EXTENDING SSMS WITH THE EXPONENTIALLY WEIGHTED SIGNATURE

Anonymous authors
Paper under double-blind review

ABSTRACT

We introduce the exponentially weighted signature (EWS), a continuous-time model that computes iterated integrals of a path, where each increment is weighted by the matrix exponential of a learnable generator over elapsed clock time. We prove that it solves a linear controlled differential equation, keeps the group-like structure and the universality of the signature, and satisfies a modified Chen identity, enabling a parallel scan. At depth one the EWS is a state-space model (SSM), and we map linear time-invariant SSMs, Mamba channels and Mamba-2 heads to it in closed form. The EWS extends SSMs through an arbitrary matrix generator, a clock that generalises the step size to causal functionals of the input, and higher truncation depths that are non-linear in the path within a single layer. Empirically, the EWS achieves the highest average accuracy and rank on six long time-series classification datasets, where depth generally helps. Learned clocks prove necessary for state tracking on formal language tasks, and at depth one, the EWS matches or exceeds competing SSMs on regression and forecasting with far fewer parameters.

1 INTRODUCTION

State-space models (SSMs) are a leading architecture for learning on long sequences (Gu et al., 2022b; Gu & Dao, 2024; Smith et al., 2023; Boyer et al., 2025). Their update is linear in the state, so they train in parallel yet run recurrently at inference with constant memory per step (Martin & Cundy, 2018; Dao & Gu, 2024; Smith et al., 2023). Most SSMs are obtained by discretising a continuous-time linear dynamical system forced by the input. In linear time-invariant (LTI) SSMs, including S4, S5, LinOSS and D-LinOSS, the step size and discretised system matrices are fixed, such that each layer is a convolution of its inputs, a structure that performs strongly on long continuous signals such as audio, images and physiological recordings (Rusch & Rus, 2025; Nguyen et al., 2022; Goel et al., 2022). In selective SSMs, notably the Mamba family, these quantities depend on the input, letting the model choose what to retain and making it competitive on discrete, information-dense data such as text, which has led large language models to interleave selective SSM layers with attention (Gu & Dao, 2024; NVIDIA et al., 2025; Lahoti et al., 2026; Dao & Gu, 2024).

Working with the continuous-time system, rather than a fixed discretisation of it, has practical and theoretical benefits. The hidden state evolves between observations and can be queried at any time, so irregularly sampled and asynchronous data need not be forced onto a regular grid (Kidger et al., 2020; Walker et al., 2024; 2026), and elapsed time becomes a modelling choice, measured by timestamps, observation counts or a quantity learned from the data, as the step size is in selective SSMs (Gu & Dao, 2024). On the theoretical side, since an SSM treats its input as samples of a latent continuous signal, SSMs can be studied as linear controlled differential equations (CDEs) (Cirone et al., 2024), a framework that characterises what their recurrences can express and shows that structured linear CDEs are maximally expressive on paths (Kidger, 2022; Walker, 2025; Walker et al., 2025).

Behind every linear CDE lies the signature of its driving path, a collection of iterated integrals graded by depth (Chen, 1954; Lyons et al., 2007). Any linear readout of a linear CDE is a series in these integrals, with coefficients set by the matrices of its vector fields (Friz & Victoir, 2010; Lyons, 1998); thus the terms a model can use depend on its driving path and these matrices. An LTI SSM is a linear CDE driven by time with an additive input, such that only terms linear in the input appear. Its readouts are fixed linear filters, whereas a selective SSM is driven by a gate computed from the

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

input, letting its filters adapt to context. With diagonal matrices, as in Mamba, the filters adapt only through increments of that gate (Cirone et al., 2024). The signature itself solves a linear CDE with fixed vector fields, and its linear functionals are universal on suitable compact sets of paths, making it a popular feature map (Graham, 2013; Gyurkó et al., 2014; Levin et al., 2016; Király & Oberhauser, 2019; Salvi et al., 2021). However, it has no learnable parameters, and treats the path's history uniformly, offering no mechanism for fading memory (Boyd & Chua, 1985), whereas SSMs forget at rates set by their learned dynamics and step size, a mechanism regarded as essential for long-sequence training (Gu et al., 2020; 2023; Zucchet & Orvieto, 2024; Morita, 2025).

We introduce the exponentially weighted signature (EWS), whose iterated integrals weight each increment, after a learnable lift, by the matrix exponential of a learnable generator scaled by elapsed clock time. At depth one, the EWS is an SSM, and trained LTI SSMs convert to it in closed form, as does each channel of Mamba and head of Mamba-2, with its accumulated step size as the clock. The EWS extends SSMs in three ways. Its deeper levels weight non-commutative products of increments, making them non-linear in the path within the layer. Its clock generalises the step size to causal functions of the input path, including the accumulated norm of its increments and learned gates. Its generator may be any matrix, including Jordan blocks whose kernels carry a polynomial–exponential factor that none of these SSMs have. These extensions preserve the algebraic structure of the signature, as the EWS solves a linear CDE whose vector fields are those of the signature together with one built from the generator and driven by the clock. Like the signature, its state is graded and group-like (Reutenauer, 1993; Chen, 1957), and it satisfies a modified Chen identity (Chen, 1954) that permits a parallel associative scan (Blelloch, 1990; Smith et al., 2023).

Contributions

1. The exponentially weighted signature (EWS) is introduced as a continuous-time layer with a learnable lift, generator and clock. It contains the signature, the EFM signature (Abi Jaber & Sotnikov, 2025) and depth-one SSMs as special cases.
2. Trained LTI SSMs, each channel of Mamba and each head of Mamba-2 are shown to convert to depth-one EWS in closed form (Section 4). For any matrix generator, the EWS solves a linear CDE, satisfies a modified Chen identity (Proposition 1, Theorem 1(ii)), and is group-like and maximally expressive (when it separates paths), while its Jordan blocks give polynomial–exponential memory kernels at depth one that no diagonalisable SSM generator reaches (Proposition 2).
3. A comprehensive empirical evaluation showing that truncation depth and learned clocks each extend what SSMs can learn, with learned clocks needed for state tracking, depth necessary on synthetic targets and beneficial on the long time-series classification, enabling a single EWS layer to achieve the highest average accuracy and rank.

Related work Memory has been added to the signature through sliding or expanding windows, which fix the horizon in advance (Kidger et al., 2019; Morrill et al., 2021a; Cohen et al., 2025; Drobac et al., 2025), through exponential weighting at prescribed rates (Abi Jaber & Sotnikov, 2025) and through general memory kernels in the Volterra signature (Harang & Tindel, 2021; Harang et al., 2023; 2024; Hager et al., 2026). Appendix D details how these last two differ from the EWS. The EWS is itself a structured linear CDE (Walker et al., 2025), in the line of neural CDEs (Walker, 2025; Kidger, 2022; Kidger et al., 2020; Morrill et al., 2021b; Walker et al., 2024; 2026) and the theory of linear CDEs (Cirone et al., 2024). Input-dependent recurrences with non-diagonal transitions (Yang et al., 2024; 2025; Siems et al., 2025; Terzić et al., 2025; 2026) are compared to empirically in Section 5.

Background Let $X \in \mathcal{V}^1([0,T];\mathbb{R}^d)$ be a path of bounded variation, with $\|X\|_{1,[s,t]}$ its total variation over $[s,t]$. Notation for path spaces and the tensor algebra is collected in Appendix B, where paths of finite p-variation with $p<2$ are also treated. The signature of X over $[s,t]\subseteq[0,T]$ is $S(X)_{s,t} = \big(1, S(X)^{(1)}_{s,t}, S(X)^{(2)}_{s,t}, \dots\big) \in T((\mathbb{R}^d)) := \bigoplus_{n\geq 0}(\mathbb{R}^d)^{\otimes n}$ where

$$S(X)^{(n)}_{s,t} = \int_{s<t_1<\cdots<t_n<t} dX_{t_1}\otimes\cdots\otimes dX_{t_n}. \tag{1}$$

The signature is graded, with depth n holding the iterated integrals of order n, and $S(X)^{i_1,\dots,i_n}_{s,t}$ denoting the coefficient of $e_{i_1} \otimes \cdots \otimes e_{i_n} \in (\mathbb{R}^d)^{\otimes n}$ (Chen, 1954; McLeod & Lyons, 2025). The universality of its linear functionals is recalled in Appendix B.3.

A linear CDE driven by X evolves a state $h_t \in \mathbb{R}^m$ by $dh_t = \sum_{i=1}^d M_i h_t dX^i_t$, where the matrices $M_i \in \mathbb{R}^{m\times m}$ are linear vector fields. When X is linear on $[s,t]$ its solution is $h_t = \exp(\sum_{i=1}^d (X^i_t - X^i_s)M_i)h_s$; for general X see Appendix B.4. With learnable matrices, an initial state computed from the first observation, and a linear readout, we have a linear neural CDE (Kidger et al., 2020; Cirone et al., 2024), and with structured matrices a SLiCE (Walker et al., 2025), for which the states at observation times are computed by a parallel scan over the flows.

An SSM maps an input signal x to a state $h \in \mathbb{R}^m$ and an output y through $\dot{h}(t) = -Ah(t) + Bx(t)$ with $y(t) = Ch(t)$, where we write $-A$ rather than A to match the EWS convention in Section 2. On a grid it is computed by $h_j = \bar{A}_j h_{j-1} + \bar{B}_j x_j$, where $\bar{A}_j$ and $\bar{B}_j$ are determined by A, B, a step size Δ_j and the discretisation rule (Gu et al., 2022b; Gu & Dao, 2024). In LTI SSMs the step size and matrices are fixed, so $\bar{A}_j$ and $\bar{B}_j$ do not depend on j. With $X_t = \int_0^t x_s ds$, the SSM is the affine CDE $dh_t = -Ah_t dt + BdX_t$, driven by time with an additive input; appending a constant coordinate to the state puts it in the form of a linear CDE (Appendix B.4); Appendix E specifies the models considered.

2 The Exponentially Weighted Signature

Definition 1 (Clock)**.** A clock is a causal map $\theta : \mathcal{V}^1([0,T];\mathbb{R}^d) \to \mathcal{V}^1([0,T];\mathbb{R})$, that is, $\theta(X)_t = \theta(X_{\cdot\wedge t})_t$ for all $X \in \mathcal{V}^1([0,T];\mathbb{R}^d)$ and $t \in [0,T]$, where $X_{\cdot\wedge t}$ is the path stopped at t.

The time clock $\theta(X)_t = t$ and the variation clock $\theta(X)_t = \|X\|_{1,[0,t]}$ are both examples. When the path is obvious, we write θ_t for $\theta(X)_t$.

Definition 2 (EWS)**.** Consider a matrix pair $\mathbb{A} = (A, B)$, with terms identified as the *generator* $A \in \mathbb{R}^{m\times m}$ and *lift* $B \in \mathbb{R}^{m\times d}$. The exponentially weighted signature (EWS) of X over $[s,t] \subseteq [0,T]$ with clock θ is $S_{\mathbb{A}}(X;\theta)_{s,t} = \big(1, S_{\mathbb{A}}(X;\theta)^{(1)}_{s,t}, S_{\mathbb{A}}(X;\theta)^{(2)}_{s,t}, \dots\big) \in T((\mathbb{R}^m))$ where

$$S_{\mathbb{A}}(X;\theta)^{(n)}_{s,t} = \int_{s<t_1<\cdots<t_n<t} \big(e^{-(\theta_t-\theta_{t_1})A} BdX_{t_1}\big) \otimes \cdots \otimes \big(e^{-(\theta_t-\theta_{t_n})A} BdX_{t_n}\big). \tag{2}$$

Each increment of the lifted path BX is weighted by a matrix exponential of the generator over the clock time elapsed until t. Truncation at depth N keeps the levels $n \le N$, denoted $S_{\mathbb{A}}(X;\theta)^{\le N}_{s,t} \in T^N(\mathbb{R}^m)$; Appendix C.1 writes the EWS in coordinates, gives a recursive form and works through an explicit example at depth two. When $A = 0$ and B is the identity, the EWS is the signature $S(X)$. When A is a diagonal matrix with positive entries, B is the identity and θ is the time clock, it is the exponentially fading memory signature (EFM) of Abi Jaber & Sotnikov (2025) (Appendix D). The EWS therefore extends the EFM through its lift, its clock, its generator and the fact that its parameters are learnable; Section 2.1 compares the two through the spectrum of the generator.

For $r \in \mathbb{R}$, the flow operator $D^r_A : T((\mathbb{R}^m)) \to T((\mathbb{R}^m))$ acts by $D^r_A(v_1 \otimes \cdots \otimes v_n) = (e^{-rA}v_1) \otimes \cdots \otimes (e^{-rA}v_n)$, and the derivation induced by A, $\Lambda_A = -\frac{d}{dr}\big|_{r=0} D^r_A : T((\mathbb{R}^m)) \to T((\mathbb{R}^m))$, acts by $\Lambda_A(v_1 \otimes \cdots \otimes v_n) = \sum_{j=1}^n v_1 \otimes \cdots \otimes Av_j \otimes \cdots \otimes v_n$. Both are extended linearly, and Appendix C.2 discusses their algebraic properties.

Proposition 1 (Linear CDE form – Appendix C.3)**.** *The EWS is the unique solution to the linear controlled differential equation*

$$dS_{\mathbb{A}}(X;\theta)_{s,t} = -\Lambda_A S_{\mathbb{A}}(X;\theta)_{s,t} d\theta_t + S_{\mathbb{A}}(X;\theta)_{s,t} \otimes BdX_t, \qquad S_{\mathbb{A}}(X;\theta)_{s,s} = 1. \tag{3}$$

Without the first term, (3) is the CDE satisfied by the signature of the lifted path BX (Appendix B.3). The first term is an additional vector field built from the generator and integrated against the clock, which applies the exponential weighting. Since Λ_A is grade-preserving and BdX_t lies at depth one, depth $n \ge 1$ evolves as $dS_{\mathbb{A}}(X;\theta)^{(n)}_{s,t} = -\Lambda_A S_{\mathbb{A}}(X;\theta)^{(n)}_{s,t} d\theta_t + S_{\mathbb{A}}(X;\theta)^{(n-1)}_{s,t} \otimes BdX_t$, where the first term weights depth n in place and the second adds depth $n-1$ tensored with the lifted increment.

Identifying $T^N(\mathbb{R}^m)$ with $\mathbb{R}^K$ for $K = \sum_{n=0}^{N} m^n$ turns the equation for the truncated EWS into a linear CDE driven by (θ, X), with matrices built from A and B. These structured matrices make the EWS a SLiCE, and Appendix C.3 constructs them explicitly.

Theorem 1 (Structure – Appendices C.4, C.6 and C.7)**.** *Let θ be a clock and $0 \le s \le t \le T$.*

(i) *The EWS over the interval $[s,t]$ is the signature of a re-weighted path, $S_{\mathbb{A}}(X;\theta)_{s,t} = S(Z^{[t]})_{s,t}$ where $Z^{[t]}_\tau = \int_s^\tau e^{-(\theta_t-\theta_u)A} B dX_u$ for $\tau \in [s,t]$. In particular, the product of a coordinate at depth n_1 and a coordinate at depth n_2 is a linear combination of coordinates at depth n_1+n_2.*

(ii) *For every $u \in [s,t]$, the EWS satisfies the modified Chen identity*

$$S_{\mathbb{A}}(X;\theta)_{s,t} = D_A^{\theta_t-\theta_u}\big(S_{\mathbb{A}}(X;\theta)_{s,u}\big) \otimes S_{\mathbb{A}}(X;\theta)_{u,t}. \tag{4}$$

(iii) *For a continuous clock and a compact set of paths, linear functionals of $S_{\mathbb{A}}(X;\theta)_{0,T}$ are dense in the continuous functions that take equal values on paths with the same EWS.*

Since the path in (i) depends on the interval, (ii) does not follow from the classical Chen identity. However, on a fixed interval, (i) makes the EWS a signature, hence group-like (Chen, 1957; Reutenauer, 1993). By (iii), the EWS is universal on paths up to the equivalence it induces (similarly to the signature), and hence on all paths whenever it separates them. For instance, it separates time-augmented paths with a common start when the lift and generator satisfy mild structural conditions (Corollary 4).

If $\theta(X \circ \varphi) = \theta(X) \circ \varphi$ for every increasing homeomorphism φ of $[0,T]$, as holds for the variation clock but not the time clock, then $S_{\mathbb{A}}(X \circ \varphi;\theta)_{s,t} = S_{\mathbb{A}}(X;\theta)_{\varphi(s),\varphi(t)}$, and hence the EWS is invariant under reparametrisation of X (Proposition 4). An LTI SSM lacks this invariance, and the signature has it but no fading memory, whereas the EWS with the variation clock can have both.

2.1 Spectral properties at depth one

At depth one, for a given clock, (2) is a linear functional of the increments of the lifted path, whose weights e^{-rA} depend on the spectrum of A and elapsed clock time r:

$$S_{\mathbb{A}}(X;\theta)^{(1)}_{s,t} = \int_s^t e^{-(\theta_t-\theta_u)A} B dX_u \in \mathbb{R}^m. \tag{5}$$

If $A = \operatorname{diag}(\lambda_1, \dots, \lambda_m)$ with real entries, the k-th coordinate of (5) is $\int_s^t e^{-\lambda_k(\theta_t-\theta_u)}(BdX_u)^k$, a finite-window Laplace transform in clock time, and depth one is then a bank of m Laplace modes, each lifted channel forgotten at its own monotone rate.

If instead $A = P\Gamma P^{-1}$ is diagonalisable over $\mathbb{C}$ with complex eigenvalues, it applies a diagonal bank to the transformed path and mixes the result via P. A conjugate pair $a \pm i\omega$, carried by the real block $\left(\begin{smallmatrix} a & -\omega \\ \omega & a \end{smallmatrix}\right)$, contributes the damped oscillatory kernels $e^{-ar}\cos(\omega r)$ and $e^{-ar}\sin(\omega r)$. Depth one thus becomes a bank of damped Fourier modes, their number fixed by the lift; that is, a learned STFT (Mallat, 2009) whose window is the exponential memory set by A. Deeper levels combine these filtered components across time and channels. A real diagonal generator can only approximate these kernels with much larger width (Appendix C.8). A diagonalisable generator gives weights that are combinations of pure exponentials and damped oscillations. A polynomial factor must instead come from a Jordan block.

Proposition 2 (Polynomial memory – Appendix C.8)**.** *Let $\theta_t = t$, let X^i be a channel of X and $\lambda \ge 0$. If A has a Jordan block of size $K+1$, with eigenvalue λ, and B maps X^i into the last coordinate of that block and no other channel into the block, then $\int_0^t \frac{(t-u)^\ell}{\ell!} e^{-\lambda(t-u)} dX^i_u$ is, up to sign, a depth-one coordinate for every $\ell \le K$. For $\ell \ge 1$, it is not a linear readout of depth one for any diagonalisable A, of any size.*

Up to normalisation the kernels in Proposition 2 are the gamma kernels of the gamma memory of de Vries & Príncipe (1992). When $\lambda = 0$ each of these functionals is a linear functional of the signature of (t, X) requiring at least depth $\ell + 1$, whereas the EWS gives it at depth one. For $\lambda > 0$, no finite depth of the signature suffices, and for any λ, linear functionals of the EFM only approximate it.

2.2 THE EWS LAYER

Up to this point, the generator, lift, and clock are fixed. An EWS layer trains them and applies the transform to a time series, with A and the clock drawn from Table 1 in our experiments, and B free. Writing $X \in \mathcal{V}^1([0,T];\mathbb{R}^d)$ for a path built from observations, the layer is the map $X|_{[0,t]} \mapsto S_{\mathbb{A}}(X;\theta)^{\leq N}_{0,t}$, defined for every $t \in [0,T]$. Read at the observation times through a linear readout,

$$X|_{[0,t]} \xrightarrow{S_{\mathbb{A}}(\,\cdot\,;\theta)} S_{\mathbb{A}}(X;\theta)^{\leq N}_{0,t} \xrightarrow{\text{linear readout}} y_t, \tag{6}$$

it is a sequence-to-sequence model, and by Theorem 1(iii) such models are maximally expressive in the sense of Walker et al. (2025) as truncation depth grows. The clocks used are built from

Table 1: Some generator families and clocks. A clock is: invariant (i) when the layer does not depend on the path's parametrisation; stopping (s) when f can vanish on non-zero increments.

Family	Spectrum	Kernels
Zero	$\{0\}$	1
Diagonal	real	$e^{-\lambda r}$
Rotation	$a \pm i\omega$	$e^{-ar}\cos(\omega r),\ e^{-ar}\sin(\omega r)$
Dense	general	mixtures
Jordan, real	λ, repeated	$r^\ell e^{-\lambda r}$
Jordan, complex	$a \pm i\omega$, repeated	$r^\ell e^{-ar}\cos(\omega r),\ r^\ell e^{-ar}\sin(\omega r)$

Clock	$f(v)$	i	s
Time	Δt		
Variation	$\|v\|$	✓	
Blended	$\alpha\Delta t + (1-\alpha)\|v\|$		
Gated	$\mathrm{ReLU}(w \cdot v - b)$	$b{=}0$	✓
Clamped	$\frac{1}{c}\sum_{k=1}^{K} 1 \wedge \mathrm{ReLU}(w \cdot v - b_k)$		✓

path increments. Fix $f\colon \mathbb{R}^d \to \mathbb{R}$ and set $\theta_{t_j} = \sum_{t_i \leq t_j} f(\Delta X_i)$, interpolated linearly between observation times, with choices of f in Table 1. Here Δt is the time coordinate of v, the increment of the time channel of X, or one when X has none. Taking f non-negative makes the clock non-decreasing, so that increments are damped as clock time passes. When f vanishes the clock stops and nothing decays in the state; a run of uninformative observations thus costs the model none of its memory. When $f(cv) = cf(v)$ for $c > 0$, the clock is unchanged by reparametrisation of the path (Proposition 5), and thus the layer is invariant by Proposition 4. This is a choice: invariance suits data whose observation times carry no information; a clock without it may be better when the timing is itself a signal.

The EWS layer can also be incorporated into an EWS block, similarly to how SSM layers are used in deep architectures (Gu et al., 2022b; Rusch & Rus, 2025; Terzić et al., 2026). Each block applies an EWS layer, whose state is normalised at each depth separately and read down to the stream width by a linear map, followed by a non-linear transformation (Appendix F.1). After several blocks, the sequence is decoded to produce the output.

Proposition 1 makes the EWS layer a SLiCE with vector fields generated by A and B and driven by the clock. Three properties follow from this structure rather than from training: *(a) Parameter count.* An EWS layer has $\mathcal{O}(m^2 + dm)$ generator and lift parameters at any depth, and $\mathcal{O}(m + dm)$ for the diagonal, rotation and Jordan families of Table 1. Among the maximally expressive structures in Walker et al. (2025), the cheapest has $\mathcal{O}(dK)$ parameters at state size $K = \sum_{n \leq N} m^n$, whereas the EWS generates all K coordinates from A and B; readout parameters are linear in K for either model. *(b) Closure.* By Theorem 1(i) every degree N polynomial in depth-one coordinates is linear in the EWS truncated at depth N, whereas the coordinates of a free linear CDE carry no such relation. *(c) Scale.* When the generator is positive definite, and the clock dominates the path (e.g. variation clock), depth n of the EWS is bounded by $C^n/n!$ on every interval, with C independent of the length of the path (Corollary 1), whereas a free linear CDE can grow exponentially in the path's total variation. Together with the meaning of its coordinates, the spectral filters of Section 2.1 and their products, these properties are the inductive bias of the layer.

The layer is computed as a parallel associative scan over the output grid. Each interval $[t_j, t_{j+1})$ contributes the pair $(\theta_{t_{j+1}} - \theta_{t_j}, S_{\mathbb{A}}(X;\theta)_{t_j,t_{j+1}})$ of its clock increment and EWS, and Theorem 1(ii) makes $(r_1, S_1) \star (r_2, S_2) = \big(r_1 + r_2,\ D_A^{r_2}(S_1) \otimes S_2\big)$ associative, so the output state is obtained by scanning $\star$ over these pairs. The EWS over each interval is the signature of some re-weighted path (Theorem 1(i)), which can be computed exactly at any depth. The path $X_t = \int_0^t x_s ds$ is taken linear between observations, which is the continuous-time counterpart of the zero-order hold used in state-space models (Gu & Dao, 2024), and the layer is causal, as the state depends only on prior observations.

3 EXPERIMENTS AGAINST THE SIGNATURE & EFM

Before comparing the EWS with SSMs, we verify that the properties established in Section 2 are achieved on synthetic tasks, each of which has the same form: a path X is generated, a target path $Y_t = F(X)_t$ is computed, and a model is trained to regress Y_t from observations of $X|_{[0,t]}$. Since F is known exactly, Section 2 tells us which generators and truncation depths represent it.

We compare the EWS with the simpler signature and EFM. All three are the layer of (6) with the time clock and linear readout. Within a task, they differ only in the generator structure, which is zero for the signature, positive diagonal for EFM (with learned parameters rather than fixed ones as in Abi Jaber & Sotnikov (2025)), and from the rotation family or a free matrix for the EWS. Each model (A, B) is identically trained over the same grid of truncation depths and widths; we report test error of the best validation at each depth.

We consider two tasks, with different F. Full details, and a higher-order expressivity task, are given in Appendix F.3. The first is oscillatory, and has F be one component of an SDE driven by $X_t = (t, W_t)$ for W a 2D Brownian path. The second tests whether the complex spectrum and deeper levels of the EWS are necessary; its path is $X_t = (t, \int_0^t x_s ds)$, where x is a sum of one or two persistent low-frequency tones and up to two high-frequency tones that switch on in bursts. We estimate three targets: (i) a causal spectrogram with frequencies $\{\omega_j\}_{j=1}^4$ and rate a, $F_1^j(t) = (\int_0^t e^{-a(t-u)} \cos(\omega_j(t-u)) x_u du, \int_0^t e^{-a(t-u)} \sin(\omega_j(t-u)) x_u du)$. By Section 2.1, F_1 is a linear readout of depth one of an EWS, but cannot be given by a linear readout at depth one with a real diagonal generator. (ii) $F_2(t) = \sum_{j \text{ high}} \|F_1^j(t)\|^2$, so F_2 is the energy in the high band. This is a quadratic in the depth-one coordinates so, by Theorem 1, a linear combination of level-two coordinates; no linear readout of depth one works. (iii) $F_3(t)$ is binary, equal to one if a high-frequency tone is on at time t.

On the oscillator, the EWS is $4.6\times$ and $6.3\times$ better than the EFM and signature respectively, and its learned generator carries a complex pair on every seed, which the EFM cannot represent. On F_1 the EWS reaches the target at depth one, whereas the EFM recovers the low band but not the high one at any depth or width. On F_2 and F_3 the EWS fails at depth one at every width and succeeds at depth two, while the EFM and signature fail at every depth. In addition, thresholding the energy F_2 recovers F_3 with only 81% accuracy. Linear readouts therefore need depth on these targets.

Table 2: Synthetic targets evaluated at depths $n = 1$ and $n = 2$, mean $\pm$ standard deviation at the best validation width. Oscillator task reports test RMSE on targets scaled to $[0, 1]$ over ten seeds, F_1 and F_2 normalised test MSE, and F_3 test accuracy over three seeds (majority-class accuracy 69.5%).

	Signature		EFM		EWS	
Target	n=1	n=2	n=1	n=2	n=1	n=2
Oscillator	0.135±0.005	0.127±0.005	0.0922±0.0034	0.0936±0.0046	**0.0201±0.0014**	0.0257±0.0026
F_1 coeff.	0.940±0.000	0.881±0.000	0.512±0.002	0.448±0.004	$(\mathbf{3.5}\pm\mathbf{1.2}) \times \mathbf{10^{-5}}$	$(6.6\pm3.7) \times 10^{-5}$
F_2 high-band	0.990±0.001	0.906±0.001	0.908±0.001	0.909±0.004	0.921±0.022	$(\mathbf{7.3}\pm\mathbf{3.4}) \times \mathbf{10^{-4}}$
F_3 burst (%)	69.5±0.0	69.2±0.0	68.7±0.0	69.0±0.0	69.5±0.0	**94.7±0.2**

4 SSMS ARE DEPTH-ONE EWS

We now demonstrate a close connection between SSMs and the depth-one terms in the EWS. Consider first the continuous-time dynamics of an LTI SSM, $\dot{h}_t = -Ah_t + Bx_t$ with $h_0 = 0$ and $y_t = Ch_t$. Let $X_t = \int_0^t x_s ds$. The solution $h_t = \int_0^t e^{-(t-s)A} B x_s ds = S_{\mathbb{A}}(X;\theta)_{0,t}^{(1)}$ is the depth-one component of the EWS of X with the time clock. Thus, the dynamics underlying the SSM are those of a depth-one EWS with generator A and lift B. At depth one, the modified Chen identity of Theorem 1(ii) reads

$$S_{\mathbb{A}}(X;\theta)_{0,t}^{(1)} = e^{-(t-s)A} S_{\mathbb{A}}(X;\theta)_{0,s}^{(1)} + S_{\mathbb{A}}(X;\theta)_{s,t}^{(1)}, \qquad 0 \le s \le t. \tag{7}$$

On a grid $t_j = j\Delta$, with X linear between observations, writing $G_\Delta(A) = \int_0^\Delta e^{-sA} ds$ and $S_j = S_{\mathbb{A}}(X;\theta)^{(1)}_{0,t_j}$, we see that (7) becomes

$$S_j = e^{-\Delta A} S_{j-1} + G_\Delta(A) B x_j. \tag{8}$$

This is the same form as the recurrence $h_j = \bar{A} h_{j-1} + \bar{B} x_j$ that SSMs compute, with $\bar{A} = e^{-\Delta A}$ and $\bar{B} = G_\Delta(A)B$, the zero-order hold. Matching coefficients in the other direction, a trained LTI SSM storing $(\bar{A}, \bar{B})$ and step Δ is reproduced at every observation by an EWS with

$$A = -\tfrac{1}{\Delta} \log(\bar{A}), \qquad B = G_\Delta(A)^{-1} \bar{B}, \tag{9}$$

whenever $\bar{A}$ has a real matrix logarithm. This holds for every model in Table 3, except some D-LinOSS blocks (Appendix E). Appendix E.4 verifies this numerically, with random and trained parameters. Unlike an SSM, the EWS never discretises its dynamics at depth one. The input x is observed at discrete times, yet the second term of (7) requires it on the whole interval. The EWS holds x constant between observations, the rectilinear embedding (Morrill et al., 2022; Walker et al., 2026), so that X is piecewise linear. The zero-order hold of S4D and S5 yields the same formula, whereas every other rule used by popular SSMs approximates it to 2nd order for S4 and Mamba-3 and to 1st order for LinOSS, Mamba and Mamba-2 (Lahoti et al., 2026). We note that Theorem 1(i) demonstrates that deeper levels of the EWS are not the signature of the latent state of the SSM $t \mapsto h_t$. A selective SSM computes its step Δ_t from its input. The accumulated step $\theta_t = \sum_{s \le t} \Delta_s$

Table 3: State-space models as depth-one EWS, with the block and spectrum of the EWS generator reproducing each trained model (unit clock for LTI models, accumulated step for Mamba). The μ_k are the eigenvalues of $\Lambda - PQ^*$, and γ_k, ω_k, ψ_k, φ_k are defined in Appendix E.

Model	Stores	Discretisation	Block	Spectrum of A	Route
S4D (Gu et al., 2022a)	a_k, step Δ_c	ZOH	2×2 rotation	$\Delta_c(\gamma_k \pm i\omega_k)$	read-off
S5 (Smith et al., 2023)	Λ_k, step Δt_k	ZOH	2×2 rotation	$\Delta t_k(\gamma_k \pm i\omega_k)$	read-off
S4 (Gu et al., 2022b)	$\Lambda - PQ^*$, step Δ_c	bilinear	dense, diagonalisable	$-2\,\mathrm{artanh}(\Delta_c \mu_k/2)$	(9)
LRU (Orvieto et al., 2023)	$e^{-\nu_k + i\vartheta_k}$	none	2×2 rotation	$\nu_k \pm i\vartheta_k$	(9)
LinOSS-IM (Rusch & Rus, 2025)	$\mathsf{A}_k \ge 0$, step Δt_k	implicit	2×2 oscillator	$\frac{1}{2}\log(1 + \Delta t_k^2 \mathsf{A}_k) \pm i\psi_k$	(9)
D-LinOSS (Boyer et al., 2025)	A_k, G_k, step Δt_k	IMEX	2×2 oscillator	$\frac{1}{2}\log(1 + \Delta t_k \mathsf{G}_k) \pm i\varphi_k$	(9)
Mamba-1 (Gu & Dao, 2024)	a_k, step Δ_t	exp. Euler	scalar	a_k, per channel	clocked
Mamba-2 (Dao & Gu, 2024)	a, step Δ_t	exp. Euler	scalar	a, per head	clocked

is thus a clock as in Definition 1. The selective transition $e^{-\Delta_t A}$ can be rewritten as $e^{-(\theta_t - \theta_{t-1})A}$, and so (7) applies, with $\theta_t - \theta_s$ instead of $t - s$. With this clock, the state of each channel of Mamba-1 is a depth-one EWS with a real diagonal generator, driven by a path with increments $G_{\Delta_t}(A)^{-1} \Delta_t^2 B_t x_t$. The state of each head of Mamba-2 is the same construction, with generator aI and increments $G_{\Delta_t}(a)^{-1} \Delta_t^2 B_t x_t^\top$. A selective layer is thus a family of depth-one EWS layers, one per channel or head, over paths built by its selection mechanism.

The state of every model in Table 3 is contained in the EWS state, which extends via depth and flexibility in the clock and generator. Deeper levels hold weighted products of increments at different times, which SSMs reach only by stacking layers with non-linearities between them (Orvieto et al., 2024; Wang & Xue, 2023). The clock may be any causal functional, including path variation. The generator may be any matrix; up to a change of basis absorbed by the lift and readout, the diagonal and rotation families of Table 1 cover every model in Table 3 and the block of Mamba-3, while its Jordan blocks, and the gamma memories they carry (Proposition 2), are more general (reached by SSMs only in the non-learnable HiPPO-LagT matrix (Gu et al., 2020) and edge cases of LinOSS and D-LinOSS, Appendix E). Retention (Sun et al., 2023) and the Mamba-2 SSD layer are also depth-one EWS whose path is a sequence of key–value outer products; a rotary embedding (Su et al., 2024) is a rotation block with angle set by the clock.

5 Experiments

5.1 Long time-series classification

We consider the long-range benchmark of Walker et al. (2024), six datasets from the UEA multivariate time-series classification archive (Bagnall et al., 2018; Middlehurst et al., 2026) with sequence

lengths from 405 to 17 984 observations. We follow the training procedure of Walker et al. (2024), with the per-dataset selection of Appendix F.4. The SSM baselines are stacks of two to six blocks with non-linearities between them. The EWS is a single layer with linear readout (6), so that its only non-linearity is from depth and the clock. In Table 4, the EWS performs best on accuracy and rank, while using on average $3.8k$ parameters per dataset, fewer than any baseline with a known count. The largest margin is EthanolConcentration, where the selected configuration is a rotation generator with the time clock and depth two. Excluding this dataset, the EWS keeps the best average rank and is within 0.1 points of the best average accuracy. Validation selects a truncation depth of two or three on every dataset, which raises the selected configuration's test accuracy over depth one on five of the six datasets, including by eight points on EigenWorms, and the ablation of Appendix F.4 largely supports the benefits of depth. Training time per step is within the range of baselines on five datasets (Appendix F.4).

Table 4: Test accuracy (%) on six UEA datasets, mean ± standard deviation over five seeds, with the average and average rank. Baselines taken from Walker et al. (2024; 2025); Boyer et al. (2025); Terzić et al. (2026).

Model	EigenWorms	Ethanol	Heartbeat	Motor	SCP1	SCP2	Avg.	Rank
Log-NCDE	85.6±5.1	34.4±6.4	75.2±4.6	53.7±5.3	83.1±2.8	53.7±4.1	64.3	7.5
S5	81.1±3.7	24.1±4.3	77.7±5.5	47.7±5.5	**89.9±4.6**	50.5±2.6	61.8	7.83
S6	85.0±16.1	26.4±6.4	76.5±8.3	51.3±4.7	82.8±2.7	49.9±9.4	62.0	9.17
Mamba	70.9±15.8	27.9±4.5	76.2±3.8	47.7±4.5	80.7±1.4	48.2±3.9	58.6	10.33
DE-SLiCE	88±4	26±5	74±5	50±6	82±4	50±3	61.6	9.67
D-DE-SLiCE	86±6	27±7	74±4	55±4	85±4	52±5	63.0	7.83
BD-SLiCE	86±4	29±7	77±6	53±3	85±2	54±8	64.0	6.50
LinOSS-IM	**95.0±4.4**	29.9±0.6	75.8±3.7	60.0±7.5	87.8±2.6	58.2±6.9	67.8	3.67
D-LinOSS	93.9±3.2	29.9±0.6	75.8±4.9	**61.1±2.0**	88.9±3.0	**58.6±2.3**	68.0	3.17
PD-SSM	90.0±5.7	34.7±4.0	**80.0±2.6**	60.0±3.7	80.9±2.0	56.1±8.6	67.0	4.50
FLASH PD-SSM	93.8±1.4	43.7±2.9	77.2±6.6	57.9±6.0	86.1±2.8	56.1±3.1	69.1	3.83
EWS	93.3±1.4	**71.9±4.8**	**80.0±4.5**	**61.1±5.0**	87.1±2.5	56.5±5.8	**75.0**	**2.33**

5.2 Regression & forecasting

We consider a regression and a forecasting task at depth one, where the EWS is a selective SSM on the first (clamped clock) and an LTI SSM on the second (time clock). The first is PPG-DaLiA (Reiss et al., 2019), in which heart rate is regressed from six channels, in windows of 49 920 observations. The second is the weather task of Zhou et al. (2021), in which one month of hourly climatological data (720 observations) is predicted from the previous month, framed as a masked sequence-to-sequence task as in Rusch & Rus (2025); Boyer et al. (2025); Gu et al. (2022b), with our hyperparameter search given in Appendix F.5. Table 5 displays the results. On PPG-DaLiA, with $13\times$ fewer parameters than D-LinOSS, the EWS has the lowest mean error reported. On weather it also has the lowest reported error, using a single EWS block with a Jordan generator where the next most accurate, LinOSS-IMEX, stacks five. PPG-DaLiA uses four blocks; both configurations, with training time and memory for weather, are in Appendix F.5.

Table 5: Regression and forecasting. Left, PPG-DaLiA test MSE ($\times 10^{-2}$), mean ± standard deviation over five seeds. Right, weather test error at a horizon of 720 steps. Baselines are taken from Walker et al. (2024); Rusch & Rus (2025); Boyer et al. (2025); Terzić et al. (2026); Zhou et al. (2021); Gu et al. (2022b). On weather, EWS and D-LinOSS are five seed mean ± standard deviation; others are single seed.

PPG-DaLiA

Model	MSE	Params.
NCDE	13.54±0.69	66k
Mamba	10.65±2.20	797k
Log-NCDE	9.56±0.59	76k
LinOSS-IMEX	7.5±0.46	26k
FLASH PD-SSM	6.5±0.35	–
LinOSS-IM	6.4±0.23	29k
D-LinOSS	6.16±0.73	151k
EWS	**5.90±0.73**	**11**k

Weather

Model	MAE	Params.
Informer	0.731	19.5M
LSTnet	0.757	–
S4	0.578	≈ 100k
LinOSS-IM	0.528	136k
LinOSS-IMEX	0.508	1.3M
D-LinOSS	0.538 ± 0.052	301k
EWS	**0.473 ± 0.002**	**19**k

5.3 State Tracking

The formal language benchmark of Delétang et al. (2023) organises language tasks by the Chomsky hierarchy (Chomsky, 1956). We use the four regular tasks, which can be solved by sequentially updating a fixed set of internal states without external memory, i.e. by state tracking; the tasks are described in Appendix F.6. Recurrent models such as LSTMs (Hochreiter & Schmidhuber, 1997) and xLSTMs (Beck et al., 2024) perform well, while parallel-in-time models generally struggle (Merrill et al., 2024) unless their transition matrices can have negative eigenvalues (Grazzi et al., 2025). Following Walker et al. (2025), models are trained on lengths 3–40 and evaluated on 40–256, using two stacked blocks as in Beck et al. (2024). The EWS uses depth one in both blocks, with a gated or clamped clock in the first and variation in the second. Generators come from the rotation family, with some parametrised as skew-symmetric matrices. The lower rows of Table 6 ablate these choices by replacing the clocks with the time clock and the generators with real diagonal ones, keeping other hyperparameters fixed; replacing both gives the learned EFM. Among parallel-in-time models, EWS achieves the highest average, solving three tasks and trailing only D-DE-SLiCE on modular arithmetic. Replacing the clocks with the time clock reduces all tasks to the level of the LTI models. When the gate returns zero the state is not rotated, and otherwise the rotation is applied once, so the total rotation depends on which tokens occur rather than on sequence length alone. For parity, a rotation by π at each one tracks whether the number of ones is odd, which no function of length can recover. Unlike the rotations of Mamba-3, which always come with a decay set by the same step, an EWS rotation block can turn without decaying and keeps this parity over long inputs. With real diagonal generators, whose transitions have positive real eigenvalues, three tasks fail while even pairs is largely unaffected, consistent with Grazzi et al. (2025).

Table 6: Formal language tasks. Validation accuracy (%) on sequence lengths 40–256, over five seeds for EWS and three for its ablations. Baselines are taken from Walker et al. (2025); Terzić et al. (2026).

Model	Cycle nav.	Even pairs	Mod. arith.	Parity	Avg.
LSTM	100.0±0.0	100.0±0.0	99.9±0.1	100.0±0.0	100.0
xLSTM[1:1]	53.5±5.6	99.0±1.9	29.3±1.6	100.0±0.0	70.5
Transformer	24.4±0.5	90.4±10.4	23.6±0.7	52.2±0.4	47.7
S4D	23.7±1.1	68.7±4.7	21.7±0.4	51.2±1.0	41.3
Mamba	48.4±2.2	**100.0±0.0**	33.1±6.6	54.2±2.1	58.9
Mamba-3	69.0±11.7	**100.0±0.0**	46.8±5.5	76.5±7.8	73.0
DeltaNet$[-1, 1]$	46.7±6.1	**100.0±0.0**	66.4±8.8	97.7±2.0	77.7
Gated DeltaProduct$[-1, 1]$	46.3±6.6	**100.0±0.0**	78.4±10.9	98.0±1.4	80.7
BD-SLiCE	99.8±0.2	85.9±11.3	54.0±12.5	95.3±3.9	83.8
D-DE-SLiCE	73.3±29.4	84.8±8.5	**98.4±0.7**	83.8±11.3	85.1
PD-SSM	99.5±0.7	99.7±0.3	96.2±3.4	99.9±0.1	98.8
FLASH PD-SSM	92.7±1.9	98.8±0.8	94.7±2.9	99.0±1.5	96.3
EWS	**99.99±0.01**	**100.0±0.0**	96.7±0.8	**99.99±0.01**	**99.2**
EWS, time clock	28.9±1.6	58.1±1.3	20.7±0.4	53.1±0.8	40.2
EWS, real diagonal generator	52.1±3.2	99.6±0.3	39.7±3.2	54.9±1.0	61.6
EWS, both (learned EFM)	50.2±9.5	90.9±2.0	*diverged*	54.7±0.6	–
Random	20.0	50.0	20.0	50.0	35.0

6 Conclusions

This paper introduced the exponentially weighted signature, a continuous-time model that generalises both the signature and state-space models. It keeps the group-like structure and universality of the signature and satisfies a modified Chen identity that enables a parallel associative scan. At depth one, it is an SSM to which trained LTI SSMs, and each channel of Mamba and head of Mamba-2, convert in closed form. The EWS extends these models through its truncation depth, clock and generator, each of which was shown to matter empirically. A complex spectrum was necessary on oscillatory targets, while rotation and Jordan generators were selected by validation on several datasets. State tracking was lost when the learned clocks were replaced by the time clock, and depth above one was selected on every classification dataset, where the EWS achieved the highest average accuracy and rank. Its limitations, including the memory cost of deeper levels, are discussed in Appendix A.

REPRODUCIBILITY STATEMENT

Our JAX implementation of the EWS layer and block, together with the code for all experiments and ablations, is provided as supplementary material. Proofs of all EWS related statements are given in Appendix C, while Appendix E derives each SSM-to-EWS conversion and verifies it numerically. Appendix F describes each experiment, including data sources, preprocessing procedures, data splits and hyperparameter configurations. All datasets are freely available for use, TUAB on request under its data use agreement.

AI USE STATEMENT

We used generative AI tools as set out below, all of whose output has been reviewed. **Code.** LLMs were used to clean, refactor and speed up our code. They were also used to generate the synthetic datasets required for the experiments of Section 3. All AI generated code was tested, and its outputs compared to the original code. **Research.** LLMs were used to find some relevant literature and adjacent work. **Writing.** LLMs were used to correct typographical errors, grammar and to format tables and figures. These tools were also used to help cut the text down to fit within the page limit and to give feedback on the readability of drafts of the paper. **Mathematics.** LLMs were used to help make proofs more accessible to readers without a background in signatures, check that their logic holds, and identify edge cases. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

REFERENCES

Eduardo Abi Jaber and Dimitri Sotnikov. Exponentially fading memory signature, 2025. URL https://arxiv.org/abs/2507.03700.

Takuya Akiba, Shotaro Sano, Toshihiko Yanase, Takeru Ohta, and Masanori Koyama. Optuna: A next-generation hyperparameter optimization framework. In *The 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 2623–2631, 2019.

Anthony Bagnall, Hoang Anh Dau, Jason Lines, Michael Flynn, James Large, Aaron Bostrom, Paul Southam, and Eamonn Keogh. The UEA multivariate time series classification archive, 2018, 2018. URL https://arxiv.org/abs/1811.00075.

Maximilian Beck, Korbinian Pöppel, Markus Spanring, Andreas Auer, Oleksandra Prudnikova, Michael Kopp, Günter Klambauer, Johannes Brandstetter, and Sepp Hochreiter. xLSTM: Extended long short-term memory. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 107547–107603. Curran Associates, Inc., 2024. doi: 10.52202/079017-3417. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/c2ce2f2701c10a2b2f2ea0bfa43cfaa3-Paper-Conference.pdf.

James Bergstra, Rémi Bardenet, Yoshua Bengio, and Balázs Kégl. Algorithms for hyper-parameter optimization. In J. Shawe-Taylor, R. Zemel, P. Bartlett, F. Pereira, and K. Weinberger (eds.), *Advances in Neural Information Processing Systems*, volume 24. Curran Associates, Inc., 2011. URL https://proceedings.neurips.cc/paper_files/paper/2011/file/86e8f7ab32cfd12577bc2619bc635690-Paper.pdf.

Guy E. Blelloch. Prefix sums and their applications. Technical Report CMU-CS-90-190, School of Computer Science, Carnegie Mellon University, November 1990.

Horatio Boedihardjo, Xi Geng, Terry Lyons, and Danyu Yang. The signature of a rough path: Uniqueness. *Advances in Mathematics*, 293:720–737, 2016.

S. Boyd and L. Chua. Fading memory and the problem of approximating nonlinear operators with Volterra series. *IEEE Transactions on Circuits and Systems*, 32(11):1150–1161, 1985. doi: 10.1109/TCS.1985.1085649.

Jared Boyer, T. Konstantin Rusch, and Daniela Rus. Learning to dissipate energy in oscillatory state-space models, 2025. URL https://arxiv.org/abs/2505.12171.

James Bradbury, Roy Frostig, Peter Hawkins, Matthew James Johnson, Yash Katariya, Chris Leary, Dougal Maclaurin, George Necula, Adam Paszke, Jake VanderPlas, Skye Wanderman-Milne, and Qiao Zhang. JAX: composable transformations of Python+NumPy programs, 2018. URL http://github.com/jax-ml/jax.

Kuo Tsai Chen. Iterated integrals and exponential homomorphisms. *Proceedings of the London Mathematical Society*, s3-4(1):502–512, 1954.

Kuo-Tsai Chen. Integration of paths, geometric invariants and a generalized Baker–Hausdorff formula. *Annals of Mathematics*, 65:163–178, 1957.

N. Chomsky. Three models for the description of language. *IRE Transactions on Information Theory*, 2(3):113–124, 1956. doi: 10.1109/TIT.1956.1056813.

Nicola Muca Cirone, Antonio Orvieto, Benjamin Walker, Cristopher Salvi, and Terry Lyons. Theoretical foundations of deep selective state-space models. In *Proceedings of the 38th Conference on Neural Information Processing Systems (NeurIPS)*, 2024.

Samuel N. Cohen, Giulia Mantoan, Lars Nesheim, Áureo de Paula, Arthur Turrell, and Lingyi Yang. Nowcasting using regression on signatures, 2025. URL https://arxiv.org/abs/2305.10256.

Walter J. Culver. On the existence and uniqueness of the real logarithm of a matrix. *Proceedings of the American Mathematical Society*, 17(5):1146–1151, 1966.

Tri Dao and Albert Gu. Transformers are SSMs: Generalized models and efficient algorithms through structured state space duality. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=ztn8FCR1td.

Bert de Vries and José C. Príncipe. The gamma model—a new neural model for temporal processing. *Neural Networks*, 5(4):565–576, 1992. ISSN 0893-6080. doi: 10.1016/S0893-6080(05)80035-8. URL https://www.sciencedirect.com/science/article/pii/S0893608005800358.

Grégoire Delétang, Anian Ruoss, Jordi Grau-Moya, Tim Genewein, Li Kevin Wenliang, Elliot Catt, Chris Cundy, Marcus Hutter, Shane Legg, Joel Veness, and Pedro A Ortega. Neural networks and the Chomsky hierarchy. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WbxHAzkeQcn.

Nina Drobac, Margaux Brégère, Joseph de Vilmarest, and Olivier Wintenberger. Sliding-window signatures for time series: Application to electricity demand forecasting, 2025. URL https://arxiv.org/abs/2510.12337.

Peter K. Friz and Nicolas B. Victoir. *Multidimensional Stochastic Processes as Rough Paths: Theory and Applications*. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2010. ISBN 9780521876070.

Karan Goel, Albert Gu, Chris Donahue, and Christopher Ré. It's raw! Audio generation with state-space models. In Kamalika Chaudhuri, Stefanie Jegelka, Le Song, Csaba Szepesvári, Gang Niu, and Sivan Sabato (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 7616–7633. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/goel22a.html.

Benjamin Graham. Sparse arrays of signatures for online character recognition, 2013. URL https://arxiv.org/abs/1308.0371.

Riccardo Grazzi, Julien Siems, Jörg K.H. Franke, Arber Zela, Frank Hutter, and Massimiliano Pontil. Unlocking state-tracking in linear RNNs through negative eigenvalues. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=UvTo3tVBk2.

Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. In *Proceedings of the 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 H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 1474–1487. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/102f0bb6efb3a6128a3c750dd16729be-Paper.pdf.

Albert Gu, Karan Goel, Ankit Gupta, and Christopher Ré. On the parameterization and initialization of diagonal state space models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 35971–35983. Curran Associates, Inc., 2022a. doi: 10.52202/068431-2607. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/e9a32fade47b906de908431991440f7c-Paper-Conference.pdf.

Albert Gu, Karan Goel, and Christopher Ré. Efficiently modeling long sequences with structured state spaces. In *Proceedings of The 10th International Conference on Learning Representations (ICLR)*, 2022b.

Albert Gu, Isys Johnson, Aman Timalsina, Atri Rudra, and Christopher Ré. How to train your HiPPO: State space models with generalized orthogonal basis projections. In *International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=klK17OQ3KB.

Lajos Gergely Gyurkó, Terry Lyons, Mark Kontkowski, and Jonathan Field. Extracting information from the signature of a financial data stream. *arXiv preprint arXiv:1307.7244*, 2014.

Paul P. Hager, Fabian N. Harang, Luca Pelizzari, and Samy Tindel. The Volterra signature, 2026. URL https://arxiv.org/abs/2603.04525.

B. Hambly and T. Lyons. Uniqueness for the signature of a path of bounded variation and the reduced path group. *Annals of Mathematics*, 171:109–167, 2010.

Fabian Harang, Samy Tindel, and Xiaohua Wang. Volterra equations driven by rough signals 3: Probabilistic construction of the Volterra rough path for fractional Brownian motions. *Journal of Theoretical Probability*, 37(1):307–351, 2024. ISSN 1572-9230. doi: 10.1007/s10959-023-01251-y. URL https://doi.org/10.1007/s10959-023-01251-y.

Fabian A. Harang and Samy Tindel. Volterra equations driven by rough signals. *Stochastic Processes and their Applications*, 142:34–78, 2021. ISSN 0304-4149. doi: 10.1016/j.spa.2021.08.001. URL https://www.sciencedirect.com/science/article/pii/S0304414921001253.

Fabian A. Harang, Samy Tindel, and Xiaohua Wang. Volterra equations driven by rough signals 2: Higher-order expansions. *Stochastics and Dynamics*, 23(01):2350002, 2023. doi: 10.1142/S0219493723500028. URL https://doi.org/10.1142/S0219493723500028.

Gyuryang Heo, Timothy Ngotiaoco, Kazuki Irie, Samuel J. Gershman, and Bernardo L. Sabatini. Why depth matters in parallelizable sequence models: A Lie algebraic view, 2026. URL https://arxiv.org/abs/2603.05573.

Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural Computation*, 9(8):1735–1780, 1997. doi: 10.1162/neco.1997.9.8.1735.

Wei-Bang Jiang, Li-Ming Zhao, and Bao-Liang Lu. Large brain model for learning generic representations with tremendous EEG data in BCI. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=QzTpTRVtrP.

Jin Jing, Wendong Ge, Shenda Hong, Marta Bento Fernandes, Zhen Lin, Chaoqi Yang, Sungtae An, Aaron F. Struck, et al. Development of expert-level classification of seizures and rhythmic and periodic patterns during EEG interpretation. *Neurology*, 100(17):e1750–e1762, 2023. doi: 10.1212/WNL.0000000000207127.

Patrick Kidger. *On Neural Differential Equations*. PhD thesis, University of Oxford, 2022.

Patrick Kidger and Cristian Garcia. Equinox: neural networks in JAX via callable PyTrees and filtered transformations. *Differentiable Programming workshop at Neural Information Processing Systems 2021*, 2021.

Patrick Kidger, Patric Bonnier, Imanol Perez Arribas, Cristopher Salvi, and Terry Lyons. Deep signature transforms. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/d2cdf047a6674cef251d56544a3cf029-Paper.pdf.

Patrick Kidger, James Morrill, James Foster, and Terry Lyons. Neural Controlled Differential Equations for Irregular Time Series. In *Proceedings of the 34th Conference on Neural Information Processing Systems (NeurIPS)*, 2020.

Franz J. Király and Harald Oberhauser. Kernels for sequentially ordered data. *Journal of Machine Learning Research*, 20(31):1–45, 2019.

Aakash Lahoti, Kevin Li, Berlin Chen, Caitlin Wang, Aviv Bick, J Zico Kolter, Tri Dao, and Albert Gu. Mamba-3: Improved sequence modeling using state space principles. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=HwCvaJOiCj.

Serge Lang. *Algebra*, volume 211 of *Graduate Texts in Mathematics*. Springer, 3rd edition, 2002. ISBN 978-0-387-95385-4.

Vernon J Lawhern, Amelia J Solon, Nicholas R Waytowich, Stephen M Gordon, Chou P Hung, and Brent J Lance. EEGNet: a compact convolutional neural network for EEG-based brain–computer interfaces. *Journal of Neural Engineering*, 15(5):056013, 2018. ISSN 1741-2552. doi: 10.1088/1741-2552/aace8c. URL http://dx.doi.org/10.1088/1741-2552/aace8c.

Daniel Levin, Terry Lyons, and Hao Ni. Learning from the past, predicting the statistics for the future, learning an evolving system. *arXiv preprint arXiv:1309.0260*, 2016.

Silvia Lopez, Gabriella Suarez, David Jungreis, Iyad Obeid, and Joseph Picone. Automated identification of abnormal adult EEGs. In *2015 IEEE Signal Processing in Medicine and Biology Symposium (SPMB)*, pp. 1–5. IEEE, 2015. doi: 10.1109/SPMB.2015.7405423.

T. Lyons, M. Caruana, and T. Lévy. *Differential Equations Driven by Rough Paths: École D'été de Probabilités de Saint-Flour XXXIV-2004*, volume 1908 of *Lecture Notes in Mathematics*. Springer, 2007.

Terry J. Lyons. Differential equations driven by rough signals. *Revista Matemática Iberoamericana*, 14(2):215–310, 1998. ISSN 0213-2230.

Stéphane Mallat. *A Wavelet Tour of Signal Processing: The Sparse Way*. Academic Press, Burlington, MA, 3rd edition, 2009. ISBN 978-0-12-374370-1.

Eric Martin and Chris Cundy. Parallelizing linear recurrent neural nets over sequence length. In *International Conference on Learning Representations*, 2018. URL https://openreview.net/forum?id=HyUNwulC-.

Andrew McLeod and Terry Lyons. Signature methods in machine learning. *EMS Surveys in Mathematical Sciences*, 2025.

William Merrill, Jackson Petty, and Ashish Sabharwal. The illusion of state in state-space models. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org, 2024.

Matthew Middlehurst, Aiden Rushbrooke, Ali Ismail-Fawaz, Maxime Devanne, Germain Forestier, Angus Dempster, Geoffrey I. Webb, Christopher Holder, and Anthony Bagnall. The multiverse of time series machine learning: an archive for multivariate time series classification, 2026. URL https://arxiv.org/abs/2603.20352.

Takashi Morita. Emergence of the primacy effect in structured state-space models. In *The 17th Asian Conference on Machine Learning (Conference Track)*, 2025. URL https://openreview.net/forum?id=TaKBnslxnS.

James Morrill, Adeline Fermanian, Patrick Kidger, and Terry Lyons. A generalised signature method for multivariate time series feature extraction, 2021a. URL https://arxiv.org/abs/2006.00873.

James Morrill, Cristopher Salvi, Patrick Kidger, James Foster, and Terry Lyons. Neural rough differential equations for long time series. In *Proceedings of the 38th International Conference on Machine Learning (ICML)*, 2021b.

James Morrill, Patrick Kidger, Lingyi Yang, and Terry Lyons. On the choice of interpolation scheme for neural CDEs. *Transactions on Machine Learning Research*, 2022. ISSN 2835-8856. URL https://openreview.net/forum?id=caRBFhxXIG.

Eric Nguyen, Karan Goel, Albert Gu, Gordon Downs, Preey Shah, Tri Dao, Stephen Baccus, and Christopher Ré. S4ND: Modeling images and videos as multidimensional signals with state spaces. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho (eds.), *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=5WuQNQwy56M.

NVIDIA, Aaron Blakeman, Aaron Grattafiori, Aarti Basant, Abhibha Gupta, et al. NVIDIA Nemotron 3: Efficient and open intelligence, 2025. URL https://arxiv.org/abs/2512.20856.

Iyad Obeid and Joseph Picone. The Temple University Hospital EEG data corpus. *Frontiers in Neuroscience*, 10, 2016. ISSN 1662-453X. doi: 10.3389/fnins.2016.00196. URL https://www.frontiersin.org/journals/neuroscience/articles/10.3389/fnins.2016.00196.

Antonio Orvieto, Samuel L Smith, Albert Gu, Anushan Fernando, Çağlar Gülçehre, Razvan Pascanu, and Soham De. Resurrecting recurrent neural networks for long sequences. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 26670–26698. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/orvieto23a.html.

Antonio Orvieto, Soham De, Çağlar Gülçehre, Razvan Pascanu, and Samuel L Smith. Universality of linear recurrences followed by non-linear projections: Finite-width guarantees and benefits of complex eigenvalues. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 38837–38863. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/orvieto24a.html.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Kopf, Edward Yang, Zachary DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. PyTorch: An imperative style, high-performance deep learning library. In H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019. URL https://proceedings.neurips.cc/paper_files/paper/2019/file/bdbca288fee7f92f2bfa9f7012727740-Paper.pdf.

Rimhak Ree. Lie elements and an algebra associated with shuffles. *Annals of Mathematics*, 68(2): 210–220, 1958.

Attila Reiss, Ina Indlekofer, Philip Schmidt, and Kristof Van Laerhoven. Deep PPG: Large-scale heart rate estimation with convolutional neural networks. *Sensors*, 19(14), 2019. ISSN 1424-8220. doi: 10.3390/s19143079. URL https://www.mdpi.com/1424-8220/19/14/3079.

C. Reutenauer. *Free Lie Algebras*. London Mathematical Society monographs series: London Mathematical Society. Clarendon Press, 1993. ISBN 9780198536796.

Walter Rudin. *Functional Analysis*. McGraw-Hill, New York, 2nd edition, 1991.

T. Konstantin Rusch and Daniela Rus. Oscillatory state-space models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=GRMfXcAAFh.

Cristopher Salvi, Thomas Cass, James Foster, Terry Lyons, and Weixin Yang. The signature kernel is the solution of a Goursat PDE. *SIAM Journal on Mathematics of Data Science*, 3(3):873–899, January 2021. ISSN 2577-0187. doi: 10.1137/20m1366794. URL http://dx.doi.org/10.1137/20M1366794.

Julien Siems, Timur Carstensen, Arber Zela, Frank Hutter, Massimiliano Pontil, and Riccardo Grazzi. DeltaProduct: Improving state-tracking in linear RNNs via Householder products. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=SoRiaijTGr.

Jimmy T.H. Smith, Andrew Warrington, and Scott Linderman. Simplified state space layers for sequence modeling. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=Ai8Hw3AXqks.

Yonghao Song, Xueyu Jia, Lie Yang, and Longhan Xie. Transformer-based spatial-temporal feature learning for EEG decoding, 2021. URL https://arxiv.org/abs/2106.11170.

Jianlin Su, Murtadha Ahmed, Yu Lu, Shengfeng Pan, Wen Bo, and Yunfeng Liu. RoFormer: Enhanced transformer with rotary position embedding. *Neurocomputing*, 568(C), 2024. ISSN 0925-2312. doi: 10.1016/j.neucom.2023.127063. URL https://doi.org/10.1016/j.neucom.2023.127063.

Yutao Sun, Li Dong, Shaohan Huang, Shuming Ma, Yuqing Xia, Jilong Xue, Jianyong Wang, and Furu Wei. Retentive network: A successor to transformer for large language models, 2023. URL https://arxiv.org/abs/2307.08621.

Aleksandar Terzić, Nicolas Menet, Michael Hersche, Thomas Hofmann, and Abbas Rahimi. Structured sparse transition matrices to enable state tracking in state-space models. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=RDbuSCWhad.

Aleksandar Terzić, Francesco Carzaniga, Nicolas Menet, Yannick Biehl, Michael Hersche, Thomas Hofmann, and Abbas Rahimi. Flash PD-SSM: Memory-optimized structured sparse state-space models, 2026. URL https://arxiv.org/abs/2605.19150.

Benjamin Walker. *Advances in Neural Controlled Differential Equations*. PhD thesis, University of Oxford, 2025.

Benjamin Walker, Andrew D. McLeod, Tiexin Qin, Yichuan Cheng, Haoliang Li, and Terry Lyons. Log neural controlled differential equations: The Lie brackets make a difference. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024.

Benjamin Walker, Lingyi Yang, Nicola Muca Cirone, Cristopher Salvi, and Terry Lyons. Structured linear CDEs: Maximally expressive and parallel-in-time sequence models. In *Proceedings of the 39th Conference on Neural Information Processing Systems (NeurIPS)*, 2025.

Benjamin Walker, Alexandre Bloch, Lingyi Yang, Sam Morley, and Terry Lyons. Faithful embeddings of irregular and asynchronous data for online Log-NCDEs, 2026. URL https://arxiv.org/abs/2605.30213.

Jiquan Wang, Sha Zhao, Zhiling Luo, Yangxuan Zhou, Haiteng Jiang, Shijian Li, Tao Li, and Gang Pan. CBraMod: A criss-cross brain foundation model for EEG decoding. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=NPNUHgHF2w.

Shida Wang and Beichen Xue. State-space models with layer-wise nonlinearity are universal approximators with exponential decaying memory. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 74021–74038. Curran Associates, Inc., 2023. doi: 10.52202/075280-3239. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/ea8608c6258450e75b3443ec8022fb2e-Paper-Conference.pdf.

Chaoqi Yang, M Westover, and Jimeng Sun. BIOT: Biosignal transformer for cross-data learning in the wild. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 78240–78260. Curran Associates, Inc., 2023. doi: 10.52202/075280-3420. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/f6b30f3e2dd9cb53bbf2024402d02295-Paper-Conference.pdf.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=y8Rm4VNRPH.

Songlin Yang, Jan Kautz, and Ali Hatamizadeh. Gated delta networks: Improving Mamba2 with delta rule. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=r8H7xhYPwz.

L. C. Young. An inequality of the Hölder type, connected with Stieltjes integration. *Acta Mathematica*, 67:251–282, 1936.

Haoyi Zhou, Shanghang Zhang, Jieqi Peng, Shuai Zhang, Jianxin Li, Hui Xiong, and Wancai Zhang. Informer: Beyond efficient transformer for long sequence time-series forecasting. In *The Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Virtual Conference*, volume 35, pp. 11106–11115. AAAI Press, 2021.

Nicolas Zucchet and Antonio Orvieto. Recurrent neural networks: vanishing and exploding gradients are not the end of the story. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 139402–139443. Curran Associates, Inc., 2024. doi: 10.52202/079017-4425. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/fbb07254ef01868967dc891ea3fa6c13-Paper-Conference.pdf.

Appendix B collects the relevant mathematical tools needed to understand the main text: the tensor algebra, paths and their integration, the signature, and controlled differential equations. The theory underlying the EWS is somewhat involved, and Appendix C presents it in full, from the operators through which the generator acts on each tensor depth to the linear controlled differential equation it solves and the proofs of the results of Section 2. It also writes out the EWS in coordinates for small examples, which makes concrete what the EWS actually computes. This appendix is not essential for following the main text, and also works in greater generality. Where the main text takes paths in $\mathbb{R}^d$, the theory takes them in a general Banach space V, which adds no particular complexity over the Euclidean case. Where the main text takes paths of bounded variation, the theory takes paths of finite p-variation, which requires more care in both the statements and their proofs.

Appendix A discusses briefly the limitations of the EWS: its memory, its cost on densely sampled series, the automata a single layer can represent, and its scalar clock. Appendix D compares the EWS with two related objects, the exponentially fading memory signature and the Volterra signature, and in particular explains why the EWS is group-like while the Volterra signature is not. Appendix E gives the conversions of Section 4 in full, writing each SSM, from S4 to Mamba-2, as an EWS, and verifying the conversions numerically. Appendix F gives the details of every experiment, as well as an additional experiment on EEG abnormality detection and one on synthetic data.

A Limitations

Depth comes at the cost of state size, as depth n of a layer with width m has m^n coordinates. Thus, while the generator and lift parameter counts do not grow with depth, memory does, and

Table 17 shows the cost, most of which likely comes from our implementation rather than from the model. This is evidenced by the MotorImagery UEA task whose EWS's output state occupies roughly 1.7GB, an order of magnitude below the measured peak of 34.5GB. Our implementation is far from the hardware-aware, device-specific engineering behind the selective scan of selective SSMs such as Mamba (Gu & Dao, 2024), and so there is clear room for improvement. The state can also be compressed: since the EWS is group-like, much of this state is redundant, and log-signature coordinates would reduce depth two from m^2 to $m(m-1)/2$ coordinates and depth three from m^3 to $(m^3-m)/3$. Similarly to other SSMs, the EWS also suffers from having to compute its state at every observation, and is therefore costly on densely sampled time series. However, the EWS is the solution of a linear CDE; for such systems, the log-ODE method (Walker et al., 2024) can be used to compute updates over intervals containing several observations, saving both time and memory. An analogue of this method for the EWS is thus a natural next step.

The vector fields of the EWS's linear CDE generate a solvable Lie algebra (Remark 1), and a single layer therefore cannot represent a non-solvable automaton, although stacking layers approximates such dynamics (Heo et al., 2026). Finally, the clock is a single scalar, shared by all rotation and Jordan blocks of the generator, whereas Mamba-3 gives each rotation its own data-dependent angle. Extending to a vector-valued clock is future work. Together with a path and readout computed from the input, it would let a single EWS layer reproduce a full Mamba-1 or Mamba-2 layer.

B Background

B.1 The Tensor algebra

We first define the tensor algebra, the space in which the signature is defined. This requires recalling the definition of the tensor product of two vector spaces and establishing what norms are considered "admissible" on these spaces.

Definition 3 (Tensor Product (Lang, 2002))**.** Let U and V be vector spaces over a field $\mathbb{F}$. A tensor product of U and V is defined as a vector space $U \otimes V$ equipped with a bilinear map $\tau : U \times V \to U \otimes V$. This space is uniquely characterised (up to isomorphism) by the following universal property: for any $\mathbb{F}$-vector space W and any bilinear map $\kappa : U \times V \mapsto W$, there exists a unique linear map $\iota : U \otimes V \to W$ such that $\kappa = \iota \circ \tau$.

The tensor product is thus the space on which bilinear maps become linear, and we write $u \otimes v$ for the image of (u, v) under τ. For $U = \mathbb{R}^a$ and $V = \mathbb{R}^b$ it is the outer product, $(u \otimes v)_{ij} = u_i v_j$, so that $U \otimes V$ is the space $\mathbb{R}^{a \times b}$ of matrices and any bilinear $\kappa(u, v)$ is the linear map $\iota(u \otimes v)$. Iterating the construction gives the tensor powers $V^{\otimes n} = V \otimes V^{\otimes n-1}$, with $V^{\otimes 1} = V$ and $V^{\otimes 0} = \mathbb{R}$ by convention.

Definition 4 (Admissible Tensor Norms (McLeod & Lyons, 2025))**.** Given a Banach space V, a family of norms on $\{V^{\otimes n}\}_{n=1}^{\infty}$ is said to be admissible if for all integers $n \geq 1$ we have chosen a norm on $V^{\otimes n}$ such that the following conditions are satisfied:

1. For all $n \geq 1$, the norm $|| \cdot ||_{V^{\otimes n}}$ is invariant under the action of the symmetric group S_n on $V^{\otimes n}$. That is,
$$||\rho v||_{V^{\otimes n}} = ||v||_{V^{\otimes n}}, \quad \forall v \in V^{\otimes n}, \forall \rho \in S_n,$$
where $\rho(v_1 \otimes \cdots \otimes v_n) = v_{\rho(1)} \otimes \cdots \otimes v_{\rho(n)}$ for $v_i \in V$.

2. For all $n, m \geq 1$, $||v \otimes w||_{V^{\otimes(n+m)}} \leq ||v||_{V^{\otimes n}} ||w||_{V^{\otimes m}}$, $\forall v \in V^{\otimes n}$, $w \in V^{\otimes m}$. That is, the norm is sub-multiplicative.

3. For all $n, m \geq 1$ and for any dual elements $\phi \in (V^{\otimes n})^\star$ and $\sigma \in (V^{\otimes m})^\star$, we have $||\phi \otimes \sigma||_{(V^{\otimes(n+m)})^\star} \leq ||\phi||_{(V^{\otimes n})^\star} ||\sigma||_{(V^{\otimes m})^\star}$.

We are now able to define the tensor algebra, the space in which path signatures live.

Definition 5 (Tensor Algebra (Lyons et al., 2007))**.** Let $\{V^{\otimes n}\}_{n=0}^{\infty}$ be equipped with admissible norms in the sense of the above definition, and let $V^{\otimes 0} = \mathbb{R}$ by convention. The tensor algebra is the space
$$T((V)) = \{\mathbf{a} = (a_0, a_1, \dots) | \forall n \geq 0, a_n \in V^{\otimes n}\},$$

For two elements of $T((V))$, $\mathbf{a} = (a_0, a_1, \dots)$ and $\mathbf{b} = (b_0, b_1, \dots)$, addition is defined as

$$\mathbf{a} + \mathbf{b} = (a_0 + b_0, a_1 + b_1, \dots),$$

and product defined as

$$\mathbf{a} \otimes \mathbf{b} = (c_0, c_1, \dots),$$

where for each $n \geq 0$

$$c_n = \sum_{k=0}^{n} a_k \otimes b_{n-k}.$$

Given the natural action of $\mathbb{R}$ by $\lambda\mathbf{a} = (\lambda a_0, \lambda a_1, \dots)$, the space $T((V))$ is a real non-commutative algebra with unit $\mathbf{1} = (1, 0, 0, \dots)$. We denote by $\tilde{T}((V))$ the space of elements $\mathbf{a}$ with $a_0 = 1$; this space is a group with $\mathbf{a}^{-1} = 1 - (\mathbf{a} - 1) + (\mathbf{a} - 1)^{\otimes 2} - \cdots$. Finally, $T^N(V)$ is the truncated tensor algebra whose elements are of the form $\mathbf{a} = (a_0, \dots, a_N)$.

When $V = \mathbb{R}^d$ with standard basis $\{e_i\}_{i=1}^d$, the space $(\mathbb{R}^d)^{\otimes n}$ has dimension d^n with basis $\{e_w = e_{i_1} \otimes \cdots \otimes e_{i_n} : w = (i_1, \dots, i_n) \in \{1, \cdots, d\}^n\}$ indexed by the words w of length $|w| = n$ in d letters, with depth zero being spanned by the empty word. An element of $T((\mathbb{R}^d))$ is then of the form $a = \sum_{n \geq 0} \sum_{|w|=n} a^w e_w$; an element of $T(\mathbb{R}^d)$ is one with finitely many non-zero coefficients, which acts on $T((\mathbb{R}^d))$ as the linear functional $\langle l, a \rangle = \sum_w l_w a^w$; and an element of $T^N(\mathbb{R}^d)$ is one whose non-zero coefficients are for words of length at most N.

Definition 6 (Shuffle product (Ree, 1958))**.** Let $\varphi \in \left(V^{\otimes n}\right)^\star$ and $\sigma \in \left(V^{\otimes k}\right)^\star$ for $n, k \geq 1$. The shuffle of φ and σ is the linear functional $\varphi \sqcup\!\sqcup \sigma \in \left(V^{\otimes (n+k)}\right)^\star$ defined by

$$\langle \varphi \sqcup\!\sqcup \sigma, w \rangle = \sum_{\rho \in \mathrm{Sh}(n,k)} \langle \varphi \otimes \sigma, \rho(w) \rangle, \qquad w \in V^{\otimes (n+k)}, \tag{10}$$

where $\mathrm{Sh}(n, k) = \{\rho \in S_{n+k} : \rho(1) < \cdots < \rho(n) \text{ and } \rho(n+1) < \cdots < \rho(n+k)\}$ and ρ acts on $V^{\otimes (n+k)}$ by permuting the tensor factors. The shuffle is commutative and associative, and is extended bilinearly to $T(V)$.

The shuffle of a functional on depth n with one on depth k is a functional on depth $n + k$.

Definition 7 (Group-like element (Reutenauer, 1993))**.** An element $a \in \widetilde{T}((V))$ is called group-like if $\langle l_1, a \rangle \langle l_2, a \rangle = \langle l_1 \sqcup\!\sqcup l_2, a \rangle$ for all $l_1, l_2 \in T(V)$.

For a group-like element, a product of two coordinates at levels n and k is therefore a linear combination of coordinates at depth $n + k$, and more generally a polynomial of degree N in the depth-one coordinates is a linear functional of the depths up to N.

B.2 Paths and integration

With the tensor algebra established, we can now define the suitable framework for integration to be used in the definition of the signature.

Definition 8 (p-Variation (Young, 1936; Lyons et al., 2007))**.** Let V be a real Banach space and $X : [a, b] \to V$ be a path. For $p \geq 1$, the p-variation of X is defined as

$$||X||_{p,[a,b]} = \left(\sup_{\mathcal{P}} \sum_{i=0}^{n-1} ||X_{t_{i+1}} - X_{t_i}||_V^p\right)^{\frac{1}{p}},$$

where $\mathcal{P}$ is the set of all finite partitions $\{t_i\}_{i=0}^n$ such that $a = t_0 < \cdots < t_n = b$.

Paths in $\mathcal{V}^1$ are of bounded variation and $\|X\|_{1,[s,t]}$ is the total variation of X over $[s, t]$. If $p \leq q$ then $\mathcal{V}^p \subseteq \mathcal{V}^q$. Piecewise linear paths built from finitely many observations, as found in the main text, belong in $\mathcal{V}^1$.

We assume throughout that every integrand and driver is regular enough for the integrals below to be defined in the Young sense (Young, 1936). That is, for $p, q \geq 1$ with $\frac{1}{p} + \frac{1}{q} > 1$, $X \in$

$\mathcal{V}^p([0,T];V)$ and $Y \in \mathcal{V}^q([0,T];L(V,W))$, the integral $\int_s^t Y_u\,dX_u$ exists as a limit of Riemann sums independently of the partitions, and satisfies the Young estimate

$$\Big|\int_s^t Y_u\,dX_u - Y_s\big(X_t - X_s\big)\Big| \le C_{p,q}\,\|Y\|_{q,[s,t]}\,\|X\|_{p,[s,t]} \tag{11}$$

for a constant $C_{p,q}$ depending only on p and q (Lyons et al., 2007, Theorem 1.16).

For $p = 1$ this is a Riemann–Stieltjes integral, and the condition $\frac{1}{p} + \frac{1}{q} > 1$ is then satisfied by every integrand of bounded variation.

B.3 The Signature

The signature of Section 2 is actually defined for every $X \in \mathcal{V}^p([0,T];V)$ with $p < 2$, since $\frac{1}{p} + \frac{1}{p} > 1$ makes each iterated integral a Young integral of the depth below it (Lyons et al., 2007). For $p = 1$ its levels decay factorially,

$$\big\|S(X)^{(n)}_{s,t}\big\| \le \frac{\|X\|^n_{1,[s,t]}}{n!}, \tag{12}$$

with the bound for $1 < p < 2$ being the neo-classical inequality of Lyons et al. (2007).

Theorem 2 (Chen Identity (Chen, 1954))**.** *Let $X \in \mathcal{V}^p([s,t],V)$ with $p < 2$. Then for $u \in [s,t]$*

$$S(X)_{s,t} = S(X)_{s,u} \otimes S(X)_{u,t}.$$

This tells us that the signature of a path over an interval can be decomposed as the product of signatures of sub-intervals.

Theorem 3 (Shuffle Product Identity (Ree, 1958))**.** *Let $X \in \mathcal{V}^p([s,t],V)$ for $p < 2$, and $n, m \ge 0$ integers. Then for all bounded $\phi \in (V^{\otimes n})^\star$ and $\sigma \in (V^{\otimes m})^\star$, we have*

$$\langle \phi, S(X)^{(n)}_{s,t}\rangle \langle \sigma, S(X)^{(m)}_{s,t}\rangle = \langle \phi \sqcup\!\sqcup \sigma, S(X)^{(n+m)}_{s,t}\rangle.$$

This tells us that polynomial functions in the lower depth terms of the signature can be expressed as linear functions of the higher depth terms. The signature is a canonical representation of the path; as shown in Hambly & Lyons (2010), two paths of bounded variation have the same signature if and only if they are tree-like equivalent. This result was later extended to the $p < 2$ setting in Boedihardjo et al. (2016). In practice, when $V = \mathbb{R}^d$, this equivalence can be removed by augmenting the path with a strictly monotone channel, such as time, which ensures that the signature is injective on the space of paths (up to a translation constant). Another key property of the signature is its universal approximation capability.

Theorem 4 (Universality)**.** *Let $K \subset \mathcal{V}^p([0,T],V)$ be a compact set of paths with $p < 2$. Let $\mathcal{F} \subset C(K,\mathbb{R})$ be the set of continuous functions such that $F(X) = F(Y)$ whenever $S(X) = S(Y)$. Then the set of linear functionals of the signature is dense in $\mathcal{F}$. That is, for any $F \in \mathcal{F}$ and $\epsilon > 0$, there exists a linear functional l such that*

$$\sup_{X \in K} |F(X) - \langle l, S(X)\rangle| < \epsilon.$$

B.4 Controlled differential equations

Definition 9 (Controlled differential equation (Lyons et al., 2007))**.** Let V and W be normed spaces, let $f\colon W \to L(V,W)$ be a vector field and let $X \in \mathcal{V}^p([0,T];V)$ with $p < 2$. A path $h\colon [0,T] \to W$ solves the CDE driven by X with initial condition h_0 if

$$h_t = h_0 + \int_0^t f(h_u)\,dX_u \qquad \text{for all } t \in [0,T], \tag{13}$$

which we write as $dh_t = f(h_t)\,dX_t$. The CDE is linear if f is linear in the state.

In the main text, $V = \mathbb{R}^d$ and $W = \mathbb{R}^m$, such that $f(h) \in L(\mathbb{R}^d; \mathbb{R}^m)$. Writing $dX_t = \sum_{i=1}^d e_i \, dX_t^i$ gives $f(h)\, dX_t = \sum_{i=1}^d f(h)e_i \, dX_t^i$, and since each $h \mapsto f(h)e_i$ is a linear map on $\mathbb{R}^m$, setting $M_i h := f(h)e_i$ defines matrices $M_1, \dots, M_d \in \mathbb{R}^{m\times m}$ for which we have the linear CDE

$$dh_t = \sum_{i=1}^d M_i h_t \, dX_t^i. \tag{14}$$

Theorem 5 (Linear CDEs (Walker, 2025)). *Let $X \in \mathcal{V}^p([0,T]; V)$ with $p < 2$ and $M \in L(V, L(W,W))$, and consider the linear CDE $dh_t = M(dX_t)h_t$, that is, $f(h)v = M(v)h$. Let $\rho_M \colon T((V)) \to L(W,W)$ be the linear map with $\rho_M(\mathbf{1}) = \mathrm{id}_W$ and $\rho_M(v_1 \otimes \cdots \otimes v_n) = M(v_n)\cdots M(v_1)$, extended by linearity and continuity. Then the CDE has a unique solution for every initial condition, given for $s \le t$ by the absolutely convergent series*

$$h_t = \rho_M\big(S(X)_{s,t}\big)h_s = \sum_{n\ge 0} \rho_M\big(S(X)_{s,t}^{(n)}\big)h_s. \tag{15}$$

When $V = \mathbb{R}^d$ and $W = \mathbb{R}^m$, setting $M_i = M(e_i)$ gives the linear CDE of the background paragraph in Section 2, and the series reads

$$h_t = \sum_{n\ge 0} \sum_{i_1,\dots,i_n=1}^d M_{i_n} \cdots M_{i_1} \, S(X)_{s,t}^{i_1,\dots,i_n} \, h_s. \tag{16}$$

The signature is itself the unique solution of a linear CDE $dS_t = S_t \otimes dX_t \in T((V))$, with $S_s = \mathbf{1}$. On a sufficiently short interval the matrix series $\rho_M\big(\log S(X)_{s,t}\big)$ converges and the solution takes the flow form, $h_t = \exp\big(\rho_M(\log S(X)_{s,t})\big)h_s$, the Magnus expansion (Friz & Victoir, 2010). If X is linear on $[s,t]$ then every bracket term vanishes and $h_t = \exp\big(\sum_{i=1}^d (X_t^i - X_s^i)M_i\big)h_s$.

The state-space model $\dot h_t = -Ah_t + Bx_t$ is, with $X_t = \int_0^t x_u \, du$, the equation $dh_t = -Ah_t\, dt + B\, dX_t$, which is affine rather than linear in h. Appending a constant coordinate, $\bar h_t = (h_t, 1) \in \mathbb{R}^{m+1}$, gives the linear CDE driven by (t, X)

$$d\bar h_t = M_0 \bar h_t \, dt + \sum_{i=1}^d M_i \bar h_t \, dX_t^i, \qquad M_0 = \begin{pmatrix} -A & 0 \\ 0 & 0 \end{pmatrix}, \qquad M_i = \begin{pmatrix} 0 & Be_i \\ 0 & 0 \end{pmatrix}, \tag{17}$$

which has the form of (14).

C THE EWS

Throughout this appendix V and W are Banach spaces whose tensor powers carry admissible norms (Definition 4), which we assume also satisfy $\|T_1 \otimes \cdots \otimes T_n\| \le \|T_1\| \cdots \|T_n\|$ for bounded operators T_j on W, as the projective and injective norms do. The main text is the case $V = \mathbb{R}^d$ and $W = \mathbb{R}^m$, in which the Euclidean norm on coefficients satisfies these conditions. Definitions 1 and 2 are read for $1 \le p < 2$, with $\mathcal{V}^p$ in place of $\mathcal{V}^1$, V and W in place of $\mathbb{R}^d$ and $\mathbb{R}^m$, $A \in L(W,W)$ and $B \in L(V,W)$, the EWS taking values in $T((W))$ and its iterated integrals being Young integrals.

C.1 COORDINATES AND EXAMPLE

Coordinates. When $W = \mathbb{R}^m$, write $E(r) = e^{-rA}$, with entries $E_{ij}(r)$. The EWS coordinate associated with the word $(i_1, \dots, i_n)$ is,

$$S_{\mathbb{A}}(X;\theta)_{s,t}^{i_1,\dots,i_n} = \int_{s<t_1<\cdots<t_n<t} \sum_{j_1,\dots j_n=1}^m \prod_{l=1}^n E_{i_l j_l}(\theta_t - \theta_{t_l}) \, d(BX)_{t_1}^{j_1} \cdots d(BX)_{t_n}^{j_n}, \tag{18}$$

and the coefficient of a word is obtained from those of the words one letter shorter by

$$S_{\mathbb{A}}(X;\theta)_{s,t}^{i_1,\dots,i_n} = \sum_{k_1,\dots,k_{n-1}=1}^m \int_s^t \sum_{j=1}^m \Big(\prod_{l=1}^{n-1} E_{i_l k_l}(\theta_t - \theta_u) \Big) S_{\mathbb{A}}(X;\theta)_{s,u}^{k_1,\dots,k_{n-1}} E_{i_n j}(\theta_t - \theta_u) \, d(BX)_u^j, \tag{19}$$

which is the coordinate form of Lemma 3 below.

Example 1. To illustrate the structure of the weighted iterated integrals, let us consider the simple setting where $V = W = \mathbb{R}^2$, B is the identity and the clock is the time clock, $\theta_t = t$, with $A \in \mathbb{R}^{2\times 2}$. **Depth 1.** For a single-letter word (i_1), we have

$$S_{\mathbb{A}}(X;\theta)^{i_1}_{s,t} = \sum_{j_1=1}^{2} \int_s^t E_{i_1,j_1}(t-t_1)\, dX^{j_1}_{t_1}, \qquad i_1 \in \{1,2\}.$$

That is,

$$S_{\mathbb{A}}(X;\theta)^{1}_{s,t} = \int_s^t \big(E_{11}(t-u)\, dX^1_u + E_{12}(t-u)\, dX^2_u\big), \qquad S_{\mathbb{A}}(X;\theta)^{2}_{s,t} = \int_s^t \big(E_{21}(t-u)\, dX^1_u + E_{22}(t-u)\, dX^2_u\big).$$

Depth 2. For a two-letter word (i_1, i_2), we obtain

$$S_{\mathbb{A}}(X;\theta)^{i_1,i_2}_{s,t} = \sum_{j_1,j_2=1}^{2} \int_s^t \int_s^{t_2} E_{i_1,j_1}(t-t_1)\, E_{i_2,j_2}(t-t_2)\, dX^{j_1}_{t_1}\, dX^{j_2}_{t_2}.$$

Expanding the indices explicitly gives

$$\begin{aligned} S_{\mathbb{A}}(X;\theta)^{1,1}_{s,t} = \int_s^t \int_s^{t_2} \big(&E_{11}(t-t_1)E_{11}(t-t_2)\, dX^1_{t_1} dX^1_{t_2} + E_{11}(t-t_1)E_{12}(t-t_2)\, dX^1_{t_1} dX^2_{t_2} \\ &+ E_{12}(t-t_1)E_{11}(t-t_2)\, dX^2_{t_1} dX^1_{t_2} + E_{12}(t-t_1)E_{12}(t-t_2)\, dX^2_{t_1} dX^2_{t_2}\big), \end{aligned}$$

$$\begin{aligned} S_{\mathbb{A}}(X;\theta)^{1,2}_{s,t} = \int_s^t \int_s^{t_2} \big(&E_{11}(t-t_1)E_{21}(t-t_2)\, dX^1_{t_1} dX^1_{t_2} + E_{11}(t-t_1)E_{22}(t-t_2)\, dX^1_{t_1} dX^2_{t_2} \\ &+ E_{12}(t-t_1)E_{21}(t-t_2)\, dX^2_{t_1} dX^1_{t_2} + E_{12}(t-t_1)E_{22}(t-t_2)\, dX^2_{t_1} dX^2_{t_2}\big), \end{aligned}$$

and analogous formulas for $(i_1, i_2) = (2, 1)$ and $(2, 2)$.

C.2 Flow and derivation operators

The iterated integrals defining the EWS repeatedly apply the operator exponential e^{-rA} to increments of the lifted path. To express these operations compactly across all tensor levels, we extend the action of e^{-rA} from W to the entire tensor algebra $T((W))$. When A is diagonal, as for the EFM, e^{-rA} acts on each word by a scalar $e^{-r(\lambda_{i_1}+\cdots+\lambda_{i_n})}$, and every coordinate of the EWS evolves separately. For a general matrix, it mixes the coordinates within each depth (Example 2), so the proofs below rely on D^r_A being an algebra automorphism and Λ_A a derivation. Since the clock is a path of finite p-variation rather than time, they also rely on the Young chain and product rules (Lemma 4).

Recall that $T((W))$ is the free unital associative algebra generated by W. In particular, any linear map $f\colon W \to W$ admits a unique extension to an algebra homomorphism on $T((W))$, a consequence of the universal property of the tensor algebra (Lang, 2002). We apply this construction to the linear map $v \mapsto e^{-rA}v$ where, since $A \in L(W, W)$ is bounded, the operator exponential e^{-rA} is well defined for all $r \in \mathbb{R}$ and defines a uniformly continuous one-parameter group on W (Rudin, 1991).

Definition 10. Let $A \in L(W, W)$ and $r \in \mathbb{R}$. The flow operator $D^r_A\colon T((W)) \to T((W))$ is the unique algebra homomorphism extending the linear map $v \mapsto e^{-rA}v$ on W. That is,

$$D^r_A|_W = e^{-rA}. \tag{20}$$

Equivalently, for any elementary tensor $v_1 \otimes \cdots \otimes v_n \in W^{\otimes n}$,

$$D^r_A(v_1 \otimes \cdots \otimes v_n) = (e^{-rA}v_1) \otimes \cdots \otimes (e^{-rA}v_n). \tag{21}$$

Lemma 1. *The family $\{D^r_A : r \in \mathbb{R}\}$ is a one-parameter subgroup of $\mathrm{Aut}(T((W)))$, the group of algebra automorphisms of $T((W))$.*

Proof. Since $\mathrm{Aut}(T((W)))$ is a group, it suffices to show that the family is closed under composition and inverses. For any $r_1, r_2 \in \mathbb{R}$ and any $v \in W$,

$$D^{r_1}_A D^{r_2}_A(v) = e^{-r_1 A}e^{-r_2 A}v = e^{-(r_1+r_2)A}v = D^{r_1+r_2}_A(v).$$

Both $D^{r_1}_A D^{r_2}_A$ and $D^{r_1+r_2}_A$ are algebra homomorphisms on $T((W))$ that agree on the generators W, and by the universal property of the tensor algebra they coincide on all of $T((W))$. Moreover $D^r_A D^{-r}_A = D^0_A = \mathrm{id}$, and hence $(D^r_A)^{-1} = D^{-r}_A$. □

We define the derivation induced by A as the infinitesimal generator of the one-parameter group of Lemma 1, which is differentiable in r since it acts on $W^{\otimes n}$ as $(e^{-rA})^{\otimes n}$.

Definition 11. Let $A \in L(W, W)$. The operator $\Lambda_A \colon T((W)) \to T((W))$ defined by

$$\Lambda_A = -\frac{d}{dr}\Big|_{r=0} D_A^r \tag{22}$$

is called the derivation induced by A.

Lemma 2. *The operator Λ_A is a derivation, that is, $\Lambda_A(a \otimes b) = \Lambda_A(a) \otimes b + a \otimes \Lambda_A(b)$ for all $a, b \in T((W))$. It satisfies $\Lambda_A(\mathbf{1}) = 0$ and, for any elementary tensor $v_1 \otimes \cdots \otimes v_n \in W^{\otimes n}$,*

$$\Lambda_A(v_1 \otimes \cdots \otimes v_n) = \sum_{k=1}^{n} v_1 \otimes \cdots \otimes (Av_k) \otimes \cdots \otimes v_n. \tag{23}$$

Proof. Since each D_A^r is an algebra homomorphism, $D_A^r(a \otimes b) = D_A^r(a) \otimes D_A^r(b)$, and differentiating at $r = 0$ gives

$$\Lambda_A(a \otimes b) = \Big(-\frac{d}{dr}\Big|_{r=0} D_A^r(a)\Big) \otimes D_A^0(b) + D_A^0(a) \otimes \Big(-\frac{d}{dr}\Big|_{r=0} D_A^r(b)\Big) = \Lambda_A(a) \otimes b + a \otimes \Lambda_A(b).$$

For $v \in W$, $\Lambda_A(v) = -\frac{d}{dr}\big|_{r=0} e^{-rA} v = Av$ by differentiability of the operator exponential, and (23) follows from the Leibniz rule. Since $\Lambda_A(\mathbf{1}) = \Lambda_A(\mathbf{1} \otimes \mathbf{1}) = 2\Lambda_A(\mathbf{1})$, we have $\Lambda_A(\mathbf{1}) = 0$. □

C.3 The linear CDE form

Lemma 3. *For any $[s, t] \subseteq [0, T]$, the EWS satisfies*

$$S_{\mathbb{A}}(X; \theta)_{s,t} = \mathbf{1} + \int_s^t D_A^{\theta_t - \theta_u}(S_{\mathbb{A}}(X; \theta)_{s,u} \otimes B\, dX_u), \tag{24}$$

where $\mathbf{1} = (1, 0, 0, \dots)$ is the identity element in $T((W))$.

Proof. Since trivially $S_{\mathbb{A}}(X; \theta)_{s,t}^{(0)} = 1$, we just need to show that the depth n term of the integral in the claim equals the depth n term of the EWS.

$$\begin{aligned}
\Big(\int_s^t D_A^{\theta_t - \theta_u}(S_{\mathbb{A}}(X; \theta)_{s,u} \otimes B\, dX_u)\Big)^{(n)} &= \Big(\int_s^t D_A^{\theta_t - \theta_u}(S_{\mathbb{A}}(X; \theta)_{s,u}) \otimes D_A^{\theta_t - \theta_u}(B\, dX_u)\Big)^{(n)} \\
&= \int_s^t \Big(D_A^{\theta_t - \theta_u}(S_{\mathbb{A}}(X; \theta)_{s,u}) \otimes D_A^{\theta_t - \theta_u}(B\, dX_u)\Big)^{(n)} \\
&= \int_s^t \sum_{k=0}^{n} \Big(D_A^{\theta_t - \theta_u}(S_{\mathbb{A}}(X; \theta)_{s,u})\Big)^{(k)} \otimes \Big(D_A^{\theta_t - \theta_u}(B\, dX_u)\Big)^{(n-k)} \\
&= \int_s^t \Big(D_A^{\theta_t - \theta_u}(S_{\mathbb{A}}(X; \theta)_{s,u})\Big)^{(n-1)} \otimes \Big(D_A^{\theta_t - \theta_u}(B\, dX_u)\Big)^{(1)} \\
&= \int_s^t D_A^{\theta_t - \theta_u}(S_{\mathbb{A}}(X; \theta)_{s,u}^{(n-1)}) \otimes (e^{-(\theta_t - \theta_u)A} B\, dX_u),
\end{aligned}$$

where the second equality uses the algebra homomorphism property of the flow operator, and the fourth equality uses the fact that BdX_u is only non-zero at depth one. We can then plug in the iterated integral expression for the EWS to get

$$\begin{aligned}
&\Big(\int_s^t D_A^{\theta_t - \theta_u}(S_{\mathbb{A}}(X; \theta)_{s,u} \otimes B\, dX_u)\Big)^{(n)} \\
&= \int_s^t D_A^{\theta_t - \theta_u}\Big(\int_s^u \int_s^{t_{n-1}} \cdots \int_s^{t_2} e^{-(\theta_u - \theta_{t_1})A} B\, dX_{t_1} \otimes \cdots \otimes e^{-(\theta_u - \theta_{t_{n-1}})A} B\, dX_{t_{n-1}}\Big) \otimes e^{-(\theta_t - \theta_u)A} B\, dX_u \\
&= \int_s^t \int_s^u \int_s^{t_{n-1}} \cdots \int_s^{t_2} e^{-(\theta_t - \theta_{t_1})A} B\, dX_{t_1} \otimes \cdots \otimes e^{-(\theta_t - \theta_{t_{n-1}})A} B\, dX_{t_{n-1}} \otimes e^{-(\theta_t - \theta_u)A} B\, dX_u.
\end{aligned}$$

Changing notation from u to t_n gives us exactly $S_{\mathbb{A}}(X; \theta)_{s,t}^{(n)}$, as required. □

This integral form now lets us write the EWS as the solution to a linear CDE; the following lemma is Proposition 1 for paths of finite p-variation with $p < 2$.

Lemma 4. *The EWS is the unique solution of the linear controlled differential equation driven by the augmented path $(\theta, X) \in \mathcal{V}^p([0,T]; \mathbb{R} \times V)$,*

$$dS_{\mathbb{A}}(X;\theta)_{s,t} = -\Lambda_A S_{\mathbb{A}}(X;\theta)_{s,t}\, d\theta_t + S_{\mathbb{A}}(X;\theta)_{s,t} \otimes B\, dX_t, \qquad S_{\mathbb{A}}(X;\theta)_{s,s} = \mathbf{1}. \tag{25}$$

Equivalently, in Young integral form,

$$S_{\mathbb{A}}(X;\theta)_{s,t} = \mathbf{1} + \int_s^t \big(-\Lambda_A S_{\mathbb{A}}(X;\theta)_{s,u}\big)\, d\theta_u + \int_s^t S_{\mathbb{A}}(X;\theta)_{s,u} \otimes B\, dX_u. \tag{26}$$

Proof. We argue by induction on the depth $n \geq 1$, assuming that $u \mapsto S_{\mathbb{A}}(X;\theta)^{(n-1)}_{s,u}$ is continuous with finite p-variation, which holds for $n = 1$ since $S_{\mathbb{A}}(X;\theta)^{(0)}_{s,u} = 1$. Since D^r_A is an algebra homomorphism that is grade-preserving and $B\, dX_u$ lies at depth one, Lemma 3 at depth n reads

$$S_{\mathbb{A}}(X;\theta)^{(n)}_{s,t} = \int_s^t D_A^{\theta_t - \theta_u}\big(S_{\mathbb{A}}(X;\theta)^{(n-1)}_{s,u} \otimes B\, dX_u\big).$$

By Lemma 1, $D_A^{\theta_t-\theta_u} = D_A^{\theta_t} D_A^{-\theta_u}$, and taking the bounded operator $D_A^{\theta_t}$ out of the integral gives $S_{\mathbb{A}}(X;\theta)^{(n)}_{s,t} = D_A^{\theta_t} Y_t$ with

$$Y_t = \int_s^t D_A^{-\theta_u}\big(S_{\mathbb{A}}(X;\theta)^{(n-1)}_{s,u} \otimes B\, dX_u\big),$$

a Young integral whose integrand is a locally Lipschitz function of θ_u and $S_{\mathbb{A}}(X;\theta)^{(n-1)}_{s,u}$, both continuous with finite p-variation. By Lemma 1 and Definition 11, $r \mapsto D^r_A$ is continuously differentiable on each level, with $\frac{d}{dr} D^r_A = \lim_{\epsilon\to 0} \epsilon^{-1}\big(D^\epsilon_A - \mathrm{Id}\big) D^r_A = -\Lambda_A D^r_A$. Hence $t \mapsto D^{\theta_t}_A$ is continuous with finite p-variation, and by the chain rule for Young integrals and the definition of Y,

$$dD^{\theta_t}_A = -\Lambda_A D^{\theta_t}_A\, d\theta_t, \qquad dY_t = D_A^{-\theta_t}\big(S_{\mathbb{A}}(X;\theta)^{(n-1)}_{s,t} \otimes B\, dX_t\big).$$

The product rule for Young integrals then gives

$$dS_{\mathbb{A}}(X;\theta)^{(n)}_{s,t} = -\Lambda_A D^{\theta_t}_A Y_t\, d\theta_t + D^{\theta_t}_A D_A^{-\theta_t}\big(S_{\mathbb{A}}(X;\theta)^{(n-1)}_{s,t} \otimes B\, dX_t\big) = -\Lambda_A S_{\mathbb{A}}(X;\theta)^{(n)}_{s,t}\, d\theta_t + S_{\mathbb{A}}(X;\theta)^{(n-1)}_{s,t} \otimes B\, dX_t.$$

The chain and product rules hold because $p < 2$. The second-order terms of their Riemann sums, $\sum_i |\theta_{u_{i+1}} - \theta_{u_i}|^2$ and $\sum_i \|D_A^{\theta_{u_{i+1}}} - D_A^{\theta_{u_i}}\|\, \|Y_{u_{i+1}} - Y_{u_i}\|$, vanish as the mesh tends to zero for continuous paths of finite p-variation (Friz & Victoir, 2010). In particular $t \mapsto S_{\mathbb{A}}(X;\theta)^{(n)}_{s,t} = D^{\theta_t}_A Y_t$ is continuous with finite p-variation, which completes the induction. At depth zero $S_{\mathbb{A}}(X;\theta)^{(0)}_{s,t} = 1$ and $\Lambda_A(\mathbf{1}) = 0$, and the equation holds trivially. Reassembling the levels gives (25), with $S_{\mathbb{A}}(X;\theta)_{s,s} = \mathbf{1}$. For uniqueness, let S' be any solution of (25). Since Λ_A preserves levels and the level-n part of $S' \otimes B\, dX$ involves S' only through depth $n-1$, $\pi_N(S')$ solves the equation of Lemma 5 for every N, and hence $\pi_N(S') = \pi_N\big(S_{\mathbb{A}}(X;\theta)_{s,\cdot}\big)$ for every N. □

Lemma 5. *Let $X \in \mathcal{V}^p([0,T]; V)$ for $p < 2$ and fix an integer $N \geq 1$. Define $f\colon T^N(W) \to L(\mathbb{R} \times V, T^N(W))$ by*

$$f(a)(\vartheta, v) = -\Lambda_A(a) \cdot \vartheta + \pi_N(a \otimes Bv),$$

where $\pi_N\colon T((W)) \to T^N(W)$ denotes projection onto the first $N+1$ levels (including the 0-th level). Then the unique solution to the CDE

$$dS_t = f(S_t)\, d(\theta, X)_t, \qquad S_s = (1, 0, \ldots, 0),$$

is the truncated EWS $S_t = \pi_N(S_{\mathbb{A}}(X;\theta)_{s,t})$.

Proof. Since $p < 2$, by Theorem 5, existence and uniqueness hold. To verify that $\pi_N(S_{\mathbb{A}}(X;\theta)_{s,t})$ solves this equation, we project the EWS dynamics from Lemma 4. Applying π_N to both sides gives

$$d\pi_N(S_{\mathbb{A}}(X;\theta)_{s,t}) = -\Lambda_A \pi_N(S_{\mathbb{A}}(X;\theta)_{s,t})\, d\theta_t + \pi_N(S_{\mathbb{A}}(X;\theta)_{s,t} \otimes B\, dX_t).$$

As the tensor product $S_{\mathbb{A}}(X;\theta)_{s,t} \otimes B\, dX_t$ only affects levels up to N through components of $S_{\mathbb{A}}(X;\theta)_{s,t}$ at levels up to $N-1$, and Λ_A preserves tensor degrees (it acts level-wise), we have

$$\pi_N(S_{\mathbb{A}}(X;\theta)_{s,t} \otimes B\, dX_t) = \pi_N(\pi_N(S_{\mathbb{A}}(X;\theta)_{s,t}) \otimes B\, dX_t).$$

Thus, $\pi_N(S_{\mathbb{A}}(X;\theta)_{s,t})$ satisfies the CDE with vector field f and initial condition $\pi_N(S_{\mathbb{A}}(X;\theta)_{s,s}) = \pi_N((1,0,\dots,0)) = (1,0,\dots,0)$. By uniqueness, it is the solution. □

Proposition 1 shows that, just like the signature, the EWS is the solution to a linear CDE which takes values in the tensor algebra. However, to make contact with standard finite-dimensional models, we let $V = \mathbb{R}^d$ and $W = \mathbb{R}^m$, and truncate the tensor algebra at depth N to work with the truncated EWS. Applying the canonical grade–lexicographic flattening

$$\Phi : T^N(\mathbb{R}^m) \xrightarrow{\cong} \mathbb{R}^K, \qquad K = \sum_{n=0}^{N} m^n, \tag{27}$$

the linear CDE satisfied by the truncated EWS becomes a finite-dimensional linear CDE in $\mathbb{R}^K$. Writing $dX_t = \sum_{i=1}^d e_i dX_t^i$, one obtains the flattened CDE

$$dS_{\mathbb{A}}(X;\theta)_{s,t}^{\le N} = -\,L\, S_{\mathbb{A}}(X;\theta)_{s,t}^{\le N}\, d\theta_t + \sum_{i=1}^{d} \rho(Be_i)\, S_{\mathbb{A}}(X;\theta)_{s,t}^{\le N}\, dX_t^i, \tag{28}$$

where

$$L = \mathrm{blockdiag}\left(L^{(0)}, L^{(1)}, \dots, L^{(N)}\right), \qquad L^{(0)} = [0], \tag{29}$$

is the matrix representation of Λ_A, and each $\rho(v) \in \mathbb{R}^{K\times K}$ is the sparse strictly lower-triangular matrix representing right tensor-multiplication by $v \in \mathbb{R}^m$, so that $\rho(Be_i)$ maps depth $n-1$ into depth n. Equivalently, writing M_i for the coefficient matrices, the system is a linear CDE driven by the path $(\theta_t, X_t) \in \mathbb{R}^{1+d}$ in the standard form

$$dS_{\mathbb{A}}(X;\theta)_{s,t}^{\le N} = M_0 S_{\mathbb{A}}(X;\theta)_{s,t}^{\le N}\, d\theta_t + \sum_{i=1}^{d} M_i S_{\mathbb{A}}(X;\theta)_{s,t}^{\le N}\, dX_t^i, \qquad \begin{matrix} M_0 = -L, \\ M_i = \rho(Be_i). \end{matrix} \tag{30}$$

When the clock is a linear functional of the path, $\theta_t = \sum_i \ell_i X_t^i$, the first driver may be removed and the system is driven by X alone, with $M_i = \rho(Be_i) - \ell_i L$.

This representation makes explicit that the truncated EWS evolves according to a linear CDE with a highly structured collection of coefficient matrices. All $d+1$ of them are generated by the $m^2 + dm$ entries of A and B. Hence when the parameters of $\mathbb{A} = (A, B)$ are learnable, the truncated EWS defines a structured linear NCDE within the SLiCE framework that encodes a specific inductive bias. In particular, its hidden state has dimension K, which is determined by the truncation depth N and the dimension m of the lifted path. We illustrate this structure with a simple example.

Example 2. We consider the case of a path X taking values in $\mathbb{R}^2$ and compute explicitly the CDE matrices for the EWS truncated at depth $N = 2$ with B the identity, so that $m = d = 2$. Let $A = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}$. The truncated tensor algebra has dimension $K = 1 + 2 + 4 = 7$, with grade–lexicographic basis $\{1,\ e_1,\ e_2,\ e_1 \otimes e_1,\ e_1 \otimes e_2,\ e_2 \otimes e_1,\ e_2 \otimes e_2\}$. The derivation matrix $L = \mathrm{blockdiag}(L^{(0)}, L^{(1)}, L^{(2)})$ has graded blocks

$$L^{(0)} = [0], \qquad L^{(1)} = \begin{pmatrix} \alpha & \beta \\ \gamma & \delta \end{pmatrix}, \qquad L^{(2)} = \begin{pmatrix} 2\alpha & \beta & \beta & 0 \\ \gamma & \alpha+\delta & 0 & \beta \\ \gamma & 0 & \alpha+\delta & \beta \\ 0 & \gamma & \gamma & 2\delta \end{pmatrix}.$$

Right tensor multiplication by e_1 and e_2 is represented by

$$\rho(e_1) = \begin{pmatrix} 0&0&0&0&0&0&0 \\ 1&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0 \\ 0&1&0&0&0&0&0 \\ 0&0&0&0&0&0&0 \\ 0&0&1&0&0&0&0 \\ 0&0&0&0&0&0&0 \end{pmatrix}, \qquad \rho(e_2) = \begin{pmatrix} 0&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0 \\ 1&0&0&0&0&0&0 \\ 0&0&0&0&0&0&0 \\ 0&1&0&0&0&0&0 \\ 0&0&0&0&0&0&0 \\ 0&0&1&0&0&0&0 \end{pmatrix}.$$

The instantaneous linear operator appearing in the flattened EWS CDE, $dS_{\mathbb{A}}(X;\theta)^{\leq 2}_{s,t} = \mathcal{M}_t\, S_{\mathbb{A}}(X;\theta)^{\leq 2}_{s,t}$, is therefore

$$\mathcal{M}_t = -L\, d\theta_t + \rho(e_1)\, dX_t^1 + \rho(e_2)\, dX_t^2,$$

which explicitly equals

$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ dX_t^1 & -\alpha\, d\theta_t & -\beta\, d\theta_t & 0 & 0 & 0 & 0 \\ dX_t^2 & -\gamma\, d\theta_t & -\delta\, d\theta_t & 0 & 0 & 0 & 0 \\ 0 & dX_t^1 & 0 & -2\alpha\, d\theta_t & -\beta\, d\theta_t & -\beta\, d\theta_t & 0 \\ 0 & dX_t^2 & 0 & -\gamma\, d\theta_t & -(\alpha+\delta)\, d\theta_t & 0 & -\beta\, d\theta_t \\ 0 & 0 & dX_t^1 & -\gamma\, d\theta_t & 0 & -(\alpha+\delta)\, d\theta_t & -\beta\, d\theta_t \\ 0 & 0 & dX_t^2 & 0 & -\gamma\, d\theta_t & -\gamma\, d\theta_t & -2\delta\, d\theta_t \end{pmatrix}.$$

Remark 1 (Solvability). The Lie algebra $\mathfrak{g}$ generated by the vector fields $M_0, \ldots, M_d$ of (30) is solvable. Let $\mathfrak{n}$ be the Lie algebra of strictly block lower-triangular matrices that map each depth into higher depths, and which is nilpotent. For $i = 1, \ldots, d$, each $M_i = \rho(Be_i)$ lies in $\mathfrak{n}$, and $M_0 = -L$ is block diagonal, which gives $[M_0, \mathfrak{n}] \subseteq \mathfrak{n}$. Thus, $\mathfrak{g} \subseteq \mathbb{R}M_0 \oplus \mathfrak{n}$ and $[\mathfrak{g}, \mathfrak{g}] \subseteq \mathfrak{n}$, and $\mathfrak{g}$ is solvable.

C.4 Equivalence to the signature of a re-weighted path

We now show that, over a fixed interval, EWS admits an equivalent representation as the classical signature of a linearly transformed path. We consider $X \in \mathcal{V}^p([0,T];V)$ and the bounded linear operators $A \in L(W,W)$ and $B \in L(V,W)$ as before. We define the exponentially re-weighted path over $[s,t] \subseteq [0,T]$ by

$$Z_\tau^{[t]} := \int_s^\tau e^{-(\theta_t - \theta_u)A} B\, dX_u, \qquad \tau \in [s,t], \tag{31}$$

where the integral is understood in the Young sense. The superscript $[t]$ denotes the horizon from which each increment is weighted, and different horizons yield different paths. For a given clock, this construction defines a linear memory transform of X up to the horizon t, with operator-valued kernel $r \mapsto e^{-rA}B$. The main observation of this subsection is that the EWS of X over the interval $[s,t]$ coincides exactly with the classical signature of the weighted path $Z^{[t]}$ over the same interval.

Proposition 3. *Let $[s,t] \subseteq [0,T]$ and let $Z^{[t]}$ be defined as above. Then,*

$$S_{\mathbb{A}}(X;\theta)_{s,t} = S(Z^{[t]})_{s,t} \tag{32}$$

as elements of the tensor algebra.

Proof. We show that this holds for the depth n term by substituting $dZ_u^{[t]} = e^{-(\theta_t - \theta_u)A} B\, dX_u$ into the classical definition of the signature:

$$\begin{aligned} S(Z^{[t]})^{(n)}_{s,t} &= \int_s^t \int_s^{t_n} \cdots \int_s^{t_2} dZ_{t_1}^{[t]} \otimes \cdots \otimes dZ_{t_n}^{[t]} \\ &= \int_s^t \int_s^{t_n} \cdots \int_s^{t_2} \left(e^{-(\theta_t - \theta_{t_1})A} B\, dX_{t_1}\right) \otimes \cdots \otimes \left(e^{-(\theta_t - \theta_{t_n})A} B\, dX_{t_n}\right) \\ &= S_{\mathbb{A}}(X;\theta)^{(n)}_{s,t}. \end{aligned}$$

□

Remark 2. The identity $S_A(X;\theta)_{s,t} = S(Z^{[t]})_{s,t}$ holds for each terminal time t separately, with a path $Z^{[t]}$ that depends on t. Therefore, the trajectory $t \mapsto S_A(X;\theta)_{s,t}$ is not the signature of a single evolving path, so the classical Chen identity does not apply, and Lemma 7 proves the modified form directly.

The identity reduces the proofs of group-likeness, the factorial bound, continuity and universality to classical signature results for the single path $Z^{[t]}$. It is not used for the linear CDE form (Lemma 4) or the modified Chen identity (Lemma 7), and it does not help to compute the EWS over time, since $S(Z^{[t]})_{s,t}$ and $S(Z^{[t+h]})_{s,t+h}$ are related only through the flow operator.

Remark 3. For the time clock and $X_t = \int_0^t x_s\,ds$, the latent state of the SSM of Section 4 is $h_\tau = \int_0^\tau e^{-(\tau-s)A}B\,dX_s$, whereas $Z_\tau^{[t]} = \int_0^\tau e^{-(t-s)A}B\,dX_s$. For $\tau < t$ the two paths differ, as h decays relative to the moving time τ and $Z^{[t]}$ relative to the fixed horizon t. Both start at zero and meet at $\tau = t$, and hence agree at depth one, but $S_{\mathbb{A}}(X;\theta)_{0,t} = S(Z^{[t]})_{0,t}$ is in general not the signature of h at deeper levels.

C.5 Analytic properties

We now establish analytic well-posedness of the EWS for paths $X \in \mathcal{V}^p([0,T];V)$ with $p < 2$, meaning that the defining tensor series converges. Since the EWS coincides with the classical signature of a suitably re-weighted path, well-posedness for general $p < 2$ follows directly from standard signature theory. For completeness, we demonstrate the factorial decay bound explicitly in the finite variation case, which makes the convergence mechanism and the dependence on the operator A transparent.

Lemma 6. *For $X \in \mathcal{V}^1([0,T];V)$, the EWS is well-defined, and for all $[s,t] \subseteq [0,T]$, the following bound holds*

$$\|S_{\mathbb{A}}(X;\theta)_{s,t}^{(n)}\|_{W^{\otimes n}} \le \frac{\big(C_{A,\theta}\|B\|_{\mathrm{op}}\|X\|_{1,[s,t]}\big)^n}{n!}, \tag{33}$$

where $C_{A,\theta} = \sup_{u\in[s,t]}\|e^{-(\theta_t-\theta_u)A}\|_{\mathrm{op}} < \infty$ since A is bounded and θ is continuous.

Proof. By Proposition 3 and the factorial decay of the signature, $\|S_{\mathbb{A}}(X;\theta)_{s,t}^{(n)}\|_{W^{\otimes n}} = \|S(Z^{[t]})_{s,t}^{(n)}\|_{W^{\otimes n}} \le \|Z^{[t]}\|_{1,[s,t]}^n/n!$. Since $\|e^{-(\theta_t-\theta_u)A}B\|_{\mathrm{op}} \le C_{A,\theta}\|B\|_{\mathrm{op}}$ for all $u \in [s,t]$, we have $\|Z^{[t]}\|_{1,[s,t]} \le C_{A,\theta}\|B\|_{\mathrm{op}}\|X\|_{1,[s,t]}$, and the bound follows. □

Lemma 6 establishes the existence of the EWS for paths of finite variation ($p = 1$). To extend this result to the case of $1 < p < 2$, note that the continuous map $u \mapsto e^{-(\theta_t-\theta_u)A}$ is of finite p-variation since $\theta \in \mathcal{V}^p([0,T];\mathbb{R})$, and since $\frac{1}{p} + \frac{1}{p} > 1$ the existence of $Z^{[t]}$ is guaranteed by the Young estimate

$$\|Z^{[t]}\|_{p,[s,t]} \le C_{\theta,A}\|X\|_{p,[s,t]},$$

where the constant $C_{\theta,A}$ depends on $\sup_{u\in[s,t]}\|e^{-(\theta_t-\theta_u)A}\|_{\mathrm{op}}$, on $\|B\|_{\mathrm{op}}$ and on the p-variation of the map $u \mapsto e^{-(\theta_t-\theta_u)A}$. By Proposition 3, the existence of the EWS for $p < 2$ then follows directly from the existence of the signature of Young paths.

Corollary 1 (Scale)**.** *Let $X \in \mathcal{V}^1([0,T];V)$ and $[s,t] \subseteq [0,T]$. Suppose that $\|e^{-rA}\|_{\mathrm{op}} \le e^{-\mu r}$ for all $r \ge 0$ and some $\mu > 0$. Assume that the clock is non-decreasing and that $\|X\|_{1,[u,v]} \le L(\theta_v - \theta_u)$ for all $s \le u \le v \le t$. Then for all $n \ge 1$, we have*

$$\|S_{\mathbb{A}}(X;\theta)_{s,t}^{(n)}\|_{W^{\otimes n}} \le \frac{1}{n!}\Big(\frac{L\|B\|_{\mathrm{op}}}{\mu}\Big)^n,$$

independently of the length of $[s,t]$ and of the total variation of X.

Proof. From Lemma 6, $\|S_{\mathbb{A}}(X;\theta)_{s,t}^{(n)}\|_{W^{\otimes n}} \le \|Z^{[t]}\|_{1,[s,t]}^n/n!$. Then,

$$\|Z^{[t]}\|_{1,[s,t]} \le \|B\|_{\mathrm{op}}\int_s^t e^{-\mu(\theta_t-\theta_u)}\,d\|X\|_{1,[s,u]} \le L\|B\|_{\mathrm{op}}\int_s^t e^{-\mu(\theta_t-\theta_u)}\,d\theta_u = \frac{L\|B\|_{\mathrm{op}}}{\mu}\big(1-e^{-\mu(\theta_t-\theta_s)}\big) \le \frac{L\|B\|_{\mathrm{op}}}{\mu}.$$

□

When $V = \mathbb{R}^d$ and $W = \mathbb{R}^m$ as in the main text, the condition on A holds whenever $A + A^\top \geq 2\mu I$, and the condition on the clock holds with $L = 1$ for the variation clock. By contrast, an unrestricted linear CDE has no bound independent of the length of the path. On an interval where X is linear its solution is $h_t = \exp\big(\sum_{i=1}^d (X_t^i - X_s^i) M_i\big) h_s$. If $\sum_i v_i M_i$ has an eigenvalue λ with $\operatorname{Re}\lambda > 0$ for some $v \in \mathbb{R}^d$, then the norm $\|h_t\|$ grows at least like $e^{\operatorname{Re}\lambda\,(t-s)}$ along the path $X_t = tv$ for generic h_s, that is, exponentially in the total variation $\|X\|_{1,[s,t]} = (t-s)\|v\|$.

A clock is continuous if it is continuous for the p-variation topologies on $\mathcal{V}^p([0,T];V)$ and $\mathcal{V}^p([0,T];\mathbb{R})$. The time clock is continuous for every $p < 2$, and the variation clock for $p = 1$.

Corollary 2. *For each $p \in [1,2)$, each integer $N \geq 0$ and each continuous clock θ, the truncated EWS defines a continuous mapping*

$$\pi_N \circ S_{\mathbb{A}}(\,\cdot\,;\theta)_{s,t} \colon \mathcal{V}^p([0,T];V) \to T^N(W)$$

with respect to the p-variation topology on the domain.

Proof. By Proposition 3, $\pi_N(S_{\mathbb{A}}(X;\theta)_{s,t}) = \pi_N(S(Z^{[t]})_{s,t})$, where $Z_\tau^{[t]} = \int_s^\tau K_u\, dX_u$ with $K_u = e^{-(\theta_t - \theta_u)A} B$. Let $X^{(k)} \to X$ in p-variation. Since θ is continuous, $\theta(X^{(k)}) \to \theta(X)$ in p-variation, and since $r \mapsto e^{-rA}B$ is smooth, the corresponding integrands $K^{(k)}$ converge to K in p-variation. The Young estimate (11) makes the Young integral continuous in its integrand and its driver, and hence the re-weighted paths of $X^{(k)}$ converge to $Z^{[t]}$ in p-variation. The truncated signature is continuous on $\mathcal{V}^p([s,t];W)$ for $p < 2$ (Lyons et al., 2007), which gives the claim. □

C.6 Algebraic properties

We now show that the EWS satisfies the key algebraic properties of the classical path signature. In particular, it is multiplicative with respect to path concatenation (Chen's identity), and linear functionals of the EWS are closed under multiplication, forming a commutative algebra via the shuffle product.

Chen's identity. As for the signature (Lyons et al., 2007), we first establish the modified Chen identity for $X \in \mathcal{V}^1([0,T];V)$ and then extend it to the $p \in (1,2)$ regime. The proof for $p = 1$ works directly with the weighted iterated integrals and does not use Proposition 3; the extension to $p < 2$ uses it only through the continuity of Corollary 2.

Lemma 7. *Let $X \in \mathcal{V}^p([0,T];V)$ for $p = 1$. Then for any $0 \leq s \leq u \leq t \leq T$, the exponentially-weighted signature satisfies*

$$S_{\mathbb{A}}(X;\theta)_{s,t} = \big(D_A^{\theta_t - \theta_u} S_{\mathbb{A}}(X;\theta)_{s,u}\big) \otimes S_{\mathbb{A}}(X;\theta)_{u,t}. \tag{34}$$

Proof. We will show that the statement holds at an arbitrary depth n. For all $s \leq u \leq t$, we can split the region of integration as

$$\begin{aligned}
S_{\mathbb{A}}(X;\theta)_{s,t}^{(n)} &= \int_s^t \int_s^{t_n} \cdots \int_s^{t_2} \big(e^{-(\theta_t - \theta_{t_1})A} B\, dX_{t_1}\big) \otimes \cdots \otimes \big(e^{-(\theta_t - \theta_{t_n})A} B\, dX_{t_n}\big) \\
&= \underbrace{\int_s^u \int_s^{t_n} \cdots \int_s^{t_2} \big(e^{-(\theta_t - \theta_{t_1})A} B\, dX_{t_1}\big) \otimes \cdots \otimes \big(e^{-(\theta_t - \theta_{t_n})A} B\, dX_{t_n}\big)}_{(\mathbf{I})} \\
&\quad + \underbrace{\int_u^t \int_u^{t_n} \cdots \int_u^{t_2} \big(e^{-(\theta_t - \theta_{t_1})A} B\, dX_{t_1}\big) \otimes \cdots \otimes \big(e^{-(\theta_t - \theta_{t_n})A} B\, dX_{t_n}\big)}_{(\mathbf{II})} \\
&\quad + \underbrace{\sum_{k=1}^{n-1} \int_u^t \int_u^{t_n} \cdots \int_u^{t_{k+2}} \int_s^u \int_s^{t_k} \cdots \int_s^{t_2} \big(e^{-(\theta_t - \theta_{t_1})A} B\, dX_{t_1}\big) \otimes \cdots \otimes \big(e^{-(\theta_t - \theta_{t_n})A} B\, dX_{t_n}\big)}_{(\mathbf{III})}
\end{aligned}$$

where $(\mathbf{I}) = D_A^{\theta_t-\theta_u} S_{\mathbb{A}}(X;\theta)_{s,u}^{(n)}$ and $(\mathbf{II}) = S_{\mathbb{A}}(X;\theta)_{u,t}^{(n)}$. We then expand $(\mathbf{III})$ as

$$(\mathbf{III}) = \sum_{k=1}^{n-1} \underbrace{\Big(\int_s^u \int_s^{t_k} \cdots \int_s^{t_2} \big(e^{-(\theta_t-\theta_{t_1})A} B\, dX_{t_1}\big) \otimes \cdots \otimes \big(e^{-(\theta_t-\theta_{t_k})A} B\, dX_{t_k}\big)\Big)}_{(\mathbf{IV})}$$
$$\otimes \underbrace{\Big(\int_u^t \int_u^{t_n} \cdots \int_u^{t_{k+2}} \big(e^{-(\theta_t-\theta_{t_{k+1}})A} B\, dX_{t_{k+1}}\big) \otimes \cdots \otimes \big(e^{-(\theta_t-\theta_{t_n})A} B\, dX_{t_n}\big)\Big)}_{(\mathbf{V})},$$

where $(\mathbf{V}) = S_{\mathbb{A}}(X;\theta)_{u,t}^{(n-k)}$. For $(\mathbf{IV})$, it remains to introduce the correct factors involving u:

$$\begin{aligned}(\mathbf{IV}) &= \int_s^u \int_s^{t_k} \cdots \int_s^{t_2} e^{-(\theta_t-\theta_u)A} e^{-(\theta_u-\theta_{t_1})A} B\, dX_{t_1} \otimes \cdots \otimes e^{-(\theta_t-\theta_u)A} e^{-(\theta_u-\theta_{t_k})A} B\, dX_{t_k} \\ &= \big(e^{-(\theta_t-\theta_u)A}\big)^{\otimes k} \int_s^u \int_s^{t_k} \cdots \int_s^{t_2} e^{-(\theta_u-\theta_{t_1})A} B\, dX_{t_1} \otimes \cdots \otimes e^{-(\theta_u-\theta_{t_k})A} B\, dX_{t_k} \\ &= D_A^{\theta_t-\theta_u} S_{\mathbb{A}}(X;\theta)_{s,u}^{(k)}.\end{aligned}$$

Putting everything together, we get

$$S_{\mathbb{A}}(X;\theta)_{s,t}^{(n)} = \sum_{k=0}^{n} D_A^{\theta_t-\theta_u} S_{\mathbb{A}}(X;\theta)_{s,u}^{(k)} \otimes S_{\mathbb{A}}(X;\theta)_{u,t}^{(n-k)}.$$

□

The above proof will not hold for $p \in (1,2)$ since it makes use of Fubini's theorem. We will extend the statement by using the continuity of the truncated EWS in the p-variation topology (Corollary 2), which allows us to pass to the limit from bounded variation approximations.

Corollary 3. *Let $X \in \mathcal{V}^p([0,T];V)$ for $p < 2$, and let θ be a continuous clock on $\mathcal{V}^{p'}([0,T];V)$ for some $p' \in (p,2)$. Then for any $0 \le s \le u \le t \le T$, the exponentially-weighted signature satisfies*

$$S_{\mathbb{A}}(X;\theta)_{s,t} = \big(D_A^{\theta_t-\theta_u} S_{\mathbb{A}}(X;\theta)_{s,u}\big) \otimes S_{\mathbb{A}}(X;\theta)_{u,t}.$$

Proof. Fix $s \le u \le t$ and let p' be as in the statement. By the density of smooth (hence finite variation) paths in $\mathcal{V}^p([0,T];V)$ for the p'-variation topology (Friz & Victoir, 2010), there exists a sequence $(X^{(m)})_{m\ge 0}$ of finite variation paths such that $X^{(m)} \to X$ in p'-variation as $m \to \infty$.

By Lemma 7, Chen's identity holds for each finite variation path $X^{(m)}$:

$$S_{\mathbb{A}}(X^{(m)};\theta)_{s,t} = \Big(D_A^{\theta_t^{(m)}-\theta_u^{(m)}} S_{\mathbb{A}}(X^{(m)};\theta)_{s,u}\Big) \otimes S_{\mathbb{A}}(X^{(m)};\theta)_{u,t},$$

where $\theta^{(m)} := \theta(X^{(m)})$ denotes the clock evaluated along $X^{(m)}$. Fix an arbitrary depth $n \ge 0$. Projecting Chen's identity to depth n gives

$$S_{\mathbb{A}}(X^{(m)};\theta)_{s,t}^{(n)} = \sum_{k=0}^{n} \Big(D_A^{\theta_t^{(m)}-\theta_u^{(m)}} S_{\mathbb{A}}(X^{(m)};\theta)_{s,u}\Big)^{(k)} \otimes \Big(S_{\mathbb{A}}(X^{(m)};\theta)_{u,t}\Big)^{(n-k)}.$$

We now pass to the limit as $m \to \infty$. By Corollary 2, the truncated EWS is continuous in the p'-variation topology, so

$$\begin{aligned}\pi_n\big(S_{\mathbb{A}}(X^{(m)};\theta)_{s,t}\big) &\to \pi_n\big(S_{\mathbb{A}}(X;\theta)_{s,t}\big), \\ \pi_k\big(S_{\mathbb{A}}(X^{(m)};\theta)_{s,u}\big) &\to \pi_k\big(S_{\mathbb{A}}(X;\theta)_{s,u}\big) \quad \text{for each } k \le n, \\ \pi_{n-k}\big(S_{\mathbb{A}}(X^{(m)};\theta)_{u,t}\big) &\to \pi_{n-k}\big(S_{\mathbb{A}}(X;\theta)_{u,t}\big) \quad \text{for each } k \le n.\end{aligned}$$

Since the clock is continuous, $\theta^{(m)} \to \theta$ in p'-variation, and in particular uniformly: $\theta_r^{(m)} \to \theta_r$ for all $r \in [0,T]$. In particular, $\theta_t^{(m)} - \theta_u^{(m)} \to \theta_t - \theta_u$. The flow operator D_A^r depends continuously

on r: for any fixed $\mathbf{a} \in T^n(W)$, the map $r \mapsto D_A^r \mathbf{a}$ is continuous since $r \mapsto e^{-rA}$ is continuous in operator norm. Combined with continuity in the tensor algebra argument, we have

$$D_A^{\theta_t^{(m)} - \theta_u^{(m)}} \pi_k\big(S_{\mathbb{A}}(X^{(m)};\theta)_{s,u}\big) \to D_A^{\theta_t - \theta_u} \pi_k\big(S_{\mathbb{A}}(X;\theta)_{s,u}\big)$$

for each $k \le n$. Since the tensor product $\otimes\colon T^k(W) \times T^{n-k}(W) \to T^n(W)$ is continuous, taking limits in the projected Chen identity yields

$$S_{\mathbb{A}}(X;\theta)_{s,t}^{(n)} = \sum_{k=0}^{n} \Big(D_A^{\theta_t - \theta_u} S_{\mathbb{A}}(X;\theta)_{s,u}\Big)^{(k)} \otimes \big(S_{\mathbb{A}}(X;\theta)_{u,t}\big)^{(n-k)} = \Big(\big(D_A^{\theta_t - \theta_u} S_{\mathbb{A}}(X;\theta)_{s,u}\big) \otimes S_{\mathbb{A}}(X;\theta)_{u,t}\Big)^{(n)}.$$

Since $S_{\mathbb{A}}(X;\theta)_{s,t}$ and $\big(D_A^{\theta_t - \theta_u} S_{\mathbb{A}}(X;\theta)_{s,u}\big) \otimes S_{\mathbb{A}}(X;\theta)_{u,t}$ agree at every depth $n \ge 0$, they are equal as elements of $T((W))$. □

Remark 4. Beyond its algebraic significance, the modified Chen identity produces the associative operation of the parallel associative scan of Section 2.2, which aggregates local EWS pieces in $O(\log J)$ parallel steps rather than $O(J)$ sequential ones, for J observations.

Lemma 8. *On $\mathbb{R} \times T((W))$ define $(r_1, S_1) \star (r_2, S_2) = \big(r_1 + r_2,\, D_A^{r_2}(S_1) \otimes S_2\big)$. Then $\star$ is associative with identity $(0, \mathbf{1})$, and $\pi_N\big(D_A^{r_2}(S_1) \otimes S_2\big) = \pi_N\big(D_A^{r_2}(\pi_N S_1) \otimes \pi_N S_2\big)$ for every N. For times $0 \le t_0 < t_1 < \cdots < t_J \le T$ and $g_j = \big(\theta_{t_j} - \theta_{t_{j-1}},\, S_{\mathbb{A}}(X;\theta)_{t_{j-1},t_j}\big)$,*

$$g_1 \star \cdots \star g_j = \big(\theta_{t_j} - \theta_{t_0},\, S_{\mathbb{A}}(X;\theta)_{t_0,t_j}\big), \qquad j = 1, \dots, J.$$

Proof. By Lemma 1 and the homomorphism property of $D_A^{r_3}$,

$$\big((r_1, S_1) \star (r_2, S_2)\big) \star (r_3, S_3) = \big(r_1 + r_2 + r_3,\, D_A^{r_2 + r_3}(S_1) \otimes D_A^{r_3}(S_2) \otimes S_3\big) = (r_1, S_1) \star \big((r_2, S_2) \star (r_3, S_3)\big),$$

and $(0, \mathbf{1})$ is an identity since $D_A^0 = \mathrm{Id}$ and $D_A^r(\mathbf{1}) = \mathbf{1}$. The truncation identity holds because $D_A^{r_2}$ is level-preserving and the level-n part of a product involves only levels up to n of each factor. The last claim follows simply by induction on j. □

Linearisation of shuffle product. We now establish the fact that products of linear functionals of the EWS can be linearised via the shuffle product (Theorem 1(i)). This follows immediately by reducing the EWS over an interval to the signature.

Lemma 9. *Let $X \in \mathcal{V}^p([0,T]; V)$ for $p < 2$ and $A \in L(W, W)$ bounded. Then for all $[s,t] \subseteq [0,T]$ and $l_1, l_2 \in T(W^\star)$,*

$$\langle l_1, S_{\mathbb{A}}(X;\theta)_{s,t} \rangle \langle l_2, S_{\mathbb{A}}(X;\theta)_{s,t} \rangle = \langle l_1 \sqcup\!\sqcup l_2, S_{\mathbb{A}}(X;\theta)_{s,t} \rangle. \tag{35}$$

Proof. By Proposition 3, $S_{\mathbb{A}}(X;\theta)_{s,t} = S(Z^{[t]})_{s,t}$. The claim is then Theorem 3 applied to the re-weighted path $Z^{[t]}$. □

Reparametrisation. The signature is invariant under reparametrisation of the path, while for the EWS, this property depends on the clock.

Proposition 4. *Let $X \in \mathcal{V}^p([0,T]; V)$ for $p < 2$ and φ be an increasing homeomorphism of $[0,T]$. Suppose that $\theta(X \circ \varphi) = \theta(X) \circ \varphi$. Then for all $[s,t] \subseteq [0,T]$,*

$$S_{\mathbb{A}}(X \circ \varphi;\theta)_{s,t} = S_{\mathbb{A}}(X;\theta)_{\varphi(s),\varphi(t)},$$

and in particular $S_{\mathbb{A}}(X \circ \varphi;\theta)_{0,T} = S_{\mathbb{A}}(X;\theta)_{0,T}$.

Proof. Denote $\tilde{X} = X \circ \varphi$ and $\tilde{\theta} = \theta(\tilde{X}) = \theta(X) \circ \varphi$, and let $\tilde{Z}^{[t]}$ be the re-weighted path of $\tilde{X}$. For $\tau \in [s,t]$,

$$\tilde{Z}_\tau^{[t]} = \int_s^\tau e^{-(\tilde{\theta}_t - \tilde{\theta}_v)A} B\, d\tilde{X}_v = \int_s^\tau e^{-(\theta_{\varphi(t)} - \theta_{\varphi(v)})A} B\, dX_{\varphi(v)} = \int_{\varphi(s)}^{\varphi(\tau)} e^{-(\theta_{\varphi(t)} - \theta_u)A} B\, dX_u = Z_{\varphi(\tau)}^{[\varphi(t)]},$$

where the third equality uses the substitution $u = \varphi(v)$. Thus, $\tilde{Z}^{[t]} = Z^{[\varphi(t)]} \circ \varphi$ on $[s,t]$. By Proposition 3 and the invariance of the signature under reparametrisation (Lyons et al., 2007),

$$S_{\mathbb{A}}(\tilde{X};\theta)_{s,t} = S(\tilde{Z}^{[t]})_{s,t} = S(Z^{[\varphi(t)]})_{\varphi(s),\varphi(t)} = S_{\mathbb{A}}(X;\theta)_{\varphi(s),\varphi(t)}.$$

□

The variation clock satisfies the required condition for all φ, as $\|X \circ \varphi\|_{1,[0,t]} = \|X\|_{1,[0,\varphi(t)]}$. However, the time clock does not, since trivially $\theta(X \circ \varphi)_t = t$ holds whereas $(\theta(X) \circ \varphi)_t = \varphi(t)$. Therefore, the EWS with time clock depends on the parametrisation of the input path. For instance, with $V = W = \mathbb{R}$, $A = B = 1$, $T = 1$, $X_u = u$ and $\varphi(u) = u^2$, depth one gives $S_{\mathbb{A}}(X;\theta)^{(1)}_{0,1} = \int_0^1 e^{-(1-u)}\, du = 1 - e^{-1}$, whereas $S_{\mathbb{A}}(X \circ \varphi;\theta)^{(1)}_{0,1} = \int_0^1 e^{-(1-v)}\, 2v\, dv = 2e^{-1}$.

LTI SSMs, which are depth-one EWS with the time clock (Section 4), thus depend on the parametrisation of their input, whereas the EWS is invariant under reparametrisation whenever its clock satisfies the condition of Proposition 4. Among the clocks of Table 1, the variation clock and the gated clock with $b = 0$ satisfy it, while the time, blended and clamped clocks do not. The following proposition shows that every clock built from path increments with $f(cv) = cf(v)$ satisfies the condition.

Proposition 5. *Let X be a piecewise linear path with observation times $t_0 < \cdots < t_J$, let θ be the clock of the form $\theta_{t_j} = \sum_{i=1}^{j} f(\Delta X_i)$, and suppose that $f\colon V \to \mathbb{R}$ satisfies $f(cv) = cf(v)$ for all $c > 0$. Then $\theta(X \circ \varphi) = \theta(X) \circ \varphi$ for every piecewise linear increasing homeomorphism φ of $[0, T]$.*

Proof. Write $\tilde{X} = X \circ \varphi$. Define $0 = s_0 < s_1 < \cdots < s_K = T$ to be the points of the set $\{\varphi^{-1}(t_0), \ldots, \varphi^{-1}(t_J)\} \cup \{\text{breakpoints of } \varphi\text{'s piecewise linear interpolation}\}$. Thus, φ is linear on each $[s_{k-1}, s_k]$ with image inside a single interval $[t_{j-1}, t_j]$, on which X is linear. Thus, $\tilde{X}$ is linear on $[s_{k-1}, s_k]$, with increment $c_k \Delta X_j$, where

$$c_k = \frac{\varphi(s_k) - \varphi(s_{k-1})}{t_j - t_{j-1}} > 0.$$

Treating each s_k as an observation time of $\tilde{X}$, the assumption on f gives that the piece $[s_{k-1}, s_k]$ advances the clock of $\tilde{X}$ by $f(c_k \Delta X_j) = c_k f(\Delta X_j)$. The pieces covering a whole interval $[t_{i-1}, t_i]$ have c_k summing to one, and together advance the clock by $f(\Delta X_i)$, while the pieces covering $[t_{j-1}, \varphi(s_k)]$ advance it by $\frac{\varphi(s_k) - t_{j-1}}{t_j - t_{j-1}} f(\Delta X_j)$. Hence,

$$\theta(\tilde{X})_{s_k} = \theta(X)_{t_{j-1}} + \frac{\varphi(s_k) - t_{j-1}}{t_j - t_{j-1}}\, f(\Delta X_j) = \theta(X)_{\varphi(s_k)},$$

where the last equality holds because $\theta(X)$ is linear on $[t_{j-1}, t_j]$. Both $\theta(\tilde{X})$ and $\theta(X) \circ \varphi$ are linear on each $[s_{k-1}, s_k]$ and agree at every s_k, and are therefore equal. □

C.7 Uniqueness and universal approximation

The signature of a path uniquely determines the path up to tree-like equivalence, and when the path includes a strictly monotone component this reduces to uniqueness up to translation and reparametrisation (Boedihardjo et al., 2016; Hambly & Lyons, 2010). While the EWS is the signature of a re-weighted path, distinguishing two re-weighted paths requires them not to be tree-like equivalent, for instance by having monotone components, and it is not obvious that this property is inherited from the input paths, since re-weighting mixes channels through $\mathbb{A}$. This section characterises when two paths yield the same EWS for a given $\mathbb{A}$ and clock, establishes uniqueness under structural assumptions on A and B, and proves universal approximation in full generality.

Characterisation of EWS equivalence. We begin by characterising the equivalence relation induced by the EWS. Two paths are $(\mathbb{A};\theta)$-equivalent if they have the same EWS for a given $\mathbb{A}$ and clock θ. Since the EWS of a path can be viewed as the classical signature of its corresponding re-weighted path, this equivalence is precisely the pull-back of tree-like equivalence through the re-weighting map $X \mapsto Z^{[\cdot]}$.

Definition 12. Let $X, Y \in \mathcal{V}^p([0,T];V)$ for $p < 2$, let $\mathbb{A} = (A, B)$ with $A \in L(W, W)$ and $B \in L(V, W)$ bounded linear operators, and let θ be a clock. We say X and Y are $(\mathbb{A};\theta)$-equivalent over $[s, t] \subseteq [0, T]$, denoted $X \sim_{\mathbb{A};\theta} Y$, if

$$S_{\mathbb{A}}(X;\theta)_{s,t} = S_{\mathbb{A}}(Y;\theta)_{s,t}.$$

By Proposition 3, $X \sim_{\mathbb{A};\theta} Y$ if and only if the re-weighted paths $Z^{X,[t]}$ and $Z^{Y,[t]}$ of X and Y are tree-like equivalent (Hambly & Lyons, 2010). For $A = 0$ and B the identity this is tree-like equivalence of X and Y; below we give structural conditions under which it determines the path.

Uniqueness of the EWS for structured A. As for the signature of a path with a time channel, uniqueness follows if the re-weighted path inherits a monotone channel from the input path. This is a sufficient condition only. We thus consider paths $X \in \mathcal{V}^p([0,T];V)$ whose first component $c = \pi_1(X)$ is strictly monotone. Such a component can be obtained by augmenting a path with any clock of Table 1 whose increments are strictly positive, such as time or the blended clock (the clock that performs the re-weighting remains arbitrary). The re-weighted path then inherits a monotone channel under the following structure on A and B.

Lemma 10. *Let $X \in \mathcal{V}^p([0,T];V)$ for $p < 2$ with $c = \pi_1(X)$ strictly monotone. Let $[s,t] \subseteq [0,T]$ and $\pi\colon W \to \mathbb{R}$ be a coordinate projection. Suppose that $A \in L(W,W)$ and $B \in L(V,W)$ satisfy*

$$\pi \circ e^{-rA} = e^{-\alpha r} \cdot \pi \quad \forall r \in \mathbb{R}, \qquad \pi \circ B = \beta \cdot \pi_1, \tag{36}$$

with $\alpha \in \mathbb{R}$ and $\beta \neq 0$. Then $\zeta = \pi(Z^{[t]})$ is strictly monotone, with

$$\zeta_\tau = \beta \int_s^\tau e^{-\alpha(\theta_t - \theta_u)}\, dc_u.$$

Proof. For any $v \in V$, the conditions (36) say that this coordinate of $e^{-rA}Bv$ depends only on the first component of v:

$$\pi(e^{-rA}Bv) = e^{-\alpha r} \cdot \beta\, \pi_1(v).$$

Applying this to the re-weighted path gives

$$\zeta_\tau = \pi\Big(\int_s^\tau e^{-(\theta_t-\theta_u)A} B\, dX_u\Big) = \int_s^\tau \pi\big(e^{-(\theta_t-\theta_u)A} B\, dX_u\big) = \beta \int_s^\tau e^{-\alpha(\theta_t-\theta_u)}\, dc_u.$$

Since $e^{-\alpha(\theta_t-\theta_u)} > 0$ for all $u \in [s,\tau]$, $\beta \neq 0$ and c is strictly monotone, ζ is strictly monotone in τ. □

Remark 5. The condition on A holds whenever $\pi \circ A = \alpha \cdot \pi$. For $W = \mathbb{R}^m$, $V = \mathbb{R}^d$ and π the k-th coordinate, the two conditions say that the k-th row of A is α in position k and zero elsewhere, and that the k-th row of B is $(\beta, 0, \dots, 0)$.

Corollary 4. *Let $X, Y \in \mathcal{V}^p([0,T];V)$ with $p < 2$ have a strictly monotone component, the same starting point, and let B be injective. Suppose that the first two coordinates of W satisfy the conditions of Lemma 10, with rates $\alpha_1 = 0$ and $\alpha_2 = \alpha \neq 0$. Then $S_{\mathbb{A}}(X;\theta)_{0,T} = S_{\mathbb{A}}(Y;\theta)_{0,T}$ implies that $X = Y \circ \varphi$ for some increasing homeomorphism φ of $[0,T]$.*

Proof. By Proposition 3, $S_{\mathbb{A}}(X;\theta)_{0,T} = S_{\mathbb{A}}(Y;\theta)_{0,T}$ implies that $S(Z^{X,[T]})_{0,T} = S(Z^{Y,[T]})_{0,T}$. By Lemma 10 applied to the first coordinate, both re-weighted paths have a strictly monotone channel, and thus $Z^{X,[T]} = Z^{Y,[T]} \circ \psi$ for some increasing homeomorphism ψ of $[0,T]$ (Boedihardjo et al., 2016), the translation vanishing since both paths start at zero. Write $\theta^X = \theta(X)$, $\theta^Y = \theta(Y)$, $c^X = \pi_1(X)$ and $c^Y = \pi_1(Y)$. The first coordinates give $c^X_\tau - c^X_0 = c^Y_{\psi(\tau)} - c^Y_0$, and since the first components start at the same point, $c^X = c^Y \circ \psi$. The second coordinates give, after substituting $u = \psi(v)$,

$$\int_0^\tau \Big(e^{-\alpha(\theta^X_T - \theta^X_v)} - e^{-\alpha(\theta^Y_T - \theta^Y_{\psi(v)})}\Big)\, dc^X_v = 0, \qquad \tau \in [0,T].$$

Since c^X is strictly monotone and the integrand is continuous, the integrand vanishes, and since $\alpha \neq 0$, $\theta^X_T - \theta^X_v = \theta^Y_T - \theta^Y_{\psi(v)}$ for every v. The weights then agree, $e^{-(\theta^Y_T - \theta^Y_{\psi(v)})A} = e^{-(\theta^X_T - \theta^X_v)A}$, and $Z^{X,[T]} = Z^{Y,[T]} \circ \psi$ gives

$$\int_0^\tau e^{-(\theta^X_T - \theta^X_v)A} B\, d\big(X - Y \circ \psi\big)_v = 0, \qquad \tau \in [0,T].$$

Since each $e^{-(\theta^X_T - \theta^X_v)A}$ is invertible and B is injective, $X - Y \circ \psi$ is constant, and it vanishes at 0. Taking $\varphi = \psi$ completes the proof. □

These conditions are mild and only prevent the other channels from influencing the two coordinates that read the monotone channel; this leaves oscillatory modes, coupled decay between data channels and clock-dependent decay available; moreover, the monotone channel may still enter every other coordinate of W through B.

Universal approximation theorem. Despite the subtleties of uniqueness, universal approximation holds in full generality. The EWS is always a universal feature map for the class of functions it can distinguish; that is, functions constant on $(\mathbb{A};\theta)$-equivalence classes. When uniqueness holds, this class is all continuous functions.

Lemma 11. *Let $K \subset \mathcal{V}^p([0,T];V)$ for $p<2$ be a compact set of paths and let θ be a continuous clock. For any continuous $F\colon K \to \mathbb{R}$ satisfying*

$$X \sim_{\mathbb{A};\theta} Y \;\Rightarrow\; F(X) = F(Y),$$

and any $\epsilon > 0$, there exists an $l \in T(W^\star)$ such that

$$\sup_{X\in K} \left|F(X) - \langle l, S_{\mathbb{A}}(X;\theta)_{0,T}\rangle\right| < \epsilon. \tag{37}$$

Proof. Let $\mathcal{A} = \{X \mapsto \langle l, S_{\mathbb{A}}(X;\theta)_{0,T}\rangle : l \in T(W^\star)\}$. Each function in $\mathcal{A}$ involves finitely many levels and is continuous on K by Corollary 2. The empty word gives the constants, and $\mathcal{A}$ is closed under multiplication by Lemma 9, making it a subalgebra of $C(K,\mathbb{R})$. If $S_{\mathbb{A}}(X;\theta)_{0,T} \neq S_{\mathbb{A}}(Y;\theta)_{0,T}$, the two differ at some depth n, and the Hahn–Banach theorem (Rudin, 1991) gives $\phi \in (W^{\otimes n})^\star$, placed at depth n in $T(W^\star)$, that separates them. The algebra $\mathcal{A}$ therefore separates the points of $K/\!\sim_{\mathbb{A};\theta}$, and the Stone–Weierstrass theorem gives (37) for every F as in the statement. □

By Corollary 4, when A and B satisfy its conditions and K consists of paths with a strictly monotone component and a common starting point, every continuous F with $F(X\circ\varphi) = F(X)$ whenever $X, X\circ\varphi \in K$ satisfies the hypothesis of Lemma 11, and is therefore approximated uniformly by linear functionals of the EWS. When the first component of every path in K is time, $X\circ\varphi$ lies in K only for φ the identity, and every continuous F on K qualifies. This mirrors the signature, whose linear functionals are universal on unparametrised paths and on all paths once a time channel is added.

C.8 Spectral properties at depth one

Throughout this subsection, we work with paths in Euclidean space as in the main text. Section 2.1 states that a real diagonal generator approximates the damped oscillatory kernels of a rotation generator only with much larger width; Proposition 6 quantifies this. We also prove Proposition 2 alongside the claims that follow it.

Proposition 6. *Let $A = \mathrm{diag}(\lambda_1,\dots,\lambda_m)$ be real and $c \in \mathbb{R}^m$. The linear readout $\langle c, S_{\mathbb{A}}(X;\theta)^{(1)}_{0,t}\rangle$ applies to each channel X^j the kernel $k_j(r) = \sum_{k=1}^m c_k B_{kj} e^{-\lambda_k r}$ in elapsed clock time r. If $\sup_{r\geq 0}|k_j(r) - e^{-ar}\cos(\omega r)| \leq \varepsilon$ for some j, with $a,\omega>0$ and $\varepsilon\in(0,1)$, then $m \geq \frac{\omega}{\pi a}\log\frac{1}{\varepsilon}$.*

Proof. Set

$$L := \frac{\omega}{\pi a}\log\frac{1}{\varepsilon} \qquad \text{and} \qquad r_n := \frac{n\pi}{\omega}.$$

For every integer $n \geq 0$ with $n < L$,

$$e^{-ar_n}\cos(\omega r_n) = (-1)^n e^{-a\pi n/\omega},$$

whose modulus is greater than ε. Since

$$\left|k_j(r_n) - (-1)^n e^{-a\pi n/\omega}\right| \leq \varepsilon,$$

it follows that $k_j(r_n)$ has sign $(-1)^n$. Let

$$M := \#\{n \in \mathbb{Z}_{\geq 0} : n < L\} = \lceil L\rceil.$$

Since k_j has opposite signs at consecutive points, by continuity it has at least $M-1$ distinct zeros on $[0,\infty)$. After combining terms with equal λ_k, k_j is a linear combination of at most m exponentials with distinct real exponents. Moreover, it is non-zero, since

$$k_j(0) \geq 1 - \varepsilon > 0.$$

A non-zero real exponential sum with at most m distinct exponents has at most $m-1$ zeros on $[0,\infty)$. This follows by induction on m, since dividing by one exponential and differentiating removes one term, while between any two distinct zeros there is a zero of the derivative. Therefore

$$M - 1 \leq m - 1,$$

and hence

$$m \geq M = \lceil L \rceil \geq L = \frac{\omega}{\pi a} \log \frac{1}{\varepsilon},$$

as claimed. □

Proposition 6 shows that a real diagonal generator can approximate a damped oscillation to within ε only with width $m \gtrsim \frac{\omega}{a} \log \frac{1}{\varepsilon}$, where ω and a are the frequency and the damping rate of the oscillation respectively. Proposition 2 states that a Jordan block generator produces depth-one coordinates with polynomial–exponential kernels, which no diagonalisable generator reaches, and the discussion that follows it compares these kernels with the signature and the EFM at deeper levels. Proposition 7 restates Proposition 2 as parts (i) and (ii), and makes this comparison precise as part (iii). For a channel X^i of X, $\lambda \geq 0$ and an integer $\ell \geq 0$, let

$$g_\ell(X)_t = \int_0^t \frac{(t-u)^\ell}{\ell!} e^{-\lambda(t-u)}\, dX_u^i.$$

Proposition 7. *Let $\theta_t = t$.*

(i) Consider the EWS $S_{\mathbb{A}}(X;\theta)$ with a generator A in Jordan normal form, and let $J = \lambda I + N \in \mathbb{R}^{(K+1)\times(K+1)}$ be one of its blocks with $\lambda \geq 0$, $Ne_{k+1} = e_k$ for $k = 1,\dots,K$ and $Ne_1 = 0$. Suppose that the rows of B corresponding to the coordinates of J form the matrix $e_{K+1}e_i^\top \in \mathbb{R}^{(K+1)\times d}$, which lifts the channel X^i into the last coordinate of the block. Then $g_\ell(X)_t$ is, up to sign, a depth-one coordinate of $S_{\mathbb{A}}(X;\theta)_{0,t}$ for every $\ell \leq K$.

(ii) For $\ell \geq 1$, there is no real generator A, of any size, that is diagonalisable over $\mathbb{C}$, and no B for which $g_\ell(X)_t$ is a linear readout of depth one of $S_{\mathbb{A}}(X;\theta)_{0,t}$.

(iii) Let $\ell \geq 1$. For $p \geq 1$, a polynomial of degree p in g_ℓ is a linear functional of the EWS of (i), with $K \geq \ell$, at depth p. When $\lambda = 0$, $g_\ell(X)_t$ is a linear functional of the signature of (t, X) at depth $\ell + 1$ and at no smaller depth, and when $\lambda > 0$, it is not one at any finite depth. For every $\lambda \geq 0$, it is not a linear functional of the EFM of (t, X) at any finite depth.

Proof. (i) Recall that $S_{\mathbb{A}}(X;\theta)^{(1)}_{0,t} = \int_0^t e^{-(t-u)A} B\, dX_u$. Since A is in Jordan normal form, e^{-rA} is block diagonal, with e^{-rJ} as the block corresponding to J. The components of depth one in the coordinates of J are therefore

$$\int_0^t e^{-(t-u)J}\, (B\, dX_u)_J,$$

where $(B\, dX_u)_J$ denotes the components of $B\, dX_u$ in these coordinates, given by the rows of B corresponding to J. These rows are $e_{K+1}e_i^\top$, and

$$(B\, dX_u)_J = e_{K+1}e_i^\top\, dX_u = e_{K+1}\, dX_u^i.$$

Hence the components of depth one in the coordinates of J are

$$\int_0^t e^{-(t-u)J} e_{K+1}\, dX_u^i.$$

Now $J = \lambda I + N$, and λI commutes with N. Therefore, for $r \geq 0$, we have that $e^{-rJ}e_{K+1} = e^{-\lambda r}e^{-rN}e_{K+1}$. Since $Ne_{k+1} = e_k$, we have $N^\ell e_{K+1} = e_{K+1-\ell}$, $0 \leq \ell \leq K$, and $N^{K+1}e_{K+1} = 0$. Hence, the exponential series terminates and gives

$$e^{-rJ}e_{K+1} = e^{-\lambda r} \sum_{\ell=0}^{K} \frac{(-r)^\ell}{\ell!} e_{K+1-\ell}.$$

Letting $r = t - u$, the $(K + 1 - \ell)$-th of these components is thus

$$(-1)^\ell \int_0^t \frac{(t-u)^\ell}{\ell!} e^{-\lambda(t-u)}\, dX_u^i = (-1)^\ell g_\ell(X)_t,$$

for every $\ell \le K$, as required.

(ii) Suppose, for a contradiction, that for some $t > 0$, some $A \in \mathbb{R}^{m\times m}$ diagonalisable over $\mathbb{C}$, some $B \in \mathbb{R}^{m\times d}$, and some $c \in \mathbb{R}^m$,

$$\big\langle c, S_{\mathbb{A}}(X;\theta)^{(1)}_{0,t} \big\rangle = g_\ell(X)_t \tag{38}$$

for every path X. By the formula for depth one,

$$\big\langle c, S_{\mathbb{A}}(X;\theta)^{(1)}_{0,t} \big\rangle = \sum_{j=1}^d \int_0^t k_j(t-u)\, dX_u^j, \qquad k_j(r) = c^\top e^{-rA} B e_j.$$

Fix $\tau \in [0, t)$ and $\delta \in (0, t - \tau)$. Let X be the path whose channels other than X^i are constant, and whose channel X^i increases linearly on $[\tau, \tau + \delta]$ with a total increase of one, and is constant elsewhere. For this path, (38) becomes

$$\frac{1}{\delta}\int_\tau^{\tau+\delta} k_i(t-u)\, du = \frac{1}{\delta}\int_\tau^{\tau+\delta} \frac{(t-u)^\ell}{\ell!} e^{-\lambda(t-u)}\, du.$$

Letting $\delta \to 0$ and using continuity gives

$$k_i(r) = \frac{r^\ell}{\ell!} e^{-\lambda r}, \qquad r \in (0, t].$$

Since A is diagonalisable over $\mathbb{C}$, we may write $A = P\Gamma P^{-1}$, where

$$\Gamma = \operatorname{diag}(\gamma_1, \dots, \gamma_m), \qquad \gamma_k \in \mathbb{C}.$$

Hence

$$k_i(r) = c^\top P e^{-r\Gamma} P^{-1} B e_i = \sum_{k=1}^m p_k e^{-\gamma_k r}$$

for suitable $p_k \in \mathbb{C}$. After combining repeated exponents, k_i is therefore a finite linear combination of exponentials with distinct exponents. For $\ell \ge 1$, $r^\ell e^{-\lambda r}/\ell!$ cannot be such a linear combination on any interval of positive length, by the linear independence of exponential monomials. This gives a contradiction and proves (ii). Both parts (i) and (ii) extend to a complex pair $a \pm i\omega$ repeated $K + 1$ times, by the same argument applied to the realified block, whose exponential has entries $r^\ell e^{-ar}\cos(\omega r)$ and $r^\ell e^{-ar}\sin(\omega r)$ up to sign and scale.

(iii) First let channel 0 denote the time channel of (t, X). By (i), $g_\ell(X)_t$ is a depth-one coordinate of the EWS, and by Lemma 9 a polynomial of degree p in it is a linear functional of the EWS at depth p. For $\lambda = 0$,

$$g_\ell(X)_t = \int_{0<u<s_1<\cdots<s_\ell<t} dX_u^i\, ds_1 \cdots ds_\ell,$$

which is the coordinate of the signature of (t, X) associated with the word $(i, 0, \dots, 0)$ with ℓ letters 0.

For the lower bounds, suppose that $g_\ell(X)_t = \langle \psi, F(X)\rangle$ for every X, where $F(X)$ is either the signature or the EFM of (t, X) truncated at depth N, and ψ is some linear functional. Take the channels of X other than X^i constant, $\varepsilon \in \mathbb{R}$ and $X^i = \varepsilon Y$ for a smooth path Y. For this path, the only non-zero coordinates of $F(X)$ are those whose words are made up only of the letters 0 and i. The coordinate of a word with n letters i and all other letters 0 equals ε^n times its value at $\varepsilon = 1$ as each letter i contributes a factor of $dX_u^i = \varepsilon\, dY_u$. The coefficient of ε in $\langle \psi, F(X)\rangle$ therefore comes only from words with exactly one letter i. Since $g_\ell(X)_t = \varepsilon\, g_\ell(Y)_t$, comparing the coefficients of ε gives

$$\int_0^t \varphi(u)\, dY_u = \int_0^t k(u)\, dY_u, \qquad \varphi(u) = \frac{(t-u)^\ell}{\ell!} e^{-\lambda(t-u)},$$

where k is a linear combination of the kernels of the words $(0, \dots, 0, i, 0, \dots, 0)$, with q_0 letters 0 before i and q_1 after, and where $q_0 + q_1 \leq N - 1$. Taking $dY_u = (\varphi - k)(u)\, du$ shows that $\varphi = k$ on $[0, t]$.

For the signature, the coordinate of the word $(0, \dots, 0, i, 0, \dots, 0)$, with q_0 letters 0 before i and q_1 after, is an iterated integral in which the letter i integrates dY_u and each letter 0 integrates time. Integrating out the q_0 time variables before u and the q_1 time variables after u gives the kernel

$$\frac{u^{q_0}}{q_0!}\,\frac{(t-u)^{q_1}}{q_1!}.$$

Each such kernel is a polynomial in u of degree $q_0 + q_1 \leq N - 1$. Since k is a linear combination of these kernels, it is thus a polynomial of degree at most $N - 1$. When $\lambda = 0$, φ is a polynomial of degree ℓ, which forces $N \geq \ell + 1$. When $\lambda > 0$, φ is not a polynomial, and so no N suffices.

For the EFM of (t, X) with generator $\operatorname{diag}(\mu_0, \dots, \mu_d)$ and $\mu_j > 0$ for all j, the coordinate of the same word is an iterated integral in which the letter i integrates $e^{-\mu_i(t-u)}\, dY_u$ and each letter 0 integrates $e^{-\mu_0(t-s)}\, ds$. Integrating out the q_0 time variables before u and the q_1 time variables after u yields the kernel

$$\frac{1}{q_0!}\Big(\frac{e^{-\mu_0(t-u)} - e^{-\mu_0 t}}{\mu_0}\Big)^{q_0} e^{-\mu_i(t-u)}\, \frac{1}{q_1!}\Big(\frac{1 - e^{-\mu_0(t-u)}}{\mu_0}\Big)^{q_1}.$$

Each such kernel is a finite linear combination of pure exponentials in u. Since k is a linear combination of these kernels, it is thus one as well. As $\ell \geq 1$, φ contains the term $u^\ell e^{\lambda u}$, which is linearly independent of pure exponentials on $[0, t]$. Hence $\varphi \neq k$, and no N suffices, for any $\lambda \geq 0$. □

Therefore, the EFM can only approximate these kernels. At fixed depth, for $\ell = 1$, $e^{-\mu_i r}(1 - e^{-\mu_0 r})/\mu_0 \to r e^{-\lambda r}$ when $\mu_0 \to 0$ and $\mu_i \to \lambda$. At fixed rates, the words $(i, 0, \dots, 0)$ give $e^{-\mu_i r}$ times polynomials in $1 - e^{-\mu_0 r}$, which approximate the kernels uniformly on $[0, T]$ as depth grows. A Jordan block instead gives them at depth one.

D THE EFM AND VOLTERRA SIGNATURE

Both the EFM and Volterra signature weight the iterated integrals of the signature, and both agree with the EWS under certain regimes.

D.1 THE EFM

Definition 13. Let $X \in \mathcal{V}^p_{\mathrm{loc}}((-\infty, T]; \mathbb{R}^d)$ for $p < 2$, and $\lambda \in \mathbb{R}^d$ with $\lambda \succ 0$, where $\prec$ denotes the entrywise order. Assume further that there exists $\rho \in \mathbb{R}^d$ with $0 \prec \rho \prec \lambda$ and $\sup_{u \leq T} |e^{-\rho(T-u)} \odot X_u| < \infty$. The λ-exponentially fading memory signature of X is $\mathbb{X}^\lambda_{s,t} = (1, \mathbb{X}^{\lambda,1}_{s,t}, \mathbb{X}^{\lambda,2}_{s,t}, \dots) \in T((\mathbb{R}^d))$, where

$$\mathbb{X}^{\lambda,n}_{s,t} = \int_s^t \int_s^{t_n} \cdots \int_s^{t_2} e^{-\lambda(t-t_1)} \odot dX_{t_1} \otimes \cdots \otimes e^{-\lambda(t-t_n)} \odot dX_{t_n} \in (\mathbb{R}^d)^{\otimes n}, \tag{39}$$

$e^{-\lambda r}$ denotes the entrywise exponential $(e^{-\lambda_1 r}, \dots, e^{-\lambda_d r})$ and $\odot$ the Hadamard product on $\mathbb{R}^d$, $x \odot y = (x_1 y_1, \dots, x_d y_d)$.

On a finite horizon the EFM is the EWS (using non-learnable parameters) with the time clock, the identity lift and $A = \operatorname{diag}(\lambda_1, \dots, \lambda_d)$. Indeed, e^{-rA} is then diagonal with entries $e^{-\lambda_1 r}, \dots, e^{-\lambda_d r}$, so that $e^{-(t-t_l)A} dX_{t_l} = e^{-\lambda(t-t_l)} \odot dX_{t_l}$ for every l, and the iterated integrals of Definition 2 become those of (39). The infinite horizon is what forces the rates to be positive, since the improper integrals converge only if the distant past is attenuated, and positivity restricts the weights to monotone decay. A spectrum confined to the positive reals is restrictive, and no choice of lift or readout produces the oscillatory kernels of Appendix C.8, which need complex eigenvalues, or the polynomial–exponential kernels of Proposition 2, which need a Jordan block, although both can be approximated (Appendix C.8). Appendix C.8 bounds how slowly a real diagonal generator approximates the oscillatory kernels. The weighting acts on each channel separately, so channels do not interact in the time propagation, and with no lift the number of modes is the number of channels.

D.2 THE VOLTERRA SIGNATURE

At depth one, with the time clock, the EWS coincides with the Volterra signature of Hager et al. (2026) for the kernel $K(t,s) = e^{-(t-s)A}$: both reduce to $\int_s^t e^{-(t-u)A} dX_u$ (under the convention $\tau = t_{n+1} = t$ in (Hager et al., 2026, Definition 2.14)). However, beyond depth one, the two objects diverge. Rewriting (Hager et al., 2026, Definition 2.14) component-wise, the depth-n term of the Volterra signature for the word $i_1 \cdots i_n$ is

$$\mathrm{VSig}(X;K)_{s,t}^{i_1\cdots i_n} = \int_s^t \int_s^{t_n} \cdots \int_s^{t_2} \sum_{j_1,\ldots,j_n \in \{1,\ldots,d\}} \left(\prod_{k=1}^{n} K^{i_k,j_k}(t_{k+1},t_k) \right) dX_{t_1}^{j_1} \cdots dX_{t_n}^{j_n}, \tag{40}$$

where the integration is over the simplex $\Delta_{s,t}^n = \{s \le t_1 \le \cdots \le t_n \le t\}$. In the EWS, the increment dX_{t_k} at each vertex t_k of the simplex is weighted by $e^{-(\theta_t - \theta_{t_k})A}$, which depends on the distance from t_k to the apex t. In the Volterra signature, the same increment is weighted by $K(t_{k+1}, t_k)$, which depends on the gap to the next vertex t_{k+1}. The EWS weighting is thus global, anchored at the terminal time, and the Volterra weighting local, anchored at the next vertex. While both the EWS and the Volterra signature extend the diagonal operator of the EFM to a general matrix-valued operator, it is precisely the global anchoring to a fixed terminal time that makes the EWS the signature of a re-weighted path evaluated at that time (Theorem 1(i)), ensuring that the EWS is group-like while the Volterra signature is not (Hager et al., 2026, Remark 2.17).

E SSM CONVERSIONS

An SSM computes the recurrence $h_j = \bar{A} h_{j-1} + \bar{B} x_j$, usually obtained as the discretisation of the continuous dynamics $\dot{h}_t = -A h_t + B x_t$ (written in the sign convention of the EWS). Trained SSMs then either store the discrete pair $(\bar{A}, \bar{B})$ directly, or the continuous pair (A, B) alongside the step size Δ and the discretisation rule that produces the matrices used in the recurrence. However, in both cases it is the recurrence that is trained and not the continuous system. Section 4 identifies the continuous dynamics of the SSM with a depth-one EWS; that is, $h_t = S_{\mathbb{A}}(X;\theta)_{0,t}^{(1)}$ for $\mathbb{A} = (A, B)$, the time clock and $X_t = \int_0^t x_s \, ds$. Therefore, if a trained SSM stores its continuous (A, B), constructing an EWS at depth one with $\mathbb{A} = (A, B)$ would produce a model that runs the continuous dynamics of the SSM exactly, rather than the discrete recurrence that it actually computes (unless the discretisation rule is the zero-order hold). Hence, the output of the SSM and the depth-one EWS would not match at the observation times. To reproduce a trained SSM as an EWS, we must instead match its recurrence with what the EWS obeys at the observation times, which we derive first. We then go through the models of Table 3 individually and check the conversions numerically.

As in Section 4, x is held constant between observations (rectilinear interpolation), equal to x_j on $[t_{j-1}, t_j]$, and so both X and θ become linear on each interval $[t_{j-1}, t_j]$ of the grid (piecewise linear interpolation). We let $\Delta\theta_j = \theta_{t_j} - \theta_{t_{j-1}}$ and $\Delta X_j = X_{t_j} - X_{t_{j-1}}$ for the increments over it, and set

$$G_r(A) = \int_0^r e^{-sA} ds = r\varphi_1(-rA), \qquad \varphi_1(Z) = \int_0^1 e^{sZ} \, ds, \tag{41}$$

where $\varphi_1(Z) = (e^Z - I)Z^{-1}$ when Z is invertible and $r \in \mathbb{R}$. Note that $G_r(A)$ integrates the weight e^{-sA} between 0 and r and $\varphi_1(-rA)$ averages it, while remaining well-defined at $r = 0$, which the stopping clocks of Table 1 reach.

E.1 THE EWS AT THE OBSERVATION TIMES

The recurrence of the EWS follows from the modified Chen identity at depth one. Between consecutive observations, with $S_j = S_{\mathbb{A}}(X;\theta)_{0,t_j}^{(1)}$, it reads

$$S_j = e^{-\Delta\theta_j A} \, S_{j-1} + S_{\mathbb{A}}(X;\theta)_{t_{j-1},t_j}^{(1)}. \tag{42}$$

By Definition 2 the last term is

$$S_{\mathbb{A}}(X;\theta)^{(1)}_{t_{j-1},t_j} = \int_{t_{j-1}}^{t_j} e^{-(\theta_{t_j}-\theta_u)A}\, B\, dX_u = \int_0^1 e^{-(1-\sigma)\Delta\theta_j A}\, B\, \Delta X_j\, d\sigma$$
$$= \Big(\int_0^1 e^{-\sigma\Delta\theta_j A}\, d\sigma\Big) B\, \Delta X_j = \varphi_1(-\Delta\theta_j A)\, B\, \Delta X_j, \tag{43}$$

where the second equality uses that X and θ are linear on the interval, parametrised by $\sigma \in [0,1]$, the third substitutes $\sigma \mapsto 1-\sigma$, and the last is the definition of φ_1. Substituting (43) into (42) gives

$$S_j = e^{-\Delta\theta_j A} S_{j-1} + \varphi_1(-\Delta\theta_j A)\, B\, \Delta X_j. \tag{44}$$

Nothing is approximated. When the clock stops, $\Delta\theta_j = 0$ and the second term is $B\,\Delta X_j$. For the time clock on the grid $t_j = j\Delta$ of Section 4, $\Delta\theta_j = \Delta$ and $\Delta X_j = \Delta x_j$, and (44) becomes

$$S_j = e^{-\Delta A} S_{j-1} + \Delta\, \varphi_1(-\Delta A)\, B x_j = e^{-\Delta A} S_{j-1} + G_\Delta(A)\, B x_j, \tag{45}$$

which is (8). The LTI models of Table 3 have a fixed step, and are matched against (45). A selective model computes its step from the input, and is matched against (44) with that step as the clock increment.

E.2 Converting a trained model

Matching the coefficients of the model's recurrence $h_j = \bar{A} h_{j-1} + \bar{B} x_j$ with those of (45), we get

$$e^{-\Delta A} = \bar{A}, \qquad G_\Delta(A) B = \bar{B}, \tag{46}$$

which gives

$$A = -\Delta^{-1} \log \bar{A}, \qquad B = G_\Delta(A)^{-1} \bar{B}. \tag{47}$$

This is (9), and the EWS with these parameters reproduces the trained model at every observation, whatever the model stores and whatever rule produced $(\bar{A}, \bar{B})$. When the model stores continuous parameters $(\hat{A}, \hat{B})$ and discretises them by the zero-order hold, $\bar{A} = e^{-\Delta\hat{A}}$ and $\bar{B} = G_\Delta(\hat{A})\hat{B}$, the matching gives

$$A = -\Delta^{-1} \log e^{-\Delta\hat{A}} = \hat{A}, \qquad B = G_\Delta(\hat{A})^{-1} G_\Delta(\hat{A})\hat{B} = \hat{B}, \tag{48}$$

up to the branch of the logarithm. Reading off is therefore exact for the zero-order hold case. Under any other rule $(\bar{A}, \bar{B})$ only approximates $(e^{-\Delta\hat{A}}, G_\Delta(\hat{A})\hat{B})$, and the matched pair differs from the stored one.

Equation (9) requires a real logarithm of $\bar{A}$ and an invertible $G_\Delta(A)$. A real logarithm exists if and only if $\bar{A}$ is invertible and, for each negative eigenvalue, the Jordan blocks of each size belonging to it occur an even number of times (Culver, 1966), which holds in particular when $\bar{A}$ has no eigenvalue on $(-\infty, 0]$.

The condition holds for every model in Table 3, except for LinOSS-IMEX on the boundary of its stability region and for D-LinOSS when its blocks have two negative eigenvalues. Such a block, with distinct eigenvalues, is still reproduced at the observation times by an EWS of twice its dimension. The logarithm is determined only up to multiples of $2\pi i$ in each eigenvalue, and every choice reproduces $\bar{A}$, which is why frequencies are recovered only modulo $2\pi/\Delta$. We take the one whose angles $\omega\Delta$ lie in $[-\pi, \pi]$. The matrix $G_\Delta(A)$, evaluated at this generator, is then invertible, since it is singular only when a mode without decay turns a whole number of times over one step, that is, when ΔA has an eigenvalue in $2\pi i\mathbb{Z} \setminus \{0\}$.

E.3 The models

Three of the models keep a complex state, which we put in real form before converting. S4D and S5 store a complex continuous A and B, and the LRU a complex $\bar{A}$ and $\bar{B}$, all three with A or $\bar{A}$ diagonal, a complex readout C and the real output $y = \mathrm{Re}(Ch)$. The generator and lift of the EWS are real valued, and we therefore realify these matrices, writing each complex coordinate h_k of the state as the pair $(\mathrm{Re}\, h_k, \mathrm{Im}\, h_k)$. The diagonal matrix then becomes block diagonal, with one 2×2

block per diagonal entry. The k-th row of the lift becomes the two rows of its real and imaginary parts, and the k-th column of the readout becomes the two columns of its real part and minus its imaginary part. These real matrices then run the same recurrence, and the conversion then proceeds as above, giving a real generator whose blocks are the rotation blocks of Table 1, one per mode, with the eigenvalues listed in Table 3. The dense complex generator of S4 is realified in the same way, entry by entry.

S4D. S4D applies to each input channel c a single-input single-output system in which $\bar{A}$ is complex diagonal, with diagonal entries $e^{\Delta_c a_k}$, the zero-order hold of the stored $a_k = -\gamma_k + i\omega_k$ and step Δ_c. Reading off and realifying gives an EWS generator that is 2×2 block diagonal, with blocks of the form $\left(\begin{smallmatrix}\gamma_k & \omega_k\\ -\omega_k & \gamma_k\end{smallmatrix}\right)$, and multiplying each channel's blocks and lift rows by Δ_c makes the whole layer a single EWS with the unit clock.

S5. S5 applies one multi-input multi-output system to all input channels, in which $\bar{A}$ is complex diagonal, with diagonal entries $e^{\Delta t_k \Lambda_k}$, the zero-order hold of the stored $\Lambda_k = -\gamma_k + i\omega_k$ and of a step Δt_k for each mode. Reading off and realifying with the unit clock gives an EWS generator that is 2×2 block diagonal, with blocks of the form $\Delta t_k \left(\begin{smallmatrix}\gamma_k & \omega_k\\ -\omega_k & \gamma_k\end{smallmatrix}\right)$ and each lift row multiplied by the step of its mode, which is exact because the modes do not interact. The conversion is the same as for S4D, and the difference lies in the lift and readout, which in S5 are dense and mix all input channels in every mode, whereas in S4D each channel is connected only to its own modes.

S4. S4 applies to each input channel c a single-input single-output system in which $\bar{A} = (I - \frac{\Delta_c}{2}\mathcal{A})^{-1}(I + \frac{\Delta_c}{2}\mathcal{A})$ is the bilinear discretisation of the stored diagonal-plus-low-rank matrix $\mathcal{A} = \Lambda - PQ^*$ and step Δ_c. Reading off would give the EWS generator $-\mathcal{A}$ of the continuous system, which the bilinear recurrence only approximates, and the trained model is instead converted by (9). This gives a dense EWS generator whose eigenvalues are $-\frac{2}{\Delta_c}\operatorname{artanh}(\frac{\Delta_c}{2}\mu_k)$ for the eigenvalues μ_k of $\mathcal{A}$. Multiplying each channel's generator and lift by Δ_c again makes the whole layer a single EWS with the unit clock.

LRU. The LRU applies one multi-input multi-output system to all input channels, in which $\bar{A}$ is complex diagonal and stored directly, with diagonal entries $e^{-\nu_k + i\vartheta_k}$, and there is no continuous system and no step. Applying (9) with $\Delta = 1$, the unit clock, and realifying gives an EWS generator that is 2×2 block diagonal, with blocks of the form $\left(\begin{smallmatrix}\nu_k & \vartheta_k\\ -\vartheta_k & \nu_k\end{smallmatrix}\right)$, each ϑ_k taken modulo 2π in $[-\pi, \pi]$, and the lift $G_1(A)^{-1}\bar{B}$. The common form shows that the LRU and S5 are the same layer, parametrised and initialised differently (the normalisation of $\bar{B}$ in the LRU being one such choice), which affects training but not what the layer can compute.

LinOSS and D-LinOSS. Both apply one multi-input multi-output system to all input channels, whose state holds the velocity $\dot{y}_k$ and position y_k of each oscillator $\ddot{y}_k + \mathsf{G}_k\dot{y}_k + \mathsf{A}_k y_k = B_k x$, with stiffness $\mathsf{A}_k \geq 0$, damping $\mathsf{G}_k \geq 0$ and step Δt_k; LinOSS is the case $\mathsf{G}_k = 0$. The models store these parameters in the block diagonal continuous A whose 2×2 blocks $\mathsf{A}^{(k)}$, acting on $(\dot{y}_k, y_k)$, are

$$\mathsf{A}^{(k)} = \begin{pmatrix} \mathsf{G}_k & \mathsf{A}_k \\ -1 & 0 \end{pmatrix}. \tag{49}$$

Discretising the oscillator implicitly for LinOSS-IM and implicitly–explicitly for D-LinOSS yields a block diagonal $\bar{\mathsf{A}}$ with blocks

$$\bar{\mathsf{A}}^{(k)}_{\mathrm{IM}} = s_k \begin{pmatrix} 1 & -\Delta t_k \mathsf{A}_k \\ \Delta t_k & 1 \end{pmatrix} \quad \text{or} \quad \bar{\mathsf{A}}^{(k)}_{\mathrm{D}} = s_k \begin{pmatrix} 1 & -\Delta t_k \mathsf{A}_k \\ \Delta t_k & 1 + \Delta t_k \mathsf{G}_k - \Delta t_k^2 \mathsf{A}_k \end{pmatrix}, \tag{50}$$

where $s_k = (1+\Delta t_k^2 \mathsf{A}_k)^{-1}$ for LinOSS-IM and $s_k = (1+\Delta t_k \mathsf{G}_k)^{-1}$ for D-LinOSS; LinOSS-IMEX is the case of D-LinOSS with $\mathsf{G}_k = 0$. With the unit clock, reading off gives blocks $\Delta t_k \mathsf{A}^{(k)}$, which reproduce the continuous system, whereas applying (9) gives blocks $-\log \bar{\mathsf{A}}^{(k)}$, which reproduce the trained model. Their eigenvalues are those of Table 3, with $\tan\psi_k = \Delta t_k \sqrt{\mathsf{A}_k}$ for LinOSS-IM and $\cos\varphi_k = (2 + \Delta t_k \mathsf{G}_k - \Delta t_k^2 \mathsf{A}_k)/(2\sqrt{1 + \Delta t_k \mathsf{G}_k})$ for D-LinOSS. The determinant of each block's transition is $e^{-\Delta t_k \mathsf{G}_k}$ when read off and s_k when converted. These differ for LinOSS-IM, where the first is 1, leaving past inputs unattenuated, and the second is $(1 + \Delta t_k^2 \mathsf{A}_k)^{-1} < 1$, attenuating them

at every step. Thus, the attenuation of a trained LinOSS-IM comes entirely from its discretisation. LinOSS and D-LinOSS reach Jordan blocks only as edge cases. LinOSS reaches a 2×2 block with eigenvalue 0 only when $\mathsf{A}_k = 0$, and D-LinOSS a 2×2 block with eigenvalue $\frac{1}{2}\log(1 + \Delta t_k \mathsf{G}_k)$ only when $\mathsf{G}_k = 2\sqrt{\mathsf{A}_k} + \Delta t_k \mathsf{A}_k$. Each needs an exact relation between the parameters, whereas the EWS reaches blocks of any size and eigenvalue on an open set of them. Similarly, the translated-Laguerre matrix of HiPPO (Gu et al., 2020) is lower triangular with constant diagonal, and hence a single Jordan block, but it is fixed rather than learned.

Mamba-1. Mamba-1 applies to each input channel a single-input single-output system with a real diagonal A of positive entries, and computes from the input a step Δ_t for each channel and a shared lift B_t, updating the state by $h_t = e^{-\Delta_t A} h_{t-1} + \Delta_t B_t x_t$. The accumulated step $\theta_t = \sum_{s \le t} \Delta_s$ is a clock, with increments $\Delta\theta_t = \Delta_t$, and matching this update with (44) at the identity lift gives the path increments

$$\Delta X_t = \varphi_1(-\Delta_t A)^{-1} \Delta_t B_t x_t, \tag{51}$$

for which depth one of the EWS equals h_t at every step. The input-dependent lift therefore enters through the path rather than through the parameters of the layer. The steps of different channels are not proportional, and a Mamba-1 layer is therefore a family of depth-one EWS, one per channel.

Mamba-2. A Mamba-2 layer splits its channels into heads, each head having its own scalar a and its own step Δ_t, both shared by the P channels of that head, and all heads sharing the lift B_t the layer computes from the input. The head updates its matrix state by $H_t = e^{-a\Delta_t} H_{t-1} + \Delta_t B_t x_t^\top$, whose channels therefore share one clock. Matching this update with (44), and identifying $\mathbb{R}^{N \times P}$ with $\mathbb{R}^{NP}$, the head is a single depth-one EWS with generator aI, identity lift, and path increments the entries of $\varphi_1(-a\Delta_t)^{-1} \Delta_t B_t x_t^\top$. The generator of a head is thus a single scalar times the identity, where Mamba-1 has a diagonal generator with a different rate for each coordinate of a channel. A Mamba-2 layer is therefore a family of depth-one EWS, one per head.

E.4 Numerical verification

Table 7 compares each model's own recurrence, as implemented in its released code (or that of Walker et al. (2024)), with depth one of the converted EWS, computed by the EWS layer used in our experiments. The comparison runs over every step and coordinate of random inputs of length 128, with parameters drawn from each model's initialisation, state sizes between 16 and 256, three seeds and double precision throughout. It reports, for the worst seed, the largest difference between the two states divided by the largest entry of the model's own state. Every row is at round-off except D-LinOSS, whose lift requires inverting an ill-conditioned $G_\Delta(A)$.

Table 7: Conversion errors on random parameters in double precision, worst of three seeds. The route is the read-off of the stored parameters, the matching (9), or the per-channel or per-head matching of a selective model.

Model	Route	Max. rel. error
S4D	read-off	9.9×10^{-15}
S5	read-off	4.5×10^{-15}
S4	(9)	2.1×10^{-14}
LRU	(9)	1.6×10^{-14}
LinOSS-IM	(9)	4.6×10^{-15}
LinOSS-IMEX	(9)	5.8×10^{-15}
D-LinOSS	(9)	5.2×10^{-10}
Mamba-1	clocked	1.5×10^{-14}
Mamba-2	clocked	7.3×10^{-15}

On trained parameters, we convert every SSM layer of LinOSS-IM models trained on Heartbeat with five seeds, leaving the rest of each network unchanged. Test accuracy is unchanged on every seed, and the logits agree to within 10^{-13}.

F EXPERIMENTAL DETAILS

Every run uses a single NVIDIA H100 GPU. Hyperparameters are chosen on validation data alone, and every test score is taken at each seed's best-validation checkpoint, with the selection rule stated for each benchmark below. Apart from the synthetic experiments of Appendix F.3, the grids given below were the final stage of the hyperparameter selection process. Prior to them, the generator family, clock, path and readout of each dataset were chosen by a Bayesian search with the tree-structured Parzen estimator of Optuna (Bergstra et al., 2011; Akiba et al., 2019) on the same validation splits, with 50 to 100 trials per dataset. Parameter counts for the EWS come from our code, and baseline counts are taken from their papers or released configurations. We use the following publicly available datasets, libraries and baseline implementations.

- UEA Multivariate Time Series Classification Archive (Bagnall et al., 2018). Licence: GPL-3.0.
  URL: `https://www.timeseriesclassification.com`, `https://github.com/time-series-machine-learning/tsml-repo`
- PPG-DaLiA (Reiss et al., 2019), UCI Machine Learning Repository, id 495. Licence: CC BY 4.0.
  URL: `https://archive.ics.uci.edu/dataset/495`
- Weather forecasting data (Zhou et al., 2021), released with the Informer code. Licence: Apache 2.0.
  URL: `https://github.com/zhouhaoyi/Informer2020`
- Formal language benchmark (Delétang et al., 2023). Licence: Apache 2.0.
  URL: `https://github.com/google-deepmind/neural_networks_chomsky_hierarchy`
- TUH Abnormal EEG Corpus v3.0.1 (Obeid & Picone, 2016; Lopez et al., 2015), used under the corpus data use agreement.
  URL: `https://isip.piconepress.com/projects/nedc/html/tuh_eeg/`
- JAX (Bradbury et al., 2018), Equinox (Kidger & Garcia, 2021) and Optax. Licence: Apache 2.0.
  URL: `https://github.com/google/jax`, `https://github.com/patrick-kidger/equinox`, `https://github.com/google-deepmind/optax`
- Signax. Licence: MIT.
  URL: `https://github.com/Anh-Tong/signax`
- Log-NCDE (Walker et al., 2024). Licence: CC BY 4.0.
  URL: `https://github.com/Benjamin-Walker/log-neural-cdes`
- LinOSS (Rusch & Rus, 2025). Licence: MIT.
  URL: `https://github.com/tk-rusch/linoss`
- D-LinOSS (Boyer et al., 2025). Licence: MIT.
  URL: `https://github.com/jaredbmit/damped-linoss`
- BIOT (Yang et al., 2023). Licence: MIT.
  URL: `https://github.com/ycq091044/BIOT`
- S4 and S4D (Gu et al., 2022b;a), Mamba and Mamba-2 (Gu & Dao, 2024; Dao & Gu, 2024). Licence: Apache 2.0.
  URL: `https://github.com/state-spaces/s4`, `https://github.com/state-spaces/mamba`
- PyTorch (Paszke et al., 2019). Licence: BSD-3-Clause.
  URL: `https://github.com/pytorch/pytorch`
- SLiCE (Walker et al., 2025). Licence: MIT.
  URL: `https://github.com/Benjamin-Walker/structured-linear-cdes`

F.1 THE EWS BLOCK

A layer acts on a path, and the input observations x are read as its values at their observation times. The path may also be preprocessed before the layer sees it, by normalising its channels to a fixed range, by prepending a zero base-point, which lets the first observation contribute an increment of its own, by augmenting it with a channel carrying the clock, or by reading the observations as increments instead, in which case the path is their cumulative sum and is the integral $X_t = \int_0^t x_s\, ds$ of a sampled signal x. These choices are stated per experiment below and leave the layer itself unchanged.

The EWS layer of (6) can be placed inside a residual block and stacked, as deep state-space architectures stack their layers. Write H for the width of the residual stream, which an encoder produces from the d channels of the data, and $u \in \mathbb{R}^{T\times H}$ for the block's input. Its layer has width m, depth N, and thus a state of $K = \sum_{n\le N} m^n$ coordinates. The block computes

$$\begin{aligned} \theta &= \mathrm{clock}(u), \\ S_t &= S_{\mathbb{A}}(P;\theta)_{0,t}^{\le N}, \\ z &= u + C\,\Gamma(S), \\ y &= z + \mathrm{FF}\big(\mathrm{RMSNorm}(z)\big), \end{aligned} \tag{52}$$

where P is the path built from $\mathrm{RMSNorm}(u)$ as above, and the lift B of $\mathbb{A}$ carries the stream into the m channels of the layer. Figure 1 shows the order of these steps.

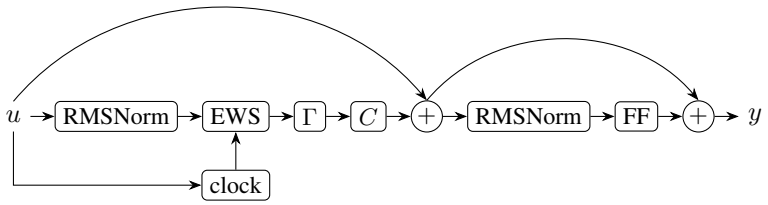


Figure 1: The EWS block. The clock reads the block's input, the state of the layer is normalised at each depth separately by Γ and read down by C, and both branches are residual.

The state S_t, which is returned at every position by a parallel associative scan, is graded, and thus its coordinates at each depth differ in their orders of magnitude (depth n is bounded by $C^n/n!$ by Lemma 6). Therefore, reading them through a single linear map would make the lower depths dominate the higher ones. The graded norm Γ normalises each depth separately,

$$\Gamma(S)^{(n)} = g_n \frac{S^{(n)}}{\max\{\rho_n, \varepsilon_n\}}, \qquad \rho_n = \Big(\tfrac{1}{m^n}\textstyle\sum_i (S_i^{(n)})^2\Big)^{1/2}, \tag{53}$$

with a learned gain g_n for each depth set from the first training batch. The read-down C linearly maps the K coordinates of $\Gamma(S)$ to the stream width, and the feed-forward branch, applied at each position separately, is $\mathrm{FF}(v) = W_1\,\mathrm{gelu}(v)\odot\sigma\big(W_2\,\mathrm{gelu}(v)\big)$ with W_1 and W_2 of width H. Dropout is applied to each residual branch.

The formal language tasks use two blocks, PPG-DaLiA four and weather one, whereas the synthetic tasks, the UEA archive and TUAB apply the layer of (6) directly, with the readout given in each subsection.

Computing the local EWS. The EWS layer is computed by a parallel associative scan that combines the local EWS states over each interval through the modified Chen identity. Unlike an SSM, the EWS never discretises its dynamics at depth one, where the local EWS is exactly $\varphi_1(-\Delta\theta A)\,B\Delta X$ (Appendix E.1), a matrix function that we evaluate to floating-point precision. Every layer of depth one in this paper is therefore exact. Deeper levels also have an exact expression, as the exponential of the structured matrix of equation 28, but it is expensive to evaluate for a general generator. However, the rotation and Jordan families have closed forms at every depth, which we implement at depth two. For other generators we instead approximate the deeper terms of the local EWS numerically yielding the small errors of Table 8.

Table 8: Median relative error of the numerically approximated local EWS against an accurate reference, over ten validation sequences at the initial parameters of each configuration.

Dataset	Depth	Level 1	Level 2	Level 3
SelfRegulationSCP1	2	1.0×10^{-4}	1.5×10^{-4}	–
Heartbeat	3	2.9×10^{-5}	4.7×10^{-5}	6.4×10^{-5}
MotorImagery	3	2.1×10^{-4}	2.9×10^{-4}	3.6×10^{-4}
SelfRegulationSCP2	3	4.9×10^{-5}	7.0×10^{-5}	8.6×10^{-5}

F.2 Additional experiment

The TUH Abnormal EEG Corpus (TUAB) (Lopez et al., 2015) is a subset of the TUH EEG Corpus (Obeid & Picone, 2016) in which each recording is labelled normal or abnormal by a neurologist, with training and evaluation sets drawn from disjoint groups of patients. We follow the preprocessing and protocol of Yang et al. (2023): we form 16 channels as differences between neighbouring electrodes, resample them to 200 Hz without filtering, split each recording into non-overlapping windows of 10 seconds (2 000 steps), and give each window the label of its recording. This gives $316\,854/55\,656/36\,945$ train/val/test windows. Each window is classified on its own, and we report the area under the ROC curve and the balanced accuracy on the test windows. Balanced accuracy, the mean of the accuracies on abnormal and on normal windows, needs a threshold on the model's score, which we set on the validation windows to make the fraction predicted abnormal equal to the fraction that are abnormal as in Yang et al. (2023).

Inspired by the multi-head structure of Mamba-2, which Appendix E shows to be a family of depth-one EWS layers, one per head, the model runs eight EWS layers (6) side by side at depth two, without blocks or residual connections. Each has its own lift from the 16 channels to $m = 8$, a generator of two real Jordan blocks of size four, each a complex pair of multiplicity two, and a blended variation clock whose weight is learned. The path is the integral of the scaled signal, with the base-point augmentation. The outputs of the eight layers are concatenated, 584 features in all, and passed through dropout to a two-layer perceptron. The configuration was selected on three-seed mean validation AUROC from a grid of dropout in $\{0.1, 0.2, 0.3\}$, learning rate in $\{2, 3, 4\} \times 10^{-3}$ and MLP's width in $\{32, 64\}$, with the other settings fixed by earlier searches. The selected model has dropout 0.1, learning rate 4×10^{-3}, hidden width 32, and has 20 250 parameters.

Training uses the cross-entropy loss and AdamW with weight decay 10^{-3}, gradients clipped at a global norm of one, a constant learning rate and a batch of 64 windows, for 5 000 steps. We evaluate every 1 000 steps, keep the checkpoint with the best validation AUROC, and report the mean over five seeds.

Table 9 compares the EWS both with models trained from scratch and with models pre-trained on several EEG datasets, with results and parameter counts taken from Yang et al. (2023), Jiang et al. (2024) and Wang et al. (2025). It also includes two variants of our model. One is the same model at depth one, and the other freezes the trained model and fits 600 gradient-boosted trees to its 584 features in place of the MLP.

Table 9: TUAB balanced accuracy and AUROC; mean ± standard deviation, over five seeds for the EWS. CBraMod's balanced accuracy uses a fixed threshold of 0.5.

Model	Params.	Balanced acc. (%)	AUROC
EEGNet (Lawhern et al., 2018)	0.003M	76.42±0.36	0.8412±0.0031
SPaRCNet (Jing et al., 2023)	0.79M	78.96±0.18	0.8676±0.0012
ST-Transformer (Song et al., 2021)	3.5M	79.66±0.23	0.8707±0.0019
BIOT (Yang et al., 2023)	3.2M	79.25±0.35	0.8691±0.0033
EWS, depth one	0.003M	62.47±3.48	0.6665±0.0469
EWS, depth two	0.020M	79.90±0.45	0.8824±0.0028
EWS, depth two + boosting	0.020M + 600 trees	80.38±0.35	0.8903±0.0021
BIOT, pretrained	3.2M	79.59±0.57	0.8815±0.0043
CBraMod, without TUAB	4.0M	82.49±0.25	0.9156±0.0017
CBraMod (Wang et al., 2025)	4.0M	82.89±0.22	0.9227±0.0011

The EWS exceeds every model trained from scratch while only using $100\times$ fewer parameters than any transformer, and matches the pretrained BIOT model which makes use of six EEG datasets prior to training on TUAB. CBraMod remains ahead, as expected of a foundation model with $200\times$ as many parameters, pretrained on large EEG corpora. Depth is essential for this task, as the same model at depth one falls below every baseline, and even with gradient boosting in place of its MLP ($64.05 \pm 3.55\%$ balanced accuracy and 0.6912 ± 0.0495 AUROC) it remains below the depth-two model with an MLP; hence, a stronger readout of depth-one features does not recover what depth two provides. Depth three, with m^3 coordinates per head, does not fit in memory at eight heads. Using gradient boosting instead raises the test scores on each seed.

F.3 Synthetic experiments

The task and the three models used are those from Section 3, all of which are trained with AdamW, a linear warmup followed by cosine decay, and the test score is taken at the best-validation checkpoint.

Expressivity. A third task, not reported in the main text, verifies whether each model can fit targets from its own class and from the other two. Its paths are 750 time-augmented Brownian paths (t, W^1, W^2) on $[0, 5]$ at $10\,000$ steps, split $70/15/15$. Each target is a depth-two cross term $S_{\mathbb{A}}(X;\theta)^{i,j}$, $i \neq j$, of the EWS of that path under a fixed generator, which is either zero, positive diagonal or has complex eigenvalues, placing the target in the class of the signature, EFM and EWS respectively. By the shuffle identity, $S^{i,j} = \frac{1}{2}S^iS^j + \frac{1}{2}(S^{i,j} - S^{j,i})$, and the second term, the Lévy area of the re-weighted path, is not a function of the depth-one coordinates of the same EWS. Every model is trained at depth two, without a lift, for $30\,000$ steps at a batch size of 32 on ten seeds. Table 10 reports the results. As expected, each model learns its own target and the signature target is learnt by all three. The EFM target is reached by the EWS but the EFM is unable to learn the EWS target due to its complex spectrum.

Table 10: Expressivity. Test RMSE on targets scaled to $[0, 1]$, mean $\pm$ standard deviation over ten seeds at depth two.

Target \ Model	Signature	EFM	EWS
Signature	$(2.00\pm0.08) \times 10^{-5}$	$(1.99\pm0.08) \times 10^{-5}$	$(2.02\pm0.17) \times 10^{-5}$
EFM	$(4.28\pm0.53) \times 10^{-2}$	$(2.66\pm0.28) \times 10^{-5}$	$(2.77\pm0.28) \times 10^{-5}$
EWS	$(2.41\pm0.26) \times 10^{-2}$	$(2.43\pm0.26) \times 10^{-2}$	$(9.18\pm8.80) \times 10^{-4}$

Oscillator. The path is (t, W^1, W^2) and the target is the first component of

$$\begin{aligned} dY_t^1 &= (3\sin Y_t^2 - \tfrac{1}{2}Y_t^1)\,dt + \tfrac{2}{5}\,dW_t^1, \\ dY_t^2 &= (3\cos Y_t^1 - \tfrac{1}{2}Y_t^2)\,dt + \tfrac{2}{5}\,dW_t^2, \end{aligned} \tag{54}$$

We generate 750 trajectories on $[0, 4]$ with $10\,000$ Euler–Maruyama steps, split $70/15/15$, scale the inputs and the target to $[0, 1]$, prepend a basepoint, and train for $15\,000$ steps at batch 32 on ten seeds. The learned generator of the EWS carries a complex pair on every seed, with real part 1.86 ± 0.15 and imaginary part 10.06 ± 0.84 at depth two, and 1.98 ± 0.01 and 10.68 ± 0.04 at depth one.

Fourier targets. The signal is a sum of tones on a grid of 256 samples within the range $[0, 1]$. One or two low tones at frequencies drawn uniformly from 1 to 5 Hz are present throughout, and one or two high tones from 20 to 30 Hz switch on for a burst of 38 to 102 samples, with no high tone at all in 15% of samples; amplitudes are drawn from $[0.5, 1.5]$, with Gaussian noise of standard deviation 0.05 added. We generate $10\,000$ samples and split them into training/validation/test with a ratio of $80/10/10$. We use the path with increments $(\Delta t,\, x_t\Delta t)$, whose second channel is the integral of the signal. The targets have four frequencies $\{2, 4, 22, 27\}$ Hz and a rate of $a = 5$, and are computed for every sample whatever tones it carries. Models are trained for $12\,000$ steps at batch 64 on three seeds, with a logistic readout for F_3, and at each depth we report the width with the best validation score, over a grid capped at $5\,000$ features (a model of depth n and width m has $\sum_{1\leq k\leq n} m^k$ features).

Table 11 gives the EWS on F_2 and F_3 at every width. Depth one is flat in width on both targets, and depth two improves steeply with it. The EFM and the signature reach at best 0.908 and 0.906 on F_2, and 69.5% on F_3, the majority class, at every depth and width of their grids.

Splitting F_1 by band shows where each model fails. At the selected width at depth one, the EWS reaches 1.2×10^{-5} and 5.8×10^{-5} on the low and high bands, the EFM 0.105 and 0.918, and the signature 0.880 and 0.999. Even the EFM's lowest test error anywhere in its grid, at depth three, is 0.798 on the high band. At the smallest width the EWS is unreliable, one seed of three reaching 0.32 against 0.005 and 0.029 for the other two, since its rotation blocks can miss the high band.

Table 11: The EWS on F_2 and F_3 by depth and width, mean $\pm$ standard deviation over three seeds.

		F_2, normalised MSE		F_3, accuracy (%)	
m	D at $N{=}2$	$N{=}1$	$N{=}2$	$N{=}1$	$N{=}2$
8	72	0.939 ± 0.042	0.285 ± 0.173	69.4 ± 0.1	77.2 ± 10.4
16	272	0.921 ± 0.023	0.046 ± 0.040	69.4 ± 0.1	92.9 ± 0.5
32	1 056	0.921 ± 0.022	$(7.3 \pm 3.4) \times 10^{-4}$	69.5 ± 0.0	94.7 ± 0.2

F.4 UEA MULTIVARIATE TIME SERIES CLASSIFICATION ARCHIVE

We use the six datasets of the long-range benchmark of Walker et al. (2024), whose dimension, length and number of classes are given in Table 12, with their 70/15/15 splits and their five seeds, and with each channel scaled to a fixed range as in their pipeline. The model is the layer of (6) applied directly, with a linear readout and no block. All six use Adam at a constant learning rate, cross-entropy loss and a best-validation checkpoint. The final grid selected the configuration with the highest five-seed mean best-validation accuracy, and the test accuracy of Table 4 is that configuration's mean over the five seeds. The selected configurations are given in Table 13, where the path column records how the observations are read and any augmentation, with base. for a basepoint and clock for a clock channel, the generator and clock columns use the families of Table 1, and EigenWorms' clamped clock has four units. The feature normalisation is a learned gain per level, except on EthanolConcentration, which uses a BatchNorm evaluated with training-set statistics.

On EthanolConcentration the EWS reads the derivative of the input as its path, rather than the input itself or its integral. To test whether this preprocessing explains its margin, we trained S5, LRU and LinOSS-IM models using the same derivative path, scaled and normalised as their original inputs. Over five seeds, they reach $32.2 \pm 12.6\%$, $26.6 \pm 6.8\%$ and $26.3 \pm 2.5\%$ respectively. Therefore, the derivative path alone does not explain the margin of the EWS. Conversely, changing the input path of the EWS model to the raw observations yields $63.2 \pm 11.4\%$, which is already much higher than the other baselines. The selected EWS at depth one is an LTI SSM that a single $S5$ layer can represent, and it already reaches $66.3 \pm 8.9\%$, so the gap reflects parametrisation and training strengths of the EWS, rather than expressivity.

Table 12: The six datasets.

Dataset	Dimension	Observations	Classes
EigenWorms	6	17 984	5
EthanolConcentration	3	1 751	4
Heartbeat	61	405	2
MotorImagery	64	3 000	2
SelfRegulationSCP1	6	896	2
SelfRegulationSCP2	7	1 152	2

Table 13: Selected configurations on the UEA archive. Skew denotes a dense skew-symmetric generator.

Dataset	Path	Gen.	Clock	Readout	N	m	Batch	Steps	LR	Params
EigenWorms	value, clock	Jordan	clamped	mean	2	8	4	6 000	4.5×10^{-3}	446
Ethanol	derivative	rotation	time	terminal	2	24	8	12 000	2.5×10^{-3}	2 508
Heartbeat	value, base.	rotation	blended	mean	3	10	8	25 000	3.0×10^{-4}	2 847
Motor	value	skew	time	terminal	3	16	16	12 000	4.3×10^{-3}	9 887
SCP1	value	skew	variation	mean	2	17	32	25 000	2.5×10^{-3}	856
SCP2	value	skew	blended	terminal	3	14	8	25 000	5.0×10^{-4}	6 104

The grids were, for EigenWorms, the learning rate, the width and the damping at initialisation over 27 cells; for EthanolConcentration, the learning rate, the width and the frequency scale over 27 cells; for Heartbeat, the initialisation variance of the generator, the learning rate and the width over 18 cells; for MotorImagery, the initialisation variance, the scale of the readout and the width over 18 cells; for SelfRegulationSCP1, the width, the initialisation variance and the scale of the readout over 18 cells; and for SelfRegulationSCP2, the scale of the readout, a penalty on the readout weights and the initialisation variance over 18 cells.

Table 14 gives each selected configuration trained at depths one, two and three, with everything else unchanged. Validation selects depth two or three on every dataset, and the selected improves the test accuracy of that configuration over depth one on five of the six. Table 15 widens this to 33 configurations per dataset (a reference and 32 drawn from a wider grid around it; 32 for EigenWorms, where one did not complete), each trained at the three depths on three seeds, at the same width for 6 000 steps. It counts how many configurations improve, tie and fall in accuracy, alongside the median change.

Table 14: Selected configuration at each depth, mean ± standard deviation of validation and test accuracy (%) over five seeds.

	$N = 1$		$N = 2$		$N = 3$	
Dataset	Val.	Test	Val.	Test	Val.	Test
EigenWorms	90.9±4.9	85.6±4.8	**96.0±2.9**	**93.3±1.4**	93.7±4.6	92.2±4.1
Ethanol	69.1±13.7	66.3±8.9	**76.0±7.7**	**71.9±4.8**	73.4±6.5	71.4±3.7
Heartbeat	77.1±3.7	76.1±5.9	74.4±5.6	77.4±2.7	**77.7±5.6**	**80.0±4.5**
Motor	57.2±3.6	50.5±6.8	58.3±6.0	52.6±7.4	**66.0±2.1**	**61.1±5.0**
SCP1	90.0±2.5	86.1±3.8	**90.2±3.4**	**87.1±2.5**	84.8±3.8	81.7±2.1
SCP2	58.3±3.9	58.6±5.3	58.2±3.9	47.4±6.2	**61.4±2.7**	**56.5±5.8**

Table 15: Paired depth comparison over the 33 configurations of each dataset. Each entry gives the number that improve/tie/fall, followed by the median change in accuracy (%).

	$N = 1 \to 2$		$N = 1 \to 3$	
Dataset	Val.	Test	Val.	Test
EigenWorms	29/0/3, +10.7	23/0/9, +9.1	28/1/3, +11.1	20/0/12, +8.2
Ethanol	22/1/10, +3.8	27/0/6, +4.2	22/0/11, +3.4	26/0/7, +4.6
Heartbeat	25/3/5, +1.1	23/2/8, +1.1	28/2/3, +1.6	23/4/6, +1.1
Motor	25/0/8, +1.8	20/5/8, +1.2	28/0/5, +4.7	19/2/12, +1.2
SCP1	17/2/14, +0.0	3/2/28, −1.6	6/1/26, −2.4	5/1/27, −2.4
SCP2	24/1/8, +2.3	10/0/23, −3.5	29/0/4, +4.3	14/0/19, −0.9

The EWS uses 3.8k parameters per dataset on average, the fewest of any model in Table 16 with an available count.

Table 16: Parameter counts at the configuration each paper reports, as published by Rusch & Rus (2025) and Boyer et al. (2025) where available and otherwise computed from the released code. A dash marks a model with no released configuration for that dataset.

Model	EigenWorms	Ethanol	Heartbeat	Motor	SCP1	SCP2	Mean
Log-NCDE	37 977	31 452	168 320	81 391	91 557	36 379	74 513
S5	22 007	76 214	158 310	17 496	226 328	5 652	84 335
S6	15 045	5 780	6 674	52 802	24 898	26 018	21 870
Mamba	27 381	1 032 772	1 034 242	228 226	184 194	356 290	477 184
DE-SLiCE	116 357	16 964	16 914	17 458	25 154	115 970	51 470
D-DE-SLiCE	4 245	1 796	16 914	17 458	2 402	3 858	7 779
BD-SLiCE	5 253	1 604	5 010	5 170	2 114	4 866	4 003
LinOSS-IM	134 279	6 728	10 936	91 844	991 240	399 112	272 357
D-LinOSS	134 279	71 112	4 356	20 598	992 776	399 496	270 436
PD-SSM	94 853	63 236	366 722	21 954	167 810	94 594	134 862
FLASH PD-SSM	–	–	–	–	–	–	–
EWS	446	2 508	2 847	9 887	856	6 104	**3 775**

Table 17 reports training time and memory. Every model was timed on the same card with no other process on it, at a batch of 32, reduced to 16 on EthanolConcentration and 8 on EigenWorms for memory, taking the median of the steady intervals of a run of 4 000 steps with the first interval dropped. Memory is the peak of the process as reported by the driver. The EWS is thus parameter-efficient but not uniformly compute-efficient as its peak memory exceeds that of all other models on three of the six UEA datasets.

Table 17: Training time in seconds per 1 000 steps and peak memory in MB.

Model	EigenWorms	Ethanol	Heartbeat	Motor	SCP1	SCP2
Log-NCDE	1955/2 744	2116/880	785/2 938	779/4 744	616/904	586/882
S5	11.4/3 036	4.6/1 430	7.5/1 694	23.9/5 040	8.7/2 462	6.2/1 180
S6	25.2/2 774	5.3/1 012	7.3/918	38.3/4 372	3.2/1 178	8.7/1 540
Mamba	34.7/3 786	80.5/3 022	37.2/1 958	28.7/3 498	5.2/1 556	31.7/2 800
LRU	22.7/4 768	10.9/1 406	5.4/1 654	38.5/8 846	5.1/1 140	4.7/1 146
BD-SLiCE	10.2/2 742	10.1/1 136	33.5/2 932	28.2/4 738	4.9/892	4.8/1 136
LinOSS-IM	18.5/4 798	10.7/1 148	9.1/1 672	21.7/4 732	26.8/5 268	18.0/3 206
D-LinOSS	15.3/3 708	20.3/3 590	6.6/888	17.4/4 740	28.1/5 070	18.3/3 210
EWS	19.7/4 788	27.8/5 964	14.6/2 764	83.6/34 506	10.2/2 752	29.5/9 540

F.5 PPG-DaLiA and weather

Both tasks follow the training protocol of Boyer et al. (2025), which on PPG-DaLiA is that of Walker et al. (2024), a budget of 100 000 steps with evaluation every 1 000 steps and early stopping after ten evaluations without an improvement; the best-validation checkpoint is reported on five seeds. On both, our configuration was the best on five-seed mean validation error among sixteen drawn at random from a grid over the learning rate, H, m and the number of blocks. The path is taken to be the integral of the input stream with the basepoint augmentation.

PPG-DaLiA. Each subject's recording is split in time into training, validation and test parts in the ratio $70/15/15$, and the test set therefore measures generalisation to later times within the same subjects, not to new subjects. The grid was the learning rate in $\{2.5, 5, 10\} \times 10^{-4}$, H in $\{8, 16, 32\}$, m in $\{32, 64\}$ and the number of blocks in $\{2, 4, 6\}$. Our configuration has four EWS blocks at depth one with $H = 16$ and $m = 64$, a time channel augmentation, a Jordan generator of 16 complex Jordan blocks and the clamped clock with four units; it is trained with the squared error and Adam at a constant rate of 2.5×10^{-4} with gradients clipped at a global norm of one and a batch of 8.

Weather. We use the matching loaders of Gu et al. (2022b) and Boyer et al. (2025), which split the series in time into $70/10/20$ and window it at a stride of one. The model's grid was the learning rate in $\{2.5, 5, 10\} \times 10^{-5}$, H in $\{32, 64, 128\}$, m in $\{16, 32, 64\}$ and the number of blocks in $\{1, 2\}$. Our configuration has one EWS block at depth one with $H = 64$ and $m = 64$, a block-diagonal generator of sixteen real Jordan blocks of size four, each a complex pair of multiplicity two, and the time clock, and is trained with the L^1 loss and Adam at a constant rate of 5×10^{-5} with a batch of 16. Table 18 compares its training time and memory with D-LinOSS, the one baseline we timed.

D-LinOSS publishes a score of 0.486 from a single seed. The mean of the five seeds committed with its code is 0.538 ± 0.052, which is the value we report in Table 5, as confirmed by the authors in private communication.

Table 18: Weather, parameters, training time in seconds per 1 000 steps and peak memory in MB, at a batch of 16.

Model	Params	Time	Memory
EWS	18 669	8.95	1 722
D-LinOSS	301 196	7.97	2 880

F.6 Regular language tasks

We use the four regular language tasks of the formal language benchmark of Delétang et al. (2023), following the protocol of Walker et al. (2025).

1. *Cycle navigation.* A walk on a cycle of length five that starts at the origin, and each step moves forward, moves back or stays; the target is the final position. A random guess gets 20%.
2. *Even pairs.* The input is a string using a two-symbol alphabet, and the target is whether it contains an even number of symbol changes, which is the case if and only if its first and last symbols agree. A random guess scores 50%.
3. *Modular arithmetic without brackets.* The input is an expression using the digits 0 to 4 with addition, subtraction and multiplication, evaluated modulo five with multiplication first. A random guess scores 20%.
4. *Parity.* The input is a string using an alphabet of two symbols, and the target is whether the second symbol occurs an even number of times. A random guess scores 50%.

Models are trained on sequences of length 3 to 40 and evaluated on $8\,192$ sequences of length 40 to 256, thus testing length generalisation. Each sequence is labelled only at its final position, where the cross-entropy loss and the accuracy are both computed. We report the best accuracy over evaluations every $5\,000$ steps on five seeds, as Walker et al. (2025) do, and a run stops early once it exceeds 99.95%. These evaluations therefore also select the checkpoint. To check that this selection does not inflate the results, we evaluate each selected checkpoint once more on $8\,192$ newly drawn sequences of lengths $40 - 256$. The EWS then reaches 99.99 ± 0.1, 100.00 ± 0.00, 96.54 ± 0.76 and 100.00 ± 0.00 on cycle navigation, even pairs, modular arithmetic and parity, all within 0.2 of Table 6.

Training runs for $100\,000$ steps at batch 256 with AdamW, weight decay 0.01 on every parameter but the embedding, a linear warmup over the first tenth of training followed by cosine decay to 10^{-5} from a peak of 4.9×10^{-3}; gradients are clipped at a norm of one within each parameter group. We use no dropout, and label smoothing of 0.05 on parity and even pairs and 0.1 on the other two; on modular arithmetic the generators have no weight decay and a penalty of 10^{-3} on the sum of their Frobenius norms. The configurations are given in Table 19.

Table 19: Configurations on the regular language tasks, using the names of Table 1. Block one has $m = 64$ and block two $m = 24$, both at depth one.

		Block 1		Block 2			
Task	H	Generator	Clock	Generator	Clock	Smoothing	Params.
Cycle nav.	128	rotation	clamped	rotation	variation	0.1	92 219
Even pairs	128	rotation	gated	rotation	variation	0.05	90 290
Mod. arith.	128	skew	gated	rotation	variation	0.1	96 165
Parity	128	rotation	gated	skew	variation	0.05	90 554