DEPTH SEPARATION FOR LEARNING HIERARCHICAL INTERACTIONS WITH NONLINEAR STATE-SPACE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

As sequence models emerge as efficient architectures for long-context modeling, it becomes important to understand the role of depth in recurrent architectures which combine recurrent memory with nonlinearity across layers. In this paper, we study the effect of depth through balanced-tree parity targets and show depth seperation phenomenon on Boolean input sequences. First, under norm constraints and bounded activation regularity, we obtain the lower bound of approximation error for deep nonlinear SSMs which means every constant-depth model cannot represent this target regardless of its hidden dimension or nonlinear width. Second, we obtain an upper bound for a logarithmic-depth SSM learned via a normalized layerwise coordinate-descent procedure when fitting this target from polynomially many noiseless samples, under the norm constraints assumptions and with constants independent of the sequence length. Third, we construct a depth-two nonlinear SSM that exactly represents a deep quadratic target, which cannot be represented by any constant-depth fully connected network of comparable size, demonstrating recurrent memory can substantially reduce the nonlinear depth needed for exact representation. These results characterize depth separation through the interplay among memory dimension, nonlinear width and network depth while extending the theoretical understanding of deep nonlinear recurrent architectures.

1 INTRODUCTION

State-space models (SSMs) process sequential input data by maintaining a recurrent hidden state and transforming this state into the output. Structured linear dynamics in each layer enable efficient computation over long sequences, as exemplified by S4 (Gu et al., 2021), while selectiveness is further introduced in input-dependent dynamics to address long-range dependence (Gu & Dao, 2023). Beyond these architectural developments, stacking recurrent layers with nonlinearity provides an additional representational advantage. This naturally raises a fundamental question: once the recurrent state is sufficiently expressive to retain the memory information from the input sequence, what advantage can the depth of nonlinearity provide? Can temporal accumulation replace nonlinear depth in forming hierarchical interactions?

Depth-separation results ask whether some functions can be represented efficiently by deep networks but require much larger budgets when the network is shallow (Eldan & Shamir, 2016; Telgarsky, 2016). More recently, Ren et al. (2026) showed a separation between constant-depth and logarithmic-depth feedforward networks from both representation and learning perspectives. They considered hierarchical target functions, such as balanced-tree parity functions and showed that a layerwise coordinate-descent method can learn the hidden interaction structure from polynomially many samples. This work provides the main theoretical starting point for our study of depth in deep nonlinear SSMs.

Our main contribution is to extend the approximation results of Ren et al. (2026) from feedforward networks to recurrent SSMs. This is the first rigorous work of analyzing depth seperation in deep nonlinear state-space models. The main difficulty is that recurrence allows information from many time steps to accumulate in the hidden state, so the feedforward argument cannot be applied directly.

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.

We handle this by tracking how perturbations propagate through the recurrent dynamics and then combining this control with a Boolean Fourier analysis.

We present our main results as follows:

Constant-Depth Lower Bound: Under norm constraints and bounded activation regularity, we obtain the lower bound of approximation error for deep nonlinear SSMs which means every constant-depth model cannot represent this target regardless of its hidden dimension or nonlinear width.

Logarithmic-Depth Upper Bound: We derive an upper bound for a logarithmic-depth SSM learned via a normalized layerwise coordinate-descent procedure when fitting this target from polynomially many noiseless samples, under the norm constraints assumptions and with constants independent of the sequence length.

Memory Beyond Depth: We construct a depth-two nonlinear SSM that exactly represents the general hierarchical target, which cannot be represented by any constant-depth fully connected network of comparable size, thereby demonstrating recurrent memory can substantially reduce the nonlinear depth needed for exact representation with a large norm budget.

2 Related Work

State-Space Models Gu et al. (2021) introduced S4, making structured linear state-space layers computationally practical for long-sequence modeling and Gu & Dao (2023) introduced input-dependent selection in Mamba, enabling content-dependent information propagation with linear scaling in sequence length. Then, Dao & Gu (2024) connected selective SSMs and attention through structured matrix representations and developed Mamba-2. Merrill et al. (2024) also established circuit-complexity upper bounds for common SSM architectures under log-precision computation. Our complementary approach studies real-valued L^2 approximation and identifies how controlled parameters and activation regularity restrict high-order interactions.

Deep Networks Bengio et al. (2006) investigated greedy layerwise pretraining followed by global fine-tuning as a way to train deep feature hierarchies. Allen-Zhu & Li (2019) proved efficient learning guarantees for structured functions using residual networks beyond what kernel methods achieve. Allen-Zhu & Li (2020) analyzed efficient learning of hierarchical quadratic targets through backward feature correction. For our Boolean tree targets, the layerwise learning procedure is simpler because intermediate parity features are orthogonal, remain available after they are selected and contribute nonzero coefficients to the target. These properties allow the hierarchy to be recovered to some extent.

Depth Seperation Eldan & Shamir (2016) exhibited a function efficiently represented by three-layer networks that requires exponential width for approximation by two-layer networks. Telgarsky (2016) established depth hierarchies using functions with increasingly complex oscillations. Cohen et al. (2016) used hierarchical tensor factorizations to establish depth efficiency for convolutional arithmetic circuits. Ren et al. (2026) proved a separation between constant-depth and logarithmic-depth networks for hierarchical Fourier targets. We build on their target family and layerwise learning procedure. Our main extension is to recurrent SSMs, where information can accumulate over time through the hidden state.

3 Problem Formulation

In this section, we define the sequence-to-scalar prediction problem studied in our paper. We first introduce the deep nonlinear SSM and specify its main architectural quantities: sequence length, depth, memory dimension and nonlinear width. We then define a hierarchical target family that allows us to study how these different quantities affect approximation and learning.

3.1 Deep Nonlinear State-Space Models

We first define our input. For an integer $q \geq 1$, let $n = 2^q$, $\mu_n = \mathrm{Unif}(\{-1, +1\}^n)$. Then our input is $X = (x(1), x(2), ..., x(n)) \sim \mu_n$, where sequence length is n and each $x(i)$ can be viewed as an independent Rademacher random variable.

We then define deep nonlinear SSM with depth L, hidden dimension m and nonlinearity width d as follows:

$$\begin{aligned} h_1(t+1) &= A_1 h_1(t) + B_1 x(t+1), \\ h_2(t+1) &= A_2 h_2(t) + B_2 f_2\big(h_1(t+1)\big), \\ &\vdots \\ h_L(t+1) &= A_L h_L(t) + B_L f_L\big(h_{L-1}(t+1)\big), \\ y(t) &= C^\top g\big(h_L(t)\big). \end{aligned} \tag{1}$$

Here, $A_1, ..., A_L \in \mathbb{R}^{m\times m}$, $B_2, ..., B_L \in \mathbb{R}^{m\times m}$, $B_1, C \in \mathbb{R}^m$. The inter-layer nonlinear mapping f_ℓ and the final nonlinear readout part g is a one-layer FFN with nonlinearity width d:

$$f_\ell(z) = V_\ell\, \sigma_\ell(U_\ell z + b_\ell) + c_\ell, \quad g(z) = V_g\, \sigma_g(U_g z + b_g) + c_g, \tag{2}$$

where $\ell = 2, \dots, L$ and each activation function σ_ℓ and σ_g acts coordinatewise. $U_\ell, U_g \in \mathbb{R}^{d\times m}$, $V_\ell, V_g \in \mathbb{R}^{m\times d}$, $b_\ell, b_g \in \mathbb{R}^d$ and $c_\ell, c_g \in \mathbb{R}^m$. We initialize $h_\ell(0) = 0,\ \ell = 1, ..., L$ and the input and output are both scalar.

3.2 Hierarchical Target Class

Assume that $n = 2^q$ for some $q \in \mathbb{N}$ and let $X = (x(1), \dots, x(n)) \in \{-1, +1\}^n$. For $S \subseteq [n] := \{1, \dots, n\}$, we define the parity function

$$\chi_S(X) = \prod_{i\in S} x(i), \qquad \chi_\varnothing(X) = 1.$$

For each level $r = 1, \dots, q$ and each $j = 1, \dots, n/2^r$, we define the dyadic block $I_{r,j} = \{(j-1)2^r + 1, \dots, j2^r\}$. Thus, each $I_{r,j}$ contains 2^r consecutive input coordinates.

For arbitrary signs $\varepsilon = (\varepsilon_{r,j})$, $\varepsilon_{r,j} \in \{-1, +1\}$, we define

$$F_{n,\varepsilon}(X) = \sum_{r=1}^{q} \sqrt{\frac{2^r}{qn}} \sum_{j=1}^{n/2^r} \varepsilon_{r,j} \chi_{I_{r,j}}(X). \tag{3}$$

The corresponding target class is

$$\mathcal{F}_n = \{F_{n,\varepsilon} : \varepsilon_{r,j} \in \{-1, +1\}\}.$$

Each $F_{n,\varepsilon}$ is a deterministic sequence-to-scalar function and our goal is to predict $F_{n,\varepsilon}(X)$ from the input sequence X.

Motivation and structure. The target family is a signed dyadic version of the hierarchical Fourier targets studied by Ren et al. (2026). At level r, each term$\chi_{I_{r,j}}(X)$ is the product of 2^r consecutive input coordinates. Therefore, the target contains interactions of different orders. At the first level, it contains products of adjacent inputs. As r increases, the interactions involve larger and larger blocks of the sequence. At the final level $r = q$, the target contains the parity of the entire input sequence.

The coefficients are chosen so that every level contributes the same amount to the squared $L^2(\mu_n)$ norm. Since the parity functions are orthonormal under μ_n, for every $r = 1, \dots, q$,

$$\sum_{j=1}^{n/2^r} \left(\sqrt{\frac{2^r}{qn}}\right)^2 = \frac{n}{2^r}\frac{2^r}{qn} = \frac{1}{q}.$$

Hence every level contributes exactly $1/q$ and therefore $\|F_{n,\varepsilon}\|^2_{L^2(\mu_n)} = \sum_{r=1}^{q} \frac{1}{q} = 1$.

This normalization ensures that the high-order interactions do not disappear as the sequence length grows. In particular, the levels $r \geq \lceil \frac{q}{2} \rceil$ contain interactions involving at least $2^r \geq 2^{q/2} = \sqrt{n}$ input coordinates. These levels together contribute at least one half of the total squared $L^2(\mu_n)$ norm. Therefore, any accurate approximation of $F_{n,\varepsilon}$ must capture a nontrivial amount of high-order interaction.

At the same time, these high-order interactions have a simple recursive structure. For $r = 2, \dots, q$,

$$\chi_{I_{r,j}}(X) = \chi_{I_{r-1,\,2j-1}}(X)\chi_{I_{r-1,\,2j}}(X).$$

Thus, each parity feature at level r can be constructed by multiplying two parity features from level $r-1$. Starting from the products of adjacent inputs, the entire hierarchy can therefore be built one level at a time.

Moreover, every intermediate parity feature appears in the target with a nonzero coefficient. As a result, these intermediate features can be detected during layerwise learning. The target family therefore has two properties that are important for our following analysis. First, a constant fraction of its $L^2(\mu_n)$ mass lies in high-order interactions, which makes approximation difficult for shallow models under the norm and regularity assumptions introduced below. Second, its recursive tree structure allows the target to be learned and represented progressively by a deeper model.

4 Depth Seperation

We present our main results in this section. First, we introduce the necessary assumptions and notations. Then, we show our main results to explain the depth seperation for deep nonlinear SSMs.

4.1 Assumptions and Notations

We now state the norm and regularity assumptions used to prove the constant-depth approximation lower bound. Their role is to limit how strongly an input perturbation can grow as it propagates through temporal recurrence and nonlinear layer mappings. We use $\|\cdot\|_{\mathrm{op}}$ for the Euclidean operator norm of a matrix and $\|\cdot\|_2$ for the Euclidean norm of a vector.

Assumption A1: For $1 \le t \le n$, we define $\mathcal{R}_t = \begin{bmatrix} A_1^{t-1}B_1 & A_1^{t-2}B_1 & \cdots & B_1 \end{bmatrix} \in \mathbb{R}^{m\times t}$. We assume

$$\sup_{1\le t\le n} \|\mathcal{R}_t\|_{\mathrm{op}} \le K. \tag{4}$$

This condition limits how much the input history can be amplified when it is stored in the first-layer state. Importantly, it does not force the model to forget earlier inputs.

Assumption A2: For every $\ell = 2, \dots, L$, we assume $\sum_{k=0}^{n-1} \left\|A_\ell^k B_\ell V_\ell\right\|_{\mathrm{op}} \le K$ and $\|U_\ell\|_{\mathrm{op}} \le K$.

The impulse-response bound controls how much information can accumulate over time in each higher layer, while the bound on U_ℓ controls how much this information can be amplified before passing through the next activation. Importantly, we only bound these combined effects and do not require a separate norm bound on every parameter matrix.

Assumption A3: We assume $\|U_g\|_{\mathrm{op}} \le K$ and $\left\|V_g^\top C\right\|_2 \le K$.

These bounds control the final readout in the same way. The first bound limits how much the last hidden state can be amplified before the final activation, while the second limits how strongly the activated features can be combined into the output.

Assumption A4: Every scalar activation σ belongs to $C^2(\mathbb{R})$ and satisfy

$$\sup_{z\in\mathbb{R}} |\sigma'(z)| \le M, \quad \sup_{z\in\mathbb{R}} |\sigma''(z)| \le M. \tag{5}$$

Together, the regularity of activation allows us to propagate a bound on simultaneous input-coordinate perturbations through the nonlinear layers.

For $r \ge 0$, let

$$\Psi(r) = MK^2 r + \frac{MK^3}{2} r^2, \qquad R_0 = \Psi^{\circ L_0}(2K), \tag{6}$$

where $\Psi^{\circ j}$ denotes j-fold composition and $\Psi^{\circ 0}(r) = r$. In particular, R_0 depends only on L_0, K, M. The map Ψ summarizes how this sensitivity bound grows through one nonlinear stage and R_0 bounds the terminal sensitivity for models of depth at most L_0.

Let $\mathcal{N}_n(L_0, K, M)$ be the class of all terminal functions Y_n realized by models satisfying Equation 1 to 2 and Assumptions A1–A4, allowing arbitrary finite state dimension m, nonlinear width d and depth $L \le L_0$. For a real number a, write $a_+ = \max\{a, 0\}$.

4.2 LOWER BOUND FOR CONSTANT DEPTH

Next, we show the lower bound of approximation error for deep nonlinear SSMs under norm constraints and bounded activation regularity, which means every constant-depth model cannot represent this target regardless of its hidden dimension or nonlinear width.

Theorem 1 (Uniform approximation lower bound)**.** *Fix $L_0 \in \mathbb{N}$ and $K, M \geq 1$. For every $n = 2^q$, $q \geq 1$, every $F_{n,\epsilon} \in \mathcal{F}_n$, and every $Y_n \in \mathcal{N}_n(L_0, K, M)$,*

$$\|F_{n,\epsilon} - Y_n\|_2 \geq \left(\sqrt{\frac{q - \lceil q/2 \rceil + 1}{q}} - \frac{R_0}{2n^{1/4}} \right)_+ \geq \left(\frac{1}{\sqrt{2}} - \frac{R_0}{2n^{1/4}} \right)_+ . \tag{7}$$

In particular, for every dyadic horizon satisfying $n \geq 4R_0^4$ (seqeunce length is large enough), we have, for every target in $\mathcal{F}_n$,

$$\inf_{Y_n \in \mathcal{N}_n(L_0,K,M)} \mathbb{E}\,|F_{n,\epsilon}(X) - Y_n(X)|^2 \geq \frac{1}{8}. \tag{8}$$

The proof of Theorem 1 can be found in Appendix A.1. This theorem shows that increasing hidden dimension or nonlinear width cannot overcome the limitation of constant depth. For fixed L_0, K and M, every admissible constant-depth SSM has mean-squared error at least $1/8$ once the sequence length is sufficiently large.

The reason is that the target contains a large amount of high-order interaction. In particular, a constant fraction of its Fourier energy lies on parity terms involving at least $\sqrt{n}$ input coordinates. By contrast, the norm and regularity assumptions limit how strongly a constant-depth SSM can respond to simultaneous changes in many input coordinates. As a result, such a model cannot place enough Fourier energy on these high-order terms when n becomes large. Therefore, even when all input information is available in the hidden state, constant depth is still insufficient to form the required hierarchical high-order interactions. The theorem thus separates recurrent memory from nonlinear composition: remembering the input is not the same as combining it through sufficiently deep nonlinear structure.

4.3 UPPER BOUND FOR LOGARITHMIC DEPTH

We now establish a constructive upper bound for the generalized target class containing the hierarchical target family considered above. The learner is not given the interaction hierarchy: it discovers the relevant parities through a normalized variant of the layerwise coordinate-descent strategy of Ren et al. (2026). The resulting predictor can then be compiled into a logarithmic-depth SSM.

4.3.1 GENERAL TREE TARGETS

Let $n = 2^q$ and let $\mathcal{T}$ be a perfect balanced binary tree whose leaves are labeled by $[n]$. Identify each node with the set of its descendant leaves. For $r = 0, \ldots, q$, let $\mathcal{T}_r$ denote the collection of supports at level r above the leaves. Thus, $\mathcal{T}_0$ consists of the singleton supports, while $\mathcal{T}_r$ partitions $[n]$ into $n/2^r$ sets of cardinality 2^r.

We denote

$$\mathcal{S}_\mathcal{T} := \mathcal{T}_{\text{int}} = \bigcup_{r=1}^{q} \mathcal{T}_r, \quad s := |\mathcal{S}_\mathcal{T}| = n - 1.$$

For signs $\theta = (\theta_S)_{S \in \mathcal{S}_\mathcal{T}} \in \{-1, +1\}^s$, we define

$$F_{\mathcal{T},\theta}(X) = \sum_{S \in \mathcal{S}_\mathcal{T}} a_S \chi_S(X), \quad a_S = \theta_S \sqrt{\frac{|S|}{qn}}. \tag{9}$$

Let $\mathcal{F}_n^{\text{tree}} := \left\{ F_{\mathcal{T},\theta} : \mathcal{T} \text{ is a perfect balanced binary tree on } [n],\ \theta \in \{-1, +1\}^{n-1} \right\}$.

The previously defined class $\mathcal{F}_n$ is the subclass obtained from the canonical dyadic tree, whose level-r supports are $I_{r,j}$. In particular, $\mathcal{F}_n \subseteq \mathcal{F}_n^{\text{tree}}$. Every labeled perfect balanced binary tree can be obtained by permuting the leaf labels of the canonical dyadic tree. Thus, the generalization allows an

unknown arrangement of the same hierarchical interaction structure. This family is a specialization of the hierarchical staircase framework of Ren et al. (2026) and its all-positive subclass coincides with their deep-quadratic target family.

Throughout this subsection, fix an unknown target $F = F_{\mathcal{T},\theta} \in \mathcal{F}_n^{\mathrm{tree}}$. The learner knows n and the target-class specification, but neither $\mathcal{T}$ nor θ. It observes independent noiseless samples $(X, F(X))$, where $X \sim \mu_n$. By parity orthogonality, $\|F\|^2_{L^2(\mu_n)} = \sum_{S\in\mathcal{S}_\mathcal{T}} a_S^2 = 1$ and $\min_{S\in\mathcal{S}_\mathcal{T}} |a_S| = \sqrt{\frac{2}{qn}}$.

4.3.2 Normalized Layerwise Coordinate Descent

Coordinate descent gives a simple way to identify useful parity features and remove duplicate ones. Each normalized parity candidate has unit empirical squared norm, so its coordinate update is just a one-dimensional least-squares step. The update coefficient is determined by the correlation between that candidate and the current residual. After a feature is selected and fitted, its contribution is subtracted from the residual. Therefore, if another candidate represents the same parity, much less signal remains for that duplicate candidate. Since different parity features are orthogonal in the population and their empirical correlations concentrate around the population values, a simple threshold can distinguish relevant features from irrelevant or duplicate ones. Finally, we pass the selected parity features themselves to the next layer, rather than multiplying them by their fitted coefficients. This keeps every recovered feature at the same scale and prevents coefficient-estimation errors from accumulating when later layers combine these features.

The learner constructs candidates by multiplying at most two available parity features. Each candidate takes values in $\{-1, +1\}$, so its empirical squared norm is exactly one. Its coordinate update is therefore the empirical correlation with the current residual. Thresholding selects relevant features, while sequential residual updates suppress duplicate representations. Selected features are passed to the next stage as unweighted parities, so errors in estimated output coefficients do not change the scale of subsequent features. The method adapts the layerwise strategy of Ren et al. (2026).

Parameter choices. Fix an accuracy level $0 < \varepsilon \le 1$ and a failure probability $0 < \delta < 1$. Define $\kappa := \sqrt{\frac{2}{qn}},$, $e := \min\left\{\frac{\kappa}{4}, \frac{\varepsilon}{\sqrt{s}}\right\}$, $P := 1 + (n+s) + \binom{n+s}{2} = 2n^2 - n + 1$, $N := \left\lceil \frac{8s}{e^2} \log \frac{2qPs}{\delta} \right\rceil$. Here, κ is the minimum target coefficient magnitude, e is the coefficient-estimation tolerance, P bounds the number of indexed candidates per stage and N is the number of samples drawn at each stage.

The learner maintains an ordered list of selected supports $\mathcal{B}_r = (B_{r,1}, \ldots, B_{r,b_r})$, initialized by the empty list $\mathcal{B}_0 = ()$. The corresponding features are the unweighted parities $\chi_{B_{r,j}}$. We allow repeated supports in this list when defining the algorithm and the theorem below shows that no repetitions survive on the successful sampling event. The raw input features $\chi_{\{1\}}, \ldots, \chi_{\{n\}}$ are always available.

For each stage $r = 1, \ldots, q$, we perform the following normalized layerwise coordinate descent steps:

Step 1: Construct and freeze the candidate list. Let $p_r = n + b_{r-1}$ and define the available supports by

$$T_{r,i} = \begin{cases} \{i\}, & 1 \le i \le n, \\ B_{r-1,i-n}, & n < i \le p_r. \end{cases}$$

Then, the available feature vector is $v^{(r)}(x) = \big(\chi_{T_{r,1}}(x), \ldots, \chi_{T_{r,p_r}}(x)\big)$. We define the candidate index set

$$\mathcal{H}_r = \{H \subseteq [p_r] : |H| \le 2\}.$$

For each $H \in \mathcal{H}_r$, set

$$z_H^{(r)}(X) = \prod_{i\in H} v_i^{(r)}(X) = \chi_{S_H^{(r)}}(X), \quad S_H^{(r)} = \bigtriangleup_{i\in H} T_{r,i}.$$

For $H = \varnothing$, the product is one and the support is empty. Distinct candidate indices may represent the same parity. Candidate supports are computed from the available features by symmetric difference. Now, we fix a visiting order on $\mathcal{H}_r$ before drawing this stage's samples.

Step 2: Draw samples and initialize coefficients. Draw a fresh batch $\big(X_r^{(a)}, F(X_r^{(a)})\big)_{a=1}^N$ independently of all previous batches. Initialize $w_H^{(r)} = 0$ for every $H \in \mathcal{H}_r$ and define

$$\mathcal{L}_r(w) = \frac{1}{2N} \sum_{a=1}^{N} \left(F(X_r^{(a)}) - \sum_{H \in \mathcal{H}_r} w_H z_H^{(r)}(X_r^{(a)}) \right)^2 .$$

Step 3: Perform one thresholded coordinate sweep. Visit each $H \in \mathcal{H}_r$ once in the chosen order. Using the current values of all other coefficients, we define

$$R_H^{(r)}(x) = F(x) - \sum_{\substack{G \in \mathcal{H}_r \\ G \neq H}} w_G^{(r)} z_G^{(r)}(x).$$

Since H has not yet been visited, $w_H^{(r)} = 0$. Apply the unit-step coordinate update

$$w_H^{(r)} \leftarrow \frac{1}{N} \sum_{a=1}^{N} z_H^{(r)}(X_r^{(a)}) R_H^{(r)}(X_r^{(a)}), \tag{10}$$

and then set $w_H^{(r)} = 0$ if $|w_H^{(r)}| \le \kappa/2$. Every candidate is a parity, so $\frac{1}{N} \sum_{a=1}^{N} \big(z_H^{(r)}(X_r^{(a)})\big)^2 = 1$..

Consequently, Equation 10 is both a coordinate gradient step with step size one and the exact minimizer of the empirical loss before thresholding with respect to that coordinate, with all other coefficients fixed.

Step 4: Retain normalized features. Let

$$\mathcal{J}_r = \{H \in \mathcal{H}_r : w_H^{(r)} \neq 0\}.$$

If $|\mathcal{J}_r| > s$, stop and return the zero SSM. Otherwise, let $\mathcal{B}_r$ be the ordered list

$$\mathcal{B}_r = \big(S_H^{(r)} : H \in \mathcal{J}_r\big),$$

ordered according to the visiting order. Record each retained coefficient and its construction.

Final predictor: If all stages are completed, define

$$\widehat{F}(x) = \sum_{H \in \mathcal{J}_q} w_H^{(q)} z_H^{(q)}(x). \tag{11}$$

Check that the distinct supports in $\mathcal{B}_q$ form the complete internal-node family of a perfect balanced binary tree on $[n]$ and that $\left\| \big(w_H^{(q)}\big)_{H \in \mathcal{J}_q} \right\|_2 \le 2$. If either check fails, return the zero SSM. Every failure branch returns the zero softplus SSM with $L = q$, $m = 2n - 1$, and $d = 4n - 2$.

4.3.3 Upper bound with logarithmic depth

Theorem 2 (Learning with logarithmic depth)**.** *Let $n = 2^q$, $q \ge 1$ and fix an unknown $F \in \mathcal{F}_n^{\mathrm{tree}}$. For every $0 < \varepsilon \le 1$ and $0 < \delta < 1$, the preceding algorithm returns a time-homogeneous, zero-initialized SSM in $\mathcal{N}_n(q, 7, 1)$ with $L = q,\ m = 2n-1,\ d = 4n-2$. With probability at least $1 - \delta$, the learner recovers the complete internal-node support family of F and its terminal output satisfies*

$$\|F - Y_n\|_{L^2(\mu_n)} \le \varepsilon. \tag{12}$$

The algorithm uses at most

$$qN = O\Big(n^2 q (q + \varepsilon^{-2}) \log \frac{qn}{\delta}\Big)$$

noiseless labeled sequence.

The proof of Theorem 2 can be found in Appendix A.2. This theorem characterizes a constructive logarithmic-depth approximation guarantee even when the interaction tree is unknown. The learner discovers the hierarchy progressively by forming products of previously recovered parity features. Orthogonality under μ_n identifies the target coefficients at the population level, while concentration of empirical correlations supports accurate estimation from finitely many samples. Sequential

coordinate updates remove fitted contributions from the residual, allowing thresholding to suppress spurious features and duplicate representations on the successful sampling event. The learned hierarchy can then be realized by a logarithmic-depth SSM. Since $\mathcal{F}_n \subseteq \mathcal{F}_n^{\mathrm{tree}}$, this guarantee applies in particular to every target in the earlier dyadic class.

The logarithmic depth in our construction has the same asymptotic scaling as that of Ren et al. (2026). Both build the interaction hierarchy through successive pairwise compositions. In our architecture, the first recurrent layer stores the input, the inter-layer maps compute successive tree levels and the nonlinear readout completes the hierarchy. Our contribution therefore concerns the realization and limitations of this mechanism in SSMs under recurrent norm constraints, rather than an simple improvement in the asymptotic depth required by the construction.

4.4 Exact Shallow Representation with Growing Norms

We next show the shallow representation when the norm budget is allowed to grow with sequence length. Using the standard $\tanh$ activation, we construct a depth-two SSM that exactly represents every target in $\mathcal{F}_n^{\mathrm{tree}}$. The construction preserves the activation regularity required by Assumption A4 and therefore clarifies the role of the uniform norm constraints in Theorem 1.

Theorem 3 (Exact depth-two representation with $\tanh$)**.** *Let $n = 2^q$, where $q \geq 1$ and fix $F = F_{\mathcal{T},\theta} \in \mathcal{F}_n^{\mathrm{tree}}$. There exists a time-homogeneous, zero-initialized SSM with depth $L = 2$, hidden dimension $m = O(n)$ and nonlinearity width $d = O(n \log n)$, using*

$$f_2(z) = V_2 \tanh(U_2 z + b_2) + c_2, \qquad g(z) = z,$$

such that

$$Y_n(x) = F(x) \qquad \text{for every } x \in \{-1, +1\}^n.$$

The proof of Theorem 3 could be found in Appendix A.3. Our construction separates two different roles of the SSM: storing the input sequence and forming high-order interactions. The first recurrent layer stores a linear encoding of the input, from which the sum over any target subset can be recovered at the final time. For a subset S with $|S| = k$, this sum can take only the $k+1$ values, where the corresponding parity alternates between $+1$ and -1. A linear combination of k shifted $\tanh$ units can interpolate these values exactly. Therefore, the second layer can compute all required parity features in parallel and the final readout combines them with the coefficients of the target. Importantly, our construction does not contradict Theorem 1, where the norm budget is required to remain bounded independently of n.

Universality with one recurrent layer. More generally, for any fixed sequence length, a depth-one nonlinear SSM with input decoding and output decoding two part nonlinearity activations can uniformly approximate any continuous sequence-to-scalar function on a compact input domain, provided that the nonlinear width and parameter norms are unrestricted. The reason is simple. The first recurrent layer can store the finite input sequence in its hidden state. Once the whole sequence is available in this state, the nonlinear map and the final readout can approximate any continuous function of the stored inputs by the classical universal approximation theorem. Now, we allow a pointwise input MLP as input encoding before a single recurrent layer, together with an output MLP:

$$\begin{aligned} h(t+1) &= Ah(t) + B\phi\big(x(t+1)\big), \\ y(t) &= C^\top g\big(h(t)\big), \qquad h(0) = 0. \end{aligned} \tag{13}$$

Here $A \in \mathbb{R}^{m \times m}$, $B, C \in \mathbb{R}^m$, $\phi : \mathbb{R} \to \mathbb{R}$, and $g : \mathbb{R}^m \to \mathbb{R}^m$. Both ϕ and g are one-hidden-layer MLPs with coordinatewise $\tanh$ activation. We write $Y_n(x) = y(n)$ for the terminal output. We give the following proposition for universality.

Proposition 4 (Universal approximation with one recurrent layer)**.** *Fix an integer $n \geq 1$ and a nonempty compact set $\mathcal{D} \subseteq \mathbb{R}^n$. For every continuous functional $F : \mathcal{D} \to \mathbb{R}$ and every $\varepsilon > 0$, there exists a time-homogeneous, zero-initialized SSM of the form equation 13, with one recurrent layer, hidden dimension $m = n$, and finite MLP width, such that*

$$\sup_{x \in \mathcal{D}} |Y_n(x) - F(x)| < \varepsilon. \tag{14}$$

The MLP parameters may depend on n, $\mathcal{D}$, F, and ε.

The proof is given in Appendix A.6. Our universality result shows that one recurrent SSM layer, equipped with input and output nonlinearities, is sufficient to approximate any continuous sequence-to-scalar functional on a compact domain at a fixed finite horizon. In comparison, Wang & Xue (2023) provide a five-layer SSM construction for universal approximation of causal sequence-to-sequence mappings. Our construction stores the encoded sequence and the required linear projections within a single recurrent state, allowing the nonlinear readout to approximate the target functional.

5 Conclusion

We provide a rigorous analysis of how recurrent memory, network depth, nonlinear width and parameter norms affect the approximation of hierarchical sequence functions. Our analysis forms three complementary results. First, Under norm constraints and bounded activation regularity, we obtain the lower bound of approximation error for deep nonlinear SSMs which means every constant-depth model cannot represent this target regardless of its hidden dimension or nonlinear width. Second, a normalized layerwise coordinate-descent procedure learns an unknown interaction hierarchy from polynomially many noiseless samples and realizes the resulting predictor as a logarithmic-depth SSM. Third, a depth-two SSM with $\tanh$ activation represents these targets exactly when the norm budget is allowed to grow with sequence length.

Our results clarify the different roles of memory and nonlinear depth. A recurrent state can store the full input sequence, but simply having access to this information is not enough to approximate our targets under the norm and regularity constraints used in the lower bound. A logarithmic-depth SSM succeeds because it builds the required high-order interactions gradually, reusing features formed in earlier layers. A shallow SSM can also represent the same interactions, but only when it is allowed to use larger operator norms. Thus, memory mainly determines what information is available, while nonlinear depth determines how that information can be combined under a fixed norm budget. To the best of our knowledge, our paper provides the first theoretical analysis of algorithmic separation between constant-depth and logarithmic-depth in deep nonlinear SSMs.

Our results also have several limitations. We focus on hierarchical balanced-tree targets under uniform Boolean inputs and our learning result assumes noiseless samples. In addition, the learning algorithm first identifies explicit parity features and only then converts the learned predictor into an SSM, so it does not describe end-to-end training of the recurrent parameters although it is just one of the learners. We also do not prove that logarithmic depth is necessary, nor do we characterize the optimal tradeoff among depth, nonlinear width and operator norms. Extending the analysis to noisy observations, more general target functions and input-dependent recurrent dynamics is left for future work.

AI USE STATEMENT

In this work, we used generative AI tools for assist in the writing of proofs. We have not used generative AI tools for helping develop theoretical models or conceptual frameworks, formulating mathematical claims, providing critical ingredients for proving mathematical claims and generating synthetic data sets, proposing or refining hypotheses, designing or providing feedback on research methodology or experiments, implementing methods, assisting with translation, clean and reformat dataset, supporting qualitative and thematic data analysis, interpreting results are not applicable to this work. We have reviewed all AI-assisted work. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Zeyuan Allen-Zhu and Yuanzhi Li. What can resnet learn efficiently, going beyond kernels? *Advances in neural information processing systems*, 32, 2019.

Zeyuan Allen-Zhu and Yuanzhi Li. Backward feature correction: How deep learning performs deep (hierarchical) learning. *arXiv preprint arXiv:2001.04413*, 2020.

Yoshua Bengio, Pascal Lamblin, Dan Popovici, and Hugo Larochelle. Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, 19, 2006.

Nadav Cohen, Or Sharir, and Amnon Shashua. On the expressive power of deep learning: A tensor analysis. In *Conference on learning theory*, pp. 698–728. PMLR, 2016.

Tri Dao and Albert Gu. Transformers are ssms: Generalized models and efficient algorithms through structured state space duality. *arXiv preprint arXiv:2405.21060*, 2024.

Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *Conference on learning theory*, pp. 907–940. PMLR, 2016.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. *arXiv preprint arXiv:2111.00396*, 2021.

William Merrill, Jackson Petty, and Ashish Sabharwal. The illusion of state in state-space models. *arXiv preprint arXiv:2404.08819*, 2024.

Yunwei Ren, Zihao Wang, and Jason D Lee. Algorithmic separation between constant-depth and logarithmic-depth neural networks. *arXiv preprint arXiv:2607.25200*, 2026.

Matus Telgarsky. Benefits of depth in neural networks. In *Conference on learning theory*, pp. 1517–1539. PMLR, 2016.

Shida Wang and Beichen Xue. State-space models with layer-wise nonlinearity are universal approximators with exponential decaying memory. *Advances in Neural Information Processing Systems*, 36:74021–74038, 2023.

A Detailed proofs

A.1 Proof of Theorem 1

Theorem 1: Uniform approximation lower bound Fix $L_0 \in \mathbb{N}$ and $K, M \geq 1$. Let

$$\Psi(r) = MK^2 r + \frac{MK^3}{2} r^2, \qquad R_0 = \Psi^{\circ L_0}(2K),$$

as in equation 6. For every $n = 2^q$ with $q \geq 1$, every $F_{n,\varepsilon} \in \mathcal{F}_n$, and every $Y_n \in \mathcal{N}_n(L_0, K, M)$,

$$\begin{aligned} \|F_{n,\varepsilon} - Y_n\|_{L^2(\mu_n)} &\geq \left(\sqrt{\frac{q - \lceil q/2 \rceil + 1}{q}} - \frac{R_0}{2n^{1/4}} \right)_+ \\ &\geq \left(\frac{1}{\sqrt{2}} - \frac{R_0}{2n^{1/4}} \right)_+ . \end{aligned} \tag{15}$$

In particular, if $n \geq 4R_0^4$, then, for every $F_{n,\varepsilon} \in \mathcal{F}_n$,

$$\inf_{Y_n \in \mathcal{N}_n(L_0,K,M)} \mathbb{E}_{X \sim \mu_n} \left| F_{n,\varepsilon}(X) - Y_n(X) \right|^2 \geq \frac{1}{8}. \tag{16}$$

Proof strategy. We first control all single-coordinate input changes together, using the operator norm of a matrix of flip differences. The local nonlinear estimate uses the dimension-free Taylor argument of Ren et al. (2026, Lemma B.2); we include its proof below. We then propagate this estimate through the finite-horizon recurrent sums under A1–A3. Finally, a Boolean Fourier identity converts the terminal sensitivity bound into a bound on high-degree Fourier mass. The target has a fixed amount of mass at those degrees, which yields the approximation lower bound.

Proof. Throughout the proof, vector norms $\|\cdot\|_2$ are Euclidean, matrix norms $\|\cdot\|_{\mathrm{op}}$ are induced Euclidean operator norms, and function norms are written explicitly as $\|\cdot\|_{L^2(\mu_n)}$. Fix $n = 2^q$, a target $F_{n,\varepsilon}$, and an admissible SSM of depth $1 \leq L \leq L_0$, with arbitrary finite hidden dimension m and nonlinear width d.

For $x \in \{-1, +1\}^n$, let $x^{(i)}$ denote the input obtained by flipping coordinate i. If $H : \{-1, +1\}^n \to \mathbb{R}^p$, define

$$\mathsf{D}H(x) = \begin{bmatrix} H(x^{(1)}) - H(x) & \cdots & H(x^{(n)}) - H(x) \end{bmatrix} \in \mathbb{R}^{p \times n}. \tag{17}$$

The operator D is linear, commutes with multiplication by a fixed matrix, and annihilates functions that are constant in x. It is a finite-difference operator on the cube; it does not require a differentiable extension of H.

Step 1: A dimension-free estimate for a nonlinear map. Suppose $H : \{-1, +1\}^n \to \mathbb{R}^p$, $U \in \mathbb{R}^{d \times p}$, $b \in \mathbb{R}^d$, and σ is applied coordinatewise and satisfies A4. We show that

$$\|\mathsf{D}H(x)\|_{\mathrm{op}} \leq r, \quad \|U\|_{\mathrm{op}} \leq K \quad \Longrightarrow \quad \left\|\mathsf{D}[\sigma(UH + b)](x)\right\|_{\mathrm{op}} \leq MKr + \frac{MK^2}{2} r^2. \tag{18}$$

Here $r \geq 0$, and the estimate contains no factor depending on p, d, or n.

Set $z = UH(x) + b$ and $T = U\mathsf{D}H(x)$. Then $\|T\|_{\mathrm{op}} \leq Kr$. For each coordinate $a \in [d]$ and input flip $i \in [n]$, Taylor's theorem with integral remainder gives

$$\sigma(z_a + T_{ai}) - \sigma(z_a) = \sigma'(z_a) T_{ai} + E_{ai}, \qquad E_{ai} = T_{ai}^2 \int_0^1 (1 - u) \sigma''(z_a + u T_{ai}) \, du.$$

Thus $|E_{ai}| \leq (M/2)|T_{ai}|^2$ and

$$\mathsf{D}[\sigma(UH + b)](x) = \mathrm{diag}(\sigma'(z)) T + E.$$

The first term has operator norm at most $M\|T\|_{\mathrm{op}}$.

For the remainder, put $\alpha = (M/2)\|T\|_{\mathrm{op}}^2$. Every row and every column of T has Euclidean norm at most $\|T\|_{\mathrm{op}}$. Consequently,

$$\sum_{i=1}^{n} |E_{ai}| \le \frac{M}{2} \sum_{i=1}^{n} |T_{ai}|^2 \le \alpha, \qquad \sum_{a=1}^{d} |E_{ai}| \le \frac{M}{2} \sum_{a=1}^{d} |T_{ai}|^2 \le \alpha. \tag{19}$$

The row/column-sum (Schur) bound now gives $\|E\|_{\mathrm{op}} \le \alpha$. To verify this directly, for any $v \in \mathbb{R}^n$, weighted Cauchy–Schwarz yields

$$\begin{aligned} \|Ev\|_2^2 &\le \sum_{a=1}^{d} \left(\sum_{i=1}^{n} |E_{ai}|\,|v_i| \right)^2 \\ &\le \sum_{a=1}^{d} \left(\sum_{i=1}^{n} |E_{ai}| \right) \left(\sum_{i=1}^{n} |E_{ai}|\,|v_i|^2 \right) \\ &\le \alpha \sum_{i=1}^{n} |v_i|^2 \sum_{a=1}^{d} |E_{ai}| \le \alpha^2 \|v\|_2^2. \end{aligned}$$

This argument also covers $\alpha = 0$. Combining the linear and remainder estimates proves

$$\left\| \mathsf{D}[\sigma(UH + b)](x) \right\|_{\mathrm{op}} \le M\|T\|_{\mathrm{op}} + \frac{M}{2}\|T\|_{\mathrm{op}}^2 \le MKr + \frac{MK^2}{2} r^2,$$

which is equation 18. In particular, the estimate does not depend on the size of the bias b or the base point $H(x)$, because the derivative bounds in A4 hold on all of $\mathbb{R}$.

Step 2: Propagation through the recurrent states. Write $h_\ell(t; x)$ for the layer-ℓ state at time t on input x, and set

$$r_\ell = \Psi^{\circ(\ell-1)}(2K), \qquad 1 \le \ell \le L.$$

We prove by induction on ℓ that

$$\sup_{x \in \{-1,+1\}^n} \max_{0 \le t \le n} \left\| \mathsf{D} h_\ell(t; \cdot)(x) \right\|_{\mathrm{op}} \le r_\ell. \tag{20}$$

The first-layer recurrence and zero initialization imply, for $1 \le t \le n$,

$$h_1(t; x) = \sum_{s=1}^{t} A_1^{t-s} B_1 x(s) = \mathcal{R}_t \begin{bmatrix} x(1) & \cdots & x(t) \end{bmatrix}^\top.$$

Flipping a coordinate after time t does not change this state. For the first t coordinates, flipping $x(i)$ changes the state by $-2A_1^{t-i} B_1 x(i)$. Therefore

$$\mathsf{D} h_1(t; \cdot)(x) = -2 \begin{bmatrix} \mathcal{R}_t \operatorname{diag}(x(1), \ldots, x(t)) & 0_{m \times (n-t)} \end{bmatrix}. \tag{21}$$

The diagonal matrix is orthogonal, and adjoining zero columns does not change the operator norm. A1 consequently bounds equation 21 by $2K = r_1$. At $t = 0$ the state is identically zero, so its flip matrix is zero. This proves the induction base.

Now suppose equation 20 holds for layer $\ell - 1$, where $2 \le \ell \le L$. Unrolling only the time recurrence in layer ℓ gives

$$\begin{aligned} h_\ell(t; x) = &\sum_{s=1}^{t} A_\ell^{t-s} B_\ell V_\ell \sigma_\ell\big(U_\ell h_{\ell-1}(s; x) + b_\ell\big) \\ &+ \sum_{s=1}^{t} A_\ell^{t-s} B_\ell c_\ell. \end{aligned}$$

The second sum is independent of x and disappears under D. Although the activation bias b_ℓ need not disappear from the first sum, Step 1 controls it uniformly. Linearity of D thus gives

$$\mathsf{D} h_\ell(t; \cdot)(x) = \sum_{s=1}^{t} A_\ell^{t-s} B_\ell V_\ell \mathsf{D}\left[\sigma_\ell\big(U_\ell h_{\ell-1}(s; \cdot) + b_\ell\big) \right](x).$$

By the induction hypothesis, Step 1, and A2,

$$\begin{aligned}\left\|\mathsf{D}h_\ell(t;\cdot)(x)\right\|_{\mathrm{op}} &\le \sum_{s=1}^{t}\|A_\ell^{t-s}B_\ell V_\ell\|_{\mathrm{op}}\left(MKr_{\ell-1}+\frac{MK^2}{2}r_{\ell-1}^2\right)\\ &\le K\left(MKr_{\ell-1}+\frac{MK^2}{2}r_{\ell-1}^2\right)=\Psi(r_{\ell-1})=r_\ell.\end{aligned}$$

The finite sum is bounded by A2 because its powers range from 0 to $t-1$, and $t \le n$; no infinite-time stability condition is used. The initial state again has zero flip matrix. This proves equation 20 for every layer. When $L = 1$, the induction step is empty and the first-layer bound is already sufficient.

Step 3: A bound on the terminal flip energy. Let $a = V_g^\top C \in \mathbb{R}^d$. The readout can be written as

$$Y_n(x) = a^\top \sigma_g\big(U_g h_L(n;x)+b_g\big)+C^\top c_g.$$

The last term is independent of x. By A3, Step 1, and equation 20,

$$\begin{aligned}\|\mathsf{D}Y_n(x)\|_{\mathrm{op}} &\le \|a\|_2\left(MKr_L+\frac{MK^2}{2}r_L^2\right)\\ &\le \Psi(r_L)=\Psi^{\circ L}(2K).\end{aligned}$$

There are $L-1$ nonlinear inter-layer maps and one nonlinear readout, explaining the total of L applications of Ψ. For $r \ge 0$, the assumptions $K, M \ge 1$ imply

$$\Psi'(r) = MK^2 + MK^3 r \ge 0, \qquad \Psi(r) \ge MK^2 r \ge r.$$

Thus the iterates starting at $2K$ are nondecreasing, and $L \le L_0$ gives $\Psi^{\circ L}(2K) \le \Psi^{\circ L_0}(2K) = R_0$. For a scalar-valued output, $\mathsf{D}Y_n(x)$ is a row vector, whose squared operator norm is the sum of its squared entries. We have therefore proved the pointwise bound

$$\sum_{i=1}^{n}\big(Y_n(x^{(i)})-Y_n(x)\big)^2 \le R_0^2 \qquad \text{for every } x\in\{-1,+1\}^n. \tag{22}$$

Step 4: From flip energy to Fourier degree. For $A, B \subseteq [n]$, independence of the Rademacher coordinates and the identity $\chi_A\chi_B = \chi_{A\triangle B}$ imply

$$\mathbb{E}_{X\sim\mu_n}[\chi_A(X)\chi_B(X)] = \begin{cases}1, & A=B,\\ 0, & A\neq B.\end{cases}$$

There are 2^n such orthonormal characters, so they form a basis for all real-valued functions on the cube. For a scalar function H, write

$$\widehat{H}(S) = \mathbb{E}_{X\sim\mu_n}[H(X)\chi_S(X)], \qquad H = \sum_{S\subseteq[n]}\widehat{H}(S)\chi_S.$$

Flipping coordinate i changes χ_S precisely when $i \in S$, and hence

$$H(x^{(i)}) - H(x) = -2\sum_{S:,i\in S}\widehat{H}(S)\chi_S(x).$$

Squaring, taking expectations, and using orthonormality gives

$$\mathbb{E}\big[H(X^{(i)})-H(X)\big]^2 = 4\sum_{S:,i\in S}\widehat{H}(S)^2.$$

All sums are finite, so summing over i yields the exact identity

$$\mathbb{E}\sum_{i=1}^{n}\big(H(X^{(i)})-H(X)\big)^2 = 4\sum_{S\subseteq[n]}|S|\,\widehat{H}(S)^2. \tag{23}$$

Applying it to equation 22 gives

$$\sum_{S\subseteq[n]}|S|\,\widehat{Y}_n(S)^2 \le \frac{R_0^2}{4}. \tag{24}$$

No bound on the constant Fourier coefficient is needed.

Step 5: Comparison with the target's large scales. Set $r_* = \lceil q/2 \rceil$ and $k_* = 2^{r_*}$, and let $\Pi_{\geq k_*}$ be the orthogonal projection onto the span of characters with degree at least k_*:

$$\Pi_{\geq k_*} H = \sum_{S:,|S|\geq k_*} \widehat{H}(S)\chi_S.$$

Since $k_* \geq \sqrt{n}$, Parseval's identity and equation 24 imply

$$\begin{aligned}\|\Pi_{\geq k_*} Y_n\|^2_{L^2(\mu_n)} &= \sum_{S:,|S|\geq k_*} \widehat{Y}_n(S)^2 \\ &\leq \frac{1}{k_*} \sum_{S\subseteq[n]} |S|\, \widehat{Y}_n(S)^2 \leq \frac{R_0^2}{4\sqrt{n}}.\end{aligned}$$

Consequently,

$$\|\Pi_{\geq k_*} Y_n\|_{L^2(\mu_n)} \leq \frac{R_0}{2n^{1/4}}. \tag{25}$$

The dyadic supports $I_{r,j}$ are all distinct: supports at a fixed level are disjoint, and supports at different levels have different cardinalities. Also $|I_{r,j}| = 2^r$. Thus the target definition gives

$$\Pi_{\geq k_*} F_{n,\varepsilon} = \sum_{r=r_*}^{q} \sqrt{\frac{2^r}{qn}} \sum_{j=1}^{n/2^r} \varepsilon_{r,j}\chi_{I_{r,j}}.$$

Since $\varepsilon_{r,j}^2 = 1$, orthonormality yields

$$\|\Pi_{\geq k_*} F_{n,\varepsilon}\|^2_{L^2(\mu_n)} = \sum_{r=r_*}^{q} \frac{2^r}{qn}\frac{n}{2^r} = \frac{q - r_* + 1}{q}. \tag{26}$$

For $q = 1$, $r_* = 1$ and the last sum contains the single target level, so its value is one; no exceptional case is required.

Contractivity of orthogonal projection and the reverse triangle inequality now give

$$\begin{aligned}\|F_{n,\varepsilon} - Y_n\|_{L^2(\mu_n)} &\geq \|\Pi_{\geq k_*}(F_{n,\varepsilon} - Y_n)\|_{L^2(\mu_n)} \\ &\geq \|\Pi_{\geq k_*} F_{n,\varepsilon}\|_{L^2(\mu_n)} - \|\Pi_{\geq k_*} Y_n\|_{L^2(\mu_n)} \\ &\geq \sqrt{\frac{q - \lceil q/2\rceil + 1}{q}} - \frac{R_0}{2n^{1/4}}.\end{aligned}$$

The norm on the left is nonnegative, so the right-hand side can be replaced by its positive part. Furthermore, $q - \lceil q/2\rceil + 1 \geq q/2$. Monotonicity of the positive-part function proves both inequalities in equation 15, including horizons for which the displayed positive part is zero.

Step 6: The constant mean-squared-error bound. If $n \geq 4R_0^4$, then

$$n^{1/4} \geq \sqrt{2}\, R_0, \qquad \frac{R_0}{2n^{1/4}} \leq \frac{1}{2\sqrt{2}}.$$

Hence every admissible Y_n satisfies

$$\|F_{n,\varepsilon} - Y_n\|_{L^2(\mu_n)} \geq \frac{1}{2\sqrt{2}}, \qquad \mathbb{E}\big|F_{n,\varepsilon}(X) - Y_n(X)\big|^2 \geq \frac{1}{8}.$$

Taking the infimum over $\mathcal{N}_n(L_0, K, M)$ proves equation 16. All estimates hold for each target sign array and every finite m, d, with constants depending only on L_0, K, M. This proves the asserted uniformity. □

A.2 Proof of Theorem 2

Theorem 2: Approximation upper bound with logarithmic depth Let $n = 2^q$ with $q \geq 1$ and fix $F \in \mathcal{F}_n^{\mathrm{tree}}$. For any $0 < \varepsilon \leq 1$ and $0 < \delta < 1$, the preceding algorithm, with the prescribed parameter choices, returns a time-homogeneous, zero-initialized ReLU SSM of depth $L = q = \log_2 n$ whose terminal output Y_n satisfies

$$\|F - Y_n\|_{L^2(\mu_n)} \leq \varepsilon$$

with probability at least $1 - \delta$ over the training samples.

Parameters and notation for the proof. Fix a representation $F = F_{\mathcal{T},\theta}$ of the target, and write

$$\mathcal{S} := \mathcal{S}_{\mathcal{T}} = \bigcup_{r=1}^{q} \mathcal{T}_r, \qquad s := |\mathcal{S}| = n - 1, \qquad F(x) = \sum_{S \in \mathcal{S}} a_S \chi_S(x), \qquad a_S = \theta_S \sqrt{\frac{|S|}{qn}}.$$

The target, including its tree and signs, is fixed throughout the probability calculation. The learner need not know either of them. All probabilities below are over the independent training samples. Recall the algorithm's parameters, and define the auxiliary tolerance η by

$$\kappa := \sqrt{\frac{2}{qn}}, \qquad e := \min\left\{\frac{\kappa}{4}, \frac{\varepsilon}{\sqrt{s}}\right\}, \tag{27}$$

$$P := 1 + (n+s) + \binom{n+s}{2} = 2n^2 - n + 1, \qquad \eta := \frac{e}{2\sqrt{s}}, \tag{28}$$

$$N := \left\lceil \frac{8s}{e^2} \log \frac{2qPs}{\delta} \right\rceil. \tag{29}$$

Function norms are denoted explicitly by $L^2(\mu_n)$; $\|\cdot\|_2$ denotes the Euclidean norm of a coefficient vector. For a batch of samples at stage r, put

$$\langle u, v \rangle_r := \frac{1}{N} \sum_{a=1}^{N} u(X_r^{(a)}) v(X_r^{(a)}).$$

An ordered list $\mathcal{B}_r$ can contain repeated supports when the algorithm is defined on an arbitrary sample outcome. We will prove that repetitions do not survive on the successful event; we do not assume this in advance.

Proof. The proof first controls every coordinate sweep conditional on the earlier stages, then shows that one additional tree level becomes accessible at each stage. It concludes by constructing the required ReLU SSM explicitly. In addition to the stated approximation guarantee, the argument establishes exact support recovery on the successful event and polynomial bounds on the number of samples, arithmetic operations, and stored quantities.

Step 1: Population orthogonality and target normalization. For $A, B \subseteq [n]$, the parity identity

$$\chi_A(x)\chi_B(x) = \chi_{A \triangle B}(x)$$

follows by canceling the squared coordinates in $A \cap B$. Since the coordinates of $X \sim \mu_n$ are independent Rademacher random variables, every nonempty parity has mean zero. Consequently,

$$\mathbb{E}_{\mu_n}[\chi_A(X)\chi_B(X)] = \mathbf{1}_{\{A=B\}}. \tag{30}$$

In particular, these are population orthogonality identities; distinct parities need not be orthogonal on a finite sample.

The level-r supports partition $[n]$ into $n/2^r$ sets of cardinality 2^r. Therefore

$$\|F\|^2_{L^2(\mu_n)} = \sum_{S \in \mathcal{S}} a_S^2 = \sum_{r=1}^{q} \frac{n}{2^r} \frac{2^r}{qn} = 1. \tag{31}$$

Writing $a = (a_S)_{S \in \mathcal{S}}$, we obtain

$$\|a\|_2 = 1, \qquad \|a\|_1 \le \sqrt{s}, \qquad \min_{S \in \mathcal{S}} |a_S| = \kappa. \tag{32}$$

The middle inequality is Cauchy–Schwarz, and the last follows because the smallest internal supports have cardinality two. The parameter choices give the two bounds that will be used separately for selection and prediction:

$$e \le \frac{\kappa}{4}, \qquad \sqrt{s}\, e \le \varepsilon \le 1. \tag{33}$$

Step 2: Conditional concentration for the adaptively selected candidates. The candidate family at a later stage depends on earlier samples. To account for this dependence, condition on the full history before the current batch is drawn. For the probability argument, we may generate all q independent batches in advance and leave unused the batches following an early stop. This does not change the distribution of the algorithm's output.

Let $\mathcal{G}_{r-1}$ denote the information available immediately before drawing the stage-r batch, including all earlier batches, the selected lists, and the frozen stage-r visiting order. The order is chosen using only this prior information. If an independent random choice is used to select that order, include it in $\mathcal{G}_{r-1}$ as well. The fresh variables $X_r^{(1)}, \dots, X_r^{(N)}$ are independent of $\mathcal{G}_{r-1}$ and have law μ_n.

On every history for which stage r is reached, the previous survivor check gives $b_{r-1} \le s$. Hence

$$p_r = n + b_{r-1} \le n + s, \qquad |\mathcal{H}_r| = 1 + p_r + \binom{p_r}{2} \le P.$$

This bound counts candidate indices, including different indices that represent the same parity, and holds even if earlier selections were incorrect. Conditional on $\mathcal{G}_{r-1}$, each support $S_H^{(r)}$ is fixed and $z_H^{(r)} = \chi_{S_H^{(r)}}$ exactly.

Define $\mathcal{C}_r$ to be the event

$$\left| \langle \chi_{S_H^{(r)}}, \chi_S \rangle_r - \mathbf{1}_{\{S_H^{(r)} = S\}} \right| \le \eta \quad \text{for every } H \in \mathcal{H}_r \text{ and } S \in \mathcal{S}. \tag{34}$$

If the algorithm has stopped before stage r, define $\mathcal{C}_r$ to hold automatically. The unknown target support family occurs only in this analysis event; the algorithm never tests membership in $\mathcal{S}$.

Fix a reached history and one pair (H, S). If $S_H^{(r)} = S$, the empirical product is identically one, and its deviation in equation 34 is zero. If $S_H^{(r)} \ne S$, the variables

$$Z_a := \chi_{S_H^{(r)} \triangle S}(X_r^{(a)}), \qquad a = 1, \dots, N,$$

are conditionally independent Rademacher random variables. For example, fixing all but one coordinate in the nonempty support $S_H^{(r)} \triangle S$ shows that this parity takes each sign with probability $1/2$.

For a Rademacher variable Z and $\lambda \in \mathbb{R}$,

$$\mathbb{E} \exp(\lambda Z) = \cosh \lambda \le \exp(\lambda^2/2).$$

One way to verify the inequality is to compare the power series and use $(2k)! \ge 2^k k!$ for every integer $k \ge 0$. Independence and exponential Markov inequality therefore give, for $\lambda > 0$,

$$\begin{aligned} \Pr\left(\frac{1}{N} \sum_{a=1}^{N} Z_a > \eta \,\middle|\, \mathcal{G}_{r-1} \right) &\le \exp(-N\lambda\eta) \prod_{a=1}^{N} \mathbb{E}[\exp(\lambda Z_a) \mid \mathcal{G}_{r-1}] \\ &\le \exp\left(-N\lambda\eta + \frac{N\lambda^2}{2} \right). \end{aligned}$$

Taking $\lambda = \eta$ and applying the same argument to $-Z_a$ yields

$$\Pr\left(\left| \frac{1}{N} \sum_{a=1}^{N} Z_a \right| > \eta \,\middle|\, \mathcal{G}_{r-1} \right) \le 2 \exp(-N\eta^2/2). \tag{35}$$

There are at most Ps candidate–target pairs at this stage. A conditional union bound, followed by equation 29, gives

$$\Pr(\mathcal{C}_r^c \mid \mathcal{G}_{r-1}) \le 2Ps \exp(-N\eta^2/2) \le \frac{\delta}{q}, \tag{36}$$

on every reached history. The bound is also true on stopped histories because then $\mathcal{C}_r^c$ is empty. The last inequality uses $\eta^2 = e^2/(4s)$.

Set

$$\mathcal{E} := \bigcap_{r=1}^{q} \mathcal{C}_r.$$

Taking expectations in equation 36 and then taking a union bound over stages gives

$$\Pr(\mathcal{E}^c) \le \sum_{r=1}^{q} \mathbb{E}[\Pr(\mathcal{C}_r^c \mid \mathcal{G}_{r-1})] \le \delta. \tag{37}$$

No independence between different empirical correlations is required. In particular, equation 36 holds conditional on every valid successful prefix. We will show below that on $\mathcal{E}$ all prefixes are successful, so no early stop occurs.

Step 3: A deterministic guarantee for one thresholded coordinate sweep. Consider a reached stage r for which $\mathcal{C}_r$ holds. The following claim does not require the candidate supports to have any particular tree structure: the sweep retains exactly one indexed representative of each target support appearing among its candidates, retains no nontarget candidate, and estimates each retained target coefficient with error at most e.

We first verify the coordinate-descent interpretation precisely. When index H is visited, all other coefficients are fixed, and its own coefficient is still zero because each index is visited only once. If it were set to a scalar u, the loss as a function of u would be

$$Q_H(u) = \frac{1}{2N} \sum_{a=1}^{N} \left(R_H^{(r)}(X_r^{(a)}) - u z_H^{(r)}(X_r^{(a)}) \right)^2.$$

Every candidate is a parity, including the constant candidate, so its squared value is one on every input. Consequently,

$$Q'_H(u) = u - \langle z_H^{(r)}, R_H^{(r)} \rangle_r, \qquad Q''_H(u) = 1.$$

The unthresholded update in equation 10 is therefore the unique coordinate minimizer and also the unit gradient step from zero. Hard thresholding is a subsequent selection operation; we do not assert that the thresholded sweep minimizes the full empirical loss jointly over all coordinates.

We now prove the claim by induction along the frozen visiting order. Just before visiting the current index, let $\mathcal{A} \subseteq \mathcal{S}$ be the set of supports with nonzero coefficients already retained in this sweep. The induction invariant is that every $S \in \mathcal{A}$ has exactly one retained index, with coefficient w_S satisfying

$$|w_S - a_S| \le e,$$

and every other previously visited coordinate has coefficient zero. It holds before the first visit because all coefficients were initialized to zero. All unvisited coefficients are also zero.

Under this invariant the current residual, viewed as a function on the entire Boolean cube, has the expansion

$$R(x) = F(x) - \sum_{S \in \mathcal{A}} w_S \chi_S(x) = \sum_{S \in \mathcal{S}} c_S \chi_S(x), \qquad c_S = \begin{cases} a_S - w_S, & S \in \mathcal{A}, \\ a_S, & S \notin \mathcal{A}. \end{cases} \tag{38}$$

Its coefficients obey

$$\sum_{S \in \mathcal{S}} |c_S| \le \|a\|_1 + se \le \sqrt{s} + \varepsilon\sqrt{s} \le 2\sqrt{s}. \tag{39}$$

Let $U = S_H^{(r)}$ be the current candidate support and let $\widetilde{w}_H$ be its proposed value before thresholding. Put $c_U = 0$ when $U \notin \mathcal{S}$. Using equation 34 for this candidate and every target support, we have

$$\begin{aligned} \widetilde{w}_H &= \langle \chi_U, R \rangle_r = c_U + E_H, \\ E_H &= \sum_{S \in \mathcal{S}} c_S \left(\langle \chi_U, \chi_S \rangle_r - \mathbf{1}_{\{U=S\}} \right), \\ |E_H| &\le \eta \sum_{S \in \mathcal{S}} |c_S| \le 2\eta\sqrt{s} = e. \end{aligned} \tag{40}$$

Although the c_S depend on earlier updates using this same batch, this inequality is deterministic on $\mathcal{C}_r$. The simultaneous correlation bounds apply to every coefficient vector satisfying equation 39; no independence between the residual and the current batch is being assumed.

There are three cases.

1. If $U \notin \mathcal{S}$, then $c_U = 0$ and
$$|\widetilde{w}_H| \le e \le \kappa/4 < \kappa/2.$$
The algorithm sets this coefficient to zero. This case includes the constant parity, because the target has no empty support.

2. If $U \in \mathcal{S} \setminus \mathcal{A}$, then $c_U = a_U$. Thus
$$|\widetilde{w}_H - a_U| \le e, \qquad |\widetilde{w}_H| \ge |a_U| - e \ge 3\kappa/4 > \kappa/2.$$
The coefficient is retained and estimates a_U within e. It is the first retained index for this support.

3. If $U \in \mathcal{A}$, then the current index is a duplicate representation of a previously retained support. The corresponding residual coefficient satisfies $|c_U| = |a_U - w_U| \le e$, and hence
$$|\widetilde{w}_H| \le |c_U| + |E_H| \le 2e \le \kappa/2.$$
The duplicate is discarded. The algorithm discards at equality as well as below the threshold; this is needed when $2e = \kappa/2$.

Each case preserves the induction invariant. Therefore the claim holds after the complete sweep. In particular, every target support that occurs at least once among the candidates is retained exactly once, irrespective of the number or order of its indexed representations. At most s candidates survive, so the survivor-count check does not trigger at this stage.

Step 4: Successive stages recover successive levels of the tree. For $r = 0, \dots, q$, define
$$\mathcal{S}_{\le r} := \bigcup_{j=1}^{r} \mathcal{T}_j, \qquad \mathcal{S}_{\le 0} := \varnothing.$$

We claim that on $\bigcap_{j=1}^{r} \mathcal{C}_j$, the first r stages complete without an early stop and the ordered list $\mathcal{B}_r$ contains exactly one copy of each support in $\mathcal{S}_{\le r}$ and no other support. Moreover, the coefficient recorded at stage r for each such support approximates its true coefficient within e.

The statement for $r = 0$ follows from $\mathcal{B}_0 = ()$. Assume it holds through stage $r-1$. The available features at stage r are then exactly the leaf parities and one unweighted copy of every parity indexed by $\mathcal{S}_{\le r-1}$. This is where the normalization step is used: the available feature associated with S is χ_S itself, not an estimated multiple or a sign-adjusted version of that parity.

Every available support has size at most 2^{r-1}. This includes $r = 1$, when only singleton leaves are available. A candidate is the product of at most two available parities, so its support is their symmetric difference and has size at most 2^r. Thus no target support above level r can occur as a stage-r candidate.

Conversely, every target support strictly below level r already occurs as an available feature and is represented by a singleton candidate index. For any $S \in \mathcal{T}_r$, its two children $S^{(1)}, S^{(2)} \in \mathcal{T}_{r-1}$ are available: they are leaves when $r = 1$ and previously recovered internal supports when $r > 1$. The two children are disjoint, so their pair-product candidate has support
$$S^{(1)} \triangle S^{(2)} = S^{(1)} \cup S^{(2)} = S.$$

It follows that
$$\{S_H^{(r)} : H \in \mathcal{H}_r\} \cap \mathcal{S} = \mathcal{S}_{\le r}. \tag{41}$$

On $\mathcal{C}_r$, Step 3 now proves the induction claim. In particular, earlier target features are re-estimated in the current sweep through singleton candidates, rather than being lost when the stage coefficients are reset to zero.

We have therefore proved that on $\mathcal{E}$,

$$b_r = |\mathcal{S}_{\le r}| = \sum_{j=1}^{r} \frac{n}{2^j} = n(1 - 2^{-r}) \le s, \qquad r = 1, \dots, q, \tag{42}$$

and every retained list has no repeated support. At stage q, the list contains exactly the complete target support family $\mathcal{S}$.

Step 5: Coefficient accuracy, prediction accuracy, and the final checks. On $\mathcal{E}$, let $\widehat{a}_S$ denote the coefficient of the unique final-stage index representing $S \in \mathcal{S}$. Steps 3–4 give

$$|\widehat{a}_S - a_S| \le e \quad (S \in \mathcal{S}), \qquad \widehat{F}(x) = \sum_{S \in \mathcal{S}} \widehat{a}_S \chi_S(x). \tag{43}$$

By equation 30,

$$\begin{aligned} \|\widehat{F} - F\|^2_{L^2(\mu_n)} &= \mathbb{E}\left[\left(\sum_{S \in \mathcal{S}} (\widehat{a}_S - a_S)\chi_S(X) \right)^2 \right] \\ &= \sum_{S \in \mathcal{S}} |\widehat{a}_S - a_S|^2 \le s e^2 \le \varepsilon^2. \end{aligned} \tag{44}$$

The final coefficient vector also obeys

$$\|\widehat{a}\|_2 \le \|a\|_2 + \|\widehat{a} - a\|_2 \le 1 + \varepsilon \le 2. \tag{45}$$

The distinct final supports are exactly the internal-node family of the target tree, and there is one retained index per support. Both final checks therefore pass on $\mathcal{E}$. Once the exact compilation in Step 7 is established, equation 44 will give the claimed terminal-output error. The event has probability at least $1 - \delta$ by equation 37. This probability statement holds for each fixed target; no single event simultaneously covering all unknown targets is asserted.

Step 6: Sample and coordinate-update bounds. Regardless of the sampling outcome, any stage that is reached has at most s previously selected features. Therefore it has at most P indexed candidates, including all duplicate representations. There are at most q stages, so the algorithm uses at most qN labeled sequences and at most qP scalar coordinate updates. The inequalities are stated as upper bounds because an early stop may use fewer samples and updates.

The definition of e gives

$$e^{-2} = \max\{16\kappa^{-2}, s\varepsilon^{-2}\} = \max\{8qn, s\varepsilon^{-2}\}.$$

Using $s \le n$, $P \le 2n^2$, and the ceiling in equation 29, we obtain

$$\begin{aligned} N &\le 1 + 8s(8qn + s\varepsilon^{-2}) \log \frac{2qPs}{\delta} \\ &= O\Big(n^2(q + \varepsilon^{-2}) \log \frac{qn}{\delta} \Big). \end{aligned}$$

Here $n \ge 2$, $q = \log_2 n \ge 1$, and $0 < \delta < 1$, so $\log(2qPs/\delta) = O(\log(qn/\delta))$. Consequently,

$$qN = O\Big(n^2 q(q + \varepsilon^{-2}) \log \frac{qn}{\delta} \Big). \tag{46}$$

These are additional quantitative conclusions of the proof; they are not extra assumptions on the main theorem.

Step 7: Exact compilation into a logarithmic-depth ReLU SSM. We construct an SSM for any final candidate family that passes the algorithm's checks, whether or not the sampling event $\mathcal{E}$ holds. Write $\widehat{\mathcal{T}}$ for the perfect balanced tree identified by the final check and $\widehat{\mathcal{S}}$ for its internal supports. The family contains exactly s distinct internal supports, whereas the survivor check allows at most s retained indices. Thus every internal support has exactly one retained index. Denote its coefficient by $\widehat{a}_S$; then

$$\widehat{F}(x) = \sum_{S \in \widehat{\mathcal{S}}} \widehat{a}_S \chi_S(x), \qquad \|\widehat{a}\|_2 \le 2.$$

On $\mathcal{E}$, the recovered tree has the same support family as the target tree, and these coefficients agree with the notation in Step 5.

Set

$$L = q, \qquad m = n + s = 2n - 1, \qquad d = 2m = 4n - 2. \tag{47}$$

Index the state coordinates by all nodes of the recovered tree, placing the singleton leaves $\{1\}, \ldots, \{n\}$ first, in the input order. Let $\mathbf{e}_D \in \mathbb{R}^m$ denote the basis vector of node D; for a leaf we also write $\mathbf{e}_i := \mathbf{e}_{\{i\}}$. Write $\widehat{\mathcal{T}}_r$ for the recovered level-r supports and $D^{(1)}, D^{(2)}$ for the two children of an internal support D.

Let $\rho(u) = \max\{u, 0\}$ denote ReLU. The elementary identities used by the construction are

$$\rho(u) - \rho(-u) = u \quad (u \in \mathbb{R}), \qquad \rho(u+v) + \rho(-u-v) - 1 = uv \quad (u, v \in \{-1, +1\}). \tag{48}$$

For the second identity, the sum of the two ReLU terms is $|u + v|$: it equals two when $u = v$ and zero when $u = -v$. Subtracting one gives the product in both cases.

For each level $r = 1, \ldots, q$, define $\Phi_r : \mathbb{R}^m \to \mathbb{R}^m$ by

$$(\Phi_r(z))_D = \begin{cases} \rho(z_{D^{(1)}} + z_{D^{(2)}}) + \rho(-z_{D^{(1)}} - z_{D^{(2)}}) - 1, & D \in \widehat{\mathcal{T}}_r, \\ z_D, & D \notin \widehat{\mathcal{T}}_r. \end{cases} \tag{49}$$

Each map computes the current tree level and copies every other coordinate. We next give matrices showing that it has exactly the required one-hidden-layer form.

Index the $2m$ hidden units by $(D, +)$ and $(D, -)$, one pair for every state coordinate. For a fixed level r, put

$$\alpha_D^{[r]} := \begin{cases} \mathbf{e}_{D^{(1)}} + \mathbf{e}_{D^{(2)}}, & D \in \widehat{\mathcal{T}}_r, \\ \mathbf{e}_D, & D \notin \widehat{\mathcal{T}}_r. \end{cases}$$

Define $\overline{U}_r \in \mathbb{R}^{2m \times m}$ by

$$(\overline{U}_r)_{(D,+),:} = (\alpha_D^{[r]})^\top, \qquad (\overline{U}_r)_{(D,-),:} = -(\alpha_D^{[r]})^\top,$$

and define $\overline{V}_r \in \mathbb{R}^{m \times 2m}$ by

$$(\overline{V}_r)_{D,(D,+)} = 1, \qquad (\overline{V}_r)_{D,(D,-)} = \begin{cases} 1, & D \in \widehat{\mathcal{T}}_r, \\ -1, & D \notin \widehat{\mathcal{T}}_r, \end{cases}$$

with all other entries zero. Take $(\overline{c}_r)_D = -\mathbf{1}_{\{D \in \widehat{\mathcal{T}}_r\}}$ and set the hidden bias to zero. Then

$$\Phi_r(z) = \overline{V}_r \rho(\overline{U}_r z) + \overline{c}_r, \tag{50}$$

where ReLU acts coordinatewise. This equality holds for all $z \in \mathbb{R}^m$; the multiplication interpretation is needed only when the two child values are Boolean.

The first recurrent layer is the time-homogeneous shift register

$$A_1 = \sum_{i=1}^{n-1} \mathbf{e}_i \mathbf{e}_{i+1}^\top, \qquad B_1 = \mathbf{e}_n, \qquad h_1(0) = 0. \tag{51}$$

All matrix entries involving internal-node coordinates are zero. The recurrence implies, for $0 \le t \le n$,

$$h_1(t) = \sum_{i=1}^{t} x(i) \mathbf{e}_{n-t+i},$$

as is verified by induction on t. In particular,

$$h_1(n) = \sum_{i=1}^{n} x(i) \mathbf{e}_i. \tag{52}$$

Thus at the terminal time the leaf coordinates contain the input in its original order and all internal coordinates are zero.

For $\ell = 2, \dots, q$, choose

$$A_\ell = 0, \qquad B_\ell = I_m, \qquad f_\ell = \Phi_{\ell-1}, \qquad h_\ell(0) = 0, \tag{53}$$

using the matrices in equation 50. Set $g = \Phi_q$ with the same ReLU form. All these choices are independent of the time index. The model's same-time update convention and $A_\ell = 0$ give

$$h_\ell(n) = \Phi_{\ell-1}(h_{\ell-1}(n)), \qquad 2 \le \ell \le q.$$

To verify the computation, start from $z^{(0)} = h_1(n)$ and define $z^{(r)} = \Phi_r(z^{(r-1)})$ for $r = 1, \dots, q$. Inductively, $z^{(r)}$ contains the correct parity $\chi_D(x)$ at every leaf and every internal support through level r, and has zero coordinates at all higher internal levels. The base case is equation 52. For the induction step, each node at level r has two children at level $r-1$, whose coordinates are already their correct Boolean parity values. Their product is $\chi_D(x)$ by disjointness, and equation 48 implements that product. Every other coordinate is copied unchanged. Consequently,

$$h_\ell(n) = z^{(\ell-1)} \quad (1 \le \ell \le q), \qquad g(h_q(n)) = z^{(q)}.$$

The higher-layer states at earlier times need not have the same Boolean interpretation. This causes no problem: their recurrence matrices vanish, so they do not enter the terminal higher-layer computation.

Finally, define $C \in \mathbb{R}^m$ by

$$C_D = \begin{cases} \widehat{a}_D, & D \in \widehat{\mathcal{S}}, \\ 0, & D \text{ is a leaf.} \end{cases}$$

We have $\|C\|_2 = \|\widehat{a}\|_2 \le 2$ and, for every Boolean input sequence,

$$Y_n(x) = C^\top g(h_q(n)) = \sum_{S \in \widehat{\mathcal{S}}} \widehat{a}_S \chi_S(x) = \widehat{F}(x). \tag{54}$$

When $q = 1$, there are no inter-layer maps: the first layer stores the two inputs, and $g = \Phi_1$ computes their parity. Thus the same construction covers $n = 2$. In the depth convention of the main model there are q recurrent layers, $q-1$ inter-layer nonlinear maps, and one nonlinear readout.

For completeness, the construction also has uniform operator gains in A1–A3. The columns of the first-layer reachability matrix are distinct coordinate vectors, so $\|\mathcal{R}_t\|_{\mathrm{op}} = 1$ for $1 \le t \le n$. For each level map,

$$\begin{aligned} \|\overline{U}_r z\|_2^2 &= 2 \sum_{D \notin \widehat{\mathcal{T}}_r} z_D^2 + 2 \sum_{D \in \widehat{\mathcal{T}}_r} (z_{D^{(1)}} + z_{D^{(2)}})^2 \\ &\le 2\|z\|_2^2 + 4 \sum_{D \in \widehat{\mathcal{T}}_r} (z_{D^{(1)}}^2 + z_{D^{(2)}}^2) \le 6\|z\|_2^2. \end{aligned}$$

The final inequality uses the fact that each child belongs to exactly one parent at that level. The output rows have disjoint pairs of hidden units and squared norm two, so $\overline{V}_r \overline{V}_r^\top = 2I_m$. Hence

$$\|\overline{U}_r\|_{\mathrm{op}} \le \sqrt{6}, \qquad \|\overline{V}_r\|_{\mathrm{op}} = \sqrt{2}, \qquad \|V_g^\top C\|_2 = \sqrt{2}\|C\|_2 \le 2\sqrt{2}.$$

For a higher layer, $A_\ell = 0$ and $B_\ell = I_m$ imply

$$\sum_{k=0}^{n-1} \|A_\ell^k B_\ell V_\ell\|_{\mathrm{op}} = \|V_\ell\|_{\mathrm{op}} = \sqrt{2},$$

where the $k = 0$ term uses $A_\ell^0 = I_m$. Thus $K = 3$ suffices for A1–A3 independently of n. ReLU, however, is not $C^2(\mathbb{R})$ and does not satisfy A4. This compilation proves the ReLU upper bound as stated; it does not assert membership in the smooth class $\mathcal{N}_n(q, 3, M)$ for any M.

Step 8: Finite resources and a well-defined output on every sample outcome. The probabilistic argument is not needed for the algorithm to terminate or return an SSM. A reached stage has at most P candidates and one finite sweep. Each support can be stored as a length-n incidence vector. Symmetric difference, equality, and containment of supports can then be computed using $O(n)$

elementary operations. Even a naive evaluation of all stage candidates on all its samples uses at most $O(PNn)$ arithmetic operations. Maintaining the current residual on the batch makes each coordinate update and residual adjustment cost $O(N)$ additional operations. Every input and label is finite, and the updates involve only finite sums and multiplication by $1/N$, so the coefficients remain finite on every outcome. No division by an empirical variance or a fitted coefficient is required.

The final tree check and reconstruction are polynomial as well. One explicit implementation first forms the family $\mathcal{Q}$ of distinct retained supports and verifies that it contains exactly s members, all of size at least two, including $[n]$. Add all singleton supports to obtain $\mathcal{Q}^+$. Check the following properties:

1. $\mathcal{Q}^+$ is laminar: any two members are disjoint or one is contained in the other.
2. For each $D \in \mathcal{Q}$, the maximal proper members of $\mathcal{Q}^+$ contained in D are exactly two sets. They are disjoint, have equal cardinality, and their union is D.

These conditions are necessary for a perfect balanced binary tree. They are also sufficient. Give each internal set its two verified maximal proper subsets as children. Laminarity and maximality identify a unique parent for every node except $[n]$, while the singleton leaves and the covering condition ensure that every coordinate is represented. Every split halves the support cardinality, so a root of size 2^q has exactly q levels above its singleton leaves. This reconstructs a perfect balanced binary tree with precisely the prescribed internal supports. The required containment and maximality checks use polynomially many comparisons of length-n incidence vectors.

If either an early survivor check or a final check fails, return the zero SSM with the dimensions in equation 47, every initial state zero, ReLU activations, and all matrices, vectors, and biases zero. It is time homogeneous and has terminal output identically zero. If the final checks pass, at most s retained indices represent exactly s distinct supports, so duplicates cannot remain. Step 7 therefore applies to every passing output, including a passing output outside $\mathcal{E}$. The final coefficient check supplies $\|C\|_2 \le 2$ in that case too.

Using dense storage, each compiled layer has $O(n^2)$ parameters and there are $q = \log_2 n$ layers. All matrices can be constructed in $O(qn^2)$ arithmetic operations once the tree is known. Storing batches, candidate values, supports, residuals, retained records, and network parameters requires polynomial storage in n, ε^{-1}, and $\log(\delta^{-1})$. Together with Step 6 and the direct implementation just described, this also gives polynomially many arithmetic operations. These statements concern real arithmetic and storage of real numbers; no bit-complexity claim for an encoding of the noiseless real labels is needed.

The algorithm thus always returns a time-homogeneous, zero-initialized ReLU SSM of depth q. On the event $\mathcal{E}$, which has probability at least $1 - \delta$, the checks pass and equation 54 and equation 44 imply $\|F - Y_n\|_{L^2(\mu_n)} \le \varepsilon$. This proves Theorem 2. □

A.3 Proof of Theorem 3

Theorem 3: Exact depth-two representation with $\tanh$. Let $n = 2^q$, where $q \ge 1$, and fix $F = F_{\mathcal{T},\theta} \in \mathcal{F}_n^{\mathrm{tree}}$. There exists a time-homogeneous, zero-initialized SSM with depth $L = 2$, hidden dimension $m = O(n)$ and nonlinearity width $d = O(n \log n)$, using

$$f_2(z) = V_2 \tanh(U_2 z + b_2) + c_2, \qquad g(z) = z,$$

such that

$$Y_n(x) = F(x) \qquad \text{for every } x \in \{-1, +1\}^n.$$

The construction satisfies Assumptions A1–A4 with $M = 1$ and a norm budget $K = O(n)$.

Proof. We give explicit dimensions

$$m = n - 1, \qquad d = nq,$$

and verify the assumptions with $K = 2n$ and $M = 1$. The construction depends on the fixed target tree and its signs; it is a representation result, not a procedure for learning an unknown tree from

samples. Write

$$\mathcal{S} = \mathcal{S}_{\mathcal{T}}, \qquad F(x) = \sum_{S \in \mathcal{S}} a_S \chi_S(x), \qquad a_S = \theta_S \sqrt{\frac{|S|}{qn}}.$$

The tree has $n-1$ internal supports, every internal support has even cardinality, and parity orthogonality gives

$$\sum_{S \in \mathcal{S}} a_S^2 = \sum_{r=1}^{q} \frac{n}{2^r} \frac{2^r}{qn} = 1. \tag{55}$$

Step 1: Store a sufficient linear encoding of the input. Let $j_\star \in \{2, \dots, n\}$ be the label of the leaf that shares a parent with leaf 1. Every internal support contains either both of these leaves or neither:

$$1 \in S \quad \Longleftrightarrow \quad j_\star \in S \qquad (S \in \mathcal{S}). \tag{56}$$

Thus, for recovering sums over internal supports, it is sufficient to store $x(1) + x(j_\star)$ together with the remaining inputs.

Let $e_1, \dots, e_m$ denote the standard basis of $\mathbb{R}^m$. Set $B_1 = e_1$ and define the time-independent matrix A_1 on this basis by

$$A_1 e_a = e_{a+1} \quad (1 \le a < m), \qquad A_1 e_m = e_{n-j_\star+1}. \tag{57}$$

The first rule is empty when $n = 2$; in that case $m = 1$, $j_\star = 2$, and $A_1 e_1 = e_1$. Starting with $h_1(0) = 0$, the recurrence $h_1(t+1) = A_1 h_1(t) + B_1 x(t+1)$ gives

$$h_1(t) = \sum_{i=1}^{t} A_1^{t-i} e_1 x(i).$$

For $0 \le a \le n-2$ we have $A_1^a e_1 = e_{a+1}$, whereas $A_1^{n-1} e_1 = e_{n-j_\star+1}$. Consequently,

$$h_1(n) = \sum_{i=2}^{n} e_{n-i+1} x(i) + e_{n-j_\star+1} x(1). \tag{58}$$

For each $S \in \mathcal{S}$, define

$$b_S := \sum_{i \in S \setminus \{1\}} e_{n-i+1} \in \mathbb{R}^m. \tag{59}$$

Combining equation 56 and equation 58 yields

$$b_S^\top h_1(n) = \sum_{i \in S \setminus \{1\}} x(i) + \mathbf{1}_{\{j_\star \in S\}} x(1) = \sum_{i \in S} x(i). \tag{60}$$

This first layer also satisfies the reachability bound uniformly over the specified horizon. For $1 \le t \le n-1$, the columns of $\mathcal{R}_t$ are distinct standard basis vectors, and hence $\|\mathcal{R}_t\|_{\mathrm{op}} = 1$. At $t = n$, one basis vector occurs twice and every other basis vector occurs once. Therefore,

$$\sup_{1 \le t \le n} \|\mathcal{R}_t\|_{\mathrm{op}} = \sqrt{2}. \tag{61}$$

In particular, the compressed memory itself requires no norm budget growing with n.

Step 2: Interpolate an even-order parity using shifted $\tanh$ units. Fix an even integer $k \ge 2$ and write $u_i = -k + 2i$ for $i = 0, \dots, k$. If k Boolean inputs have sum u_i, exactly i of them are $+1$, so their product equals $(-1)^{k-i} = (-1)^i$. We construct a scalar function

$$p_k(u) = c_k + \sum_{j=1}^{k} w_j^{(k)} \tanh(u + k - 2j + 1) \tag{62}$$

such that $p_k(u_i) = (-1)^i$ at all $k+1$ possible sums.

Define the $k \times k$ matrix $D^{(k)}$ by

$$D_{ij}^{(k)} = \tanh\big(2(i-j)+1\big) - \tanh\big(2(i-j)-1\big), \qquad i,j \in [k]. \tag{63}$$

Every entry is positive, and every diagonal entry equals $2\tanh(1)$. The sum along row i telescopes to

$$\sum_{j=1}^{k} D_{ij}^{(k)} = \tanh(2i-1) - \tanh(2i-2k-1) < 2.$$

It follows that the diagonal-dominance margin satisfies

$$D_{ii}^{(k)} - \sum_{j \neq i} D_{ij}^{(k)} > 4\tanh(1) - 2 =: \gamma > 1. \tag{64}$$

Here, $\tanh(1) > 3/4$ follows, for example, from $\exp(2) > 7$.

For completeness, strict diagonal dominance gives the needed inverse estimate directly. For any $v \in \mathbb{R}^k$, choose i with $|v_i| = \|v\|_\infty$. Then

$$\|D^{(k)}v\|_\infty \geq |(D^{(k)}v)_i| \geq \left(D_{ii}^{(k)} - \sum_{j \neq i} D_{ij}^{(k)} \right) \|v\|_\infty \geq \gamma \|v\|_\infty.$$

Thus, $D^{(k)}$ is invertible and its inverse, in the induced ℓ_∞ norm, has norm at most γ^{-1}. Choose $w^{(k)}$ to be the unique solution of

$$D^{(k)}w^{(k)} = \big(2(-1)^i\big)_{i=1}^{k}, \qquad c_k = 1 - \sum_{j=1}^{k} w_j^{(k)} \tanh(1-2j). \tag{65}$$

The coefficient bound is uniform in k:

$$\|w^{(k)}\|_\infty \leq \frac{2}{\gamma} < 2, \qquad \|w^{(k)}\|_2 \leq \frac{2\sqrt{k}}{\gamma} < 2\sqrt{k}. \tag{66}$$

By construction, $p_k(u_0) = 1$. Also, for $i = 1, \ldots, k$,

$$p_k(u_i) - p_k(u_{i-1}) = \sum_{j=1}^{k} D_{ij}^{(k)} w_j^{(k)} = 2(-1)^i = (-1)^i - (-1)^{i-1}.$$

Induction therefore gives $p_k(u_i) = (-1)^i$ for every i. In particular,

$$p_{|S|}\left(\sum_{i \in S} x(i) \right) = \chi_S(x) \quad \big(S \in \mathcal{S},\ x \in \{-1,+1\}^n\big). \tag{67}$$

Step 3: Compute all internal-node parities in parallel. Choose any bijection $\iota : \mathcal{S} \to [m]$ to index the coordinates of the second hidden state. Index its nonlinear units by the pairs

$$\mathcal{I} = \{(S,j) : S \in \mathcal{S},\ j \in [|S|]\}.$$

Their number is

$$d = |\mathcal{I}| = \sum_{S \in \mathcal{S}} |S| = \sum_{r=1}^{q} \frac{n}{2^r}\, 2^r = nq. \tag{68}$$

Define $U_2 \in \mathbb{R}^{d \times m}$, $b_2 \in \mathbb{R}^d$, and $V_2 \in \mathbb{R}^{m \times d}$ by their rows, entries, and columns as follows:

$$(U_2)_{(S,j),:} = b_S^\top, \qquad (b_2)_{(S,j)} = |S| - 2j + 1, \qquad (V_2)_{:,(S,j)} = w_j^{(|S|)} e_{\iota(S)}. \tag{69}$$

Set

$$c_2 = \sum_{S \in \mathcal{S}} c_{|S|} e_{\iota(S)}, \qquad A_2 = 0, \qquad B_2 = I_m.$$

The second-layer map is precisely $f_2(z) = V_2 \tanh(U_2 z + b_2) + c_2$, and the second-layer recurrence, initialized at $h_2(0) = 0$, satisfies

$$h_2(t+1) = f_2(h_1(t+1)).$$

All parameters are independent of t. By equation 60 and equation 67, its terminal coordinates are

$$(h_2(n))_{\iota(S)} = p_{|S|}(b_S^\top h_1(n)) = \chi_S(x) \qquad (S \in \mathcal{S}). \tag{70}$$

Choose

$$C = \sum_{S \in \mathcal{S}} a_S e_{\iota(S)}, \qquad g(z) = z.$$

Then equation 70 gives exact equality on every Boolean input:

$$Y_n(x) = C^\top g(h_2(n)) = \sum_{S \in \mathcal{S}} a_S \chi_S(x) = F(x). \tag{71}$$

To place the identity readout in the prescribed one-hidden-layer form with the same width d, take the coordinatewise identity activation $\sigma_g(u) = u$ and set

$$U_g = \begin{bmatrix} I_m \\ 0_{(d-m)\times m} \end{bmatrix}, \qquad V_g = \begin{bmatrix} I_m & 0_{m\times(d-m)} \end{bmatrix}, \qquad b_g = 0, \qquad c_g = 0.$$

This padding is possible because $d = nq \geq m$.

Step 4: Verify the norm and regularity assumptions. Assumption A1 follows from equation 61. For A2, the input matrix of the nonlinear map satisfies

$$U_2^\top U_2 = \sum_{S \in \mathcal{S}} |S|\, b_S b_S^\top = \sum_{r=1}^{q} 2^r \sum_{S \in \mathcal{T}_r} b_S b_S^\top. \tag{72}$$

At each fixed level r, the vectors b_S, $S \in \mathcal{T}_r$, have pairwise disjoint supports, and $\|b_S\|_2^2 = |S \setminus \{1\}| \leq 2^r$. Consequently,

$$\left\| 2^r \sum_{S \in \mathcal{T}_r} b_S b_S^\top \right\|_{\mathrm{op}} \leq 4^r,$$

and therefore

$$\|U_2\|_{\mathrm{op}}^2 \leq \sum_{r=1}^{q} 4^r = \frac{4(n^2-1)}{3}. \tag{73}$$

The root support $[n]$ alone contributes $n\, b_{[n]} b_{[n]}^\top$ to equation 72, and $\|b_{[n]}\|_2^2 = n - 1$. Hence,

$$\sqrt{n(n-1)} \leq \|U_2\|_{\mathrm{op}} \leq \sqrt{\frac{4(n^2-1)}{3}}. \tag{74}$$

Thus, this construction uses an input operator norm of order n at the nonlinear stage.

Different rows of V_2 have disjoint column supports. It follows from equation 66 that

$$\|V_2\|_{\mathrm{op}} = \max_{S \in \mathcal{S}} \|w^{(|S|)}\|_2 \leq \frac{2\sqrt{n}}{\gamma} < 2\sqrt{n}.$$

Since $A_2 = 0$ and $B_2 = I_m$, with the usual convention $A_2^0 = I_m$, we have

$$\sum_{a=0}^{n-1} \|A_2^a B_2 V_2\|_{\mathrm{op}} = \|V_2\|_{\mathrm{op}} < 2\sqrt{n}. \tag{75}$$

Both bounds required by A2 are therefore at most $K = 2n$.

For A3, the padded identity readout satisfies

$$\|U_g\|_{\mathrm{op}} = 1, \qquad \|V_g^\top C\|_2 = \|C\|_2 = 1,$$

where the last equality uses equation 55. Thus A3 also holds with $K = 2n$.

Finally, both activations are C^2 on $\mathbb{R}$. For $\sigma(u) = \tanh(u)$,

$$\sigma'(u) = 1 - \tanh^2(u), \qquad \sigma''(u) = -2\tanh(u)(1 - \tanh^2(u)).$$

Therefore

$$\sup_u |\sigma'(u)| = 1, \qquad \sup_u |\sigma''(u)| = \max_{0 \le v \le 1} 2v(1 - v^2) = \frac{4}{3\sqrt{3}} < 1.$$

The identity activation has first derivative one and second derivative zero. Assumption A4 thus holds with $M = 1$. The biases b_2, c_2 are finite; A1–A4 do not impose bounds on these biases, and no uniform bound on all parameter norms is asserted here.

We have constructed a time-homogeneous, zero-initialized model with $L = 2$, $m = n - 1$, $d = nq$, exact terminal output, and the claimed gain and regularity bounds. This proves the theorem. □

Scope of the construction. The operator-norm estimate equation 74 describes this construction and is not a lower bound on the norms of all shallow representations. In particular, the result does not contradict the constant-depth lower bound, which fixes the gain budget independently of n. Depth two is sufficient here, but is not claimed to be necessary: because the architecture permits a nonlinear readout, the map f_2 can instead be placed in that readout above the same first recurrent layer, giving a depth-one representation with a growing norm budget as well.

A.4 A SMOOTH REALIZATION UNDER THE SAME NORM AND REGULARITY ASSUMPTIONS

Smooth compilation of the predictor in Theorem 2 . Let $n = 2^q$, $q \ge 1$, and let $\mathcal{T}$ be a perfect balanced binary tree on $[n]$. For any real coefficient vector $\alpha = (\alpha_S)_{S \in \mathcal{S}_\mathcal{T}}$ with $\|\alpha\|_2 \le 2$, define

$$H(x) = \sum_{S \in \mathcal{S}_\mathcal{T}} \alpha_S \chi_S(x).$$

There is a time-homogeneous, zero-initialized SSM whose activations are all the standard softplus function $\sigma(u) = \log(1 + e^u)$, with

$$L = q, \qquad m = 2n - 1, \qquad d = 4n - 2,$$

whose terminal output equals $H(x)$ for every $x \in \{-1, +1\}^n$. It satisfies Assumptions A1–A4 with $K = 7$ and $M = 1$. Consequently, the learner in Theorem 2 can also be compiled into an output in $\mathcal{N}_n(q, 7, 1)$ on every sampling outcome, with the same probability of success, approximation guarantee, and sample bound.

Purpose of this additional result. The main-text statement of Theorem 2 explicitly specifies ReLU. Its proof above proves that statement as written. ReLU does not satisfy A4, so that realization alone does not establish an upper bound within the smooth model class used in Theorem 1. The following separate construction resolves this issue using a standard smooth activation. It changes only the compilation, and does not change the learner, its selected features, or its statistical analysis.

Proof. We first construct exact copy and product gates, then verify their layerwise operator norms and the resulting finite-horizon SSM bounds.

Step 1: Two exact identities for softplus. For every real u,

$$\sigma(u) - \sigma(-u) = u, \qquad \sigma(u) + \sigma(-u) = 2\log\big(2\cosh(u/2)\big). \tag{76}$$

Indeed, $(1 + e^u)/(1 + e^{-u}) = e^u$ and $(1 + e^u)(1 + e^{-u}) = 4\cosh^2(u/2)$. Set

$$\lambda = \frac{1}{\log\cosh(1)}, \qquad \beta = -1 - 2\lambda\log 2.$$

The denominator is strictly positive. If $u, v \in \{-1, +1\}$, then $u + v = 0$ when $uv = -1$, and $u + v \in \{-2, 2\}$ when $uv = 1$. The second identity in equation 76 therefore gives

$$uv = \lambda\{\sigma(u + v) + \sigma(-u - v)\} + \beta \qquad (u, v \in \{-1, +1\}). \tag{77}$$

For completeness, at $u + v = 0$ the right-hand side is $2\lambda\log 2 + \beta = -1$. At $u + v = \pm 2$ it is $2\lambda\log(2\cosh 1) + \beta = 1$. The product identity is only asserted for Boolean children; the copy identity in equation 76 holds on all of $\mathbb{R}$.

Step 2: A map for one level of the tree. Index the m state coordinates by the leaves and internal supports of $\mathcal{T}$. For $r \in [q]$, define $\widetilde{\Phi}_r$ by copying every coordinate except the nodes at level r, and by replacing each such node's coordinate with the expression in equation 77 applied to its two children. We now specify matrices for a one-hidden-layer representation

$$\widetilde{\Phi}_r(z) = V_r \sigma(U_r z) + c_r.$$

Allocate two distinct hidden units to each output coordinate, giving common width $2m$. For a copied coordinate a, take input rows $e_a^\top, -e_a^\top$, output weights $1, -1$ into coordinate a, and output bias zero. For a newly computed coordinate a with children b, c, take input rows $(e_b + e_c)^\top, -(e_b + e_c)^\top$, output weights λ, λ into coordinate a, and output bias β. All unspecified entries are zero. The preactivation bias is zero. This construction uses the same scalar activation at every hidden unit.

Step 3: Dimension-independent bounds for these matrices. At a fixed level, the pairs of child coordinates are disjoint. For any $z \in \mathbb{R}^m$, the input matrix obeys

$$\begin{aligned}\|U_r z\|_2^2 &= 2 \sum_{a \text{ copied}} z_a^2 + 2 \sum_{(b,c) \text{ child pairs}} (z_b + z_c)^2 \\ &\le 2\|z\|_2^2 + 4 \sum_{(b,c) \text{ child pairs}} (z_b^2 + z_c^2) \\ &\le 6\|z\|_2^2.\end{aligned}$$

Thus $\|U_r\|_{\mathrm{op}} \le \sqrt{6}$. The rows of V_r have disjoint hidden-unit supports. Their squared Euclidean norms are either 2 or $2\lambda^2$, so

$$\|V_r\|_{\mathrm{op}} \le \sqrt{2}\max\{1, \lambda\}.$$

Here $1 < \lambda < 7/3$. One fully explicit verification of the upper bound is as follows. The power series gives $\cosh 1 > 1 + 1/2 + 1/24 + 1/720 = 1111/720$. For $0 < t < 3$, the inequality $j! \ge 2\,3^{j-2}$ for $j \ge 2$ gives

$$e^t \le 1 + t + \frac{t^2/2}{1 - t/3}.$$

At $t = 3/7$, its right-hand side is $43/28 < 1111/720$. Hence $e^{3/7} < \cosh 1$, so $\log\cosh 1 > 3/7$ and $\lambda < 7/3$. Also $\cosh 1 < e$ implies $\lambda > 1$. In particular,

$$\|U_r\|_{\mathrm{op}} \le \sqrt{6} < 7, \qquad \|V_r\|_{\mathrm{op}} \le \sqrt{2}\lambda < 7. \tag{78}$$

Step 4: Recurrent implementation and exact terminal output. Use the same length-n shift register as in the ReLU compilation, with the internal-node coordinates initially zero. Its terminal state is

$$h_1(n) = (x(1), \ldots, x(n), 0, \ldots, 0),$$

and its finite-horizon reachability norm is one. For $\ell = 2, \ldots, q$, set

$$A_\ell = 0, \qquad B_\ell = I_m, \qquad f_\ell = \widetilde{\Phi}_{\ell-1}, \qquad g = \widetilde{\Phi}_q.$$

If $q = 1$, there are no inter-layer maps and only the last choice is used. Induction on the tree level, using the exact copy and Boolean product identities, shows that $g(h_q(n))$ contains all internal-node parities. Set $C_S = \alpha_S$ at an internal node and $C_{\{i\}} = 0$ at a leaf. Then $C^\top g(h_q(n)) = H(x)$ on the entire Boolean cube. All matrices are time independent and all states are initialized at zero. Earlier-time higher-layer states have no effect at time n because $A_\ell = 0$ for $\ell \ge 2$.

Step 5: Verification of A1–A4. The shift register proves A1 with bound one. For $\ell \ge 2$, the only nonzero term of the impulse-response sum is the term with exponent zero; therefore

$$\sum_{k=0}^{n-1} \|A_\ell^k B_\ell V_\ell\|_{\mathrm{op}} = \|V_\ell\|_{\mathrm{op}} \le \sqrt{2}\lambda < 7.$$

Equation equation 78 bounds all inter-layer input matrices and the readout input matrix. For the readout output gain,

$$\|V_g^\top C\|_2 \le \|V_g\|_{\mathrm{op}} \|C\|_2 \le 2\sqrt{2}\lambda < \frac{14\sqrt{2}}{3} < 7.$$

Finally,

$$\sigma'(u) = \frac{e^u}{1+e^u} \in (0,1), \qquad \sigma''(u) = \frac{e^u}{(1+e^u)^2} \in (0, 1/4].$$

Thus A4 holds with $M = 1$, and all norm inequalities hold with $K = 7$. The output biases c_r need no bound because A1–A4 do not constrain them.

Step 6: Applying the construction to every learner output. On a successful sampling outcome, use the recovered tree and $\alpha = \widehat{a}$. Theorem 2's proof gives $\|\widehat{a}\|_2 \le 2$ and the desired approximation error. If the final checks pass on an unsuccessful sampling outcome, they still provide a valid balanced tree and coefficient norm at most two, so the same construction is admissible. If any check fails, return a zero SSM with the specified depth and dimensions and with all parameter matrices and biases zero. Softplus activations are allowed in this zero model, since the output is zero when $C = 0$. The samples and selected coefficients are unchanged by recompilation. This proves the asserted membership in $\mathcal{N}_n(q, 7, 1)$ on every outcome and the probability and approximation claims on the same successful event as before. □

Scope of the corollary. Together with Theorem 1 at $K = 7, M = 1$, this is a comparison under the same norm and regularity constraints. It establishes a constant-depth approximation lower bound for sufficiently long sequences and logarithmic-depth sufficiency for the target family. It does not prove that logarithmic depth is minimal. The softplus realization is an additional result; an assertion in the main text that the upper bound satisfies A1–A4 should explicitly reference this corollary or state the smooth realization.

A.5 Universal approximation at a fixed horizon

Universal approximation at depth two. Fix an integer $n \ge 1$ and a nonempty compact set $\mathcal{D} \subseteq \mathbb{R}^n$. For every continuous function $F : \mathcal{D} \to \mathbb{R}$ and every $\varepsilon > 0$, there exists a time-homogeneous, zero-initialized SSM of the stated architecture, with depth $L = 2$, hidden dimension $m = n$ and a common finite nonlinearity width d, such that

$$\sup_{x \in \mathcal{D}} |Y_n(x) - F(x)| < \varepsilon.$$

Both f_2 and g can be implemented as one-hidden-layer FFNs with coordinatewise $\tanh$ activation. The width and parameter norms may depend on n, $\mathcal{D}$, F, and ε.

Proof. The proof has three components: an exact linear memory for the finite input sequence, a scalar feedforward approximation, and its embedding into the two nonlinear maps required by the architecture. No dyadic assumption on n is used in this proposition.

Step 1: A time-homogeneous shift register. Let $e_1, \dots, e_n$ be the standard basis of $\mathbb{R}^n$. Set $B_1 = e_n$ and define

$$A_1 e_1 = 0, \qquad A_1 e_i = e_{i-1} \quad (2 \le i \le n).$$

For $n = 1$, this means $A_1 = 0$ and $B_1 = 1$. The zero-initialized first-layer recurrence satisfies, for $1 \le t \le n$,

$$\begin{aligned} h_1(t; x) &= \sum_{j=1}^{t} A_1^{t-j} e_n \, x(j) \\ &= \sum_{j=1}^{t} e_{n-t+j} \, x(j). \end{aligned}$$

In particular,

$$h_1(n; x) = (x(1), \dots, x(n))^\top = x. \tag{79}$$

The memory is exact on all of $\mathbb{R}^n$, not only on Boolean inputs. The columns of every $\mathcal{R}_t$ are distinct standard basis vectors, so $\|\mathcal{R}_t\|_{\mathrm{op}} = 1$ for $1 \le t \le n$. The memory itself thus needs no growing reachability gain.

Step 2: A bounded inverse of the final activation. Since F is continuous on the nonempty compact set $\mathcal{D}$, the number $B = \sup_{x\in\mathcal{D}} |F(x)|$ is finite. Set $a = 1 + B$ and define

$$H(x) = \operatorname{arctanh}\big(F(x)/a\big).$$

For all $x \in \mathcal{D}$, $|F(x)/a| \le B/(1+B) < 1$. Hence H is well defined, continuous, and bounded on $\mathcal{D}$. It obeys the exact identity

$$a \tanh(H(x)) = F(x) \qquad (x \in \mathcal{D}). \tag{80}$$

This inverse transformation is what allows both nonlinear blocks to use $\tanh$; we do not replace the readout by the identity map.

Step 3: Apply scalar universal approximation. The activation $\tanh$ is continuous, bounded, and nonconstant. The uniform-on-compact universal approximation theorem (**?**) therefore gives a finite integer $d \ge 1$, vectors $u_j \in \mathbb{R}^n$, and scalars v_j, b_j, c, such that

$$p(z) = c + \sum_{j=1}^{d} v_j \tanh(u_j^\top z + b_j)$$

satisfies

$$\sup_{x\in\mathcal{D}} |p(x) - H(x)| < \varepsilon/a. \tag{81}$$

The theorem is used only as an existence result: no uniform bound on d, the coefficients, or their dependence on accuracy is inferred. It applies to an arbitrary compact subset of $\mathbb{R}^n$; the domain need not be a box and need not have nonempty interior.

Step 4: Explicit embedding into both FFNs. Use $m = n$. Define $U_2 \in \mathbb{R}^{d\times n}$ to have row $u_j^\top$, take $b_2 = (b_j)_{j=1}^d$, and set

$$V_2 = e_1(v_1, \dots, v_d), \qquad c_2 = ce_1.$$

Then $f_2(z) = e_1 p(z)$. Set $A_2 = 0$ and $B_2 = I_n$. To implement $g(z) = e_1 \tanh(z_1)$ at the same width d, let U_g have first row $e_1^\top$ and all other rows zero, and let V_g have first column e_1 and all other columns zero. Take $b_g = 0$, $c_g = 0$, and $C = ae_1$. These choices give exactly the model's required matrix dimensions and use $\tanh$ in both FFNs. All parameters are independent of time. Initialize both states at zero. Since $A_2 = 0$, the terminal state and output are, by equation 79,

$$h_2(n; x) = e_1 p(x), \qquad Y_n(x) = a \tanh(p(x)).$$

No constraint is needed on what the second layer computes before the terminal time, because its previous states are not carried forward.

Step 5: Uniform error bound. The derivative of $\tanh$ has magnitude at most one, so the mean-value theorem gives $|\tanh u - \tanh v| \le |u - v|$ for all real u, v. Combining equation 80 and equation 81 yields

$$\begin{aligned}\sup_{x\in\mathcal{D}} |Y_n(x) - F(x)| &= a \sup_{x\in\mathcal{D}} |\tanh(p(x)) - \tanh(H(x))| \\ &\le a \sup_{x\in\mathcal{D}} |p(x) - H(x)| < \varepsilon.\end{aligned}$$

This proves the proposition. □

Relation to the constrained results. The horizon and target are fixed before the approximating model is chosen. The proposition does not assert approximation on an unbounded domain, exact representation of every continuous function, one common model for all horizons, or a polynomial resource bound. Although $\tanh$ satisfies A4 with $M = 1$, no common norm budget K is imposed on the entire family of approximants. It therefore does not conflict with Theorem 1.

Depth convention. The proposition establishes sufficiency of two recurrent layers with two $\tanh$ blocks. Under the paper's general convention permitting a nonlinear readout, depth one is already sufficient for universality: retain the same shift register and approximate F directly by the scalar readout $C^\top g$. Thus this proposition is a finite-horizon embedding of classical feedforward universality, rather than a lower bound on the number of nonlinear blocks required.

A.6 Proof of Proposition 4

Proof. Fix $n \geq 1$, a nonempty compact set $\mathcal{D} \subseteq \mathbb{R}^n$, a continuous functional $F : \mathcal{D} \to \mathbb{R}$, and $\varepsilon > 0$. We construct the input map and recurrent memory explicitly, and then apply the classical universal approximation theorem to the output MLP. Throughout the proof, $x = (x(1), \ldots, x(n)) \in \mathcal{D}$.

Step 1: An information-preserving nonlinear input map. Set

$$\phi(u) = \tanh(u), \qquad T(x) = \big(\tanh(x(1)), \ldots, \tanh(x(n))\big).$$

The scalar function $\tanh$ is a continuous, strictly increasing bijection from $\mathbb{R}$ to $(-1, 1)$, with continuous inverse $\operatorname{arctanh}$. Consequently, $T : \mathbb{R}^n \to (-1, 1)^n$ is a homeomorphism, with inverse

$$T^{-1}(z) = \big(\operatorname{arctanh}(z_1), \ldots, \operatorname{arctanh}(z_n)\big).$$

The encoded domain

$$\mathcal{E} := T(\mathcal{D})$$

is nonempty and compact, and T restricts to a homeomorphism between $\mathcal{D}$ and $\mathcal{E}$. Indeed, compactness of $\mathcal{D}$ gives some $R > 0$ such that $\mathcal{D} \subseteq [-R, R]^n$, and therefore

$$\mathcal{E} \subseteq [-\tanh(R), \tanh(R)]^n \subset (-1, 1)^n.$$

Thus the inverse is continuous on a neighborhood of $\mathcal{E}$.

Define

$$H : \mathcal{E} \to \mathbb{R}, \qquad H(z) := F\big(T^{-1}(z)\big).$$

The function H is continuous, and

$$H(T(x)) = F(x) \qquad \text{for every } x \in \mathcal{D}.$$

Hence the nonlinear input encoding loses no information relevant to the target.

Step 2: Store the encoded sequence in one recurrent layer. Take $m = n$, let $e_1, \ldots, e_n$ denote the standard basis of $\mathbb{R}^n$, and set $B = e_n$. Define the shift matrix A by

$$Ae_1 = 0, \qquad Ae_j = e_{j-1} \quad (2 \leq j \leq n).$$

For $n = 1$, these definitions give $A = 0$ and $B = 1$. All parameters are independent of time.

Starting from $h(0) = 0$, repeated substitution of the recurrence gives, for $1 \leq t \leq n$,

$$\begin{aligned} h(t; x) &= \sum_{j=1}^{t} A^{t-j} B \, \phi(x(j)) \\ &= \sum_{j=1}^{t} e_{n-t+j} \tanh(x(j)). \end{aligned} \tag{82}$$

The second equality uses $A^k e_n = e_{n-k}$ for $0 \leq k \leq n-1$. In particular,

$$h(n; x) = \sum_{j=1}^{n} e_j \tanh(x(j)) = T(x). \tag{83}$$

Thus the terminal state contains the entire encoded sequence in its original order.

Step 3: Approximate the transformed target. The activation $\tanh$ is continuous, bounded, and nonconstant. The uniform-on-compact universal approximation theorem (**?**) therefore gives an integer $d \geq 1$, vectors $u_1, \ldots, u_d \in \mathbb{R}^n$, and real coefficients $\alpha_1, \ldots, \alpha_d$, $\beta_1, \ldots, \beta_d$, and c, such that

$$\rho(z) = c + \sum_{j=1}^{d} \alpha_j \tanh(u_j^\top z + \beta_j)$$

satisfies

$$\sup_{z \in \mathcal{E}} |\rho(z) - H(z)| < \varepsilon. \tag{84}$$

The theorem applies to the compact set $\mathcal{E}$; this set need not be a box or have nonempty interior.

Step 4: Implement the prescribed output MLP. Let $U_g \in \mathbb{R}^{d\times n}$ have rows $u_j^\top$, and define

$$\alpha = (\alpha_1, \dots, \alpha_d)^\top, \qquad b_g = (\beta_1, \dots, \beta_d)^\top,$$
$$V_g = e_1\alpha^\top, \qquad c_g = ce_1.$$

Choose

$$g(z) = V_g \tanh(U_g z + b_g) + c_g = e_1\rho(z), \qquad C = e_1.$$

Then $g : \mathbb{R}^n \to \mathbb{R}^n$ is a one-hidden-layer MLP of width d, and

$$C^\top g(z) = \rho(z).$$

The input map $\phi(u) = \tanh(u)$ is a one-hidden-layer MLP of width one. To implement it with the same width d, write

$$\phi(u) = v_{\text{in}}^\top \tanh(a_{\text{in}}u + b_{\text{in}}) + c_{\text{in}},$$

where

$$a_{\text{in}} = v_{\text{in}} = (1, 0, \dots, 0)^\top \in \mathbb{R}^d, \qquad b_{\text{in}} = 0, \qquad c_{\text{in}} = 0.$$

The added units have zero output weights, so padding does not change the input map. Both nonlinear blocks therefore have the required MLP form.

Step 5: Verify the approximation error. By equation 83, the terminal prediction is

$$Y_n(x) = C^\top g(h(n; x)) = \rho(T(x)).$$

Since T maps $\mathcal{D}$ onto $\mathcal{E}$,

$$\begin{aligned}\sup_{x\in\mathcal{D}} |Y_n(x) - F(x)| &= \sup_{x\in\mathcal{D}} |\rho(T(x)) - H(T(x))| \\ &= \sup_{z\in\mathcal{E}} |\rho(z) - H(z)| \\ &< \varepsilon.\end{aligned}$$

The construction has one recurrent layer, hidden dimension $m = n$, zero initial state, and parameters independent of time. This proves the proposition. □

Memory, nonlinearity, and the scope of the result. The construction uses the same input map and shift register for all targets at a given horizon. Only the output MLP needs to be chosen for the target and approximation accuracy. Moreover, the columns of

$$\mathcal{R}_t = [A^{t-1}B \; A^{t-2}B \; \cdots \; B]$$

are distinct standard basis vectors for $1 \le t \le n$. Hence

$$\|\mathcal{R}_t\|_{\text{op}} = 1.$$

Universality therefore follows from preserving the sequence and applying an unrestricted nonlinear readout; it does not require amplification by the recurrent dynamics.

The result concerns uniform approximation on each fixed compact domain and finite horizon. It does not claim exact representation of every continuous target, one common model for all horizons, or a uniform bound on the output MLP width or parameter norms. It also does not claim that both nonlinear blocks are necessary: with $\phi(u) = u$, the same shift register stores x itself, and a universal output MLP again suffices. The proposition shows that a nonlinear input encoder can be included while retaining universality.