = 0$ was already established. Conversely, $N_T > 1$ forces $B > 0$ and $n_{\mathrm{cap}} \geq 2$. Taking logarithms of $N_T \leq n_{\mathrm{cap}}^d$ and dividing by $\log n_{\mathrm{cap}} > 0$ gives the dimension bound involving n_{cap}. For every $B > 0$, the denominator $\log(1 + B/a_T)$ is positive, so the same argument applied to the volume bound gives $d \geq \log N_T / \log(1 + B/a_T)$, including when $N_T = 1$. □

A.22 FINITE-HORIZON REALIZATION

Proof of Proposition 1. Use one symbolic type and a singleton choice space for each transition label, so the choice argument may be suppressed. Throughout the construction T, d, B, and σ are fixed, and $a_T = T\sigma > 0$.

When $B = 0$, the cardinality assumption implies $|\bar{S}| = 1$. More generally, if $|\bar{S}| = 1$, encode its only state by $0 \in [-B, B]^d$, let δ be constant, and let every T_r be the identity map. The identity is a finite ReLU map because $f = \mathrm{ReLU}(f) - \mathrm{ReLU}(-f)$ coordinatewise. Its exact reachable set at time t is $\mathbb{B}_\infty(0, t\sigma)$, by the Minkowski-sum identity for norm balls. All assertions follow in this case. It remains to consider $|\bar{S}| \geq 2$, which implies $B > 0$ and $n_{\mathrm{cap}} \geq 2$.

Take the grid Γ_{cap} from the preceding proof and select $|\bar{S}|$ distinct points from Γ_{cap}^d. Define η as any bijection from $\bar{S}$ to the selected points. The construction gives

$$\|\eta(\bar{s}) - \eta(\bar{s}')\|_\infty > 2a_T \text{ whenever } \bar{s} \neq \bar{s}'.$$

Because the codebook is finite, the following positive number is well defined:

$$\lambda_{\mathrm{cap}} = \frac{1}{2} \min_{\bar{s} \neq \bar{s}'} \big(\|\eta(\bar{s}) - \eta(\bar{s}')\|_\infty - 2a_T \big) > 0.$$

For each $\bar{s} \in \bar{S}$, define

$$w_{\mathrm{cap},\bar{s}}(f) = \mathrm{ReLU}\left(1 - \frac{\max\{0, \|f - \eta(\bar{s})\|_\infty - a_T\}}{\lambda_{\mathrm{cap}}}\right).$$

The function $w_{\mathrm{cap},\bar{s}}$ equals 1 throughout $\mathbb{B}_\infty(\eta(\bar{s}), a_T)$. If $f \in \mathbb{B}_\infty(\eta(\bar{s}), a_T)$ and $\bar{s}' \neq \bar{s}$, then

$$\|f - \eta(\bar{s}')\|_\infty \geq \|\eta(\bar{s}) - \eta(\bar{s}')\|_\infty - \|f - \eta(\bar{s})\|_\infty \geq a_T + 2\lambda_{\mathrm{cap}}.$$

Thus $w_{\mathrm{cap},\bar{s}'}(f) = 0$ throughout that same ball.

For every $r \in \mathcal{R}_{\mathrm{cap}}$, set

$$T_r(f) = f + \sum_{\bar{s} \in \bar{S}} \big(\eta(\bar{r}(\bar{s})) - \eta(\bar{s})\big) w_{\mathrm{cap},\bar{s}}(f).$$

This formula defines a finite feed-forward ReLU map on all of $\mathbb{R}^d$. Indeed, $|u| = \mathrm{ReLU}(u) + \mathrm{ReLU}(-u)$, and $\max\{u, v\} = \mathrm{ReLU}(u - v) + v$. Finite repetition therefore implements the ℓ_∞ norm and the maximum in each $w_{\mathrm{cap},\bar{s}}$. The remaining operations are affine maps and ReLU activations. The products in the displayed definition of T_r multiply a scalar activation by a constant vector, so they are affine output operations. The identity summand is implemented by the coordinatewise identity representation above. Parallel copies of the finitely many subnetworks, with identity representations to align their depths if necessary, followed by an affine output layer give an ordinary finite feed-forward ReLU network.

On $\mathbb{B}_\infty(\eta(\bar{s}), a_T)$ exactly the coefficient $w_{\mathrm{cap},\bar{s}}$ is 1 and all other coefficients are 0. Hence

$$T_r(f) = f + \eta(\bar{r}(\bar{s})) - \eta(\bar{s}) \text{ for every } f \in \mathbb{B}_\infty(\eta(\bar{s}), a_T).$$

In particular, $T_r(\eta(\bar{s})) = \eta(\bar{r}(\bar{s}))$. For every $0 \leq u \leq a_T$, translation of the smaller ball gives the exact image identity

$$T_r\big(\mathbb{B}_\infty(\eta(\bar{s}), u)\big) = \mathbb{B}_\infty(\eta(\bar{r}(\bar{s})), u).$$

This verifies local openness with coefficient 1 up to a_T. Different source states may have the same successor: their disjoint source balls are then translated onto the same target ball. This does not affect the definition or continuity of T_r.

Fix any total ordering of $\bar{S}$. For each $f \in \mathbb{R}^d$, let $\delta(f)$ be a state minimizing $\|f - \eta(\bar{s})\|_\infty$, using the fixed ordering to resolve ties. Finiteness of $\bar{S}$ ensures that a minimizer exists, and this rule defines a deterministic decoder on the whole space. If $f \in \mathbb{B}_\infty(\eta(\bar{s}), a_T)$ and $\bar{s}' \neq \bar{s}$, strict code separation gives

$$\|f - \eta(\bar{s}')\|_\infty > 2a_T - \|f - \eta(\bar{s})\|_\infty \geq \|f - \eta(\bar{s})\|_\infty.$$

Therefore the minimizer is uniquely $\bar{s}$ on this ball. We have both code consistency and

$$\mathbb{B}_\infty(\eta(\bar{s}), a_T) \subseteq \delta^{-1}(\{\bar{s}\}).$$

Now set $I(\bar{s}) = \{\eta(\bar{s})\}$ and $\mathcal{D}_r(f) = \mathbb{B}_\infty(T_r(f), \sigma)$ for every f and every label r. These maps give serial execution semantics and allow every sequence of disturbances with per-step ℓ_∞ norm at most σ. Fix an arbitrary symbolic trace π of length $m \leq T$. We prove its exact reachable-set formula by induction on t. At $t = 0$, the initial singleton is the asserted radius-zero ball. Suppose that $0 \leq t < m$ and that

$$\mathsf{Tube}_t^\pi = \mathbb{B}_\infty(\eta(\bar{s}_t), t\sigma).$$

Since $t\sigma \le a_T$, the next nominal map is the prescribed translation on this entire predecessor ball. The successor relation and the exact Minkowski-sum identity for norm balls give

$$\mathsf{Tube}^{\pi}_{t+1} = T_{r_{t+1}}(\mathsf{Tube}^{\pi}_{t}) + \mathbb{B}_{\infty}(0,\sigma) = \mathbb{B}_{\infty}(\eta(\bar{s}_{t+1}),(t+1)\sigma).$$

This is an equality of reachable sets, since each element of the Minkowski sum is an admissible successor and every admissible successor belongs to that sum. Induction proves the formula at every prefix. Because $t \le T$, each such ball lies in $\mathbb{B}_{\infty}(\eta(\bar{s}_t), a_T)$ and therefore in the corresponding decoder cell. The argument applies to every trace and all admissible disturbance sequences, with the same encoder, decoder, and transition maps fixed throughout.

Finally, suppose $|\bar{S}| = n^d_{\mathrm{cap}}$ and the transition family includes the identity. For every $\bar{s} \in \bar{S}$, the length-T identity trace starting at $\bar{s}$ ends at that same state. Thus the horizon-T terminal state set contains all n^d_{cap} states. Each terminal reachable set is the radius-$T\sigma$ ball around its encoding, and all these balls are decoded correctly. The exact upper bound in Theorem 4 is attained. This also covers $B = 0$, where the attained count is 1. □

Remark 7 (Scope of the realization)**.** *The proposition is an existence result for a fixed finite horizon. It constructs the encoding, decoder, and ReLU parameters; it does not constrain network size, parameter magnitudes, or training. Finite partial transitions may be included by extending each one arbitrarily to a total map on $\bar{S}$ and restricting correctness to the original valid traces. The construction applies to every finite deterministic transition family, but an upper bound on the total number $|\bar{S}|$ is necessary from horizon-T decoding only when all those states occur at that horizon. An identity transition ensures this coverage. Without coverage, the necessary bound applies to N_T, the number of distinct terminal states, and may be smaller than $|\bar{S}|$.*

A.23 Full-ball disturbance capacity

Corollary 9 (Full-ball disturbance capacity)**.** *Let $T, d \in \mathbb{N}$ with $T \ge 1$, let $B \ge 0$, and let $S_T \subseteq \bar{S}_\tau$ be a finite set satisfying $|S_T| = N_T$. Assume that every state in S_T is realized at horizon T by a trace family $\mathcal{K}_T$ whose traces have exactly T transitions. Let $F_\tau = \mathbb{R}^d$ be equipped with the ℓ_∞ norm and its induced metric $d_\tau(x,y) = \|x-y\|_\infty$. Assume that $\eta_\tau(\bar{s}) \in [-B,B]^d$ for every $\bar{s} \in S_T$, and that the deterministic decoder $\delta_\tau : F_\tau \to \bar{S}_\tau$ is terminally robust for $\mathcal{K}_T$ at horizon T.*

Fix $\sigma > 0$. For every $\bar{s} \in S_T$, assume that there exists a trace $\pi_{\bar{s}} \in \mathcal{K}_T$ ending at $\bar{s}$ such that the hypotheses of Corollary 4 hold along $\pi_{\bar{s}}$ with $\sigma_t = \sigma$ for every $1 \le t \le T$. Equivalently, along each selected trace, the initial tube is the singleton canonical encoding, every nominal transition is exact on the canonical encoding, the nominal transition at time t is locally open with coefficient 1 up to radius $(t-1)\sigma$, and the successor map at time t satisfies full-ball disturbance inclusion with radius σ on the corresponding predecessor ball. Then $N_T \le \left(1 + \frac{B}{T\sigma}\right)^d$.

Proof. Fix $\bar{s} \in S_T$. By assumption, there exists a trace $\pi_{\bar{s}} \in \mathcal{K}_T$ ending at $\bar{s}$ to which Corollary 4 applies with $\sigma_t = \sigma$ at every step. That corollary gives $a_T = \sum_{t=1}^{T} \sigma_t = T\sigma$ and $\mathbb{B}_\infty\left(\eta_\tau(\bar{s}), T\sigma\right) \subseteq \mathsf{Tube}^{\pi_{\bar{s}}}_T$.

Since $\pi_{\bar{s}}$ ends at $\bar{s}$, the definition of the terminal reachable set gives $\mathsf{Tube}^{\pi_{\bar{s}}}_T \subseteq A_T(\bar{s})$. Therefore, $\mathbb{B}_\infty\left(\eta_\tau(\bar{s}), T\sigma\right) \subseteq A_T(\bar{s})$.

Because $\bar{s} \in S_T$ was arbitrary, the number $T\sigma$ is a common inner reachability radius for S_T. Theorem 4, applied with $a_T = T\sigma$, now yields $N_T \le \left(1 + \frac{B}{T\sigma}\right)^d$. □

A.24 Static ℓ_∞ grid construction for decoder cells

Proposition 12 (Static ℓ_∞ grid construction for decoder cells)**.** *Let $B > 0$, $a > 0$, and $\gamma > 0$. Let $d \in \mathbb{N}$ satisfy $d \ge 1$. Define $q = \left\lfloor \frac{2B}{2a+\gamma} \right\rfloor$ and $M_\gamma = (q+1)^d$. Then there exist a quotient state set $\bar{S}$ with $|\bar{S}| = M_\gamma$, an encoder $\eta : \bar{S} \to [-B,B]^d$, and a deterministic decoder $\delta : \mathbb{R}^d \to \bar{S}$ such that $\mathbb{B}_\infty\left(\eta(\bar{s}), a\right) \subseteq \delta^{-1}(\{\bar{s}\})$ for every $\bar{s} \in \bar{S}$.*

Proof. Let $G = \{-B + (2a+\gamma)k : k \in \{0,\dots,q\}\}$. We first verify that $G^d \subseteq [-B,B]^d$. By the definition of q, $q \le \frac{2B}{2a+\gamma}$. Hence, for every $k \in \{0,\dots,q\}$, $-B \le -B + (2a+\gamma)k \le -B + (2a+\gamma)q \le B$. Thus $G^d \subseteq [-B,B]^d$. Moreover, $|G^d| = (q+1)^d = M_\gamma$.

Let $x, y \in G^d$ be distinct. Then there exists a coordinate $j \in \{1,\dots,d\}$ such that $x_j \neq y_j$. For some distinct $k_j, \ell_j \in \{0,\dots,q\}$, $x_j = -B + (2a+\gamma)k_j$ and $y_j = -B + (2a+\gamma)\ell_j$. Therefore, $|x_j - y_j| = (2a+\gamma)|k_j - \ell_j| \ge 2a+\gamma$, and hence $\|x-y\|_\infty \ge 2a+\gamma > 2a$.

We claim that the closed ℓ_∞ balls of radius a centered at the points of G^d are pairwise disjoint. Indeed, suppose that distinct $x, y \in G^d$ and some $z \in \mathbb{R}^d$ satisfy $z \in \mathbb{B}_\infty(x,a) \cap \mathbb{B}_\infty(y,a)$. Then $\|x-y\|_\infty \le \|x-z\|_\infty + \|z-y\|_\infty \le 2a$, contradicting $\|x-y\|_\infty > 2a$.

Choose a set $\bar{S}$ with $|\bar{S}| = M_\gamma$ and a bijection $\eta : \bar{S} \to G^d$. Since $G^d \subseteq [-B,B]^d$, we may regard η as an encoder into $[-B,B]^d$.

Let $U = \bigcup_{\bar{s}\in\bar{S}} \mathbb{B}_\infty(\eta(\bar{s}), a)$. For every $\bar{s} \in \bar{S}$, define $\delta(f) = \bar{s}$ for $f \in \mathbb{B}_\infty(\eta(\bar{s}), a)$. This defines δ unambiguously on U because these balls are pairwise disjoint.

The set $\bar{S}$ is nonempty because $M_\gamma \ge 1$. Fix $\bar{s}_* \in \bar{S}$ and define $\delta(f) = \bar{s}_*$ for every $f \in \mathbb{R}^d \setminus U$. This extends δ to a deterministic map $\delta : \mathbb{R}^d \to \bar{S}$. By construction, $\mathbb{B}_\infty(\eta(\bar{s}), a) \subseteq \delta^{-1}(\{\bar{s}\})$ for every $\bar{s} \in \bar{S}$. □

Remark 8 (Scope of the grid construction)**.** *Proposition 12 establishes only a static decoder-cell packing construction. For fixed d and fixed ratio γ/a, its cardinality satisfies $M_\gamma = \Theta\left((B/a)^d\right)$ as $B/a \to \infty$, matching the polynomial dependence of the static upper bound in Theorem 4. The proposition does not establish dynamical tightness of the horizon-capacity theorem, because it does not construct terminal reachable sets or prove that the displayed balls are reachable under a transition semantics.*

A.25 Quantization-aware outer-radius propagation

Proof. Let $z \in \mathsf{Tube}_0^{\pi,\mathsf{Q}}$. Then $z = \mathsf{Q}_{\tau_0}(y)$ for some $y \in I_{\tau_0}(\bar{s}_0)$, and therefore $d_{\tau_0}(z, \eta_{\tau_0}(\bar{s}_0)) \le d_{\tau_0}(z,y) + d_{\tau_0}(y, \eta_{\tau_0}(\bar{s}_0)) \le \kappa_0 + b_0 = B_0^{\mathsf{Q}}$.

Now fix a trusted predecessor state f and let $z \in \mathcal{D}_t^{\mathsf{Q}}(f)$. There exists $y \in \mathcal{D}_t(f)$ such that $z = \mathsf{Q}_{\tau_t}(y)$. Conditions *(iv)* and *(v)* give $d_{\tau_t}(z, T_t(f)) \le d_{\tau_t}(z,y) + d_{\tau_t}(y, T_t(f)) \le \kappa_t + u_t$. Hence $\mathcal{D}_t^{\mathsf{Q}}(f) \subseteq \mathbb{B}_{\tau_t}(T_t(f), u_t + \kappa_t)$. Theorem 2 now applies to the quantized successor maps with initial radius $b_0 + \kappa_0$ and effective successor uncertainty $u_t + \kappa_t$. It yields the stated recurrence and tube inclusions. The decoder-radius condition then implies decoded correctness for the state-quantized execution by applying Theorem 1 to the serial semantics with initial map I^{Q} and successor maps $\mathcal{D}_t^{\mathsf{Q}}$. □

A.26 Digital address counting bound

Proof. The address set has cardinality $|\mathcal{A}_{p,d}| = (2^p)^d = 2^{pd}$. Since images cannot have larger cardinality than their domains, $|\delta_\tau(\mathcal{Q}_{p,d})| \le |\mathcal{Q}_{p,d}| \le |\mathcal{A}_{p,d}| = 2^{pd}$. The assumption $S_T \subseteq \delta_\tau(\mathcal{Q}_{p,d})$ therefore gives $N_T \le 2^{pd}$. Taking base-2 logarithms and using the integrality of pd proves the claim. □

A.27 Address requirement for quantized terminal execution

Proof. Fix $\bar{s} \in S_T$ and choose a trace $\pi \in \mathcal{K}_T$ ending at $\bar{s}$. The quantized execution semantics is serial, so an admissible quantized trajectory exists. Its terminal state belongs to $\mathcal{Q}_\tau \subseteq \mathcal{Q}_{p,d}$ and decodes to $\bar{s}$. Hence $S_T \subseteq \delta_\tau(\mathcal{Q}_{p,d})$, and Proposition 5 applies. □

A.28 PRE-QUANTIZATION GEOMETRIC PACKING BOUND

Proposition 13 (Pre-quantization geometric packing bound)**.** *Let $d, N_T \in \mathbb{N}$ satisfy $d \geq 1$ and $N_T \geq 1$. Let $B \geq 0$, and equip $F_\tau = \mathbb{R}^d$ with the ℓ_∞ norm and its induced metric. Assume that N_T distinct terminal quotient states have encodings in $[-B, B]^d$, that the deterministic decoder is terminally robust at horizon T for the unquantized execution, and that $a_T > 0$ is a common inner reachability radius for their unquantized terminal reachable sets. Then $d \log_2 \left(1 + \frac{B}{a_T}\right) \geq \log_2 N_T$. If $B = 0$, then $N_T \leq 1$. If $B > 0$, then $d \geq \frac{\log_2 N_T}{\log_2(1+B/a_T)}$.*

Proof. Theorem 4 gives $N_T \leq \left(1 + \frac{B}{a_T}\right)^d$. Taking base-2 logarithms proves the first inequality. The cases $B = 0$ and $B > 0$ follow exactly as stated in Theorem 4. □

A.29 LAYERWISE OPERATOR-NORM LIPSCHITZ CERTIFICATE

Proof. Let $x, x' \in E_0$. Repeated application of the boundedness of each linear layer and the 1-Lipschitz property of each activation gives $\|T(x) - T(x')\|_m \leq \left(\prod_{j=1}^m \|W_j\|\right) \|x - x'\|_0$. This proves the global Lipschitz bound.

Now fix $i \in \{1, \ldots, k\}$ and let $x, x' \in E_0$ differ only in their ith block. The first-layer pre-activation difference is $W_1 x - W_1 x' = W_{1,i}(x_i - x_i')$. Because $W_{1,i}$ is bounded, $\|W_{1,i}(x_i - x_i')\|_1 \leq \|W_{1,i}\| \|x_i - x_i'\|_{0,i}$. Propagating this difference through the remaining layers yields $\|T(x) - T(x')\|_m \leq \left(\prod_{j=2}^m \|W_j\|\right) \|W_{1,i}\| \|x_i - x_i'\|_{0,i}$. Thus the stated value of L_i is a valid childwise Lipschitz constant. □

A.30 EXACT RELU INTERPOLATION OF A FINITE OPERATOR TABLE

Proof. Fix $i \in \{1, \ldots, M\}$. Since $\|x_i - x_i\|_\infty = 0$, we have $\phi_i(x_i) = 1$. If $j \neq i$, then the choice of r gives $\|x_i - x_j\|_\infty > 2r > r$, and hence $\phi_j(x_i) = 0$. When $M = 1$, this second statement is vacuous. Therefore, $G(x_i) = \sum_{j=1}^M \phi_j(x_i) y_j = y_i$.

Let $\mathrm{ReLU}(z) = \max\{z, 0\}$ denote the scalar ReLU function. The identities $|z| = \mathrm{ReLU}(z) + \mathrm{ReLU}(-z)$ and $\max\{a, b\} = \mathrm{ReLU}(a - b) + b$ show that scalar absolute value and binary maximum are exactly realizable using affine maps and ReLU units. Repeated application of the binary maximum identity therefore gives an exact realization of $\|x - x_i\|_\infty = \max_{1 \leq \ell \leq p} |x_\ell - (x_i)_\ell|$. It follows that every ϕ_i is exactly realizable by a finite ReLU network.

The functions $\phi_1, \ldots, \phi_M$ can be computed in parallel, and the affine output map $(c_1, \ldots, c_M) \mapsto \sum_{i=1}^M c_i y_i$ then produces $G(x)$. Thus G is exactly realizable by a finite feed-forward ReLU network with an affine output layer. □

A.31 POLYTOPE CERTIFICATES AND SEPARATION FROM GLOBAL BOUNDS

A.31.1 CERTIFICATE DATA AND DECODER CONTAINMENT

The statement of Theorem 5 concerns a finite collection of types, symbolic states, operators, and permitted input tuples. All neural spaces in this collection are finite-dimensional real vector spaces, and all checked modules are finite feed-forward ReLU networks. For each state $\bar{s} \in \bar{S}_\tau$, the certificate supplies nonempty compact polytopes $P_{\bar{s},1}, \ldots, P_{\bar{s},m_{\bar{s}}}$ and their union $K_{\bar{s}}$. Each polytope must be contained in $\mathcal{C}_\tau(\bar{s})$. Every admissible leaf or initial state for $\bar{s}$ must belong to $K_{\bar{s}}$.

For a finite nearest-code decoder with at least two states, put

$$\Delta_{\bar{s}} = \min_{\bar{t} \neq \bar{s}} \|\eta_\tau(\bar{s}) - \eta_\tau(\bar{t})\|_\infty.$$

When there is more than one state, injectivity gives $\Delta_{\bar{s}} > 0$. Choose $0 \le q_{\bar{s}} < \Delta_{\bar{s}}/2$. If $\|x - \eta_\tau(\bar{s})\|_\infty \le q_{\bar{s}}$, then, for every competing state $\bar{t}$,

$$\|x - \eta_\tau(\bar{t})\|_\infty \ge \Delta_{\bar{s}} - q_{\bar{s}} > q_{\bar{s}} \ge \|x - \eta_\tau(\bar{s})\|_\infty.$$

Thus a component contained in $\mathbb{B}_\infty(\eta_\tau(\bar{s}), q_{\bar{s}})$ lies in the intended decoder cell independently of tie breaking. Containment in this box can be checked by maximizing each positive and negative coordinate displacement over the component. This is a sufficient construction; the theorem also allows other polytopes whose decoder containment has been established. It does not assume that a nearest-code decoder cell in ℓ_∞ is a convex polytope. For a single-state decoder, decoder containment is automatic.

For each permitted input tuple, enumerate all products

$$P_{\bar{s}_1,j_1} \times \cdots \times P_{\bar{s}_k,j_k}.$$

Take a finite polyhedral subdivision of each product, and let Q_ι range over all cells. The unsubdivided product is always an admissible choice. The cells together cover $K_{\bar{s}_1} \times \cdots \times K_{\bar{s}_k}$, including their boundaries. The certificate specifies one target component $P_{\bar{s}_0,j_\iota}$ for each cell. This assignment is part of the certificate, not a property inferred from finite trajectory samples. A target union may require finer source subdivisions. For a fixed subdivision and assignment, a failed inclusion does not rule out a different polytope certificate.

For the input tuple $\bar{\boldsymbol{s}} = (\bar{s}_1, \ldots, \bar{s}_k)$, the additive disturbance set $D_{\alpha,\bar{\boldsymbol{s}}}$ is a nonempty compact polytope. It can be an outer approximation: soundness requires every admissible output from an input x in the declared source product to belong to $T_\alpha(x) + D_{\alpha,\bar{\boldsymbol{s}}}$. No independence assumption is needed for disturbances across nodes or time. If the execution uses deterministic nominal outputs, take $D_{\alpha,\bar{\boldsymbol{s}}} = \{0\}$.

A.31.2 Proof of Theorem 5

Proof. A finite feed-forward ReLU network has finitely many binary activation patterns. For a fixed pattern, recursively replacing an active ReLU by the identity and an inactive ReLU by zero gives an affine output map $A_\beta x + b_\beta$. The inequalities requiring the resulting preactivations to be nonnegative at active units and nonpositive at inactive units define a closed polyhedron $\mathcal{V}_\beta$. Some such polyhedra are empty. Their union covers the input space. Whenever a preactivation is zero, either pattern gives the same activation value, so the affine representation remains valid on shared boundaries.

Fix a source polytope Q_ι and its assigned target $\{y : H_\iota y \le h_\iota\}$. On a nonempty intersection $Q_\iota \cap \mathcal{V}_\beta$, membership of every disturbed output in the target is equivalent to

$$H_{\iota,i}(A_\beta x + b_\beta + \xi) \le h_{\iota,i}$$

for every target row i, every $x \in Q_\iota \cap \mathcal{V}_\beta$, and every $\xi \in D_{\alpha,\bar{\boldsymbol{s}}}$. The feasible set is a nonempty compact polytope and the objective is linear, so its maximum exists. These pointwise inequalities therefore hold exactly when the linear-programming inequalities in the theorem hold. Taking the union over activation regions proves the stated equivalence with $T_\alpha(Q_\iota) + D_{\alpha,\bar{\boldsymbol{s}}} \subseteq P_{\bar{s}_0,j_\iota}$.

Taking the union over source cells proves

$$T_\alpha(K_{\bar{s}_1} \times \cdots \times K_{\bar{s}_k}) + D_{\alpha,\bar{\boldsymbol{s}}} \subseteq K_{\bar{s}_0}.$$

This inclusion applies to every admissible module output by the disturbance assumption. At a leaf, the neural state belongs to the set associated with its symbolic value. Suppose all children of an internal node have this property. Their actual input tuple belongs to the corresponding product, so the inclusion places every admissible output in the set associated with the symbolic parent value. Induction on the finite derivation tree establishes membership at every node. Decoder containment then gives correct decoding at every node.

For a serial transition label and fixed symbolic choice, the map $f \mapsto T_r(f, \chi_r(c))$ is a ReLU map with constant choice coordinates. Apply the same argument to the unary source set, using a distur-

bance polytope that encloses all admissible successors relative to this nominal map. Initial membership and induction on the number of transitions establish correctness along every checked finite trace. The certificate is shared across transitions and does not depend on the length of a particular trace. Hence no common finite horizon bound is required. □

Each linear program can equivalently be checked on the vertices of its feasible polytope. If the network parameters and certificate data, including fixed choice embeddings for serial transitions, are rational, the activation constraints and linear programs have rational coefficients. Exact rational feasibility tests and dual upper-bound certificates can therefore verify the stated inclusions. A primal feasible point whose objective exceeds the target bound certifies a failed assigned inclusion; an arbitrary primal feasible point does not certify an upper bound. Polyhedral subdivision and coverage must also be justified; taking the full component products avoids a separate subdivision-coverage claim. Ordinary floating-point optimization without sound error bounds does not provide exact verification. The number of activation patterns and source-component products can be exponential, so finite reducibility alone does not establish computational scalability.

A.31.3 What the certificate says about training

For a module family $\mathcal{T}_{\mathrm{mod}} = (T_\alpha)_\alpha$ and fixed source polytopes, disturbance sets, target assignments, and target inequalities, define the closure margin by

$$\gamma_{\mathrm{cl}}(\mathcal{T}_{\mathrm{mod}}) = \min_{\iota,\beta,i} \left[h_{\iota,i} - \max_{\substack{x \in Q_\iota \cap \mathcal{V}_\beta \\ \xi \in D_{\alpha,\bar{s}}}} H_{\iota,i}(A_\beta x + b_\beta + \xi) \right],$$

where empty intersections are omitted and the minimum also ranges over the finite permitted operator tuples. If there are no constraints to check, define this margin as $+\infty$. The affine maps and activation regions are those of the modules being checked. Nonnegative margin is equivalent to all assigned inclusions. To compare numerical margins across trained systems, the certificate sets and target-row normalizations must be fixed; for example, each nonzero target row can be normalized to have ℓ_1 norm one. The sign test does not depend on multiplying an individual inequality by a positive constant.

The margin depends on module outputs within the declared source sets. It can improve while a global norm bound worsens, since the latter also constrains irrelevant input regions and may discard affine cancellations. This gives a precise property to test when comparing canonical-input and neighborhood supervision. It does not establish that either training procedure increases the margin. Neighborhood samples alone do not verify the universal inequalities. The 128-dimensional diagnostics use conservative enclosures. The four-dimensional tests verify assigned inclusions by activation-region analysis without requiring exact values of the closure margin.

If a margin is negative, a linear-programming optimizer witnesses failure of the assigned target-polytope inclusion. That output may still belong to another component for the same symbol, or to the decoder cell outside the selected components. Consequently, a negative margin is not by itself a decoding-error witness. It need not be reachable from the declared initial states either. Proposition 2 supplies a separate construction in which the rejected local inclusion corresponds to a reachable decoding failure.

Combining all representations of one symbolic state also introduces requirements absent from a trace-specific analysis. In particular, the source products include combinations of child states that may never occur together. The certificate remains sound under this stronger requirement. For a restricted trace family, one may instead supply separate polytopes for a finite collection of history or control-state indices and check only the permitted index transitions. Every allowed execution must then admit such an index sequence. The same induction applies after adding these indices; sampling a few source combinations is insufficient to establish coverage.

A.31.4 Proof of Proposition 2

Proof. Let the symbolic states be $\bar{S} = \{-1, 1\}$, use the encoder $\eta(s) = s$, and let the symbolic transition be $s \mapsto -s$. The decoder selects the nearest of the two codes, with any fixed convention

at zero. Fix $0 < r < 1/2$, put $u = r/2$, and set $P_s = [s - r, s + r]$. Each interval lies in its intended decoder cell. Any ρ satisfying $r \le \rho < 1$ is a valid common decoding radius containing these intervals.

Write $[z]_+ = \max\{z, 0\}$ and define

$$G_r(x) = 1 - \frac{[x + 1 - r]_+ - [x - 1 + r]_+}{1 - r},$$

$$E_{\mathrm{sep}}(x) = L_{\mathrm{sep}}\big([x - 3]_+ - 2[x - 4]_+ + [x - 5]_+\big),$$

$$V_r(x) = \frac{4}{r}\big([x - 1]_+ - 2[x - 1 - r/2]_+ + [x - 1 - r]_+\big).$$

The two networks are

$$T_{\mathrm{good}} = G_r + E_{\mathrm{sep}},$$

$$T_{\mathrm{bad}} = G_r + E_{\mathrm{sep}} + V_r.$$

These formulas are finite sums of scalar ReLUs followed by an affine output map, so they give finite feed-forward ReLU networks directly.

The map G_r is constant at 1 on $(-\infty, -1 + r]$, affine with slope $-1/(1 - r)$ on $[-1 + r, 1 - r]$, and constant at -1 on $[1 - r, \infty)$. The function E_{sep} is supported on $[3, 5]$ and has slopes L_{sep} and $-L_{\mathrm{sep}}$ on its two nonconstant intervals. Its support is disjoint from both P_{-1} and P_1. The function V_r is supported on $[1, 1 + r]$, vanishes at both endpoints, has value 2 at $1 + r/2$, and has slopes $4/r$ and $-4/r$ on its two nonconstant intervals.

In particular, $T_{\mathrm{good}}(s) = T_{\mathrm{bad}}(s) = -s$ at each canonical code. The nonzero slopes of G_r, E_{sep}, and V_r occur on disjoint open intervals. Since $L_{\mathrm{sep}} \ge 4/r > 1/(1 - r)$, both networks have all slope magnitudes at most L_{sep} and attain a slope of magnitude L_{sep}. Their exact global Lipschitz constants are therefore both L_{sep}.

On the complete source interval P_s, the good network is constant at $-s$. Hence

$$T_{\mathrm{good}}(P_s) + [-u, u] = [-s - u, -s + u] \subseteq P_{-s}.$$

The two intervals, with these inclusions, form a certificate in Theorem 5. Thus every admissible execution starting anywhere in the appropriate interval is correct at every finite time, for arbitrary disturbance choices in $[-u, u]$.

For the bad network, initialize at the canonical code $x_0 = -1$, choose the first disturbance to be u, and choose the second disturbance to be zero. Then

$$x_1 = T_{\mathrm{bad}}(-1) + u = 1 + r/2,$$

$$x_2 = T_{\mathrm{bad}}(1 + r/2) = -1 + 2 = 1.$$

The symbolic states are -1, 1, and -1. The first neural state decodes correctly, while x_2 decodes as 1 and is incorrect. Both disturbances are admissible. The offending second-step input belongs to P_1, so the interval-inclusion test also detects the failure of $T_{\mathrm{bad}}(P_1) + [-u, u] \subseteq P_{-1}$.

Finally, canonical point error is zero for both networks. Applying the scalar outer-radius recurrence with the exact global Lipschitz constant and a canonical initial state gives $B_0 = 0$, $B_1 = u$, and

$$B_2 = L_{\mathrm{sep}} u + u \ge 2 + r/2 > 1.$$

Thus it cannot certify the second step for any common decoding radius $\rho < 1$, even for the good network. A larger global bound, including any valid layerwise product bound, cannot improve this conclusion. The local constant on each good source interval is zero, so a sufficiently precise local Lipschitz calculation would also certify this example. The separation is from global bounds, not from every local verification method. □

A.32 ADDITIONAL EXPERIMENTAL REPORTS

We examine four consequences of the theory using trained arithmetic modules and analytic ReLU controls: correctness on canonical inputs versus noncanonical representations, the usefulness of outer-radius certificates, capacity restrictions from inner reachability, and the effect of state quantization. The learned modules implement addition, subtraction, and multiplication modulo 64, with fixed encodings in $\{-0.5, 0.5\}^{128}$, nearest-code decoding in the ℓ_∞ norm, and certified decoding radius $\rho = 0.49$. Each module has four ReLU hidden layers of width 1024 and $3{,}543{,}168$ parameters. We compare canonical-input supervision with neighborhood supervision, which perturbs both inputs in half of each training batch. Ten paired training seeds are used, and frozen checkpoints are reused across the composition, outer-radius, and quantization experiments.

These experiments use FP64 on a single NVIDIA GeForce RTX 4090. Reported uncertainty is the sample standard deviation of seed-level statistics, not of individual nodes or overlapping trajectory prefixes. The numerical bounds implement constructions valid in exact arithmetic, but do not use outward rounding. Consequently, numerical agreement is not a formal floating-point verification result. Protocols, additional results, and the distinction between analytic and sampled checks appear in Appendix A.33.

A.32.1 CANONICAL COMPOSITION AND REPRESENTATION STRESS

All 60 trained modules decode the complete $4{,}096$-entry operator table correctly. In the unperturbed composition test (Exp1), all $15{,}360$ evaluations of balanced and chain-shaped trees with 8, 32, 128, or 512 internal nodes are correct at every internal node, both with direct execution and with external canonicalization. The largest observed intermediate deviation is 0.0651, below ρ. Thus these unperturbed evaluations do not themselves exhibit a learned direct-composition failure; the affine construction in Example 1 supplies the corresponding analytic counterexample.

Exp1b separately tests sensitivity to noncanonical representations. After a nonroot internal node produces f, its transmitted state is replaced by $\widetilde{f} = \eta(\delta(f)) + \xi$, where $\|\xi\|_\infty \le r < \rho$. Perturbations are sampled either uniformly from the box or at Rademacher corners. The comparison condition applies $P = \eta \circ \delta$ to $\widetilde{f}$ before the parent module receives it. This intervention is centered on the model's own decoded output, not the reference label. It preserves that output's decoding, but changes direct neural execution; in particular, it is not additive noise applied to the unmodified continuous output, and its $r = 0$ case already transmits canonical codes.

The numerical results are reported in Table 1.

The stress test contains $215{,}040$ tree evaluations. All $31{,}150{,}080$ nonzero-radius interventions preserve the preceding prediction. Nevertheless, the point-trained modules fail extensively on the intervened inputs (Table 1). For 512-node balanced trees with corner perturbations of radius 0.20, whole-tree accuracy is $1.41 \pm 2.49\%$ under point supervision and $98.28 \pm 1.72\%$ under neighborhood supervision. The paired improvement is 96.88 ± 2.55 percentage points. Canonicalization after the intervention yields perfect root and whole-tree accuracy in all $71{,}680$ corresponding evaluations, consistent with Theorem 7.

These comparisons show that complete canonical-table accuracy does not establish robustness to decoder-preserving changes in representation. They do not show that the same trained modules fail under unmodified direct execution. Neighborhood supervision improves robustness over much of the sweep but supplies no general closure guarantee: under radius-0.48 corner perturbations, both trained variants have zero whole-tree accuracy on the largest balanced trees. Root accuracy can also conceal intermediate failures. On 512-node chains with radius-0.48 uniform perturbations, point-trained modules achieve $87.66 \pm 4.74\%$ root accuracy but zero whole-tree accuracy.

A.32.2 OUTER-RADIUS CERTIFICATES AND THEIR CONSERVATISM

Exp2 composes the learned operators serially, holds the encoded right operand fixed at each step, and adds bounded disturbances after each update. We evaluate the recurrence in Theorem 2 using either the layerwise induced-ℓ_∞ norm bound or a local bound obtained by interval propagation and absolute-weight sensitivity propagation. The local construction removes ReLUs certified inactive over the entire predecessor box. Its trusted radius is $q = 0.25$. Local propagation is stopped permanently when the preceding radius exceeds q, and path certification additionally requires every preceding decoding-radius test to pass.

Table 4: Outer-radius diagnostics (Exp2). Learned-network entries average 32 trajectories within each of ten seeds. Both learned bounds certify only the first step, although all tested trajectories remain pathwise correct through 512 steps. Control failure times are the earliest failure in each 32-trajectory batch; the displayed values agree across all ten disturbance seeds. The control's local trusted radius is $q = 1/16$ and its certified decoding radius is $\rho = 0.1225$.

Learned modules and analytic controls				
Trained operators, no execution disturbance, second step				
Training	Mean e_2	Mean B_2^{global}	Mean B_2^{local}	Last certified step
Point	0.00905	53,678	11,527	1
Neighborhood	0.01877	202,388	29,080	1
Analytic counter controls, tested through step 512				
L	u	Last certified step	First observed failure	Path accuracy at 512 (%)
0.5	1/1024	512	None observed	100.00
1.0	0	65	None observed	100.00
1.0	1/1024	33	40	6.56
1.02	1/1024	25	31	3.12

All 2,560 full-length learned executions remain pathwise correct through 512 steps over disturbance radii 0, 0.001, 0.005, and 0.02. Both bound constructions certify only step 1. At step 2, the local assumptions still apply, but the resulting radius already exceeds both q and ρ; from step 3 onward the local recurrence is marked invalid. Table 4 illustrates the discrepancy between observed error and propagated bounds without disturbances. Averaged within each training-mode/operator group, the global-to-maximum-local Lipschitz-bound ratio ranges from 4.1 to 7.2, while the maximum local Lipschitz bounds remain of order 10^6 (Table 8). This improvement reflects both absolute-weight sensitivity propagation instead of a product of layerwise maxima and the inactive-unit masks; it should not be attributed solely to pruning inactive ReLUs.

A continuous piecewise-linear ReLU counter provides a complementary control with prescribed local slope L. Within the trusted neighborhoods, positive-corner disturbances attain the propagated radius to within 2.3×10^{-15} over all 17,160 valid seed-step records. The $L = 0.5$ controls satisfy the uniform contraction condition and remain certified through 512 steps. The noncontractive controls separate certificate loss from observed failure: for $L = 1$ without disturbances, certification ends at step 65, but no pathwise failure is observed through step 512. No reported finite valid bound is exceeded beyond the numerical tolerance, and no numerically certified prefix contains a decoding error. These checks illustrate a sufficient condition and its conservatism; they do not establish that failure must occur when certification ends.

A.32.3 INNER REACHABILITY, COLLISIONS, AND CAPACITY

Exp3 first compares the inner-radius recurrence with directly propagated reachable boxes for affine ReLU controls. Of 6,912 parameter-step records, 3,876 retain a valid recurrence, and none reports an inner-inclusion violation. The recursive and exact inner radii agree within 10^{-12} in 3,099 valid records. When the recurrence first becomes negative, it is invalidated rather than truncated to zero.

Alternating biases also give cases in which the target later becomes reachable again without restoring the previously invalidated recurrence.

For the capacity experiment, the symbolic encodings form a full Cartesian grid in $[-1, 1]^d$, and a coordinate permutation realizes the nominal transition. Under full-ball disturbances, the terminal reachable set is exactly the radius-$T\sigma$ ball around its target encoding. Under shared-scalar disturbances $\xi_t = z_t\mathbf{1}$ with $|z_t| \leq \sigma$, the reachable set has the same outer radius $T\sigma$ but remains in a one-dimensional affine subspace and has no positive full-dimensional inner radius. This contrast holds in all 420 matched analytic settings and isolates the assumption needed for Theorem 4; a disturbance-magnitude bound alone is insufficient.

The numerical results are reported in Table 2.

Among the 420 full-ball settings, the complete grid is robust in 200 and has intersecting reachable balls in 220. Exact rational postprocessing identifies 44 first-contact settings and 176 settings strictly beyond first contact. At first contact, $N_T = (1+B/a_T)^d$ for these grids, although robustness already fails because the closed reachable balls touch. Thus satisfying the volume inequality is not sufficient for robust decoding. We construct 60 pairs of admissible trajectories with distinct terminal symbols but identical neural endpoints; every pair necessarily causes at least one decoding failure, and the saved FP64 endpoints agree exactly. Table 2 additionally compares the volume upper bound with the exact packing number of the fixed grid. The latter is a geometric postprocessing quantity and does not assert correctness under the full-grid decoder after states are removed.

A.32.4 State quantization versus address capacity

Exp4 compares unquantized execution, terminal-only quantization, and quantization of the initial state and every subsequent neural state. The weights and nominal computations remain FP64. The state quantizer has 2^p equally spaced levels per coordinate in $[-1, 1]$, with $p \in \{2, 3, 4, 6, 8, 12, 16\}$. The radius recurrence includes κ_0 at initialization and κ_t at each update. These terms account for possible saturation instead of assuming a half-step error bound everywhere.

The numerical results are reported in Table 3.

The learned networks remain pathwise correct in every tested unquantized and terminal-only condition. Repeated state quantization behaves differently (Table 3). With two-bit every-step quantization, every 512-step trajectory has an intermediate error, although neighborhood-trained modules still achieve $96.88 \pm 2.95\%$ terminal accuracy. At three bits, point-trained modules attain $62.81 \pm 17.83\%$ path accuracy at $u = 0.005$, whereas the neighborhood-trained modules are correct on all tested paths. No learned path failures are observed at four bits or more. These are empirical thresholds for the tested weights, encodings, traces, and quantizer, not universal precision requirements.

For the analytic contracting counter, terminal-only four-bit quantization is correct and numerically certified at 512 steps, whereas repeated four-bit quantization has zero path accuracy. Repeated six-bit quantization remains correct and certified through 512 steps, with final local bound 0.049324. Omitting quantization terms from the same control recurrence leaves 63.44% of the sampled final errors above the resulting radius. In the learned networks, the large Lipschitz bounds preclude even first-step certification under every-step quantization; nevertheless, all reported finite valid quantization-aware bounds pass the numerical violation checks. Thus the control, rather than the learned-network bound, provides the nonvacuous long-horizon certificate.

Complete enumeration of 48 address systems confirms the necessary constraint in Proposition 5. Nevertheless, two bits in each of 128 neural-state coordinates provide 256 address bits for 64 symbols, exceeding six required bits, while every tested two-bit 512-step every-step-quantized path fails. Address cardinality therefore does not ensure dynamic correctness. Counting saturation once per complete trajectory gives 38 events, all in low-precision point-trained executions; the quantization-aware checks retain the corresponding error bounds.

A.33 Experimental Protocols and Additional Results

This appendix records the protocols and additional results for Section 9. The evaluation includes 128-dimensional arithmetic experiments and analytic controls, high-dimensional diagnostics, re-

Table 5: Training and evaluation settings for the 128-dimensional modules. Configuration values are shared across the experiment records; only the intervention, disturbance, and state-quantization protocols vary.

Item	Setting
Symbolic operators	Addition, subtraction, multiplication modulo 64
State encoding	64 distinct codes in $\{-0.5, 0.5\}^{128}$; code seed 20260915
Decoder	Nearest code in ℓ_∞; first-index tie breaking; $\rho = 0.49$
Architecture	Input 256, four hidden layers of width 1024, output 128
Parameters	$3{,}543{,}168$ per operator; three operators per trained system
Training	Adam; 6000 updates; batch size 256; mean squared error
Learning rate	Cosine decay from 5×10^{-4} to 5×10^{-5}
Regularization	Weight decay 0; gradient norm clipping threshold 5
Neighborhood supervision	Half canonical inputs, half perturbed inputs; radius 0.15
Checkpoint selection	Full-table evaluation every 200 updates; see text
Training seeds	$0, 1, \ldots, 9$; paired across the two training modes
Composition sizes	$8, 32, 128, 512$ internal nodes; balanced and chain trees
Serial horizons	$1, 8, 32, 128, 512$; 32 trajectories per seed and setting
Lipschitz trusted radius	$q = 0.25$ for learned serial transitions
Numerical tolerance	Absolute 10^{-10} and relative 10^{-10}
Execution environment	FP64; RTX 4090; PyTorch 2.5.1; CUDA 12.4

gional verification of four-dimensional MLPs, and a fixed-gate model with learned translations. Sampling, regional verification, and scalar rechecking provide different forms of evidence and are reported separately.

A.33.1 Training protocol for the 128-dimensional arithmetic modules

The symbolic state set is the set of residues modulo 64. A fixed codebook assigns a distinct vector in $\{-0.5, 0.5\}^{128}$ to each residue. Distinct codes have ℓ_∞ distance 1, so every closed radius-0.49 ball around a code is contained in its nearest-code decoder cell. Ties outside those balls are resolved by the first code index. Each of the three operators receives two concatenated state vectors and has four width-1024 ReLU hidden layers followed by an affine output layer. There are $3{,}543{,}168$ parameters per operator and $10{,}629{,}504$ across the three operators for one training condition.

Point supervision uses canonical input pairs. Neighborhood supervision uses canonical inputs in half of each batch and independent uniform-box perturbations of radius 0.15 on both inputs in the other half; the target remains the canonical encoding of the symbolic result. Both methods minimize mean squared error. Training runs for 6000 Adam updates, with cosine learning-rate decay and no early stopping. Checkpoints are selected lexicographically by the number of full-table decoding errors, the maximum full-table point error, and the full-table mean squared error. This selection does not use test trees or long trajectories. Initialization and symbolic training-pair generation are matched across the two training modes within each seed.

The complete operator table contains $64^2 = 4096$ pairs. All 60 selected modules have zero decoding errors over these pairs, giving $245{,}760$ correctly decoded table evaluations. Table 6 reports the nonzero representation errors: correctness on canonical inputs is an equality after decoding, not equality to the canonical output vector. The Exp1b checkpoint checks agree with the selected modules' full-table statistics.

Aggregation. For a fixed condition, let $z_{s,i}$ be a trajectory-level measurement for seed s and test instance i. We first form $\bar{z}_s = n_s^{-1} \sum_{i=1}^{n_s} z_{s,i}$ and report the mean $\bar{z} = 10^{-1} \sum_{s=0}^{9} \bar{z}_s$ together with the sample standard deviation $\left(\sum_{s=0}^{9} (\bar{z}_s - \bar{z})^2 / 9\right)^{1/2}$. Paired improvements are computed by subtracting method means within each seed before computing the mean and standard deviation of the differences. These deviations describe variation across the ten seeds; they are not standard errors or confidence intervals.

Table 6: Full-table fit across ten training seeds. The error is the maximum ℓ_∞ error over all 4096 input pairs for one trained module, followed by mean $\pm$ sample standard deviation across seeds. The same selected checkpoints are used by Exp1b, Exp2, and Exp4.

Training	Operator	Decoding accuracy (%)	Maximum point error	Selected update range
Point	add	100.00	0.01486 ± 0.00082	5600–6000
Point	sub	100.00	0.01487 ± 0.00045	5800–6000
Point	mul	100.00	0.01536 ± 0.00134	5600–6000
Neighborhood	add	100.00	0.04488 ± 0.00173	5600–6000
Neighborhood	sub	100.00	0.04470 ± 0.00252	5200–6000
Neighborhood	mul	100.00	0.04206 ± 0.00209	5600–6000

Root accuracy checks only the final node of a tree. Whole-tree accuracy is the proportion of trees for which every internal node decodes correctly; node accuracy instead averages over individual internal nodes. For serial executions, terminal accuracy checks only the terminal state, whereas path accuracy includes the initial state and every preceding state. The error at a state is its ℓ_∞ distance from the correct canonical encoding. Tree size denotes the number of internal operator nodes, not a common tree depth.

The 15,360 unperturbed tree evaluations and 215,040 representation-stress evaluations count separate method/condition evaluations, some with matched symbolic trees. Serial horizons are nested prefixes of a trajectory generated to length 512. Consequently, Exp2's 12,800 saved horizon records correspond to 2560 full-length method/condition executions, and Exp4's 96,000 horizon records correspond to 19,200 full-length executions. Counts accumulated along a trajectory are therefore evaluated at the final horizon, not summed over its saved prefixes. Analytic controls reuse ten disturbance seeds; these are not ten independently trained control models.

Numerical interpretation. The logged environment uses Python 3.11.5, NumPy 1.23.5, PyTorch 2.5.1 with CUDA 12.4, and a GeForce RTX 4090. Neural computations use FP64 with deterministic algorithms enabled and TF32 disabled. Bound-exceedance diagnostics compare a finite observed error e with a finite valid radius B using the tolerance $10^{-10} + 10^{-10}|B|$. An invalid local recurrence is recorded as missing rather than continued as a certificate. Overflowed global radii are not informative finite certificates. Neither interval endpoints nor point errors are enclosed using directed outward rounding, so a successful diagnostic is not a formally verified floating-point certificate.

Numerical endpoint tests. Let $\widehat{J}$ denote the FP64 endpoint Jacobian. Numerical rank counts singular values exceeding $100 \max\{128, \dim \mathcal{W}\} 2^{-52} \max\{\|\widehat{J}\|_2, 1\}$. For a full-row-rank matrix under this threshold, an SVD-based candidate $\widehat{\Lambda}$ is accepted when the computed induced-ℓ_∞ residual $\|\widehat{J}\widehat{\Lambda} - I\|_\infty$ is at most 10^{-8}. The nominal-centered radius estimate is $h/\|\widehat{\Lambda}\|_\infty$ minus the hybrid-zonotope remainder estimate; the target-centered estimate additionally subtracts the nominal endpoint error. The positive-radius flag requires a finite estimate greater than 10^{-10}. Rejected inverse checks leave the radius entries undefined rather than zero. These are numerical acceptance rules, not validated arithmetic certificates.

A.33.2 Composition and decoder-preserving representation intervention

Exp1 uses balanced and chain-shaped trees with 8, 32, 128, and 512 internal nodes, with 64 trees per seed and condition. The three comparisons are point-trained direct execution, neighborhood-trained direct execution, and point-trained execution with exact external canonicalization. All are correct at every tested internal node. The maximum observed intermediate error is 0.035568 for point-trained direct execution and 0.065094 for neighborhood-trained direct execution. The affine counterexample in Example 1 is analytic, not a learned network failure.

Exp1b uses the same frozen modules, but changes the transmitted representation at each nonroot internal node. If its raw neural output is f_v, write $z_v = \delta(f_v)$ and form $\widetilde{f}_v = \eta(z_v) + \xi_v$. The stressed-input condition passes $\widetilde{f}_v$ to its parent. The canonicalized condition instead passes $P(\widetilde{f}_v) =$

Table 7: Exp1b results pooled equally over the two tree shapes and four operator counts within each seed, then summarized across ten seeds. Each seed-level entry averages 512 tree evaluations at a fixed perturbation distribution and radius. All percentages are mean ± sample standard deviation. Canonicalization has 100% root and whole-tree accuracy in every condition and is omitted here to reduce repetition.

Distribution	r	Point root	Point whole tree	Neighborhood root	Neighborhood whole tree
Corners	0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
Corners	0.05	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
Corners	0.10	99.65 ± 0.18	99.04 ± 0.37	100.00 ± 0.00	100.00 ± 0.00
Corners	0.20	84.10 ± 2.69	62.36 ± 3.80	99.94 ± 0.13	99.75 ± 0.28
Corners	0.30	60.96 ± 3.38	30.39 ± 2.59	92.83 ± 2.75	84.51 ± 3.74
Corners	0.40	38.63 ± 1.76	15.74 ± 1.35	73.69 ± 2.29	50.61 ± 3.98
Corners	0.48	23.81 ± 2.26	7.52 ± 1.23	58.14 ± 2.90	30.61 ± 2.50
Uniform	0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
Uniform	0.05	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
Uniform	0.10	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00	100.00 ± 0.00
Uniform	0.20	99.24 ± 0.46	97.44 ± 0.77	100.00 ± 0.00	100.00 ± 0.00
Uniform	0.30	90.94 ± 2.33	75.92 ± 4.25	100.00 ± 0.00	99.98 ± 0.06
Uniform	0.40	77.09 ± 1.92	49.71 ± 2.71	99.51 ± 0.42	98.87 ± 0.69
Uniform	0.48	66.11 ± 3.67	35.94 ± 3.11	95.88 ± 1.80	90.80 ± 2.55

$\eta(\delta(\widetilde{f}_v))$. Leaves are initialized canonically and are not perturbed; root outputs are scored without an outgoing intervention. The reference symbolic computation supplies evaluation targets only and never substitutes its labels for z_v.

The radii are $r \in \{0, 0.05, 0.10, 0.20, 0.30, 0.40, 0.48\}$. Uniform-box sampling draws coordinates independently from $[-r, r]$, whereas Rademacher-corner sampling sets each coordinate independently to r or $-r$ with equal probability. For both distributions, $\|\xi_v\|_\infty \le r < 0.49$, ensuring that the intervention preserves the current predicted symbol. This statement concerns the current prediction; after an earlier operator error, that prediction can already differ from the correct symbolic state. Across the nonzero-radius conditions, all 31,150,080 logged interventions preserve the preceding prediction and no nonfinite output is reported.

At $r = 0$, this protocol transmits $\eta(\delta(f_v))$ even in the condition named `offcode_direct` in the CSV files. That condition is therefore not unmodified direct execution. The stress experiment isolates a lack of robustness to changes in representation within a decoder cell. It does not establish a failure of the same learned modules under unmodified noiseless composition. Before the first decoding failure, each intervention preserves the correct symbolic state, so the first failure does expose an operator that is not robust to the intervened, noncanonical input tuple.

For 512-node balanced trees with radius-0.20 corner perturbations, neighborhood supervision improves root accuracy by 60.31 ± 8.50 percentage points and whole-tree accuracy by 96.88 ± 2.55 percentage points in paired seed comparisons. With radius-0.48 uniform-box perturbations, its paired root improvement is 65.63 ± 8.24 percentage points. These differences coexist with failures of both training modes in the most severe balanced-tree corner condition reported in Table 1.

The large difference between root and whole-tree accuracy on chains is also important: a later operation can return to the correct residue despite an earlier symbolic error. High root accuracy therefore does not imply that intermediate semantics were preserved. Conversely, perfect accuracy of the canonicalized control is consistent with Theorem 7, but does not transfer to direct execution, because external canonicalization changes the states supplied to later modules. Sampling the two disturbance distributions does not verify closure of certified neighborhoods over an entire ball.

A.33.3 Outer-radius constructions and diagnostic validity

At each serial step, the current neural state is the left operand, while the right operand is the fixed canonical encoding of the selected symbolic choice. The symbolic schedule is generated before neural execution. Point errors and Lipschitz bounds are looked up using the intended predecessor

Table 8: Lipschitz diagnostics, averaged over ten training seeds. Both bound columns are in units of 10^6. The local column first maximizes over the complete state–choice table within each trained module. The final column is the percentage of hidden ReLUs removed by the interval masks, averaged over table entries and seeds; the small values preclude attributing the entire global-to-local improvement to inactive-unit removal.

Training	Operator	Global bound	Maximum local bound	Inactive units (%)
Point	add	6.513	1.564	0.0235
Point	sub	6.162	1.517	0.0234
Point	mul	6.733	1.110	0.0219
Neighborhood	add	11.447	1.720	0.0464
Neighborhood	sub	11.738	1.687	0.0475
Neighborhood	mul	9.657	1.344	0.0465

state, choice, and operator for that trace. They affect only the radius calculation, not the propagated neural state. The initial radius is $b_0 = 0$, disturbance radii are $u \in \{0, 0.001, 0.005, 0.02\}$, and there are 32 trajectories per seed and condition. Two disturbance trajectories use positive and negative corners; the others use uniform-box samples.

For layer weights $W_1, \ldots, W_m$, let $W_{1,\mathrm{left}}$ be the block multiplying the left input. The global bound is $L_{\mathrm{global}} = \|W_m\|_\infty \cdots \|W_2\|_\infty \|W_{1,\mathrm{left}}\|_\infty$, where $\|A\|_\infty = \max_i \sum_j |A_{ij}|$. Proposition 6 makes this an upper bound for the state-variable Lipschitz constant with the choice fixed.

The local construction uses a box of radius $q = 0.25$ on the left input and zero radius on the right input. For an intermediate box represented by midpoint m and coordinate radii r, an affine layer gives midpoint $c = Wm + b$ and radii $h = |W|r$, hence bounds $c - h$ and $c + h$. Applying ReLU to both endpoints gives the next activation box. A hidden unit whose upper endpoint is nonpositive is identically zero on the box. Its sensitivity is set to zero; every other ReLU retains the upper slope bound 1.

More precisely, the first affine layer has a coordinatewise sensitivity bound $s_1 = |W_{1,\mathrm{left}}|\mathbf{1}$. At a ReLU layer the mask D_j is diagonal, with a zero for a nonpositive interval upper endpoint and a one otherwise. Subsequent sensitivities propagate through $|W_{j+1}|D_j s_j$. Taking the largest output-coordinate sensitivity gives the local ℓ_∞ Lipschitz upper bound in exact arithmetic. This calculation covers the entire specified box; it is not a maximum over sampled Jacobians. It is nevertheless conservative, because absolute weights discard sign cancellation and uncertain activation patterns are overapproximated. The distinction between coordinatewise propagation and multiplication of layerwise maxima also contributes to the global/local difference.

The local radius is used at step t only if the previous radius satisfies $B_{t-1} \le q$. Once this condition fails, local validity is not restored by later numerical values. Path certification additionally accumulates every test $B_j \le \rho$ through the current step. This prevents a later small radius from erasing an earlier uncertified prefix. For the learned networks, step 2 still has a valid local bound, but its radius is too large to certify decoding or to justify local propagation at step 3. No observed learned path fails through the maximum horizon. At $u = 0.02$, the mean maximum error over a complete trajectory is 0.063844 for point training and 0.058766 for neighborhood training.

The control counter has eight scalar codes from -0.875 to 0.875, with spacing $1/4$, repeated over 16 coordinates. On a radius-$1/16$ neighborhood of each code, its transition is affine with slope L and output offset $\epsilon = 1/1024$ from the successor code. Linear interpolation between these neighborhoods is implemented by ReLU hinge functions. The cycle is not made globally contractive; only the separate trusted neighborhoods have the prescribed local slopes. We test $L \in \{0.5, 1, 1.02\}$ and $u \in \{0, 1/1024, 1/128\}$ through 512 steps. The decoder radius is $0.49/4 = 0.1225$.

For $L = 0.5$, the uniform bound $(\epsilon + u)/(1 - L)$ is respectively 0.001953125, 0.00390625, and 0.017578125, each below both the trusted and decoder radii. These parameters instantiate Corollary 3 in exact arithmetic. In the numerical control, the positive-corner trajectory attains the local recurrence within 2.3×10^{-15} while that recurrence is valid. Once a local assumption fails, its formula is not used to infer later error growth. In particular, the observed failure times in Table 9 are

Table 9: Complete Exp2 counter-control results. Validity tests the predecessor radius against $q = 1/16$; certification also tests all current and preceding radii against $\rho = 0.1225$. "First failure" is the earliest sampled path failure within a seed's batch, reported as a range across ten seeds when it differs. "None by 512" is right-censoring at the tested horizon, not a failure time of 513.

L	u	Last valid bound	Last certified step	First failure
0.5	0	512	512	None by 512
0.5	1/1024	512	512	None by 512
0.5	1/128	512	512	None by 512
1.0	0	65	65	None by 512
1.0	1/1024	33	33	40
1.0	1/128	8	8	15
1.02	0	42	42	48
1.02	1/1024	25	25	31
1.02	1/128	7	7	13–14

Table 10: Inner-box diagnostics. Each row contains 1152 parameter-step records from nine bias/noise combinations and 128 steps. Equality means absolute difference at most 10^{-12} on a valid recurrence. The recovery column counts target-reachable records after the recurrence has been invalidated; it does not count reactivated certificates.

Bias pattern	Scale	Valid records	Equal radii	Reachable after invalidation	Violations
Constant	0.5	642	642	0	0
Constant	1.0	648	648	0	0
Constant	65/64	648	648	0	0
Alternating	0.5	642	387	2	0
Alternating	1.0	648	387	435	0
Alternating	65/64	648	387	435	0

measurements, not claims that the scalar local recurrence remains exact after the trajectory leaves its trusted neighborhood.

A.33.4 Exact reachable boxes, capacity, and collision witnesses

The first inner-reachability control starts from a centered box with radius $1/32$ in 16 dimensions. Its affine ReLU realization has scale $\mu \in \{0.5, 1, 65/64\}$, bias magnitude in $\{0, 1/256, 1/16\}$, and full-box disturbance radius $v \in \{0, 1/256, 1/64\}$. Both constant and alternating bias signs are tested for 128 steps. If c_t is the reachable-box center relative to the intended encoding and r_t its coordinate radius, direct propagation gives $c_t = \mu c_{t-1} + b_t$ and $r_t = \mu r_{t-1} + v$, with $c_0 = 0$ and $r_0 = 1/32$. Its exact outer radius is $r_t + \|c_t\|_\infty$. An inner ball centered at the intended encoding exists exactly when $r_t \geq \|c_t\|_\infty$, and its maximal radius is then $r_t - \|c_t\|_\infty$.

Theorem 8 is evaluated separately using $a_t = \mu a_{t-1} + v - \|b_t\|_\infty$. A negative candidate invalidates this recurrence; it is not replaced by zero. There are 24 first-negative records. Across the 3036 records with invalidated recurrences, the exact target center is reachable in 872, all with alternating bias. This recovery is a property of the directly propagated box, not a reactivation of the invalidated recurrence. Table 10 separates validity, equality, and later target reachability.

Dynamic grid experiment. For capacity, dimensions are $d \in \{2, 4, 8, 16\}$ and the number of levels per coordinate is $\ell \in \{2, 3, 5, 9, 17\}$. The encodings occupy the full Cartesian grid $G_\ell^d \subset [-1, 1]^d$ with spacing $\Delta = 2/(\ell - 1)$, so $N = \ell^d$. A coordinate permutation matrix C maps the grid onto itself and is implemented as an affine ReLU map. Disturbance radii are $\sigma \in \{1/1024, 1/256, 1/64\}$, and the analytic sweep uses $T \in \{1, 8, 32, 64, 128, 256, 512\}$.

Table 11: First contact of the closed full-ball reachable sets, independent of dimension for these grids. These times are determined by the nearest-center spacing. Witness construction includes horizons up to 1024, although the sampled capacity sweep stops at 512.

Levels/coordinate	$\sigma = 1/1024$	$\sigma = 1/256$	$\sigma = 1/64$
2	1024	256	64
3	512	128	32
5	256	64	16
9	128	32	8
17	64	16	4

For an initial code x_0, the terminal state has the form $f_T = C^T x_0 + \sum_{j=1}^{T} C^{T-j}\xi_j$. With $\xi_j \in \mathbb{B}_\infty(0,\sigma)$, coordinate permutations preserve the disturbance ball, and its repeated Minkowski sum is exactly $\mathbb{B}_\infty(0, T\sigma)$. With $\xi_j = z_j\mathbf{1}$ and $|z_j| \le \sigma$, the corresponding disturbance sum is instead $\{z\mathbf{1} : |z| \le T\sigma\}$. Here the superscripts of C denote integer powers, not transposes. These identities establish the full-ball and shared-scalar reachable sets. The accompanying sampled trajectories are numerical checks of these constructions; they do not establish a full-dimensional inner ball by sampling.

Closed full-ball reachable sets of distinct nearest grid neighbors are disjoint exactly when $2T\sigma < \Delta$. Hence first contact is at $T_{\text{contact}} = \lceil \Delta/(2\sigma) \rceil$. All ratios in this experiment are integral, so contact occurs at equality. The resulting times are shown in Table 11. The full-grid decoder is robust before contact; at or beyond contact, the two different target symbols share a reachable neural state, so no deterministic decoder can correctly decode all full-grid trajectories.

Explicit witnesses. For each of the 60 combinations of dimension, grid density, and noise radius, two distinct terminal encodings y and y' at distance Δ are selected at first contact. Let $w = (y + y')/2$, and choose initial codes with $C^T x_0 = y$ and $C^T x_0' = y'$. For the first trajectory, disturbances $\xi_j = C^{-(T-j)}(w - y)/T$ give terminal point w and satisfy $\|\xi_j\|_\infty = \Delta/(2T) \le \sigma$; using y' gives the second trajectory. Thus the terminal symbols differ while the neural endpoints agree. The saved witness data contain 60 such pairs and 10,416 paired-step records. The recorded disturbances obey their bounds, the recorded transition residuals are zero, and the two endpoints agree exactly in FP64. This witness construction is distinct from observing a collision in a random finite sample.

Exact boundary accounting. Using the rational input parameters rather than floating logarithms, the 420 full-ball settings separate into 200 robust settings, 44 exact first-contact settings, and 176 settings beyond contact. For the grids at exact contact, the volume expression $(1+1/a_T)^d$ equals ℓ^d, yet the closed balls already intersect. The recorded floating-point strict-comparison flag classifies 12 of these equalities as strict exceedances. The reported counts use exact rational comparisons.

Packing within the fixed grid. The postprocessing quantity M_T maximizes the number of pairwise disjoint closed radius-a_T balls whose centers belong to the existing grid. Set $k = \lfloor 2a_T/\Delta \rfloor + 1$ and $m = \lceil \ell/k \rceil$. Then $M_T = m^d$. To obtain an upper bound, partition each coordinate's grid indices into consecutive blocks of at most k indices. Two centers in the same product block differ by at most $(k-1)\Delta \le 2a_T$ in every coordinate, so at most one retained center can lie in that block. There are m^d blocks. Conversely, the Cartesian product of indices $0, k, 2k, \ldots$ contains exactly m^d centers, and two distinct selected centers differ by at least $k\Delta > 2a_T$ in some coordinate. Their balls are disjoint, attaining the bound.

This is an exact packing number for the specified finite grid, not an empirical accuracy estimate. Achieving robust decoding on a retained codebook requires a decoder for that codebook and a compatible restricted trace family. Merely removing test labels while leaving the full-grid decoder unchanged does not establish that guarantee. Likewise, the ratio $(1 + B/a_T)^d/M_T$ combines the looseness of the volume bound with the restriction to the fixed grid; it is not the exact capacity of arbitrary encodings in the ambient box. Integer counts such as $17^{16} = 48{,}661{,}191{,}875{,}666{,}868{,}481$ are stored without conversion to a fixed-width signed integer.

Table 12: Low-precision path failures in learned 512-step every-step-quantized executions. Path accuracy is mean ± sample standard deviation across ten seeds. Median first-failure times are computed only among the failed trajectories and are not unconditional survival medians.

Training	u	p	Path accuracy (%)	Failed paths	Conditional median
Point	0.0	2	0.00 ± 0.00	320/320	11
Point	0.0	3	63.44 ± 20.25	117/320	236
Point	0.005	2	0.00 ± 0.00	320/320	11
Point	0.005	3	62.81 ± 17.83	119/320	236
Neighborhood	0.0	2	0.00 ± 0.00	320/320	53.5
Neighborhood	0.0	3	100.00 ± 0.00	0/320	–
Neighborhood	0.005	2	0.00 ± 0.00	320/320	53.5
Neighborhood	0.005	3	100.00 ± 0.00	0/320	–

A.33.5 State quantization, omitted terms, and address enumeration

The learned-module precision sweep uses $p \in \{2, 3, 4, 6, 8, 12, 16\}$ bits per coordinate and disturbance radii $u \in \{0, 0.005\}$. The quantizer has endpoint-inclusive levels in $[-1, 1]$ with spacing $h_p = 2/(2^p - 1)$, clips to that interval, and rounds to the nearest level. Neural weights and nominal updates remain FP64. The `none` mode does not quantize; `terminal_only` quantizes the observed endpoint for each evaluated horizon but propagates the unquantized state into the next step; `every_step` quantizes the initial state and all subsequent updates. Thus the last comparison changes both the initial-state representation and the repeatedly transmitted states. It does not separately isolate those two effects.

Quantizer error bound. Suppose the pre-quantization reachable set is enclosed by a radius-R ball about target code c. For quantizer range $[\ell_Q, u_Q]$, the clipping-aware error bound is

$$\kappa(c, R) = \max\left\{ \frac{h_p}{2},\ \left(\max_j c_j + R - u_Q\right)_+,\ \left(\ell_Q - \min_j c_j + R\right)_+ \right\},$$

where $(z)_+ = \max\{z, 0\}$. This controls the whole pre-quantization ball, not just clipping observed on sampled trajectories. The initial radius includes κ_0, and each subsequent radius includes the corresponding κ_t as in Theorem 10. Quantized states, rather than reference labels, are passed to the next neural update in every-step execution.

All unquantized and terminal-only learned executions are pathwise correct at every tested condition. Table 12 gives the low-precision failures under every-step quantization. The conditional first-failure medians exclude trajectories surviving to step 512; for the three-bit point models most trajectories survive, so the reported value 236 is therefore conditional on failure rather than a median over all trajectories. At four bits and above, no learned path failure is observed in this sweep.

Nonvacuous control certificates. We use the same eight-state ReLU counter, with 16 coordinates, local slope $L = 0.5$, point offset $1/1024$, and execution-disturbance radius $1/128$. The unquantized uniform bound is 0.017578125. In every-step execution, six bits suffice for the computed bound to remain inside the trusted neighborhood through the tested horizon, whereas four-bit execution has path failures. Terminal-only four-bit quantization remains both correct and numerically certified. The local bound for a terminal-only output need not be smaller than the trusted predecessor radius, because that quantized endpoint is not propagated into a later neural update.

The six-bit every-step control has final local bound $0.0493241567 < 1/16$, and no numerically certified prefix contains a decoding error. At step 512, 63.4375% of the sampled errors exceed the control recurrence from which quantization terms were omitted. That final-step fraction differs from the cumulative indicator reported for the learned networks in Table 14. The unquantized control is saved repeatedly under the seven precision labels; these copies coincide and are not additional independent control trials.

Table 13: Quantized counter control at step 512. Path and certification columns are percentages averaged over ten disturbance seeds. The final column is the fraction whose current endpoint error exceeds the radius computed without quantization terms; it is not a cumulative event count. Unquantized execution is correct and certified throughout, with $B_{512} = 0.017578125$. The repeated unquantized baselines saved under different bit labels are identical and are not treated as independent evidence.

p	State quantization	Path (%)	Certified (%)	Local radius	Omitted-term excess (%)
2	Terminal only	50.00	0.00	0.350911	100.00
2	Every step	0.00	0.00	Invalid	25.00
3	Terminal only	100.00	0.00	0.160435	100.00
3	Every step	0.00	0.00	Invalid	12.50
4	Terminal only	100.00	100.00	0.084245	75.00
4	Every step	0.00	0.00	Invalid	0.00
6	Terminal only	100.00	100.00	0.033451	56.25
6	Every step	100.00	100.00	0.049324	63.44
8	Terminal only	100.00	100.00	0.021500	1.25
8	Every step	100.00	100.00	0.025421	3.75
12	Terminal only	100.00	100.00	0.017822	3.12
12	Every step	100.00	100.00	0.018067	3.12
16	Terminal only	100.00	100.00	0.017593	3.12
16	Every step	100.00	100.00	0.017609	3.12

Table 14: Percentage of complete learned trajectories with at least one observed error exceeding the global recurrence that omits quantization terms. Each entry uses the 320 complete trajectories at the specified training mode, noise radius, and precision, not the sum over nested horizon records. The local omitted-term indicator gives the same per-trajectory percentages on this data, but is checked only while its own predecessor-radius condition remains valid.

p	Point, $u = 0$	Point, $u = 0.005$	Neighborhood, $u = 0$	Neighborhood, $u = 0.005$
2	100.00	100.00	100.00	100.00
3	100.00	100.00	100.00	100.00
4	100.00	100.00	98.12	94.69
6	55.31	48.44	65.00	51.88
8	91.56	28.75	56.56	16.56
12	52.19	1.56	45.94	0.94
16	53.75	0.94	41.88	0.94

Omitting quantization error. The intentionally incomplete recurrence starts from b_0 and omits both initial and subsequent quantization contributions. For learned trajectories, we count whether it is exceeded at least once through step 512. An exceedance establishes that this incomplete expression is not an upper bound for that sampled execution. Its absence does not justify omission: the learned global radius quickly becomes enormous or overflows, and local comparisons are meaningful only until their predecessor-radius condition fails.

No finite valid quantization-aware comparison exceeds the stated numerical tolerance, no numerically certified prefix has a decoding error, and no nonfinite neural state is recorded in the learned sweep. Certificate coverage is nevertheless limited. Every-step quantization is uncertified even at the first learned step at all tested precisions because the initial quantization bound is multiplied by the large Lipschitz bound. The analytic control, not the learned-network certificate, supplies the long-horizon positive-certification result.

At the final horizon there are 38 saturation events: 17 for point-trained two-bit executions without disturbance, 20 for point-trained two-bit executions with $u = 0.005$, and one for point-trained three-bit execution without disturbance. Summing the same cumulative counters over all saved horizon prefixes would give 43 and would double-count earlier events. Saturation-aware error terms are retained regardless of whether a sampled trajectory clips.

Table 15: Low-dimensional complete address enumeration, grouped by symbol levels per coordinate. Each row spans dimensions $1, 2, 3$ and precisions $1, 2, 3, 4$. A violation would require full addressability despite failure of the necessary counting condition.

Symbol levels	Configurations	Count condition passes	All symbols addressable	Violations
2	12	12	12	0
3	12	9	9	0
5	12	6	6	0
9	12	3	3	0

Address enumeration. The address control completely enumerates dimensions $d \in \{1, 2, 3\}$, coordinate precisions $p \in \{1, 2, 3, 4\}$, and symbol-level counts $\ell \in \{2, 3, 5, 9\}$. This gives 48 systems, with at most 4096 addresses per system. Full addressability and satisfaction of the necessary cardinality condition coincide in 30 of these constructed systems, and both fail in the other 18 (Table 15). This agreement in the chosen finite control is not a theorem that cardinality alone guarantees addressability for arbitrary encoders or decoders.

For the learned system, $N = 64$ and $d = 128$. Even $p = 2$ gives $pd = 256 \geq \lceil \log_2 N \rceil = 6$, yet all tested two-bit 512-step every-step executions have a path failure. These results show that address cardinality does not ensure dynamic correctness, not that the counting inequality predicts the observed precision threshold.

A.33.6 High-dimensional regional and endpoint diagnostics

Regional inclusion and direct additive perturbations. We further test the 20 frozen addition modules with the right operand fixed to 1. For all 64 states, source and target boxes have the same radius $r \in \{0.05, 0.10, 0.20\}$, and additive output disturbances have radius 0.005. None of the $11{,}520$ method-specific checks using global layer norms, interval bound propagation, or hybrid-zonotope propagation establishes the required inclusion. At $r = 0.05$, hybrid zonotopes reduce the mean output-deviation bound from 3.26×10^5 to 1.56×10^3 under point supervision and from 5.72×10^5 to 1.32×10^3 under neighborhood supervision, but both remain far above the target radius. Separately, we apply positive-corner, negative-corner, and uniform-box disturbances directly to the continuous outputs, without canonicalization. All $30{,}720$ logged 512-step execution records are pathwise correct, with maximum observed error $0.0633 < \rho$. The deterministic corner conditions contain eight identical replicates per start state; counting each such trajectory once gives $12{,}800$ distinct execution traces across the three disturbance conditions. Thus substantially tighter regional enclosures still fail to establish closure, whereas the tested executions remain correct; an inconclusive enclosure test is not a witness that the underlying inclusion fails.

Multistep endpoint diagnostics. Using the same addition transition and 16 initial states per seed, we inject disturbances into successive coordinate blocks. The rank-16 sweep over horizons $m \in \{1, 2, 4, 8, 16\}$ and radii $h \in \{0.001, 0.005\}$ yields no accepted positive target-centered inner-radius estimate. We fix the total disturbance dimension at 128 and compare (per-step rank, horizon) pairs $(16, 8)$, $(32, 4)$, and $(64, 2)$ at $h \in \{0.0005, 0.001, 0.002\}$. At $(16, 8)$, the mean numerical endpoint ranks are 105.48 ± 0.66 and 97.26 ± 0.38 under point and neighborhood supervision, respectively. Both shorter configurations attain numerical rank 128 in all 160 seed–start pairs per training mode. However, no $(32, 4)$ pair passes the numerical right-inverse residual check, compared with $160/160$ point-trained and $159/160$ neighborhood-trained pairs at $(64, 2)$. These rank and residual counts are unchanged across the three disturbance radii. At $(64, 2)$ and $h = 0.0005$, three neighborhood-trained cases have positive numerical radius estimates about $\Psi(0)$, ranging from 3.40×10^{-10} to 6.60×10^{-9}, with no unstable ReLUs in the computed enclosures. Their nominal endpoint errors are between 0.0151 and 0.0237, so recentering at the intended encoding makes all three estimates negative. Neither sweep produces an accepted positive target-centered estimate. These results distinguish full numerical rank of a local linearization from quantitative reachability around the symbolic encoding. Because rank and horizon vary jointly, the comparison does not isolate an effect of injection rank alone; nor do ordinary FP64 calculations establish the theorem's exact right-inverse and inclusion assumptions.

Table 16: Regional verification at $q = 1/8$ and $v = 1/256$. Percentages are mean $\pm$ sample standard deviation across ten seeds. Complete denotes neighborhood-trained seeds with every transition of the indicated operator or family certified.

Transition	Point certified (%)	Neighborhood certified (%)	Complete
Increment	0	83.75 ± 12.22	$2/10$
Decrement	0	90.00 ± 5.27	$0/10$
Doubling	0	100.00 ± 0.00	$10/10$
All three	0	91.25 ± 4.03	$0/10$

A.33.7 Regional verification of learned four-dimensional modules

Task and training. The states are residues modulo 16, encoded by the four binary digits of each residue with $1/2$ subtracted from each coordinate. We train separate increment, decrement, and doubling modules, each a 4–64–4 network with one ReLU hidden layer and 580 trainable parameters. All weights and biases are optimized from random initialization. Ten paired seeds compare canonical-input supervision with neighborhood supervision, which perturbs half of each batch uniformly within a radius-0.25 box; targets remain canonical output codes. Training uses 12,000 Adam updates, batch size 256, zero weight decay, gradient clipping at 5, and cosine learning-rate decay from 0.003 to 0.0001. Every 200 updates, checkpoints are compared by canonical decoding errors, maximum representation error, and mean squared error, in that order. Verification outcomes do not enter checkpoint selection.

Verification protocol. For each state s and operator r, we test $T_r(K_s) + \mathbb{B}_\infty(0, v) \subseteq K_{r(s)}$, where $K_s = \mathbb{B}_\infty(\eta(s), q)$. The common source and target radius ranges over $q \in \{1/32, 1/16, 1/8, 1/4\}$, and $v \in \{1/256, 1/128\}$. These boxes lie strictly inside the nearest-code decoder cells. We compare global layer-norm bounds, interval propagation, and activation-region verification with affine ReLU relaxations. The latter uses at most 8192 search nodes and 120 seconds per state–operator–source-radius problem, with a 5-second limit for each linear program. Region calculations are shared between the two disturbance radii. Candidate dual bounds and inclusion witnesses are checked using outward-rounded elementary operations; a solver status or sampled output alone is not accepted as a certificate. Exhausted budgets and unverified bounds remain inconclusive. The verified function is the real network with stored FP64 coefficients, not its floating-point execution.

Results. All 60 selected modules decode their complete canonical tables correctly. Across all eight radius pairs, activation-region verification certifies 2432 of 3840 neighborhood-trained transition checks, finds 1050 assigned-inclusion violations, and leaves 358 inconclusive. Point-trained modules have 3840 witnessed assigned-inclusion violations. Neither baseline certifies a transition. At $q = 1/8$ and $v = 1/256$, neighborhood-trained systems have $438/480$ certified transitions across seeds, 25 assigned-inclusion violations, and 17 inconclusive checks. The seed-level certified fraction is $91.25 \pm 4.03\%$; this is not the fraction of completely certified systems.

All sixteen doubling transitions are certified in every neighborhood-trained seed at all eight radius pairs. This gives correctness for arbitrary finite repetitions of that operator under each checked disturbance model. No seed has a complete certificate for the three-operator family. A violated assigned inclusion need not be a decoding failure or a reachable failure from canonical initialization, and an inconclusive result establishes neither. Source and target radii vary together, so certification need not improve monotonically as q decreases. The results support the applicability of regional verification to these learned modules, not a general robustness guarantee from neighborhood training.

A.33.8 Capacity attainment with learned translations and fixed ReLU gates

Task and architecture. We use the same four-bit encoding $c_s = \eta(s) \in \{-0.5, 0.5\}^4$ for $s \in \{0, \dots, 15\}$ and the bijection $r(s) = (s+1) \bmod 16$. Hence the terminal-state set has size $N_T = 16$ at every horizon when all initial states are included. Let $q = 0.40$, $q_{\text{out}} = 0.49$, $\alpha = 0.82$, and $v = 0.06$. For $d_s(x) = \|x - c_s\|_\infty$, the fixed gate is $g_s(x) = 1 - \mathrm{ReLU}((d_s(x) - q)/(q_{\text{out}} - q)) + \mathrm{ReLU}((d_s(x) - q_{\text{out}})/(q_{\text{out}} - q))$. Absolute values and maxima are implemented by finite

ReLU compositions. The module is $T_\theta(x) = \alpha x + \sum_{s=0}^{15} g_s(x)\theta_s$, with $\theta_s \in \mathbb{R}^4$. Only these 64 translation parameters are trained. The codebook, gates, and coefficient α are fixed; no decoder or external canonicalizer is inserted between updates.

On $K_s = \mathbb{B}_\infty(c_s, q)$, the matching gate equals one. Every competing code is at distance at least $1 - q = 0.60 > q_{\text{out}}$, so its gate is zero. Thus $T_\theta(x) = \alpha x + \theta_s$ throughout K_s, irrespective of the learned translations. Its local Lipschitz and openness coefficients there are both α. This identity follows from the fixed architecture. The gate outputs are constant on these boxes even though internal ReLUs used to implement absolute values and maxima can change activation pattern.

Training. Translation coordinates are initialized independently from $\mathcal{N}(0, 0.05^2)$ for ten seeds. Every update combines the mean squared error on all sixteen canonical inputs with a batch of 512 uniformly selected states and independent perturbations $z \sim \mathrm{Unif}([-q, q]^4)$. For the perturbed input $c_s + z$, the target is $c_{r(s)} + \alpha z$. The loss is $4\mathcal{L}_{\text{canonical}} + \mathcal{L}_{\text{neighborhood}}$. Adam runs for 3000 updates, with zero weight decay, gradient clipping at 5, and cosine learning-rate decay from 0.01 to 0.0001. Evaluation occurs every 100 updates using the complete table and a fixed probe of 256 perturbed inputs per state. Checkpoints are ordered lexicographically by decoding errors, positive excess of maximum canonical error over $q - (\alpha q + v)$, maximum canonical error, maximum probe error, and canonical mean squared error. The selected checkpoints lie between updates 1100 and 2700. In exact arithmetic, the architecture makes the neighborhood residual independent of z on each K_s; this is fitting a finite translation table, not learning the gate geometry.

Regional correctness and inner reachability. Execution starts at c_s and uses the full successor set $\mathcal{D}(x) = T_\theta(x) + \mathbb{B}_\infty(0, v)$ with $v = 0.06$. Write $\epsilon_s = \|\alpha c_s + \theta_s - c_{r(s)}\|_\infty$. The exact worst-case target deviation on K_s is $\alpha q + v + \epsilon_s$. Therefore $\alpha q + v + \epsilon_s \le q$ for all sixteen states establishes regional closure and decoded correctness at every finite time, since $q = 0.40$ is a valid decoding radius. Here $\alpha q + v \approx 0.388$, leaving approximately 0.012 for canonical transition error.

For the trace from initial state s, inner-radius propagation starts from $a_0(s) = 0$ and uses $a_t(s) = \alpha a_{t-1}(s) + v - \epsilon_{r^{t-1}(s)}$ while the predecessor ball remains in its certified region. The common radius is $a_T = \min_s a_T(s)$. Bijectivity ensures that these paths cover all sixteen terminal states, so positive common inner reachability and regional correctness jointly instantiate Theorem 4. With $B = 0.5$ and $d = 4$, its integer bound is $\lceil 0.5/a_T \rceil^4$. The inner-radius calculation uses Theorem 8; it does not test the multistep endpoint certificate in Theorem 3. Nor does it implement the unit-openness translation construction of Proposition 1, since the local scale is 0.82, not one.

Numerical checks and conservative recheck. Training and execution use FP64. We supplement the recorded scalar checks with an exact rational recheck using a uniform error budget $\bar{\epsilon} = 10^{-12}$. Let $\widehat{\epsilon}_s$ be the recorded canonical-error value. Multiplication of stored α by a code coordinate $\pm 1/2$ is exact in binary64. For the near-code sums in these records, a conservative allowance for the remaining scalar addition and subtraction is 2^{-52}. Thus $\epsilon_s \le \widehat{\epsilon}_s + 2^{-52} < 1.17 \times 10^{-15} < \bar{\epsilon}$ under binary64 round-to-nearest. This step uses the recorded scalar errors and the stated evaluation procedure; it does not independently reload the trained checkpoints.

Using the exact rational values of the stored binary64 constants, we check $q - (\alpha q + v + \bar{\epsilon}) > 0$ and propagate $\underline{a}_0 = 0$, $\underline{a}_t = \alpha \underline{a}_{t-1} + v - \bar{\epsilon}$. The margin exceeds 0.011999999998, and the lower radii remain positive and below q through step 32. These conservative radii give the same six-decimal values and integer capacity counts as Table 17. The recheck supports the real-network scalar inclusions under the stated arithmetic assumptions; it does not certify all rounding errors of the CUDA execution. The recorded CSV files are retained, with the recheck reported separately.

Results. All ten selected models decode every canonical state correctly. Across seeds, maximum canonical error is 9.16×10^{-16} on average, and maximum error against the neighborhood target is 1.57×10^{-14}. All 160 state–seed regional checks pass, and all 5120 initial-state–step records retain positive inner-radius estimates through step 32. At every reported horizon from 8 through 32, the integer bound is 16 in every seed and equals the actual terminal-state count. Stepwise records and the conservative recheck place the first crossing of $a_t > 0.25$ at $t = 7$, with $a_7 \approx 0.250238$; $T = 8$ is the first reported horizon after this crossing.

Table 17: Capacity calculation for the fixed-gate model with learned translations. Radii are seed-level minima over all sixteen initial states, then averaged across ten seeds. Their sample standard deviations are below 4×10^{-16} and are omitted at the displayed precision. Exact scalar rechecking with $\bar{\epsilon} = 10^{-12}$ preserves the displayed radii and every capacity decision.

T	Mean a_T	N_T	$\lceil 0.5/a_T \rceil^4$	Seeds attaining bound
1	0.060000	16	6561	0/10
2	0.109200	16	625	0/10
4	0.182626	16	81	0/10
8	0.265195	16	16	10/10
12	0.302527	16	16	10/10
16	0.319405	16	16	10/10
24	0.330486	16	16	10/10
32	0.332751	16	16	10/10

Execution checks and scope. We evaluate 512-step executions from all sixteen codes under positive-corner, negative-corner, and uniform-box disturbances. The 15,360 logged executions include 32 identical replicates for each deterministic corner condition. Counting those once gives 320 distinct corner executions and 5120 uniform-box trials, or 5440 execution trials after removing deterministic replications. All are correct at every step, with no exit from the radius-0.40 regions. Corner errors approach $1/3$; for uniform disturbances, the mean of seedwise maximum errors is 0.244913 ± 0.005359. These trajectories check implementation behavior and are not the proof of universal correctness. The theoretical quantities are compatible because the prescribed local scale both contracts outer radii and propagates inner radii. The result demonstrates capacity attainment with learned translations in a selected finite architecture, not an unconstrained learning result or a general capacity law inferred from training.

A.33.9 Data provenance and scope of the evidence

The tables use trajectory, checkpoint, per-step, control, and paired-comparison CSV files. Exp3 additionally uses reachable-box, analytic-capacity, and collision-witness records with exact rational postprocessing. The regional experiment uses its certificate checks and seed-level operator summaries. The structured capacity experiment uses its training, regional, inner-radius, capacity, and trajectory records, with the conservative scalar recheck reported separately. Configuration and environment files specify the corresponding runs; files with the same basename belong to separate experiment directories.

The 128-dimensional arithmetic results concern the specified task, checkpoints, and test distributions. The regional certificates cover declared boxes of the four-dimensional MLPs, not every representation in their decoder cells. Capacity attainment is demonstrated in a fixed-gate model whose translations are learned. None of these experiments establishes generalization to other symbolic systems. Fixed-grid packing does not certify the full-grid decoder, and representation intervention changes execution.