Approximate Memory Suffices for Universal Approximation with State Space Models

Yoshihiro Maruyama
Department of Mathematical and Information Sciences
Kyoto University
maruyama@i.h.kyoto-u.ac.jp

Abstract

Sequence models in machine learning compress long input histories into finite-dimensional states, from recurrent networks and reservoir computing to recent state space models for long-context data. This raises a basic expressivity question: how simple can the recurrent memory be before it can no longer approximate stable causal sequence-to-sequence maps? We answer this question for fading-memory operators, where the distant past has uniformly diminishing influence. We show that every continuous causal time-equivariant fading-memory operator on a compact input set can be uniformly approximated by a state space model whose recurrence is a fixed diagonal linear contraction. The dynamics are time-invariant, input-independent, and contain no nonlinear hidden-state update or delay-line structure. The construction uses a bank of short exponential traces with eigenvalues $0, \delta, 2\delta, \ldots, L\delta$, where L is the maximum lag and $\delta > 0$ is the spectral spacing: although such traces do not exactly store a recent input window, a singular Vandermonde extraction approximately reconstructs the window while the remaining tail becomes negligible as δ decreases. We also prove that exact delay recovery is impossible for any fixed finite trace bank, showing that the accuracy-dependent extraction is essential. The result separates recurrent expressivity from exact memory storage, while exposing a conditioning cost in the readout. More broadly, it clarifies why highly stable, structurally simple SSM cores can still be universal approximators for stable sequence behavior.

1 Introduction

Sequence models in machine learning compress long input histories into finite-dimensional internal states, including recurrent networks, reservoir computing systems, and recent state space models. They all use this principle to turn sequence histories into features from which predictions can be read out. The same viewpoint is natural in signal processing and dynamical systems, where many tasks are causal: the output at time t may depend on the current input and the entire past. For stable input-output behavior, however, the remote past should have diminishing influence. On compact input sets, the fading-memory framework formalizes this idea by placing a topology on one-sided histories in which recent coordinates are weighted more heavily than old ones. It has long served as a foundation for approximation results for nonlinear filters, Volterra-type operators, and discrete-time input-output maps (Volterra, 1959; Wiener, 1958; Fliess, 1981; Boyd and Chua, 1985; Sandberg, 1991, 1992; Matthews, 1993; Matthews and Moschytz, 1994; Stiles et al., 1997). In machine learning, the same setting appears in neural approximation of functionals, operators, and recurrent state-space models (Chen and Chen, 1993, 1995a,b; Schäfer and Zimmermann, 2007) and in reservoir-computing models, where a stable recurrent system embeds the input history into a finite-dimensional state before a readout is applied (Jaeger, 2001; Maass et al., 2002; Jaeger and Haas, 2004; Lukoševičius and Jaeger, 2009; Grigoryeva and Ortega, 2018a,b, 2019; Gonon and Ortega, 2020).

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

Recent long-sequence architectures have renewed interest in a particularly structured version of this paradigm: state space models with linear recurrent dynamics. In their simplest discrete-time form, they evolve according to

$$s_t = As_{t-1} + Bx_t, \qquad y_t = \Phi(s_t), \tag{1}$$

where x_t, s_t, and y_t are the input, state, and output, A and B are the state-transition and input matrices, and Φ is a readout map. Architectures based on orthogonal polynomial memories, structured linear state space layers, diagonal SSMs, selective SSMs, and related long-convolutional or linear-recurrent layers exploit this form to combine stable recurrence with efficient convolutional or recurrent computation (Voelker et al., 2019; Gu et al., 2020, 2021, 2022a; Gupta et al., 2022; Gu et al., 2022b; Smith et al., 2023; Fu et al., 2023; Orvieto et al., 2023; Poli et al., 2023; Gu and Dao, 2024; Dao and Gu, 2024).

These successes raise a basic expressivity question: how simple can the recurrent memory be before it can no longer approximate stable causal sequence-to-sequence maps? Equivalently, is recurrent complexity itself needed for universality, or can a fixed collection of stable linear history features suffice once paired with an expressive readout? Classical universality arguments often rely on rich nonlinear reservoirs, input-dependent dynamics, or the ability to store a finite delay line (i.e., a shift-register state that contains $(x_t, \dots, x_{t-L})$). It is not clear a priori whether universality can survive when the recurrent core is reduced to a fixed bank of independent, strongly stable linear traces.

This paper shows that it can. We prove that every causal time-equivariant fading-memory operator on a compact input set can be uniformly approximated by an extremely restricted class of causal sequence models: the recurrent dynamics are diagonal, linear, time-invariant, input-independent, and strongly contracting, while all task-specific nonlinear approximation is delegated to the readout. The severity of these restrictions matters: the smaller the admissible recurrent model class, the stronger and more surprising a universality theorem for that class becomes. Concretely, fixing any spectral cap $\rho \in (0, 1)$, every eigenvalue of the recurrent matrix is constrained to lie in $[0, \rho]$, so each state-coordinate block retains a fixed fraction of its previous value and adds the current input:

$$s_t^{(i)} = \lambda_i s_{t-1}^{(i)} + x_t, \qquad 0 \le \lambda_i \le \rho, \tag{2}$$

and therefore

$$s_t^{(i)} = \sum_{r=0}^{\infty} \lambda_i^r x_{t-r}. \tag{3}$$

Thus a coordinate stores a single exponentially weighted trace of the past. The whole recurrent core is therefore a bank of such blurred traces. It has no nonlinear hidden-state dynamics, no input-dependent transitions, no Jordan blocks, and no coordinate-shifting structure.

A standard route to sequence universality is to reduce the problem to finite-window approximation. On a compact input set, a continuous fading-memory functional can be uniformly approximated by a function of $(x_t, x_{t-1}, \dots, x_{t-L})$, and a finite-dimensional universal approximation theorem can then be applied to this window (Cybenko, 1989; Hornik et al., 1989; Hornik, 1991; Leshno et al., 1993; Pinkus, 1999). The obstacle is that diagonal exponential memories of the form (3) do not store such a window directly. A delay line would present the coordinates $(x_t, \dots, x_{t-L})$ explicitly; here those coordinates are hidden inside overlapping exponential averages. The construction below shows that exact storage is unnecessary. By choosing several decay rates and taking suitable linear combinations of the resulting traces, one can approximately undo the blur on the recent window while canceling the remote tail to first order. Algebraically, this cancellation is implemented by an accuracy-dependent singular Vandermonde extractor. Its role is to show that the condition required for universality in stable causal sequence tasks is approximate recoverability of the relevant finite window, not exact symbolic storage of a delay line.

The main result of this paper formalizes this point. For every causal time-equivariant fading-memory operator on a compact input set, every accuracy $\varepsilon > 0$, and every spectral cap $\rho \in (0, 1)$, there is a diagonal SSM of the form (2), with a polynomial or neural readout, whose uniform operator error is at most ε. Moreover, the eigenvalues may be chosen as

$$\lambda_i = i\delta, \qquad i = 0, 1, \dots, L, \qquad L\delta \le \rho, \tag{4}$$

with $\delta > 0$ arbitrarily small. Thus universality persists even when the recurrent spectrum is confined to an arbitrarily small neighborhood of zero.

The proof relies on a singular Vandermonde recovery mechanism. Given a maximum lag L, define exponential traces

$$h_i(t) = \sum_{r=0}^{\infty} (i\delta)^r x_{t-r}, \qquad i = 0, 1, \ldots, L. \tag{5}$$

If the far-past tail were absent, then the first $L+1$ traces would be related to the window $(x_t, \ldots, x_{t-L})$ by a finite Vandermonde matrix with nodes $0, \delta, 2\delta, \ldots, L\delta$. Inverting this matrix recovers the recent window. With the infinite tail present, the same inverse produces a residual error whose leading terms begin only at degree $L + 1$. Although the Vandermonde inverse becomes ill-conditioned as $\delta \to 0$, the tail acquires enough powers of δ that the final recovery error is $O(\delta)$ uniformly over bounded input sequences. A continuous finite-window functional can then be approximated by a polynomial or multilayer perceptron applied to the recovered window.

The contributions of this paper can be summarized as follows.

- We formulate and prove a universal approximation theorem for causal fading-memory operators using diagonal linear SSMs whose spectral radius is bounded by an arbitrary prescribed $\rho \in (0, 1)$.
- We prove a singular Vandermonde window-recovery theorem showing how strongly contracting exponential traces can approximate recent input coordinates.
- We provide a quantitative approximation result separating truncation error, Vandermonde recovery error, and readout approximation error.
- We quantify a contraction–accuracy–robustness tradeoff: for fixed L, the Vandermonde extractor has uniform recovery error $O(\delta)$ and sharp worst-case noise gain $\Theta(\delta^{-L})$, making explicit how stronger contraction tightens the recovery bound while increasing sensitivity to trace perturbations. For scalar delay recovery, we characterize the optimal tradeoff between readout sensitivity and uniform error for sufficiently small spectral spacing.
- We give a separation result: no fixed finite diagonal exponential memory bank can exactly recover a nonzero delay coordinate on all bounded scalar sequences.

The separation result explains why this accuracy-dependent extraction is not merely a reparametrized delay line. Technically, the result separates memory storage from recurrent expressivity: a strongly contracting diagonal trace bank can approximate every stable causal operator without implementing exact delays. More broadly, it clarifies why highly stable, structurally simple SSM cores can still have universal approximation power for fading-memory sequence behavior, and why spectral stability, feature conditioning, and readout capacity should be treated as separate axes.

The paper is organized as follows. Section 2 fixes the setting and SSM class. Section 3 gives the universality results, Section 4 proves the Vandermonde recovery mechanism, Section 5 gives quantitative and neural-readout consequences, and Section 6 concludes. Proofs are in the appendix.

2 Preliminaries

Sequence spaces and fading-memory topology. Let $\mathcal{X} \subset \mathbb{R}^p$ be compact and nonempty. We write

$$\mathcal{X}_- := \mathcal{X}^{\mathbb{N}_0} \tag{6}$$

for the space of one-sided pasts

$$u = (u_0, u_1, u_2, \ldots), \tag{7}$$

where u_0 denotes the current input, u_1 the previous input, and so on. We equip $\mathcal{X}_-$ with the metric

$$d_-(u, v) := \sum_{j=0}^{\infty} 2^{-(j+1)} \min\{1, \|u_j - v_j\|\}. \tag{8}$$

This metric induces the product topology on $\mathcal{X}^{\mathbb{N}_0}$. Since $\mathcal{X}$ is compact, $\mathcal{X}_-$ is compact. Continuity with respect to d_- is the compact fading-memory condition used throughout the paper: perturbations in the sufficiently remote past have uniformly small effect on the output.

A causal time-equivariant sequence operator is represented by a functional on $\mathcal{X}_-$. Specifically, a map

$$F : \mathcal{X}^{\mathbb{Z}} \to (\mathbb{R}^q)^{\mathbb{Z}} \tag{9}$$

is generated by $f : \mathcal{X}_- \to \mathbb{R}^q$ if

$$F(x)_t = f(x_t, x_{t-1}, x_{t-2}, \ldots) \qquad \text{for all } x \in \mathcal{X}^{\mathbb{Z}},\ t \in \mathbb{Z}. \tag{10}$$

Every operator of the form (10) is causal and time-equivariant. Conversely, every causal time-equivariant operator on bi-infinite sequences is determined by such a generator whenever its output at time zero depends only on the one-sided past.

Definition 1 (Fading-memory operator)**.** *A causal time-equivariant operator F has fading memory if its generator $f : \mathcal{X}_- \to \mathbb{R}^q$ is continuous with respect to d_-.*

The approximation norm is the uniform sequence-operator norm

$$\|F - G\|_\infty := \sup_{x \in \mathcal{X}^{\mathbb{Z}}} \sup_{t \in \mathbb{Z}} \|F(x)_t - G(x)_t\|. \tag{11}$$

Finite-horizon reduction. The following lemma uses compactness and continuity to obtain uniform finite-window approximation.

Lemma 1 (Uniform finite-horizon approximation)**.** *Let $f : \mathcal{X}_- \to \mathbb{R}^q$ be continuous. Then for every $\varepsilon > 0$ there exist an integer $L \geq 0$ and a continuous function*

$$g_L : \mathcal{X}^{L+1} \to \mathbb{R}^q \tag{12}$$

such that

$$\sup_{u \in \mathcal{X}_-} \|f(u) - g_L(u_0, u_1, \ldots, u_L)\| < \varepsilon. \tag{13}$$

It is useful to quantify the preceding reduction. Define the truncation modulus

$$\omega_f(L) := \sup\big\{\|f(u) - f(v)\| : u, v \in \mathcal{X}_-,\ u_j = v_j \text{ for } 0 \leq j \leq L\big\}. \tag{14}$$

For continuous f on compact $\mathcal{X}_-$, we have $\omega_f(L) \to 0$ as $L \to \infty$: if two pasts agree on coordinates $0, \ldots, L$, then their d_--distance is at most $\sum_{j>L} 2^{-(j+1)} = 2^{-(L+1)}$, and uniform continuity of f gives the claim.

Strongly contracting diagonal SSMs. Fix $\rho \in (0, 1)$. A ρ-contracting diagonal SSM with polynomial readout is a model of the form

$$s_t^{(i)} = \lambda_i s_{t-1}^{(i)} + x_t, \qquad i = 0, 1, \ldots, m, \tag{15}$$

where each $s_t^{(i)} \in \mathbb{R}^p$ and

$$0 \leq \lambda_i \leq \rho, \tag{16}$$

together with an output equation

$$y_t = P(s_t^{(0)}, s_t^{(1)}, \ldots, s_t^{(m)}), \tag{17}$$

where $P : \mathbb{R}^{p(m+1)} \to \mathbb{R}^q$ is a polynomial map. Since $\rho < 1$ and $\mathcal{X}$ is bounded, the unique causal bounded solution of (15) on bi-infinite inputs is

$$s_t^{(i)} = \sum_{r=0}^{\infty} \lambda_i^r x_{t-r}. \tag{18}$$

When $\lambda_i = 0$, we use the convention $0^0 = 1$ and $0^r = 0$ for $r \geq 1$, so that $s_t^{(i)} = x_t$.

We denote by $\mathcal{S}_\rho^{\mathrm{poly}}$ the class of sequence operators realized by (15)–(17) with finitely many coordinates and eigenvalues in $[0, \rho]$. The central question is whether $\mathcal{S}_\rho^{\mathrm{poly}}$ is dense in the space of causal fading-memory operators under (11) even when every recurrent coordinate is just a strongly contracting exponential trace.

3 Universality results

Universality and window recovery. Our first result states that strongly contracting diagonal SSMs are universal approximators of causal fading-memory operators. The eigenvalues are not merely bounded away from one; they may be forced into an arbitrarily small interval near zero.

Theorem 1 (Universality without delay lines)**.** *Let $\mathcal{X} \subset \mathbb{R}^p$ be compact, and let*

$$F : \mathcal{X}^{\mathbb{Z}} \to (\mathbb{R}^q)^{\mathbb{Z}} \tag{19}$$

be a causal time-equivariant operator generated by a continuous functional $f : \mathcal{X}_- \to \mathbb{R}^q$. Then for every $\varepsilon > 0$ and every $\rho \in (0, 1)$, there exist an integer $L \geq 1$, a scalar $\delta > 0$, and a polynomial

$$P : \mathbb{R}^{p(L+1)} \to \mathbb{R}^q \tag{20}$$

such that, with

$$\lambda_i = i\delta, \qquad i = 0, 1, \ldots, L, \qquad L\delta \leq \rho, \tag{21}$$

the diagonal SSM

$$s_t^{(i)} = \lambda_i s_{t-1}^{(i)} + x_t, \qquad i = 0, 1, \ldots, L, \tag{22}$$

$$y_t = P(s_t^{(0)}, s_t^{(1)}, \ldots, s_t^{(L)}) \tag{23}$$

satisfies

$$\sup_{x \in \mathcal{X}^{\mathbb{Z}}} \sup_{t \in \mathbb{Z}} \|F(x)_t - y_t\| < \varepsilon. \tag{24}$$

Consequently, for every $\rho \in (0, 1)$, the class $\mathcal{S}_\rho^{\mathrm{poly}}$ is dense in the space of causal time-equivariant fading-memory operators on $\mathcal{X}$.

For a fixed horizon L, choose eigenvalues $\lambda_i = i\delta$ and define

$$H_\delta(u) := \big(h_0(u), h_1(u), \ldots, h_L(u)\big), \qquad h_i(u) = \sum_{r=0}^{\infty} (i\delta)^r u_r. \tag{25}$$

Let

$$V_\delta = \big((i\delta)^j\big)_{0 \leq i,j \leq L} \tag{26}$$

be the Vandermonde matrix associated with nodes $0, \delta, 2\delta, \ldots, L\delta$. Its inverse defines linear extraction maps

$$E_j^\delta(h_0, \ldots, h_L) := \sum_{i=0}^{L} (V_\delta^{-1})_{ji} h_i, \qquad j = 0, 1, \ldots, L. \tag{27}$$

These maps invert the first $L + 1$ Taylor-like coefficients of the exponential traces. The following theorem is the quantitative core of the paper.

Theorem 2 (Singular Vandermonde window recovery)**.** *Let $L \geq 1$ and let*

$$M := \sup_{x \in \mathcal{X}} \|x\| < \infty. \tag{28}$$

There exists a constant $C_L > 0$, depending only on L, such that whenever $\delta > 0$ satisfies $L\delta \leq 1/2$,

$$\max_{0 \leq j \leq L} \sup_{u \in \mathcal{X}_-} \big\| E_j^\delta(H_\delta(u)) - u_j \big\| \leq C_L M \delta. \tag{29}$$

One admissible choice is

$$C_L = 2\Gamma_L L^{L+1}, \qquad \Gamma_L := \max_{0 \leq j \leq L} \sum_{i=0}^{L} |(W^{-1})_{ji}|, \qquad W = (i^j)_{0 \leq i,j \leq L}. \tag{30}$$

Here and below, the Vandermonde convention is that the first column is identically one, so 0^0 is read as 1.

The key point is that $V_\delta = WD_\delta$, where $D_\delta = \mathrm{diag}(1, \delta, \ldots, \delta^L)$. Thus V_δ approaches a singular matrix as $\delta \to 0$, while the norm of $V_\delta^{-1} = D_\delta^{-1} W^{-1}$ diverges. Nevertheless, after the first $L+1$ window coordinates are inverted exactly, all residual terms come from powers $r \geq L+1$, leaving at least one power of δ after the worst inverse scaling. This yields the $O(\delta)$ bound in (29).

Combining finite-horizon truncation, Theorem 2, and polynomial approximation gives the following quantitative version of Theorem 1.

The anchor a fills the history beyond lag L with a fixed input, producing the finite-window map g_L. The recovered window may lie slightly outside $\mathcal{X}^{L+1}$, so we extend g_L continuously to the surrounding neighborhood K_L^+. The modulus of continuity $\omega_{\widetilde{g}_L}$ then converts a bound on the window reconstruction error into a bound on the resulting output error.

Theorem 3 (Quantitative approximation decomposition)**.** *Fix $L \geq 1$, $\rho \in (0,1)$, and an anchor $a \in \mathcal{X}$. Let $M := \sup_{x \in \mathcal{X}} \|x\|$ and let C_L be a recovery constant from Theorem 2. Define*

$$g_L(z_0, \ldots, z_L) := f(z_0, \ldots, z_L, a, a, a, \ldots). \tag{31}$$

Let $\widetilde{g}_L$ be a continuous extension of g_L to the compact one-neighborhood, measured in the maximum block norm,

$$K_L^+ := \big\{ z = (z_0, \ldots, z_L) \in (\mathbb{R}^p)^{L+1} : \inf_{w \in \mathcal{X}^{L+1}} \max_{0 \leq j \leq L} \|z_j - w_j\| \leq 1 \big\}, \tag{32}$$

and let $\omega_{\widetilde{g}_L}$ be its modulus of continuity with respect to the same norm. Suppose $\pi_L : \mathbb{R}^{p(L+1)} \to \mathbb{R}^q$ is a polynomial satisfying

$$\sup_{z \in K_L^+} \|\pi_L(z) - \widetilde{g}_L(z)\| \leq \alpha. \tag{33}$$

If $\delta > 0$ is small enough that $L\delta \leq \min\{\rho, 1/2\}$ and $C_L M \delta \leq 1$, then the diagonal SSM with eigenvalues $\lambda_i = i\delta$ and readout

$$P(h) := \pi_L(E_0^\delta(h), \ldots, E_L^\delta(h)), \qquad h = (h_0, \ldots, h_L) \tag{34}$$

realizes a sequence operator $\widehat{F}$ satisfying

$$\left\| F - \widehat{F} \right\|_\infty \leq \omega_f(L) + \omega_{\widetilde{g}_L}(C_L M \delta) + \alpha. \tag{35}$$

Readout variants. The polynomial readout in Theorem 1 can be replaced by a standard neural network readout.

Corollary 1 (Neural readouts)**.** *Let σ be an activation for which feedforward neural networks are dense in $C(K)$ on compact subsets of Euclidean space. Then the polynomial readout in Theorem 1 may be replaced by a multilayer perceptron with activation σ.*

Corollary 2 (Selective SSM classes)**.** *Let $\mathcal{A}$ be a class of causal SSM sequence operators with neural readouts. Suppose that, for every $\rho \in (0,1)$, $\mathcal{A}$ contains all diagonal models of the form*

$$s_t^{(i)} = \lambda_i s_{t-1}^{(i)} + x_t, \qquad 0 \leq \lambda_i \leq \rho, \tag{36}$$

with finitely many coordinates and arbitrary multilayer-perceptron readouts. Then $\mathcal{A}$ is dense in the space of causal time-equivariant fading-memory operators on $\mathcal{X}$.

Separation from exact delays. We also give the separation result that distinguishes the construction from exact window storage.

Theorem 4 (No exact delay recovery from fixed diagonal traces)**.** *Let $\lambda_1, \ldots, \lambda_n \in (-1, 1)$ be fixed, and define scalar traces*

$$h_i(u) = \sum_{r=0}^{\infty} \lambda_i^r u_r, \qquad u \in [-1,1]^{\mathbb{N}_0}, \qquad i = 1, \ldots, n. \tag{37}$$

For every delay $j \geq 1$, there is no function $\Phi : \mathbb{R}^n \to \mathbb{R}$ such that

$$\Phi(h_1(u), \ldots, h_n(u)) = u_j \qquad \textit{for all } u \in [-1,1]^{\mathbb{N}_0}. \tag{38}$$

The rest of the main text develops the constructive side of the theorem. Section 4 gives the coefficient-level Vandermonde mechanism used in Theorem 2, while Section 5 provides the quantitative and neural-readout consequences.

4 Singular Vandermonde recovery

This section develops the ingredients behind Theorem 2. Throughout the section we fix an integer $L \geq 1$ and a spacing $\delta > 0$ satisfying $L\delta \leq 1/2$. We also write $M := \sup_{x \in \mathcal{X}} \|x\|$.

For $u \in \mathcal{X}_-$, recall that

$$h_i(u) = \sum_{r=0}^{\infty} (i\delta)^r u_r, \qquad i = 0, 1, \ldots, L. \tag{39}$$

The series converges absolutely because $\|u_r\| \leq M$ and $i\delta \leq L\delta \leq 1/2$. Let $V_\delta = \big((i\delta)^j\big)_{0 \leq i,j \leq L}$. The extractor E_j^δ is the jth row of V_δ^{-1} applied to the trace vector $H_\delta(u)$.

Lemma 2 (Coefficient cancellation)**.** *For $0 \leq j \leq L$ and $r \geq 0$, define*

$$c_{jr}^\delta := \sum_{i=0}^{L} (V_\delta^{-1})_{ji} (i\delta)^r. \tag{40}$$

Then $c_{jr}^\delta = \mathbf{1}_{\{j=r\}}$ for $0 \leq r \leq L$. Consequently,

$$E_j^\delta(H_\delta(u)) = u_j + \sum_{r=L+1}^{\infty} c_{jr}^\delta u_r. \tag{41}$$

To bound the tail coefficients, write $W = (i^j)_{0 \leq i,j \leq L}$ and $D_\delta = \mathrm{diag}(1, \delta, \delta^2, \ldots, \delta^L)$. Then $V_\delta = W D_\delta$, so $V_\delta^{-1} = D_\delta^{-1} W^{-1}$. Thus the jth row of V_δ^{-1} scales like δ^{-j}. The crucial point is that the residual terms in (41) start only at degree $L + 1$.

Lemma 3 (Tail coefficient bound)**.** *Let*

$$\Gamma_L := \max_{0 \leq j \leq L} \sum_{i=0}^{L} |(W^{-1})_{ji}|. \tag{42}$$

Then, for every $0 \leq j \leq L$,

$$\sum_{r=L+1}^{\infty} |c_{jr}^\delta| \leq 2\Gamma_L L^{L+1} \delta. \tag{43}$$

The inverse V_δ^{-1} becomes badly conditioned as $\delta \to 0$; indeed, the jth row carries the factor δ^{-j}. However, the Vandermonde identities cancel all powers $0, 1, \ldots, L$. The remaining terms have degree at least $L + 1$, so the worst product has the form $\delta^{-L}\delta^{L+1} = \delta$. This is the mechanism by which a bank of very short exponential memories approximates a finite delay window without using a delay line.

5 Quantitative consequences and neural readouts

We now clarify the quantitative content of the construction. The proof of Theorem 1 is explicit enough to expose the relevant approximation parameters, and this section clarifies the consequences that are useful for neural sequence models. Throughout the section, set $M := \sup_{x \in \mathcal{X}} \|x\|$ and use the constants C_L and Γ_L from Theorem 2. There are three independent choices: the truncation horizon L, the spectral spacing δ, and the finite-dimensional readout accuracy. The horizon controls the fading-memory truncation error, the spacing controls the deterministic Vandermonde recovery error, and the readout class controls the static approximation error.

A constructive parameter schedule. Fix a target error $\varepsilon > 0$ and spectral cap $\rho \in (0, 1)$. The construction gives the following explicit schedule. First choose L so that $\omega_f(L) < \varepsilon/3$. Then choose a continuous finite-window representative

$$g_L(z_0, \ldots, z_L) = f(z_0, \ldots, z_L, a, a, a, \ldots) \tag{44}$$

and an extension $\widetilde{g}_L$ to K_L^+. Since $\widetilde{g}_L$ is uniformly continuous, choose $\eta \in (0,1)$ such that $\omega_{\widetilde{g}_L}(\eta) < \varepsilon/3$. If $M > 0$, finally choose

$$\delta < \min\left\{\frac{\rho}{L}, \frac{1}{2L}, \frac{\eta}{C_L M}\right\}, \tag{45}$$

while if $M = 0$, choose any $\delta < \min\{\rho/L, 1/(2L)\}$. Then choose a polynomial or neural readout with static error less than $\varepsilon/3$ on K_L^+. Theorem 3 then gives

$$\left\|F - \widehat{F}\right\|_\infty < \varepsilon. \tag{46}$$

The state dimension of this construction is $p(L+1)$, and the recurrent spectral radius is at most $L\delta$. In particular, after L is fixed, the spectral radius may be pushed below any prescribed number without losing universality; one only pays through the conditioning of the readout.

The same schedule also shows that the spectral cap ρ imposes no approximation-theoretic obstruction. For a smaller cap, one decreases δ rather than increasing the recurrent complexity. This is the precise sense in which the theorem proves universality by pinching the spectrum instead of by creating a long-memory recurrent core.

Structured neural implementation. The diagonal SSM used in the theorem can be written in block form as

$$s_t = A_\delta s_{t-1} + B x_t, \qquad A_\delta = \operatorname{diag}(0, \delta, 2\delta, \ldots, L\delta) \otimes I_p, \qquad B = \mathbf{1}_{L+1} \otimes I_p, \tag{47}$$

where $\mathbf{1}_{L+1}$ is the column vector of ones and $s_t \in (\mathbb{R}^p)^{L+1} \simeq \mathbb{R}^{p(L+1)}$. The echo-state solution is $s_t = H_\delta(x_t, x_{t-1}, \ldots)$. The construction naturally factors the readout into a linear Vandermonde extractor followed by an ordinary finite-dimensional approximator:

$$y_t = \varphi_\theta\big((V_\delta^{-1} \otimes I_p) s_t\big). \tag{48}$$

Here $(V_\delta^{-1} \otimes I_p)s_t$ is an approximate reconstruction of $(x_t, \ldots, x_{t-L})$, and φ_θ approximates the extension $\widetilde{g}_L$ on K_L^+.

Equivalently, one may absorb the linear extractor into the first layer of a multilayer perceptron and write simply $y_t = N_\theta(s_t)$. The structured form (48) is useful conceptually because it separates two roles that are often conflated in neural SSMs: the recurrent core builds a fixed set of exponential traces, while the readout converts those traces into a local window and then applies the desired nonlinear finite-window map.

This factorization also gives a natural initialization for trainable models. One can initialize A_δ with contracted eigenvalues, initialize the first readout layer as $V_\delta^{-1} \otimes I_p$, and train only the remaining finite-dimensional approximator.

Conditioning and robustness of the extractor. The recovery estimate in Theorem 2 improves as $\delta \to 0$. The price is that the Vandermonde inverse becomes large. The following bound makes this tradeoff explicit.

Theorem 5 (Noise amplification by the Vandermonde extractor)**.** *Let $L \geq 1$, let $0 < \delta \leq 1$, and let $E^\delta = (E_0^\delta, \ldots, E_L^\delta)$ be the extractor defined in* (27)*. If $h, \widetilde{h} \in (\mathbb{R}^p)^{L+1}$ satisfy*

$$\max_{0 \leq i \leq L} \left\|\widetilde{h}_i - h_i\right\| \leq \sigma, \tag{49}$$

then

$$\max_{0 \leq j \leq L} \left\|E_j^\delta(\widetilde{h}) - E_j^\delta(h)\right\| \leq \Gamma_L \delta^{-L} \sigma, \tag{50}$$

where Γ_L is defined in (30)*.*

Remark 1 (Sharpness of noise amplification)**.** *Under the hypotheses of Theorem 5,*

$$\begin{aligned} (V_\delta^{-1})_{Li} &= \frac{(-1)^{L-i}}{\delta^L i!(L-i)!}, \qquad 0 \leq i \leq L, \\ \frac{2^L}{L!}\delta^{-L} &\leq \left\|V_\delta^{-1}\right\|_{\infty \to \infty} \leq \Gamma_L \delta^{-L}, \end{aligned} \tag{51}$$

where the matrix norm is the maximum absolute row sum. The entry formula is the leading coefficient of the ith Lagrange basis polynomial. Perturbations $\widetilde{h}_i - h_i = \sigma(-1)^{L-i}v$, with $\|v\| = 1$, attain the last row's absolute sum, proving sharpness of the exponent δ^{-L} for this extractor.

Together with Theorem 2, this gives deterministic recovery error $O(\delta)$ and worst-case noise gain $\Theta(\delta^{-L})$ for fixed L.

Optimal sensitivity for delay recovery. For scalar delay recovery, the same construction yields an exact characterization over all Lipschitz readouts, including nonlinear ones. This identifies both the error floor of a fixed trace bank and the sensitivity required to approach it.

Theorem 6 (Exact sensitivity–accuracy tradeoff)**.** *Let $L \geq 1$, $M > 0$, and $0 < \delta < 1/[L(L+1)]$. Specialize to $\mathcal{X} = [-M, M]$, with the shared-input trace map H_δ from* (25)*, and define*

$$a_\delta = \frac{2^L}{L!\delta^L}, \qquad \beta_\delta = M\left(\prod_{i=1}^{L}(1-i\delta)^{-1} - 1\right). \tag{52}$$

Then $0 < \beta_\delta < M$. For $K \geq 0$, let

$$\mathcal{R}_\delta(K) := \inf_{\substack{\Phi:\mathbb{R}^{L+1}\to\mathbb{R} \\ \mathrm{Lip}_\infty(\Phi)\leq K}} \sup_{u\in[-M,M]^{\mathbb{N}_0}} |\Phi(H_\delta(u)) - u_L|, \tag{53}$$

where Lip_∞ is the Lipschitz constant in the maximum norm on the raw trace coordinates. The optimal error is

$$\mathcal{R}_\delta(K) = \max\left\{\beta_\delta, M - \frac{K}{a_\delta}(M - \beta_\delta)\right\}, \tag{54}$$

attained by $\Phi = \min\{1, K/a_\delta\}E_L^\delta$. Equivalently, for $0 \leq \varepsilon < M$, a readout with Lipschitz constant at most K and uniform error at most ε exists if and only if

$$\varepsilon \geq \beta_\delta, \qquad K \geq a_\delta \frac{M - \varepsilon}{M - \beta_\delta}. \tag{55}$$

For fixed L and $\varepsilon \in (0, M)$, the minimum sensitivity therefore scales as δ^{-L} as $\delta \to 0$. The theorem makes this cost intrinsic to accurate recovery of the scalar delay from the specified trace bank. The proof is in Appendix A; Appendix C extends the characterization to bounded trace noise.

Propagation to operator approximation. For the general compact input setting, Theorem 5 yields a noisy version of the quantitative decomposition. Suppose that instead of observing the exact trace vector $H_\delta(u)$, the readout receives a perturbed trace vector $\widetilde{H}_\delta(u)$ satisfying

$$\max_{0\leq i\leq L}\left\|\widetilde{H}_{\delta,i}(u) - H_{\delta,i}(u)\right\| \leq \sigma \qquad \text{for all } u \in \mathcal{X}_-. \tag{56}$$

If $L\delta \leq \min\{\rho, 1/2\}$ and

$$C_L M\delta + \Gamma_L \delta^{-L}\sigma \leq 1, \tag{57}$$

then the readout from Theorem 3, applied to $\widetilde{H}_\delta$, defines a sequence operator $\widehat{F}_\sigma$ satisfying

$$\left\|F - \widehat{F}_\sigma\right\|_\infty \leq \omega_f(L) + \omega_{\widetilde{g}_L}\left(C_L M\delta + \Gamma_L\delta^{-L}\sigma\right) + \alpha. \tag{58}$$

Indeed, the recovered window differs from the true window by at most the sum of the deterministic tail error $C_L M\delta$ and the amplified trace perturbation $\Gamma_L\delta^{-L}\sigma$.

If $\widetilde{g}_L$ is G_L-Lipschitz on K_L^+, then the middle term in (58) is bounded by

$$G_L C_L M\delta + G_L\Gamma_L\delta^{-L}\sigma. \tag{59}$$

Ignoring the spectral cap and constants independent of δ, and assuming $M > 0$ and $\sigma > 0$, the balance between deterministic recovery and noise amplification is obtained at

$$\delta_* = \left(\frac{L\Gamma_L\sigma}{C_L M}\right)^{1/(L+1)}. \tag{60}$$

Under the constraint $L\delta \leq \min\{\rho, 1/2\}$, the displayed Lipschitz upper bound is minimized at $\widehat{\delta} = \min\{\delta_*, \rho/L, 1/(2L)\}$. This choice is admissible provided $C_L M\widehat{\delta} + \Gamma_L\widehat{\delta}^{-L}\sigma \leq 1$.

Spectral geometry and conditioning. The preceding tradeoff also depends on the arrangement of spectral nodes. To compare at a fixed radius $0 < r < 1$, set $N = L + 1$ and let $\zeta = e^{2\pi \mathrm{i}/N}$. The circular-spectrum traces $h_k = \sum_{m \geq 0} (r\zeta^k)^m u_m$, $0 \leq k < N$, admit the Fourier decoder

$$D_j h = \frac{r^{-j}}{N} \sum_{k=0}^{N-1} \zeta^{-kj} h_k, \qquad 0 \leq j \leq L. \tag{61}$$

Writing $D = (D_0, \ldots, D_L)$, Fourier orthogonality gives the uniform recovery bound and exact perturbation gain

$$\max_{0 \leq j \leq L} \sup_{u \in \mathcal{X}_-} \|D_j h(u) - u_j\| \leq \frac{M r^{L+1}}{1 - r^{L+1}}, \qquad \|D\|_{\infty \to \infty} = r^{-L}, \tag{62}$$

where the operator norm uses the maximum Euclidean norm across complex trace and output blocks. At the same radius $r = L\delta$, (51) gives the nonnegative equally spaced grid a noise gain of at least $(2^L L^L / L!) r^{-L}$. Thus allowing circular spectra improves conditioning at fixed spectral radius and state dimension. For real inputs, conjugate trace pairs are represented by real two-dimensional rotation–contraction blocks, giving $p(L + 1)$ real state coordinates in total. Theorem 1 establishes universality within the more restricted class of nonnegative scalar diagonal dynamics. Remark 2 in Appendix B gives the recovery identity and the derivation of these bounds.

The nonnegative-grid construction provides a canonical, quantitative realization of diagonal reservoir universality. Grigoryeva and Ortega (2018b, Corollary 11(i)) establish universality for diagonal linear reservoirs with polynomial readouts. Li et al. (2024) and Fong et al. (2025) establish universality for cycle reservoirs with equal-modulus spectra and binary or complex-binary input weights. Theorem 1 establishes universality for the shared-input family $A = \mathrm{diag}(0, \delta, \ldots, L\delta) \otimes I_p$, $B = \mathbf{1}_{L+1} \otimes I_p$, under every prescribed spectral cap $\rho \in (0, 1)$ with $L\delta \leq \rho$. For each fixed window horizon L, the single spacing parameter δ determines the entire recurrent core. The singular Vandermonde construction provides an explicit window decoder with uniform recovery error $O(\delta)$ and sharp extractor noise gain $\Theta(\delta^{-L})$, quantifying the relationship between spectral contraction, approximation accuracy, and robustness.

6 Conclusion

We proved that diagonal linear SSMs with a fixed, strongly contracting recurrent core are universal approximators for causal time-equivariant fading-memory operators on compact input sets. This is a deliberately severe model class: the recurrence is time-invariant, input-independent, diagonal, and can have all eigenvalues confined to an arbitrarily small interval near zero. Universality therefore does not require nonlinear hidden-state dynamics, input-dependent transitions, or a recurrent mechanism that explicitly preserves a long delay line.

The significance of the result is that it separates two notions that are often treated as inseparable in sequence modeling: memory storage and recurrent expressivity. A strongly contracting exponential trace bank rapidly forgets the past in each coordinate and cannot exactly recover even a single nonzero delay coordinate when the bank is fixed. Nevertheless, when the target operator has fading memory, only a finite recent window must be recovered to a prescribed accuracy. The singular Vandermonde construction shows that such approximate window recovery is possible from sufficiently short traces, with the remote tail canceled to first order. Thus exact symbolic storage of the input history is not the right benchmark for universality in stable causal sequence tasks.

This perspective clarifies why very simple and highly stable SSM cores can still support rich sequence-to-sequence behavior once paired with an expressive readout. The recurrent state need only provide a structured collection of history features; the nonlinear readout can then perform the task-specific finite-dimensional approximation.

Overall, this work gives an approximation-theoretic explanation for the power of stable diagonal SSMs. It shows that recurrent complexity is not necessary for universality over fading-memory operators; what is necessary is a readout capable of extracting and recombining the finite-window information that stable causal tasks actually use. This suggests that future analyses of SSM architectures should treat spectral stability, feature conditioning, and readout capacity as separate axes rather than attributing expressivity solely to the form of the recurrent dynamics.

Acknowledgments and Disclosure of Funding

This work was supported by JST and JSPS. There are no competing financial or commercial interests; this is theoretical work on universal approximation of fading-memory operators by state space models.

References

Stephen Boyd and Leon O. Chua. Fading memory and the problem of approximating nonlinear operators with Volterra series. *IEEE Transactions on Circuits and Systems*, 32(11):1150–1161, 1985.

Tianping Chen and Hong Chen. Approximations of continuous functionals by neural networks with application to dynamic systems. *IEEE Transactions on Neural Networks*, 4(6):910–918, 1993.

Tianping Chen and Hong Chen. Approximation capability to functions of several variables, nonlinear functionals, and operators by radial basis function neural networks. *IEEE Transactions on Neural Networks*, 6(4):904–910, 1995.

Tianping Chen and Hong Chen. Universal approximation to nonlinear operators by neural networks with arbitrary activation functions and its application to dynamical systems. *IEEE Transactions on Neural Networks*, 6(4):911–917, 1995.

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

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *International Conference on Machine Learning*, pages 10041–10071, 2024.

Michel Fliess. Fonctionnelles causales non linéaires et indéterminées non commutatives. *Bulletin de la Société Mathématique de France*, 109:3–40, 1981.

Robert Simon Fong, Boyu Li, and Peter Tiňo. Universality of real minimal complexity reservoir. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(16):16622–16629, 2025.

Daniel Y. Fu, Tri Dao, Khaled Saab, Armin W. Thomas, Atri Rudra, and Christopher Ré. Hungry hungry hippos: Towards language modeling with state space models. In *International Conference on Learning Representations*, 2023.

Lukas Gonon and Juan-Pablo Ortega. Reservoir computing universality with stochastic inputs. *IEEE Transactions on Neural Networks and Learning Systems*, 31(1):100–112, 2020.

Lyudmila Grigoryeva and Juan-Pablo Ortega. Echo state networks are universal. *Neural Networks*, 108:495–508, 2018.

Lyudmila Grigoryeva and Juan-Pablo Ortega. Universal discrete-time reservoir computers with stochastic inputs and linear readouts using non-homogeneous state-affine systems. *Journal of Machine Learning Research*, 19(24):1–40, 2018.

Lyudmila Grigoryeva and Juan-Pablo Ortega. Differentiable reservoir computing. *Journal of Machine Learning Research*, 20(179):1–62, 2019.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *First Conference on Language Modeling*, 2024.

Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Ré. HiPPO: Recurrent memory with optimal polynomial projections. In *Advances in Neural Information Processing Systems*, pages 1474–1487, 2020.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *International Conference on Learning Representations*, 2022.

Albert Gu, Ankit Gupta, Karan Goel, and Christopher Ré. On the parameterization and initialization of diagonal state space models. In *Advances in Neural Information Processing Systems*, pages 35971–35983, 2022.

Albert Gu, Isys Johnson, Karan Goel, Khaled Saab, Tri Dao, Atri Rudra, and Christopher Ré. Combining recurrent, convolutional, and continuous-time models with linear state space layers. In *Advances in Neural Information Processing Systems*, pages 572–585, 2021.

Ankit Gupta, Albert Gu, and Jonathan Berant. Diagonal state spaces are as effective as structured state spaces. In *Advances in Neural Information Processing Systems*, pages 22982–22994, 2022.

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

Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural Networks*, 2(5):359–366, 1989.

Herbert Jaeger. The echo state approach to analysing and training recurrent neural networks. GMD Report 148, 2001.

Herbert Jaeger and Harald Haas. Harnessing nonlinearity: Predicting chaotic systems and saving energy in wireless communication. *Science*, 304(5667):78–80, 2004.

Moshe Leshno, Vladimir Ya. Lin, Allan Pinkus, and Shimon Schocken. Multilayer feedforward networks with a nonpolynomial activation function can approximate any function. *Neural Networks*, 6(6):861–867, 1993.

Boyu Li, Robert Simon Fong, and Peter Tiňo. Simple cycle reservoirs are universal. *Journal of Machine Learning Research*, 25(158):1–28, 2024.

Mantas Lukoševičius and Herbert Jaeger. Reservoir computing approaches to recurrent neural network training. *Computer Science Review*, 3(3):127–149, 2009.

Wolfgang Maass, Thomas Natschläger, and Henry Markram. Real-time computing without stable states: A new framework for neural computation based on perturbations. *Neural Computation*, 14(11):2531–2560, 2002.

Michael B. Matthews. Approximating nonlinear fading-memory operators using neural network models. *Circuits, Systems, and Signal Processing*, 12(2):279–307, 1993.

Michael B. Matthews and George S. Moschytz. The identification of nonlinear discrete-time fading-memory systems using neural network models. *IEEE Transactions on Circuits and Systems II: Analog and Digital Signal Processing*, 41(11):740–751, 1994.

Antonio Orvieto, Samuel L. Smith, Albert Gu, Anushan Fernando, Caglar Gulcehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences. In *International Conference on Machine Learning*, pages 26670–26698, 2023.

Allan Pinkus. Approximation theory of the MLP model in neural networks. *Acta Numerica*, 8:143–195, 1999.

Michael Poli, Stefano Massaroli, Eric Nguyen, Daniel Y. Fu, Tri Dao, Stephen Baccus, Yoshua Bengio, Stefano Ermon, and Christopher Ré. Hyena hierarchy: Towards larger convolutional language models. In *International Conference on Machine Learning*, pages 28043–28078, 2023.

Irwin W. Sandberg. Approximation theorems for discrete-time systems. *IEEE Transactions on Circuits and Systems*, 38(5):564–566, 1991.

Irwin W. Sandberg. Approximately-finite memory and input-output maps. *IEEE Transactions on Circuits and Systems I: Fundamental Theory and Applications*, 39(7):549–556, 1992.

Anton Maximilian Schäfer and Hans-Georg Zimmermann. Recurrent neural networks are universal approximators. *International Journal of Neural Systems*, 17(4):253–263, 2007.

Jimmy T. H. Smith, Andrew Warrington, and Scott W. Linderman. Simplified state space layers for sequence modeling. In *International Conference on Learning Representations*, 2023.

Bryan W. Stiles, Irwin W. Sandberg, and Joydeep Ghosh. Complete memory structures for approximating nonlinear discrete-time mappings. *IEEE Transactions on Neural Networks*, 8(6):1397–1409, 1997.

Marshall H. Stone. The generalized Weierstrass approximation theorem. *Mathematics Magazine*, 21(4):167–184 and 21(5):237–254, 1948.

Aaron R. Voelker, Ivana Kajić, and Chris Eliasmith. Legendre memory units: Continuous-time representation in recurrent neural networks. In *Advances in Neural Information Processing Systems*, pages 15544–15553, 2019.

Vito Volterra. *Theory of Functionals and of Integral and Integro-Differential Equations*. Dover Publications, 1959.

Norbert Wiener. *Nonlinear Problems in Random Theory*. MIT Press, 1958.

A Proofs

Here we present proofs omitted above, including those for the universality theorem and its corollaries and for the separation and noise-amplification results.

Proof of Lemma 1. Since $\mathcal{X}_-$ is compact and f is continuous, f is uniformly continuous. Choose $\eta > 0$ such that $d_-(u, v) < \eta$ implies

$$\|f(u) - f(v)\| < \varepsilon. \tag{63}$$

Choose L with $2^{-(L+1)} < \eta$. If two pasts agree in coordinates $0, 1, \ldots, L$, then their d_--distance is at most $2^{-(L+1)}$, so their f-values differ by less than ε. Fix an anchor point $a \in \mathcal{X}$ and define

$$g_L(z_0, \ldots, z_L) := f(z_0, \ldots, z_L, a, a, a, \ldots). \tag{64}$$

The embedding $(z_0, \ldots, z_L) \mapsto (z_0, \ldots, z_L, a, a, \ldots)$ is continuous, hence g_L is continuous. Applying the preceding estimate to u and $(u_0, \ldots, u_L, a, a, \ldots)$ gives (13). □

Proof of Lemma 2. The rth column of V_δ is the vector $((i\delta)^r)_{i=0}^L$. Therefore, for $0 \le r \le L$,

$$\left(V_\delta^{-1} V_\delta\right)_{jr} = \sum_{i=0}^{L} (V_\delta^{-1})_{ji} (i\delta)^r = c_{jr}^\delta. \tag{65}$$

Since $V_\delta^{-1} V_\delta = I$, this proves the stated coefficient cancellation. Using absolute convergence, we may interchange the finite sum over i and the infinite sum over r:

$$E_j^\delta(H_\delta(u)) = \sum_{i=0}^{L} (V_\delta^{-1})_{ji} \sum_{r=0}^{\infty} (i\delta)^r u_r \tag{66}$$

$$= \sum_{r=0}^{\infty} \left(\sum_{i=0}^{L} (V_\delta^{-1})_{ji} (i\delta)^r \right) u_r = \sum_{r=0}^{\infty} c_{jr}^\delta u_r. \tag{67}$$

Splitting the sum into $0 \le r \le L$ and $r \ge L + 1$ gives (41). □

Proof of Lemma 3. By the factorization $V_\delta^{-1} = D_\delta^{-1} W^{-1}$,

$$(V_\delta^{-1})_{ji} = \delta^{-j} (W^{-1})_{ji}. \tag{68}$$

For $r \ge L + 1$, using $0 \le i \le L$ gives

$$|c_{jr}^\delta| \le \delta^{-j} \sum_{i=0}^{L} |(W^{-1})_{ji}| (i\delta)^r \tag{69}$$

$$\le \Gamma_L \delta^{-j} (L\delta)^r = \Gamma_L L^r \delta^{r-j}. \tag{70}$$

Therefore

$$\sum_{r=L+1}^{\infty} |c_{jr}^\delta| \le \Gamma_L \sum_{r=L+1}^{\infty} L^r \delta^{r-j} \tag{71}$$

$$= \Gamma_L L^{L+1} \delta^{L+1-j} \sum_{k=0}^{\infty} (L\delta)^k. \tag{72}$$

Since $L\delta \le 1/2$, the geometric series is at most 2. Also $0 < \delta \le 1/2$ and $L + 1 - j \ge 1$, so $\delta^{L+1-j} \le \delta$. Hence (43) follows. □

Proof of Theorem 2. By Lemma 2,

$$E_j^\delta(H_\delta(u)) - u_j = \sum_{r=L+1}^{\infty} c_{jr}^\delta u_r. \tag{73}$$

Since $\|u_r\| \le M$, Lemma 3 yields

$$\left\|E_j^\delta(H_\delta(u)) - u_j\right\| \le M \sum_{r=L+1}^{\infty} |c_{jr}^\delta| \le 2\Gamma_L L^{L+1} M\delta. \tag{74}$$

Taking the maximum over j and the supremum over $u \in \mathcal{X}_-$ proves the result with $C_L = 2\Gamma_L L^{L+1}$. □

Proof of Theorem 3. Let $u \in \mathcal{X}_-$. Since g_L is obtained by anchoring the tail of f after coordinate L, the pasts

$$u = (u_0, u_1, \ldots) \qquad \text{and} \qquad (u_0, \ldots, u_L, a, a, a, \ldots) \tag{75}$$

agree on coordinates $0, 1, \ldots, L$. Hence, by the definition of $\omega_f(L)$,

$$\|f(u) - g_L(u_0, \ldots, u_L)\| \le \omega_f(L). \tag{76}$$

Let

$$R_\delta(u) := \big(E_0^\delta(H_\delta(u)), \ldots, E_L^\delta(H_\delta(u))\big). \tag{77}$$

By Theorem 2,

$$\max_{0\le j\le L} \left\|E_j^\delta(H_\delta(u)) - u_j\right\| \le C_L M\delta. \tag{78}$$

Because $C_L M\delta \le 1$, the recovered point $R_\delta(u)$ lies in K_L^+. Since $\widetilde{g}_L$ agrees with g_L on $\mathcal{X}^{L+1}$ and has modulus of continuity $\omega_{\widetilde{g}_L}$ with respect to the maximum block norm, (78) gives

$$\|g_L(u_0, \ldots, u_L) - \widetilde{g}_L(R_\delta(u))\| \le \omega_{\widetilde{g}_L}(C_L M\delta). \tag{79}$$

Finally, by the polynomial approximation assumption and the fact that $R_\delta(u) \in K_L^+$,

$$\|\widetilde{g}_L(R_\delta(u)) - \pi_L(R_\delta(u))\| \le \alpha. \tag{80}$$

The readout P satisfies $P(H_\delta(u)) = \pi_L(R_\delta(u))$. Combining (76), (79), and (80) gives

$$\|f(u) - P(H_\delta(u))\| \le \omega_f(L) + \omega_{\widetilde{g}_L}(C_L M\delta) + \alpha. \tag{81}$$

For a bi-infinite input sequence $x \in \mathcal{X}^{\mathbb{Z}}$ and a time $t \in \mathbb{Z}$, set

$$u = (x_t, x_{t-1}, x_{t-2}, \ldots). \tag{82}$$

The causal echo-state solution of the diagonal SSM is exactly $H_\delta(u)$, so the preceding estimate is precisely the pointwise bound for $\left\|F(x)_t - \widehat{F}(x)_t\right\|$. Taking the supremum over x and t proves (35). □

Proof of Theorem 1. Fix $\varepsilon > 0$ and $\rho \in (0, 1)$. Since f is continuous on compact $\mathcal{X}_-$, the truncation modulus satisfies $\omega_f(L) \to 0$. Choose $L \ge 1$ such that $\omega_f(L) < \varepsilon/3$. Fix an anchor $a \in \mathcal{X}$ and define

$$g_L(z_0, \ldots, z_L) := f(z_0, \ldots, z_L, a, a, a, \ldots). \tag{83}$$

Let

$$K_L^+ = \big\{z = (z_0, \ldots, z_L) \in (\mathbb{R}^p)^{L+1} : \inf_{w\in\mathcal{X}^{L+1}} \max_{0\le j\le L} \|z_j - w_j\| \le 1\big\}. \tag{84}$$

Since $\mathcal{X}^{L+1}$ is compact and hence closed in K_L^+, the Tietze extension theorem, applied coordinate-wise, gives a continuous extension

$$\widetilde{g}_L : K_L^+ \to \mathbb{R}^q \tag{85}$$

of g_L. The set K_L^+ is compact, so $\widetilde{g}_L$ is uniformly continuous. Hence there exists $\eta \in (0, 1)$ such that $\omega_{\widetilde{g}_L}(\eta) < \varepsilon/3$. Let

$$M := \sup_{x\in\mathcal{X}} \|x\|. \tag{86}$$

Let C_L be the constant from Theorem 2, and choose $\delta > 0$ such that

$$L\delta \le \min\{\rho, 1/2\}, \qquad C_L M\delta \le \eta. \tag{87}$$

If $M = 0$, the second condition holds automatically. By Stone–Weierstrass (Stone, 1948), applied coordinatewise on the compact set K_L^+, there exists a polynomial map $\pi_L : \mathbb{R}^{p(L+1)} \to \mathbb{R}^q$ such that

$$\sup_{z \in K_L^+} \|\pi_L(z) - \widetilde{g}_L(z)\| < \varepsilon/3. \tag{88}$$

Define the linear window extractor

$$E^\delta(h) := \big(E_0^\delta(h), \ldots, E_L^\delta(h)\big) \tag{89}$$

and define the SSM readout

$$P(h) := \pi_L(E^\delta(h)). \tag{90}$$

Since E^δ is linear and π_L is polynomial, P is polynomial. The diagonal SSM with eigenvalues $\lambda_i = i\delta$ has spectral radius at most $L\delta \le \rho$. Theorem 3, together with the choices $\omega_f(L) < \varepsilon/3$ and $\omega_{\widetilde{g}_L}(\eta) < \varepsilon/3$ and the bounds in (87) and (88), yields

$$\left\| F - \widehat{F} \right\|_\infty < \varepsilon. \tag{91}$$

This is exactly (24). □

Proof of Corollary 1. Let P be the polynomial readout obtained in Theorem 1, chosen so that its approximation error is less than $\varepsilon/2$. The set of states reached by the corresponding SSM,

$$\mathcal{K}_\delta := H_\delta(\mathcal{X}_-) \subset \mathbb{R}^{p(L+1)}, \tag{92}$$

is compact because H_δ is continuous and $\mathcal{X}_-$ is compact. By the assumed feedforward universality of the activation σ, there exists a multilayer perceptron N_θ such that

$$\sup_{h \in \mathcal{K}_\delta} \|N_\theta(h) - P(h)\| < \varepsilon/2. \tag{93}$$

The triangle inequality combines the operator approximation and neural-readout approximation, each with error less than $\varepsilon/2$, to give a total operator error less than ε. □

Proof of Corollary 2. By the containment hypothesis, $\mathcal{A}$ contains the diagonal SSMs with input-independent transitions and neural readouts supplied by Theorem 1 and Corollary 1. These models attain every prescribed approximation accuracy, which proves the claim. □

Proof of Theorem 4. Let $\Lambda \subset (-1, 1)$ be the set of distinct values among $\lambda_1, \ldots, \lambda_n$, and define the annihilating polynomial

$$Q(z) := \prod_{\mu \in \Lambda} (z - \mu). \tag{94}$$

Then $Q(\lambda_i) = 0$ for every i. Let m_0 denote the order of vanishing of Q at zero. Since the roots in Λ are distinct, m_0 is either 0 or 1. Because the target delay satisfies $j \ge 1$, we have $j - m_0 \ge 0$. Define

$$R(z) := z^{j-m_0} Q(z). \tag{95}$$

Then $R(\lambda_i) = 0$ for all i. Moreover, the coefficient of z^j in R is the lowest nonzero coefficient of Q, hence is nonzero. Write

$$R(z) = \sum_{r=0}^{d} c_r z^r, \qquad c_j \neq 0. \tag{96}$$

Choose a scalar $\beta > 0$ small enough that $|\beta c_r| \le 1$ for all $0 \le r \le d$. Define two pasts $u, v \in [-1, 1]^{\mathbb{N}_0}$ by

$$u_r := 0 \quad \text{for all } r, \tag{97}$$

and

$$v_r := \begin{cases} \beta c_r, & 0 \le r \le d, \\ 0, & r > d. \end{cases} \tag{98}$$

For every i,

$$h_i(v) = \sum_{r=0}^{d} \lambda_i^r \beta c_r = \beta R(\lambda_i) = 0 = h_i(u). \tag{99}$$

The jth delay coordinates satisfy

$$v_j = \beta c_j \neq 0 = u_j. \tag{100}$$

Thus the two pasts share the same diagonal exponential traces and have distinct jth delay coordinates. Every function $\Phi : \mathbb{R}^n \to \mathbb{R}$, including a discontinuous one, assigns a single value to the common trace vector. Its recovery error is therefore positive on at least one of these pasts, proving the exact-recovery separation. □

Proof of Theorem 5. Since $V_\delta = W D_\delta$ with $D_\delta = \mathrm{diag}(1, \delta, \ldots, \delta^L)$, we have

$$V_\delta^{-1} = D_\delta^{-1} W^{-1}. \tag{101}$$

Hence, for each row j,

$$\sum_{i=0}^{L} |(V_\delta^{-1})_{ji}| = \delta^{-j} \sum_{i=0}^{L} |(W^{-1})_{ji}| \leq \Gamma_L \delta^{-L}, \tag{102}$$

because $0 < \delta \leq 1$ and $j \leq L$. Therefore

$$\left\| E_j^\delta(\widetilde{h}) - E_j^\delta(h) \right\| \leq \sum_{i=0}^{L} |(V_\delta^{-1})_{ji}| \left\| \widetilde{h}_i - h_i \right\| \leq \Gamma_L \delta^{-L} \sigma. \tag{103}$$

Taking the maximum over j proves the claim. □

Proof of Theorem 6. Write $\lambda_i = i\delta$, $V = V_\delta$, $A = V^{-1}$, $E = E_L^\delta$, $a = a_\delta$, and $\beta = \beta_\delta$. Set

$$S = \sum_{i=1}^{L} \lambda_i = \frac{L(L+1)}{2}\delta < \frac{1}{2}, \qquad P = \prod_{i=1}^{L}(1 - \lambda_i), \qquad T = P^{-1} - 1 = \frac{\beta}{M}. \tag{104}$$

Since $P \geq 1 - S > 1/2$ and $P < 1$, we have $0 < T < 1$ and hence $0 < \beta < M$.

Extractor gain and residual. The last-row formula in (51) gives

$$A_{Li} = \frac{(-1)^{L-i}}{\delta^L i!(L-i)!}, \qquad \sum_{i=0}^{L} |A_{Li}| = a. \tag{105}$$

For $m \geq 0$, let $d_m = \sum_{i=0}^{L} A_{Li} \lambda_i^m$. The Vandermonde identities give $d_m = \mathbf{1}_{\{m=L\}}$ for $0 \leq m \leq L$. Moreover,

$$\sum_{m \geq 0} d_m z^m = \sum_{i=0}^{L} \frac{A_{Li}}{1 - \lambda_i z} = \frac{z^L}{\prod_{i=1}^{L}(1 - \lambda_i z)}. \tag{106}$$

Indeed, multiplication by the displayed denominator produces a polynomial of degree at most L; the first $L+1$ Taylor coefficients force that polynomial to be z^L. The series converges for $|z| < (L\delta)^{-1}$. Expanding its geometric factors shows that $d_m \geq 0$ for $m > L$, and evaluation at $z = 1$ gives $\sum_{m>L} d_m = T$. Consequently,

$$EH_\delta(u) = u_L + \sum_{m>L} d_m u_m, \qquad \sup_{u \in [-M,M]^{\mathbb{N}_0}} |EH_\delta(u) - u_L| = MT = \beta. \tag{107}$$

The equality is attained by setting all tail coordinates to M.

Two auxiliary interpolation vectors. Define

$$s_i = (-1)^{L-i}, \qquad b = As, \qquad g_i = \frac{\lambda_i^{L+1}}{1 - \lambda_i}, \qquad c = Ag. \tag{108}$$

We claim that

$$\|b\|_\infty = b_L = a, \qquad \|c\|_\infty = c_L = T. \tag{109}$$

The leading-coefficient formula gives $b_L = a$. For a Lagrange basis polynomial, the absolute value of its coefficient of degree j, divided by the absolute value of its leading coefficient, is the elementary

symmetric polynomial of degree $L-j$ in its nonnegative roots. Their sum is at most S, so this ratio is at most $S^{L-j}/(L-j)!$. Summing over the basis polynomials yields

$$|b_j| \le \sum_{i=0}^{L} |A_{ji}| \le a\frac{S^{L-j}}{(L-j)!} \le a, \qquad 0 \le j \le L. \tag{110}$$

This proves the assertion for b.

To prove the assertion for c, let

$$Q(z) = \prod_{i=0}^{L}(z-\lambda_i) = \sum_{k=0}^{L+1} q_k z^k, \qquad C(z) = \frac{z^{L+1} - Q(z)/P}{1-z}. \tag{111}$$

Since $Q(1) = P$, the numerator vanishes at 1, so C is a polynomial of degree at most L. At each node, $C(\lambda_i) = g_i$; hence $C(z) = \sum_{j=0}^{L} c_j z^j$. Comparing coefficients gives

$$c_j = -\frac{\sum_{k=0}^{j} q_k}{P}, \qquad c_0 = 0, \qquad c_L = P^{-1} - 1 = T. \tag{112}$$

Let e_m be the elementary symmetric polynomial of degree m in $\lambda_1, \dots, \lambda_L$, with $e_0 = 1$ and $e_m = 0$ for $m > L$. In particular, $e_1 = S$. Counting each product of $m+1$ distinct nodes in Se_m gives

$$(m+1)e_{m+1} \le Se_m. \tag{113}$$

Thus these symmetric polynomials decrease with m. For $1 \le m \le L$, set $F_m = e_m - e_{m+1} + \cdots + (-1)^{L-m} e_L$. Alternating-sum bounds give $0 \le F_m \le e_m$. For $2 \le m \le L$,

$$F_m \le e_2 \le \frac{S^2}{2}, \qquad F_1 \ge S - e_2 \ge S - \frac{S^2}{2} \ge \frac{S^2}{2}. \tag{114}$$

Using the signs of the coefficients of Q, (112) gives $|c_j| = F_{L+1-j}/P$ for $1 \le j \le L$. Therefore $|c_j| \le F_1/P = c_L = T$, proving (109).

Histories attaining the lower bounds. Put $\tau_0 = (M-\beta)/a > 0$. For $0 \le \tau \le \tau_0$, define

$$u_j^{(\tau)} = \tau b_j + Mc_j \quad (0 \le j \le L), \qquad u_m^{(\tau)} = -M \quad (m > L). \tag{115}$$

These histories lie in $[-M, M]^{\mathbb{N}_0}$ because $|u_j^{(\tau)}| \le \tau a + MT \le M$. Their finite window contributes $V(\tau b + Mc) = \tau s + Mg$ to the trace vector, while their tail contributes $-Mg$. Hence

$$H_\delta(u^{(\tau)}) = \tau s, \qquad u_L^{(\tau)} = \tau a + \beta. \tag{116}$$

At $\tau = 0$, the histories $\pm u^{(0)}$ have the same trace vector and target values $\pm\beta$. Every readout therefore has uniform error at least β. At $\tau = \tau_0$, the histories $\pm u^{(\tau_0)}$ have target values $\pm M$ and trace vectors $\pm\tau_0 s$. If a K-Lipschitz readout has uniform error R, the triangle inequality gives

$$2M \le 2R + |\Phi(\tau_0 s) - \Phi(-\tau_0 s)| \le 2R + 2K\tau_0. \tag{117}$$

Combining the two pairs of histories yields

$$R \ge \max\left\{\beta, M - \frac{K}{a}(M-\beta)\right\}. \tag{118}$$

A readout attaining the bound. Let $\theta = \min\{1, K/a\}$ and take $\Phi = \theta E$. Its Lipschitz constant is $\theta a \le K$, and (107) gives

$$|\theta E H_\delta(u) - u_L| \le (1-\theta)|u_L| + \theta|EH_\delta(u) - u_L| \le (1-\theta)M + \theta\beta. \tag{119}$$

This is exactly the preceding lower bound, proving (54) and attainment. Solving that expression for an error at most $\varepsilon < M$ yields (55). □

B Window recovery from diagonal traces

Theorem 1 constructs finite-window approximations using independent exponential traces and an accuracy-dependent Vandermonde readout. Theorem 4 complements this construction by characterizing histories that each fixed finite trace bank maps to the same state. A comparison with shift-register SSMs makes the two memory representations explicit.

Comparison with shift-register SSMs. A finite delay line stores recent inputs through a nilpotent shift transition. For block inputs $x_t \in \mathbb{R}^p$, one may define a state $z_t = (z_t^{(0)}, \dots, z_t^{(L)}) \in (\mathbb{R}^p)^{L+1}$ by

$$z_t^{(0)} = x_t, \qquad z_t^{(k)} = z_{t-1}^{(k-1)} \quad \text{for } k = 1, \dots, L. \tag{120}$$

Equivalently, $z_t = A_{\text{shift}} z_{t-1} + B_{\text{shift}} x_t$, where A_{shift} is the block matrix with identity blocks on the first subdiagonal and zeros elsewhere, and $B_{\text{shift}} = e_0 \otimes I_p$ with $e_0 = (1, 0, \dots, 0)^\top \in \mathbb{R}^{L+1}$. Hence $z_t^{(j)} = x_{t-j}$ for $0 \le j \le L$. Applying finite-horizon truncation and static feedforward approximation to this state gives universality for fading-memory operators.

The construction in Theorem 1 uses the diagonal recurrent matrix

$$A_\delta = \operatorname{diag}(0, \delta, 2\delta, \dots, L\delta) \otimes I_p. \tag{121}$$

It is diagonal, normal over $\mathbb{R}$, input-independent, and strongly contracting. Each coordinate evolves independently and stores an exponential trace

$$\sum_{r=0}^{\infty} (i\delta)^r x_{t-r}. \tag{122}$$

Applying an ε-dependent Vandermonde extractor to these traces reconstructs the finite window to the prescribed accuracy. The extractor's jth row scales as δ^{-j}, as shown in Section 4.

Accuracy-dependent window recovery. For every fixed finite diagonal trace bank satisfying the hypotheses of Theorem 4 and every delay $j \ge 1$, there are histories $u, v \in [-1, 1]^{\mathbb{N}_0}$ with $u_j \ne v_j$ that have the same image under

$$u \mapsto (h_1(u), \dots, h_n(u)). \tag{123}$$

Every readout assigns the same output to these histories, so its error in recovering the jth delay coordinate is positive on at least one of them. This applies to arbitrary readouts, including discontinuous functions, and gives the corresponding separation for every window $(u_0, \dots, u_L)$ with $L \ge 1$.

Theorem 1 chooses the trace bank as a function of the target accuracy. For each fixed window length L, the spectral nodes approach zero according to

$$0, \delta, 2\delta, \dots, L\delta, \qquad \delta \to 0. \tag{124}$$

For each accuracy level, the associated Vandermonde inverse recovers the recent coordinates from the truncated traces and leaves a uniformly controlled tail of order $O(\delta)$, as quantified in Theorem 2. This gives arbitrarily accurate window recovery along the accuracy-dependent contracting spectral family.

Remark 2 (Circular spectra and conditioning)**.** *Fix $L \ge 1$, an integer $N > L$, and $0 < r < 1$. Let $\zeta = e^{2\pi \mathrm{i}/N}$ and, for $k = 0, \dots, N-1$, define complex traces*

$$h_k = \sum_{m \ge 0} (r\zeta^k)^m u_m. \tag{125}$$

Fourier orthogonality gives, for $0 \le j \le L$,

$$D_j h := \frac{r^{-j}}{N} \sum_{k=0}^{N-1} \zeta^{-kj} h_k = u_j + \sum_{q \ge 1} r^{qN} u_{j+qN}. \tag{126}$$

Consequently, if $\|u_m\| \le M$ and $\max_k \left\| \widetilde{h}_k - h_k \right\| \le \sigma$, then

$$\max_{j \le L} \|D_j h - u_j\| \le \frac{M r^N}{1 - r^N}, \qquad \max_{j \le L} \left\| D_j(\widetilde{h} - h) \right\| \le r^{-L} \sigma. \tag{127}$$

The identity follows from $N^{-1} \sum_{k=0}^{N-1} \zeta^{k(m-j)} = \mathbf{1}_{m \equiv j \pmod N}$; the bounds follow by summing the geometric tail and the decoder's absolute coefficients. The noise bound is attained at $j = L$ by $\widetilde{h}_k - h_k = \sigma \zeta^{kL} v$ for a real unit vector v.

For real inputs, conjugate trace pairs admit real two-dimensional rotation–contraction blocks, giving Np real state coordinates in total. At $N = L + 1$ and the same spectral radius $r = L\delta$, the circular decoder has noise gain r^{-L}, whereas (51) *gives equal-grid gain at least $(2^L L^L / L!) r^{-L}$. Allowing circular spectra thus improves conditioning; Theorem 1 establishes universality under the additional restriction to nonnegative scalar diagonal dynamics.*

C Optimal delay recovery with trace noise

The histories used to prove Theorem 6 also determine the exact effect of bounded additive noise on the raw traces. We retain the scalar input interval and the small-spacing condition of that theorem, and use a_δ and β_δ from (52).

Theorem 7 (Exact sensitivity–accuracy tradeoff with trace noise)**.** *Under the hypotheses of Theorem 6, let $K, \sigma \geq 0$ and define*

$$\mathcal{R}_\delta(K,\sigma) := \inf_{\substack{\Phi:\mathbb{R}^{L+1}\to\mathbb{R} \\ \mathrm{Lip}_\infty(\Phi)\leq K}} \sup_{\substack{u\in[-M,M]^{\mathbb{N}_0} \\ \eta\in\mathbb{R}^{L+1},\ \|\eta\|_\infty\leq\sigma}} |\Phi(H_\delta(u)+\eta) - u_L|. \tag{128}$$

Then, with $(z)_+ = \max\{z, 0\}$,

$$\mathcal{R}_\delta(K,\sigma) = M - \min\left\{1, \frac{K}{a_\delta}\right\}(M - \beta_\delta - a_\delta\sigma)_+. \tag{129}$$

Allowing arbitrary readouts, including discontinuous functions, gives

$$\inf_{\Phi:\mathbb{R}^{L+1}\to\mathbb{R}} \sup_{\substack{u\in[-M,M]^{\mathbb{N}_0} \\ \eta\in\mathbb{R}^{L+1},\ \|\eta\|_\infty\leq\sigma}} |\Phi(H_\delta(u)+\eta) - u_L| = \min\{M, \beta_\delta + a_\delta\sigma\}. \tag{130}$$

Proof. Write $a = a_\delta$, $\beta = \beta_\delta$, and $\tau_0 = (M-\beta)/a$. Use the sign vector s and the histories $u^{(\tau)}$ from (108) and (115). They satisfy $\|s\|_\infty = 1$, $H_\delta(u^{(\tau)}) = \tau s$, and $u_L^{(\tau)} = \beta + a\tau$ for every $0 \leq \tau \leq \tau_0$.

First take $\tau = \min\{\sigma, \tau_0\}$. For the two histories $\pm u^{(\tau)}$, choose noises $\mp\tau s$. Both observations are zero, so every readout has worst-case error at least

$$\beta + a\min\{\sigma, \tau_0\} = \min\{M, \beta + a\sigma\}. \tag{131}$$

Next take the histories $\pm u^{(\tau_0)}$, whose targets are $\pm M$. Noises $\mp\min\{\sigma, \tau_0\}s$ yield observations $\pm(\tau_0 - \sigma)_+ s$. A K-Lipschitz readout with worst-case error R must satisfy

$$2M \leq 2R + 2K(\tau_0 - \sigma)_+, \qquad R \geq M - K(\tau_0 - \sigma)_+. \tag{132}$$

If $\sigma \geq \tau_0$, (131) already gives $R \geq M$. If $\sigma < \tau_0$, the maximum of the two lower bounds is

$$M - \min\{a, K\}(\tau_0 - \sigma) = M - \min\{1, K/a\}(M - \beta - a\sigma). \tag{133}$$

These are the two cases of (129).

For attainment, use the zero readout when $\beta + a\sigma \geq M$. Otherwise, take $\Phi = \theta E_L^\delta$ with $\theta = \min\{1, K/a\}$. By (105) and (107), this readout has Lipschitz constant at most K and error bounded by

$$|\theta E_L^\delta(H_\delta(u)+\eta) - u_L| \leq (1-\theta)M + \theta(\beta + a\sigma), \tag{134}$$

which equals the lower bound. Finally, (131) applies to arbitrary readouts, and taking $K = a$ in the construction gives the matching upper bound in (130). □

At $\sigma = 0$, the theorem recovers (54). For fixed L, $\beta_\delta = ML(L+1)\delta/2 + O_L(M\delta^2)$, while $a_\delta = 2^L/(L!\delta^L)$. Thus (130) gives the exact balance between the fixed-bank approximation floor and amplified trace noise for this scalar delay task. The normalization of the shared input gains and the maximum norm on raw trace perturbations are part of this characterization.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes].

   Justification: The claims made in the abstract and introduction are rigorously proved in the following mathematical development of the paper.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes].

   Justification: The paper discusses conditioning and robustness limits of the construction.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [Yes].

Justification: The assumptions are stated in the theorem, proposition, lemma, and corollary statements, and the appendix contains full proofs for the main result and supporting results.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [N/A].

   Justification: The paper is theoretical and does not report experimental results.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
   - If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
   - Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
   - While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
     (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
     (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
     (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
     (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

Answer: [N/A].

Justification: The paper does not include experiments, datasets, or code required to reproduce empirical claims.

Guidelines:

- The answer [N/A] means that paper does not include experiments requiring code.
- Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
- The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
- The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
- The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
- At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
- Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [N/A].

   Justification: The paper does not include experiments.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [N/A].

   Justification: The paper does not include experiments or statistical empirical comparisons.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
   - The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
   - The assumptions made should be given (e.g., Normally distributed errors).
   - It should be clear whether the error bar is the standard deviation or the standard error of the mean.

- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [N/A].

   Justification: The paper does not include experiments requiring compute resources.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes].

   Justification: The work is a theoretical approximation result and does not involve human subjects, private data, deployed systems, or released high-risk models.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [N/A].

    Justification: The work is foundational theory and does not introduce a deployed system, dataset, or model with a direct societal application.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.

- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A].

    Justification: The paper does not release data or models that pose high misuse risk.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [N/A].

    Justification: The paper does not use datasets, code packages, models, or other existing assets beyond cited scholarly literature.

    Guidelines:

    - The answer [N/A] means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.

- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [N/A].

    Justification: The paper does not introduce or release new datasets, code, models, or benchmarks.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing or human-subject research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A].

    Justification: The paper does not involve crowdsourcing or human-subject research.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A].

Justification: LLMs are not an important, original, or non-standard part of the core methodology.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.