Lag Operator SSMs: A Geometric Framework for Structured State Space Modeling

Sutashu Tomonaga
Okinawa Institute of Science and Technology

Kenji Doya
Okinawa Institute of Science and Technology

Noboru Murata
Waseda University

Abstract

Structured State Space Models (SSMs), which are at the heart of the recently popular Mamba architecture, are powerful tools for sequence modeling. However, their theoretical foundation relies on a complex, multi-stage process of continuous-time modeling and subsequent discretization, which can obscure intuition. We introduce a direct, first-principles framework for constructing discrete-time SSMs that is both flexible and modular. Our approach is based on a novel lag operator, which geometrically derives the discrete-time recurrence by measuring how the system's basis functions undergo what we call a *domain expansion* from one timestep to the next. The resulting state matrices are computed via a single inner product involving this operator, enabling a modular design space for creating novel SSMs by flexibly combining different basis functions and time-warping schemes. To validate our framework, we demonstrate that a specific instance exactly recovers the recurrence of the influential HiPPO model. Numerical simulations confirm our derivation, providing new theoretical tools for designing flexible and robust sequence models.

1 INTRODUCTION

Structured State Space Models (SSMs), which form the backbone of modern architectures like Mamba, have emerged as a powerful architecture for long-range sequence modeling (Gu and Dao, 2023). Their key innovation is imposing mathematical structure on traditional SSMs to enable effective dynamical memory compression. They achieve performance that rivals Transformers on many tasks while breaking free from the quadratic scaling bottleneck of attention mechanisms (Gu et al., 2021; Somvanshi et al., 2025).

As many real-world phenomena are inherently dynamical systems, SSMs provide a natural choice for learning a latent representation of a signal's history for downstream tasks like prediction or classification. Unlike attention-based models with direct access to the entire input sequence, SSMs operate recurrently and must compress the past into a finite-dimensional state. The efficacy of an SSM, therefore, hinges on the quality of this compression mechanism. In Structured SSMs, this mathematical structure arises from rigorous theorems of continuous-time memory updates, where the SSM's state is structured as coefficients of a signal's history projected onto an orthogonal basis whose domain expands over time. However, the theoretical foundation that enabled this resurgence, most notably the High-order Polynomial Projection Operators (HiPPO) framework (Gu et al., 2020), relies on an intricate multi-stage derivation involving continuous-time ODEs and subsequent discretization. This complexity can obscure the geometric intuition behind the state update, making principled extensions difficult.

In this paper, we introduce a novel framework for constructing discrete-time Structured SSMs based on discrete-time geometric principles. The hallmark of our framework is the **lag operator** ($\sigma_t \circ \sigma_{t+1}^{-1}$), a geometric mapping that quantifies the *domain expansion* of the structured projection basis between timesteps. This operator derives the complete discrete recurrence ($\boldsymbol{A}_t, \boldsymbol{B}_t$) directly from a single inner product. This provides a complete, **end-to-end discrete-time construction**, as it entirely bypasses the intermediate ODE and discretization steps of traditional methods. A key advantage of this direct construction is its modularity: by defining the time-warping function σ_t, one can construct varied SSMs with different memory

Proceedings of the 29th International Conference on Artificial Intelligence and Statistics (AISTATS) 2026, Tangier, Morocco. PMLR: Volume 300. Copyright 2026 by the author(s).

properties, such as models with multi-resolution memory to track long-term trends and fine-grained recent details simultaneously.

To validate our framework, we show theoretically that an exponential warping instance exactly recovers the original HiPPO recurrence (via Propositions 1 and 2), providing a more fundamental geometric foundation for HiPPO itself. Numerical simulations confirm this with (1) precise matrix equivalence to HiPPO counterparts and (2) replication of HiPPO's memory dynamics with our end-to-end recurrence.

2 CONTINUOUS-TIME ORIGINS OF STRUCTURED SSMS

The theoretical underpinnings of structured SSMs like Mamba trace back to the Legendre Memory Unit (LMU) (Voelker et al., 2019; Voelker and Eliasmith, 2018), which sought to enable continuous memory storage by approximating a continuous-time delay line $u(t-\theta)$ for a given input signal $u(t)$. Since the delay operator's transfer function in the Laplace domain, $e^{-\theta s}$, is irrational, it cannot be directly realized by a finite-dimensional state-space system. The LMU's key insight was to first approximate this function with a rational Padé approximant, which can then be converted into a state-space form. The resulting dynamical system's state matrices are intrinsically linked to the recurrence relations of Legendre polynomials, establishing them as a natural basis for approximating signal history.

The HiPPO framework (Gu et al., 2020) generalized this approach by formalizing the online compression of a signal's history into a set of orthogonal polynomial coefficients, effectively solving the problem in reverse, by first approximating the signal and then constructing a dynamic model.

Under the HiPPO framework, the LMU is considered a special case referred to as HiPPO-LegT (Legendre Truncated). A key variant, HiPPO-LegS (Legendre Scaled), uses a time-varying measure $\mu(t,s) = 1/t$ to weight the inner product integral used for projecting the signal onto an orthogonal basis. This measure is defined on the interval $[0,t]$ and ensures that the total weight remains constant as time evolves: the projection coefficients are given by

$$c_n(t) = \int_0^t u(s) L_n(s/t) \mu(t,s)\, ds, \tag{1}$$

where $u(t)$ is the input signal, and L_n is the n-th Legendre polynomial (normalized on $[0,1]$). This leads to a continuous-time Ordinary Differential Equation (ODE) for the projection coefficients $\boldsymbol{c}(t)$ that is fundamentally time-varying:

$$\boldsymbol{c}'(t) = \frac{1}{t}\boldsymbol{A}\boldsymbol{c}(t) + \frac{1}{t}\boldsymbol{B}u(t), \tag{2}$$

where $\boldsymbol{c}(t)$ is the state vector of projection coefficients, and $\boldsymbol{A}$ and $\boldsymbol{B}$ are constant matrices derived from the properties of Legendre polynomials. However, the time-dependent factor $1/t$ makes this system difficult to implement directly as a standard time-invariant convolutional or recurrent model.

A major breakthrough in making these models practical was the Structured State Space Sequence (S4) model (Gu et al., 2021), which proposed using a time-invariant version of the system,

$$\boldsymbol{c}'(t) = \boldsymbol{A}\boldsymbol{c}(t) + \boldsymbol{B}u(t). \tag{3}$$

This is equivalent to dropping the $1/t$ factor from the HiPPO-LegS ODE (eq.(2)). This simplification proved remarkably effective, achieving state-of-the-art performance on tasks involving long-range dependencies (Gu et al., 2021). However, it initially lacked a clear theoretical justification, raising the question of why this approximation worked so well.

The crucial insight connecting these two systems was provided later in (Gu et al., 2022). Corollary C.10 of that work proved that the time-invariant system used by S4 is not an arbitrary approximation, but is itself a principled HiPPO operator under a different construction. Specifically, it is equivalent to projecting the signal onto a basis of Legendre polynomials with an *exponential time warping*, i.e., basis functions of the form $L_n(e^{s-t})$ (e.g., instead of $L_n(s/t)$) with an exponential measure e^{s-t} over the infinite history $(-\infty, t]$: the projection coefficients are given by

$$c_n(t) = \int_{-\infty}^t u(s) L_n(e^{s-t}) e^{s-t}\, ds. \tag{4}$$

However, the HiPPO derivation is a multi-stage, intricate process. First, a signal's history is approximated continuously using orthonormal polynomials under a chosen measure. Second, an ODE is derived to govern the evolution of the approximation coefficients, a step that can involve complex calculus with time-varying measures and basis functions. Finally, the resulting continuous-time system must be discretized to yield a practical, discrete-time recurrence. This multi-step pipeline, while elegant and robust, obscures the direct geometric connection between the signal history and the final state update. Our work builds directly on this insight, proposing a general framework where time warping and discretization are not a post-hoc justification, but the central, first-principles mechanism for constructing the SSM recurrence.

3 OUR FRAMEWORK

This section details our lag-operator framework, which constructs a discrete-time SSM by consecutively projecting the input signal onto an evolving time-warped basis. We introduce a novel lag operator, $\sigma_t \circ \sigma_{t+1}^{-1}$, which directly yields the state transition by quantifying a *domain expansion* of the basis functions from one time step to the next.

3.1 Definitions

3.1.1 Orthonormal Basis System

Consider a canonical interval Z (e.g., $(0,1]$ for the shifted orthonormal Legendre polynomials or $[-\pi,\pi]$ for Fourier basis). For arbitrary functions U, V supported on Z, an inner product is defined as:

$$\langle U, V\rangle = \int_Z U(z)V(z)\, d\mu(z), \tag{5}$$

where μ is a measure on Z (e.g., the Lebesgue measure $d\mu(z) = dz$ for uniform weighting). Define an orthonormal basis

$$\Phi = \{\phi_n(z) \mid n = 0, 1, \dots, N-1,\ z \in Z\}, \tag{6}$$

with inner product $\langle \phi_n, \phi_m\rangle = \delta_{nm}$. For normalized Legendre polynomials on $(0,1]$, $\phi_n(z) = \sqrt{2n+1}P_n(2z-1)$ (Gu et al., 2022), where P_n are the standard Legendre polynomials.

3.1.2 Signal

Consider a discrete-time input sequence $\{u_k\} = (u_1, u_2, \dots, u_t, \dots) \in \mathbb{R}^{\mathbb{N}}$, observed sequentially up to time t (unit timestep; $t \in \mathbb{N}$).

To enable continuous-time analysis and basis transformations, interpolate u_k into a continuous function $u : \mathbb{R} \to \mathbb{R}$ on time domain s via zero-order hold (ZOH): for history up to t,

$$u(s) = u_k \quad \text{for} \quad s \in [k-1, k), \quad k = 1, \dots, t, \tag{7}$$

with $u(s) = 0$ for $s < 0$ or $s > t$. This piecewise-constant representation treats the discrete signal as continuous for projection onto orthonormal bases, preserving sample-and-hold semantics.

3.1.3 Time Warping

At time t, the interval of interest is $T_t = (-\infty, t]$, encompassing the signal history.

Define an invertible map σ_t that *warps* the time interval T_t to the canonical interval Z (Fig. 1):

$$\sigma_t : T_t \to Z, \quad s \in T_t \mapsto z = \sigma_t(s) \in Z, \tag{8}$$

analogous to HiPPO's $\sigma(t, s)$ (Gu et al., 2022).

Construct an orthonormal basis on T_t by warping ϕ_n:

$$\psi_{t,n}(s) = \phi_n(\sigma_t(s)) = \phi_n \circ \sigma_t(s), \tag{9}$$

denoted as Ψ_t:

$$\begin{aligned}\Psi_t &= \{\psi_{t,n}(s) \mid n = 0, \dots, N-1\} \\ &= \Phi \circ \sigma_t.\end{aligned} \tag{10}$$

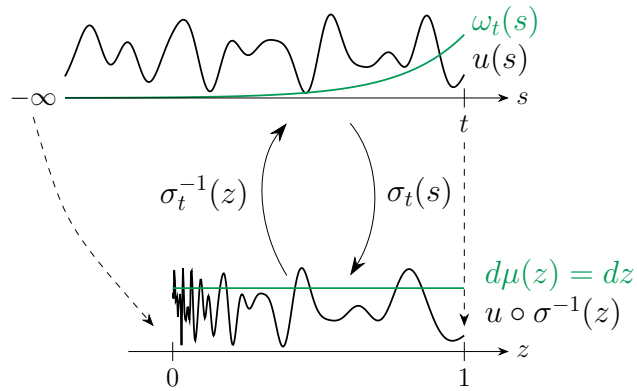

Figure 1: Illustration of the time-warping function σ_t. The infinite history interval $T_t = (-\infty, t]$ (s-axis) is mapped to the bounded canonical interval $Z = (0,1]$ (z-axis). The input signal $u(s)$ (black) is weighted by the exponentially decaying measure density $\omega_t(s) = |\sigma_t'(s)|$ (green curve, shown here for the HiPPO-LegS exponential warp $\sigma_t(s) = e^{s-t}$). In the canonical domain the induced measure becomes uniform ($d\mu(z) = dz$, density $= 1$, green line). The warped signal $u \circ \sigma_t^{-1}(z)$ (bottom black curve) shows compression of older history (dense oscillations near $z = 0$) and higher resolution of recent history (near $z = 1$).

3.1.4 Induced Inner Product and Measure

The inner product on T_t, denoted $\langle\cdot,\cdot\rangle_{\omega_t}$, is defined on the weighted Lebesgue space $L_2(T_t, \omega_t ds)$, where $\omega_t(s) = |\sigma_t'(s)|$ is the induced density ensuring orthonormality. It arises as the *pullback* of the canonical inner product on $L_2(Z)$ under the map σ_t^{-1}.

Let $u(s)$ and $v(s)$ be functions on the time interval T_t, defined via composition with the warping map from functions $U(z)$ and $V(z)$ on the canonical interval Z,

$$\begin{aligned}u(s) &= (U \circ \sigma_t)(s) = U(\sigma_t(s)), \\ v(s) &= (V \circ \sigma_t)(s) = V(\sigma_t(s)).\end{aligned} \tag{11}$$

Then the inner product on T_t can be derived from Z

via change of variables in the inner product.

$$\begin{aligned}\langle u, v\rangle_{\omega_t} &= \langle U, V\rangle \\ &= \int_Z U(z)V(z)\,dz \\ &= \int_{T_t} (U\circ\sigma_t)(s)(V\circ\sigma_t)(s)\,|\sigma_t'(s)|\,ds \\ &= \int_{T_t} u(s)\,v(s)\,\omega_t(s)\,ds,\end{aligned} \tag{12}$$

with $\omega_t(s) = |\sigma_t'(s)|$.

This ensures the warped basis Ψ_t is orthonormal with respect to this inner product:

$$\langle \psi_{t,n}, \psi_{t,m}\rangle_{\omega_t} = \langle \phi_n, \phi_m\rangle = \delta_{nm}. \tag{13}$$

As an example, for exponential forgetting in HiPPO-LegS (Gu et al., 2022), the warping function and inverse are: $z = \sigma_t(s) = e^{s-t}, s = \sigma_t^{-1}(z) = t + \log(z)$, yielding $\omega_t(s) = e^{s-t}$, which integrates to 1 over T_t (normalizing the "forgetting rate").

3.2 Approximation of the Signal

Consider approximating a signal u on T_t (i.e., the history up to time t) using the inner product induced by the image $\sigma_t^{-1}L_2(Z)$. Since $\{\psi_{t,n}\}$ are orthonormal under $\langle\cdot,\cdot\rangle_{\omega_t}$, the optimal L^2 approximation $\hat{u}_t$ is:

$$\begin{aligned}\hat{u}_t(s) &= \sum_{n=0}^{N-1} c_{t,n}\psi_{t,n}(s), \\ c_{t,n} &= \langle \psi_{t,n},\, u\rangle_{\omega_t} = \langle \phi_n,\, u\circ\sigma_t^{-1}\rangle.\end{aligned} \tag{14}$$

Equivalently, approximate the warped signal $U(z) = u(\sigma_t^{-1}(z))$ on Z:

$$\hat{U}_t(z) = \sum_{n=0}^{N-1} c_{t,n}\phi_n(z). \tag{15}$$

Note that coefficients $c_{t,n}$ are equivalent whether computed on Z or T_t. In matrix form:

$$\begin{aligned}\boldsymbol{c}_t &= (c_{t,0},\ldots,c_{t,N-1})^\mathsf{T}, \\ \boldsymbol{\Psi}_t &= (\psi_{t,0},\ldots,\psi_{t,N-1})^\mathsf{T}, \\ \boldsymbol{\Phi} &= (\phi_0,\ldots,\phi_{N-1})^\mathsf{T}.\end{aligned} \tag{16}$$

Hence:

$$\begin{aligned}\hat{u}_t &= \boldsymbol{\Psi}_t^\mathsf{T}\boldsymbol{c}_t, \\ \hat{U}_t &= \boldsymbol{\Phi}^\mathsf{T}\boldsymbol{c}_t.\end{aligned} \tag{17}$$

Geometrically, this projects the signal's entire history up to t onto an orthogonal basis, compressing it into N coefficients.

3.3 Online Updating of the Approximation

Upon receiving new input u_{t+1} at time $t{+}1$, extend the history to $T_{t+1} = (-\infty, t{+}1]$ and compute $c_{t+1,n}$.

Define an interim extended signal (Fig.2), modeling impulse addition per HiPPO (Gu et al., 2020):

$$\tilde{u}_{t+1}(s) = \hat{u}_t(s) + \mathrm{Input}_{t+1}(s)\cdot u_{t+1}, \tag{18}$$

where $\hat{u}_t(s)$ is the prior approximation on $T_t = (-\infty, t]$ (zero elsewhere), and $\mathrm{Input}_{t+1}(s)$ is a distribution or function that mixes the new sample linearly over the interval boundary (e.g., Dirac delta[1] at $s = t{+}1$, zero-order hold constant over $[t, t{+}1]$, or first-order hold linear interpolation). The theoretical implications of these choices and their convergence to a common continuous-time generator are discussed in Section 4.1 and Appendix C.

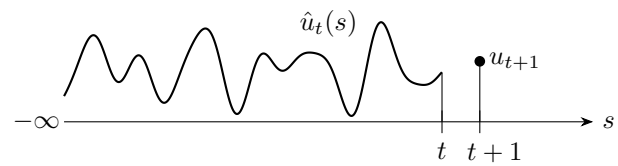

Figure 2: Conceptualization of the extended signal $\tilde{u}_{t+1}(s)$. The new input u_{t+1} is mixed over the interval boundary to the prior signal approximation $\hat{u}_t(s)$.

The updated approximation $\hat{u}_{t+1}$ (on T_{t+1}) is the projection of $\tilde{u}_{t+1}$ onto $\boldsymbol{\Psi}_{t+1}$, yielding the new coefficients $c_{t+1,n}$, we obtain:

$$\begin{aligned}\hat{u}_{t+1}(s) &= \sum_{n=0}^{N-1} \langle \psi_{t+1,n}, \tilde{u}_{t+1}\rangle_{\omega_{t+1}}\, \psi_{t+1,n}(s) \\ &= \sum_{n=0}^{N-1} c_{t+1,n}\psi_{t+1,n}(s).\end{aligned} \tag{19}$$

Substituting $\tilde{u}_{t+1}$ in eq.(18), we obtain:

$$\begin{aligned}c_{t+1,n} &= \langle \psi_{t+1,n},\, (\hat{u}_t + \mathrm{Input}_{t+1}\, u_{t+1})\rangle_{\omega_{t+1}} \\ &= \langle \psi_{t+1,n}, \hat{u}_t\rangle_{\omega_{t+1}} + \langle \psi_{t+1,n},\, \mathrm{Input}_{t+1}\, u_{t+1}\rangle_{\omega_{t+1}} \\ &= \sum_{m=0}^{N-1} c_{t,m}\,\langle \psi_{t+1,n},\, \psi_{t,m}\rangle_{\omega_{t+1}} \\ &\qquad + \langle \psi_{t+1,n},\, \mathrm{Input}_{t+1}\rangle_{\omega_{t+1}}\, u_{t+1} \\ &= \sum_{m=0}^{N-1} c_{t,m}(\boldsymbol{A}_t)_{nm} + (\boldsymbol{B}_t)_n\, u_{t+1},\end{aligned} \tag{20}$$

[1] While the Dirac delta is not in L_2, the projection is well-defined distributionally and aligns with HiPPO conventions (Gu et al., 2020). See Appendix C.1 for convergence to L_2 holds.

where

$$(\boldsymbol{A}_t)_{nm} = \langle \psi_{t+1,n}, \psi_{t,m} \rangle_{\omega_{t+1}} = \langle \phi_n, \phi_m \circ \sigma_t \circ \sigma_{t+1}^{-1} \rangle, \tag{21}$$

and

$$(\boldsymbol{B}_t)_n = \langle \psi_{t+1,n}, \mathrm{Input}_{t+1}(s) \rangle_{\omega_{t+1}}. \tag{22}$$

Here, $\boldsymbol{B}_t$ is derived from the projection of Input onto $\boldsymbol{\Psi}_{t+1}$ (see Appendix C for the derivation of $\mathbf{B}_t$ under different input-hold assumptions). Thus,

$$\begin{aligned} \hat{u}_{t+1}(s) &= \sum_{n=0}^{N-1} c_{t+1,n} \psi_{t+1,n}(s) \\ &= \sum_{n=0}^{N-1} \left(\sum_{m=0}^{N-1} c_{t,m} (\boldsymbol{A}_t)_{nm} + (\boldsymbol{B}_t)_n u_{t+1} \right) \psi_{t+1,n}(s). \end{aligned} \tag{23}$$

In vector form, this yields the HiPPO SSM recurrence (Gu et al., 2020)

$$\begin{aligned} \hat{u}_{t+1} &= \boldsymbol{\Psi}_{t+1}^{\mathsf{T}} \boldsymbol{c}_{t+1}, \\ \boldsymbol{c}_{t+1} &= \boldsymbol{A}_t \boldsymbol{c}_t + \boldsymbol{B}_t \, u_{t+1}. \end{aligned} \tag{24}$$

The composition of $\boldsymbol{A}_t$ in eq.(21), which we define as the backward **lag operator** $\sigma_t \circ \sigma_{t+1}^{-1}$, acts on the canonical basis functions to accommodate the *domain expansion* as time advances from t to $t+1$. This formulation presents a novel insight that geometrically interprets HiPPO's state transition as a "re-warping" of projections. A geometric illustration of this domain-expansion process realized by the lag operator is provided in Appendix D (Figure 5).

4 THEORETICAL VALIDATION

4.1 Recovering the Time-Invariant HiPPO-LegS System

To bridge the discrete-time framework of Section 3.3 to continuous-time limits and connect it to the time-invariant HiPPO-LegS model (Gu et al., 2022), we generalize the unit timestep $[t, t+1]$ to an arbitrarily small step $[t, t+\Delta]$. We then analyze the limit as $\Delta \to 0$. In this regime, the time-varying matrices $\boldsymbol{A}_t$ and $\boldsymbol{B}_t$ (from eq. (24)) become step-size-dependent $\boldsymbol{A}_\Delta$ and $\boldsymbol{B}_\Delta$, which approximate the identity plus a first-order correction: $\boldsymbol{A}_\Delta \approx \boldsymbol{I} + \Delta \boldsymbol{A}_{\mathrm{gen}}$ and $\boldsymbol{B}_\Delta \approx \Delta \boldsymbol{B}_{\mathrm{gen}}$. These converge to the generators of the continuous-time SSM

$$\boldsymbol{c}'(t) = \boldsymbol{A}_{\mathrm{gen}} \boldsymbol{c}(t) + \boldsymbol{B}_{\mathrm{gen}} u(t), \tag{25}$$

via the forward difference: $\boldsymbol{c}'(t) = \lim_{\Delta \to 0} \frac{\boldsymbol{c}(t+\Delta) - \boldsymbol{c}(t)}{\Delta}$. The discrete recurrence $\boldsymbol{c}_{t+\Delta} = \boldsymbol{A}_\Delta \boldsymbol{c}_t + \boldsymbol{B}_\Delta u_{t+\Delta}$ thus recovers the continuous dynamics through a first-order Taylor expansion of the lag operator $\sigma_t \circ \sigma_{t+\Delta}^{-1}$ (eq. (21)). For the stable HiPPO-LegS system, the generator $\boldsymbol{A}_{\mathrm{gen}}$ relates to the canonical HiPPO matrix via $\boldsymbol{A}_{\mathrm{HiPPO}} = -(\boldsymbol{A}_{\mathrm{gen}} + \boldsymbol{I})^{\mathsf{T}}$ (incorporating a tilting shift for measure renormalization; see Appendix B), while $\boldsymbol{B}_{\mathrm{gen}}$ matches directly. For clarity, we summarize the key matrices and their relationships (see also Table 4 in Appendix A for a tabular overview):

- $\boldsymbol{A}_t$ and $\boldsymbol{B}_t$: The discrete-time transition and input matrices at time t, as defined in eq.(24), for unit timestep ($\Delta = 1$).
- $\boldsymbol{A}_\Delta$ and $\boldsymbol{B}_\Delta$: Generalizations of $\boldsymbol{A}_t$ and $\boldsymbol{B}_t$ for arbitrary timestep Δ, used to analyze the continuous limit. Note that $\boldsymbol{A}_\Delta$ and $\boldsymbol{B}_\Delta$ depend on t via $\sigma_t \circ \sigma_{t+\Delta}$, but become time-invariant under stationary warping.
- $\boldsymbol{A}_{\mathrm{gen}}$ and $\boldsymbol{B}_{\mathrm{gen}}$: Continuous-time generators in the ODE $\boldsymbol{c}'(t) = \boldsymbol{A}_{\mathrm{gen}} \boldsymbol{c}(t) + \boldsymbol{B}_{\mathrm{gen}} u(t)$, derived as the first-order terms in the Taylor expansion of the discrete recurrence: $\boldsymbol{c}(t+\Delta) = \boldsymbol{c}(t) + \Delta[\boldsymbol{A}_{\mathrm{gen}} \boldsymbol{c}(t) + \boldsymbol{B}_{\mathrm{gen}} u(t)] + O(\Delta^2)$. Equivalently, $\boldsymbol{c}'(t) = \lim_{\Delta \to 0} [\boldsymbol{c}(t+\Delta) - \boldsymbol{c}(t)]/\Delta$.
- $\boldsymbol{A}_{\mathrm{HiPPO}}$: The stable state matrix from HiPPO-LegS, related to our $\boldsymbol{A}_{\mathrm{gen}}$ by transposition and a tilting shift: $\boldsymbol{A}_{\mathrm{HiPPO}} = -(\boldsymbol{A}_{\mathrm{gen}}^{\mathsf{T}} + \boldsymbol{I})$.

The derivations of $\boldsymbol{A}_{\mathrm{gen}}$ and $\boldsymbol{B}_{\mathrm{gen}}$ are given in Propositions 1 and 2 (with proofs in Appendices B and C.2), and we show their equivalence to HiPPO-LegS below.

The components are:

Canonical Interval Z: The interval $(0, 1]$.

Canonical Basis $\boldsymbol{\Phi}$: The Legendre polynomials $L_n(z)$ are orthonormal over $(0, 1]$.

Warping Function σ_t: The exponential map $\sigma_t(s) = e^{s-t}$, which maps the time domain $s \in (-\infty, t]$ to the canonical interval $z \in (0, 1]$.

To derive the continuous-time state generator $\boldsymbol{A}_{\mathrm{gen}}$, which governs the evolution of the state in the limit of infinitesimally small time steps, we introduce the following proposition that encapsulates this process for a stationary warping function.

Proposition 1 (Continuous Generator $\boldsymbol{A}_{\mathrm{gen}}$)**.** *For a stationary warping $\sigma_t(s) = f(s-t)$ with inverse $g = f^{-1}$, the generator matrix is*

$$(\boldsymbol{A}_{gen})_{nm} = \left\langle \phi_n, \frac{\phi_m'}{g'} \right\rangle = \int_Z \phi_n(z) \frac{\phi_m'(z)}{g'(z)} \, dz. \tag{26}$$

(See Appendix B for proof.)

Applying Proposition 1 to the exponential warp ($f(x) = e^x$, $g(z) = \ln(z)$), we obtain

$$(\boldsymbol{A}_{\text{gen}})_{nm} = \int_0^1 L_n(z) \cdot z \cdot L'_m(z)\, dz, \tag{27}$$

with diagonals n and off-diagonals $\sqrt{(2n+1)(2m+1)}$ for $n < m$ (upper triangular). This matches the transpose of the reference matrix $\boldsymbol{A}^0$ from (Gu et al., 2022) (App. C.8), defined as

$$(\boldsymbol{A}^0)_{nm} = \begin{cases} \sqrt{(2n+1)(2m+1)} & \text{if } m < n, \\ n & \text{if } n = m, \\ 0 & \text{otherwise.} \end{cases} \tag{28}$$

Thus, $\boldsymbol{A}_{\text{gen}} = (\boldsymbol{A}^0)^\mathsf{T}$. The HiPPO-LegS state matrix is $\boldsymbol{A}_{\text{HiPPO}} = -(\boldsymbol{A}^0 + \boldsymbol{I})$, related by transposition and a shift: $\boldsymbol{A}_{\text{HiPPO}} = -(\boldsymbol{A}_{\text{gen}} + \boldsymbol{I})^\mathsf{T}$, where the shift arises from a "tilting" term in the HiPPO derivation, adjusting for measure renormalization.

For the continuous-time input generator $\boldsymbol{B}_{\text{gen}}$, our framework is flexible with respect to the input model (e.g., Dirac delta, ZOH, or FOH; see Appendix C.1). Regardless of the choice, the following proposition shows convergence to the same continuous $\boldsymbol{B}_{\text{gen}}$ in the limit $\Delta \to 0$, equivalent to the Dirac boundary evaluation, reassuring that all models recover the HiPPO input matrix.

Proposition 2 (Continuous Generator for $\boldsymbol{B}_{\text{gen}}$)**.** *Let the discrete input matrix $\boldsymbol{B}_\Delta$ be the coefficient such that the state update includes the term $\boldsymbol{B}_\Delta u(t+\Delta)$. In the continuous limit $\Delta \to 0$, the generator is given by*

$$(\boldsymbol{B}_{\text{gen}})_n = \lim_{\Delta \to 0} \frac{(\boldsymbol{B}_\Delta)_n}{\Delta} = \phi_n(1) f'(0). \tag{29}$$

(See Appendix C.2 for proof.)

Applying Proposition 2 to the exponential warp ($f(x) = e^x$, $f'(0) = 1$) yields

$$(\boldsymbol{B}_{\text{gen}})_n = \phi_n(1) = L_n(1) = \sqrt{2n+1}, \tag{30}$$

which matches the HiPPO-LegS input matrix from (Gu et al., 2022) (App. C.8). This confirms consistency with the continuous-time ODE summarized at the start of this section.

Hence, our framework successfully derives generators equivalent to the HiPPO-LegS system, confirming its validity as a novel geometric construction capable of recovering and extending foundational SSMs.

4.2 Long-Term Behavior of the State Transition and its Connection to the Continuous-Time Generator

This section validates the continuous-time generator $\boldsymbol{A}_{\text{gen}}$ by showing that the discrete-time transition matrix $\boldsymbol{A}_t$, derived from a finite time step, converges to a stationary matrix $\boldsymbol{A}_\infty$ that is consistent with $\boldsymbol{A}_{\text{gen}}$ through the matrix exponential.

For stationary warping $\sigma_t(s) = f(s-t)$ with inverse $g = f^{-1}$, the backward lag operator $\sigma_t \circ \sigma_{t+\Delta}^{-1}(z) = f(\Delta + g(z))$ is time-independent. Thus, $\boldsymbol{A}_t$ is constant, and we define the stationary transition matrix $\boldsymbol{A}_\infty = \boldsymbol{A}_\Delta$ (as summarized in Section 4.1). Note that $\boldsymbol{A}_\Delta$ depends on t via $\sigma_t \circ \sigma_{t+\Delta}^{-1}$ in general, but becomes time-invariant here under stationary warping.

$$(\boldsymbol{A}_\infty)_{nm} = (\boldsymbol{A}_\Delta)_{nm} = \int_Z \phi_n(z) \phi_m\big(f(\Delta + g(z))\big)\, dz. \tag{31}$$

This provides a first-principles justification for time-invariant systems like S4. For HiPPO-LegS exponential warp ($f(x) = e^x$, $g(z) = \ln(z)$),

$$(\boldsymbol{A}_\Delta)_{nm} = \int_0^1 L_n(z) L_m(e^\Delta z)\, dz, \tag{32}$$

the exact discrete transition over Δ.

The matrix $\boldsymbol{A}_\Delta$ relates to $\boldsymbol{A}_{\text{gen}}$ by $\boldsymbol{A}_\Delta = e^{\Delta \cdot \boldsymbol{A}_{\text{gen}}}$. To match the stable HiPPO system, incorporate the tilting term: the stable generator is $\boldsymbol{A}_{\text{stable}} = -(\boldsymbol{A}_{\text{gen}} + \boldsymbol{I})$, and the named $\boldsymbol{A}_{\text{HiPPO}} = \boldsymbol{A}_{\text{stable}}^\mathsf{T} = -(\boldsymbol{A}_{\text{gen}}^\mathsf{T} + \boldsymbol{I})$. The corrected discrete matrix is

$$\boldsymbol{A}_\Delta^{\text{corrected}} = e^{\Delta \cdot \boldsymbol{A}_{\text{stable}}} = (\boldsymbol{A}_\Delta)^{-1} e^{-\Delta}. \tag{33}$$

This matches our numerical experiments and shows the exact discrete transition can be found by inverting $\boldsymbol{A}_\Delta$ and applying a scalar decay. The transpose in HiPPO is notational, and our framework derives the underlying stable dynamics.

5 NUMERICAL VALIDATION

This section provides numerical validation of our lag-operator framework. We confirm our theoretical claims by demonstrating both matrix equivalence and accurate signal reconstruction against the HiPPO-LegS system. While this study serves as a proof of concept, broader benchmarking is deferred to future work.

5.1 Experimental Setup

We numerically validate our framework by recovering the time-invariant HiPPO-LegS system, using the same components (exponential warp $\sigma_t(s) = e^{s-t}$, orthonormal Legendre basis on $(0, 1]$) detailed for the theoretical validation in Section 4.1. As input, we use a 1D signal from the chaotic Lorenz63 system ($\sigma = 10, \rho = 28, \beta = 8/3$), which provides a challenging, non-stationary signal to test memory compression.

For our simulation, we use a basis size of $N = 64$, a time step of $\Delta = 0.01$ s, and a total duration of $T = 10$ s, parameters chosen to match the original HiPPO demos. [2]

5.2 Validation of Core Theoretical Claims

This subsection compares key matrices and signal reconstructions with those of HiPPO-LegS, ensuring that our approach correctly captures its dynamics.

5.2.1 Matrix Equivalence Validation

We assess matrix equivalence by measuring relative difference using the Frobenius norm ($\mathrm{Diff}(\boldsymbol{M}_1, \boldsymbol{M}_2) = \|\boldsymbol{M}_1 - \boldsymbol{M}_2\|_F / \|\boldsymbol{M}_1\|_F$).

Evaluation of $\boldsymbol{A}_\infty$ vs. $\exp(\Delta \boldsymbol{A}_{\mathbf{gen}})$: We compare the stationary matrix $\boldsymbol{A}_\infty = \boldsymbol{A}_\Delta$ (computed via lag operator integration from Section 4.2) with the matrix exponential $\exp(\Delta \boldsymbol{A}_{\mathrm{gen}})$, where $\boldsymbol{A}_{\mathrm{gen}}$ is from Proposition 1. Table 1 shows the differences across various Δ values, indicating convergence as $\Delta \to 0$.

Table 1: Frobenius relative difference between $\boldsymbol{A}_\infty$ and $\exp(\Delta \boldsymbol{A}_{\mathrm{gen}})$ across Δ values

$\boldsymbol{\Delta}$	$\mathbf{10^{-4}}$	$\mathbf{10^{-3}}$	$\mathbf{10^{-2}}$	$\mathbf{10^{-1}}$
Diff	$7.85 \cdot 10^{-12}$	$3 \cdot 10^{-11}$	$1.52 \cdot 10^{-9}$	$4.24 \cdot 10^{-5}$

Evaluation of $\boldsymbol{A}_{\mathbf{gen}}$ vs. $\boldsymbol{A}_{\mathbf{HiPPO}}$: We confirm $\boldsymbol{A}_{\mathrm{HiPPO}} = -(\boldsymbol{A}_{\mathrm{gen}} + \boldsymbol{I})^\top$ by comparing the numerically integrated $\boldsymbol{A}_{\mathrm{gen}}$ with the analytical HiPPO-LegS matrix. Table 2 shows the differences across basis sizes N, demonstrating high precision.

Table 2: Frobenius relative difference between $\boldsymbol{A}_{\mathrm{HiPPO}}$ and $-(\boldsymbol{A}_{\mathrm{gen}} + \boldsymbol{I})^\top$ across N values

N	**10**	**30**	**50**
Diff	$2.26 \cdot 10^{-14}$	$3.69 \cdot 10^{-13}$	$5.26 \cdot 10^{-12}$

Evaluation of $\boldsymbol{A}_\Delta^{\mathbf{corrected}}$ vs. Discretized $\boldsymbol{A}_{\mathbf{HiPPO}}$: We verify that our exact discrete matrix, $\boldsymbol{A}_\Delta^{\mathrm{corrected}} = (\boldsymbol{A}_\Delta)^{-1} e^{-\Delta}$ aligns with the standard bilinear discretization of $\boldsymbol{A}_{\mathrm{HiPPO}}$. Table 3 shows the differences across Δ values, indicating convergence as $\Delta \to 0$.

These results show convergence as $\Delta \to 0$ or $N \to \infty$, validating that our lag operator framework accurately recovers HiPPO's continuous and discrete dynamics.

[2] The original HiPPO source code and demos with parameters can be found at: `https://github.com/state-spaces/s4/tree/main/notebooks`

Table 3: Frobenius relative difference between $\boldsymbol{A}_\Delta^{\mathrm{corrected}}$ and discretized $\boldsymbol{A}_{\mathrm{HiPPO}}$ across Δ values

$\boldsymbol{\Delta}$	$\mathbf{10^{-4}}$	$\mathbf{10^{-3}}$	$\mathbf{10^{-2}}$	$\mathbf{10^{-1}}$
Diff	$5.95 \cdot 10^{-5}$	0.0257	0.334	0.955

5.2.2 Geometric Lag Operator Validation

To illustrate the lag operator's geometric intuition, we apply $\boldsymbol{A}_\Delta$ as a backward shift and its inverse as a forward shift on basis functions. For stability correction, the backward shift operator is $\boldsymbol{A}_{\mathrm{back}} = \boldsymbol{A}_\Delta e^\Delta$ and the forward shift operator is $\boldsymbol{A}_{\mathrm{for}} = \boldsymbol{A}_\Delta^{\mathrm{corrected}} = (\boldsymbol{A}_\Delta)^{-1} e^{-\Delta}$, such that $\psi_{t \to t-1,n} = \sum_m (\boldsymbol{A}_{\mathrm{back}})_{nm} \cdot \psi_{t,m}$ and $\psi_{t \to t+1,n} = \sum_m (\boldsymbol{A}_{\mathrm{for}})_{nm} \cdot \psi_{t,m}$. Figure 3 shows both shifts for $n = 63$.

The backward shift compresses the basis toward older history, while the forward shift expands it, with amplitude shrinkage reflecting the stabilization correction, consistent with HiPPO's design (noted in Section 4.1).

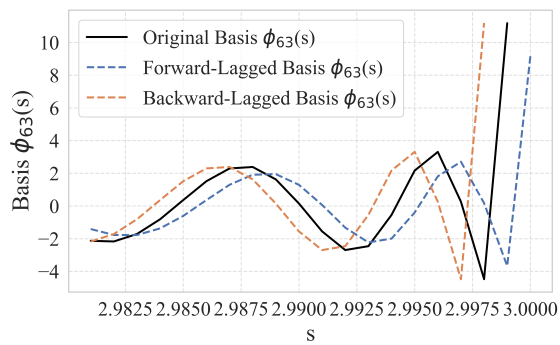

Figure 3: Example basis functions lag-shifted forward and backward using the lag operator $\boldsymbol{A}_\Delta$. The backward shift compresses toward old history, while the forward shift expands with amplitude shrinkage due to stabilization. The plot is zoomed near $t = 3.0$ ($\Delta t = 0.001$) to clearly visualize the lag shift; the apparent jaggedness is the expected high-frequency oscillation of high-degree Legendre polynomials at extreme zoom.

5.2.3 Evaluating Memory Compression and Forgetting

We evaluate the quality of our framework's memory compression by visualizing how well the state can reconstruct the signal's history. Crucially, the goal of HiPPO is not to perfectly reconstruct the entire past, but to create a compressed state that optimally represents the *recent* past while gracefully forgetting distant history. The "reconstructed signal" shown here is gen-

erated by applying the projection defined in eq. (14) to the final state vector $\boldsymbol{c}_T$, which itself is the result of the recurrent update rule. It is therefore a direct visualization of the information stored in this compressed memory.

We compare our Lag Operator HiPPO against the original HiPPO-LegS, using the x-component of the chaotic Lorenz63 system as input. Initial conditions were $[1.0, 1.0, 8.0]$; integration used forward Euler with fixed step size $\Delta = 0.01$ (no adaptive tolerances). Only the x-coordinate was used as input $u(t)$. Our reconstruction method uses the derived $\boldsymbol{A}_\Delta^{\text{corrected}}$ from eq.(33) and the ZOH input vector $\boldsymbol{B}_\Delta^{\text{ZOH}}$ from eq.(42) in Appendix C.1, while the baseline uses its standard analytical matrices (shown in eq.(28) and eq.(30)).

Figure 4 shows that our model's reconstruction almost perfectly aligns with the original HiPPO-LegS. Both models accurately capture the recent history of the signal (near $t = 10$s) while exhibiting the expected exponential decay of information from the distant past (near $t = 0$s). The mean squared error (MSE) between the two reconstructions is a negligible $5.76 \cdot 10^{-7}$, confirming that our first-principles framework successfully replicates HiPPO's memory dynamics.

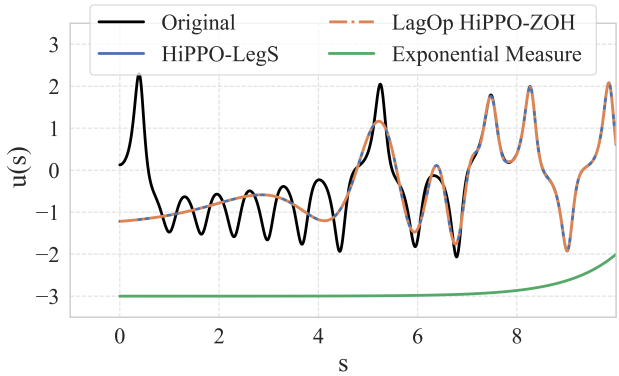

Figure 4: Visualization of the final compressed memory state $\boldsymbol{c}_T$ (reconstructed via eq. (14)) for our Lag Operator model and the HiPPO-LegS baseline. The near-perfect alignment demonstrates that both models accurately capture recent history while forgetting the distant past, a behavior governed by the plotted exponential measure ($\omega_t(s)$), which dictates the memory weighting.

6 MODULAR DESIGN: NOVEL SSM EXAMPLES

A key advantage of our framework is its modularity. While standard approaches require deriving a new ODE for every change in the measure or basis, our geometric construction allows us to independently select a warping function σ_t and a canonical basis $\boldsymbol{\Phi}$. The Lag Operator $\sigma_t \circ \sigma_{t+\Delta}^{-1}$ then automatically derives the consistent discrete-time recurrence. In this section, we outline two novel SSM designs enabled by our framework. While we defer the full implementation and downstream benchmarking of these models to future work, we provide these theoretical sketches to demonstrate the flexibility of decoupling the time-warp from the basis.

6.1 Multi-Resolution Lag-SSM (Fast/Slow Memory)

Standard HiPPO models enforce a single memory decay rate across the entire state. However, many applications require tracking both long-term trends and fine-grained recent details. Our framework enables the construction of a unified state space with heterogeneous decay rates.

Construction: Construct a hybrid basis set $\boldsymbol{\Phi}_{\text{hybrid}}$ on the canonical interval Z that combines "Fast" components (e.g., associated with a rapid decay warp $\sigma_t^{(f)}(s) = e^{2(s-t)}$) and "Slow" components (e.g., $\sigma_t^{(s)}(s) = e^{0.5(s-t)}$).

Geometric Insight: Rather than running two independent SSMs, we can enforce orthogonality across this entire mixed set using Gram–Schmidt orthogonalization within the canonical domain Z. This ensures the state variables are uncorrelated and optimally span the history space. The Lag Operator then acts on this orthogonalized hybrid basis to derive the correct recurrence matrix $\boldsymbol{A}_t$. This allows the model to track history at multiple timescales simultaneously within a single coherent state representation.

6.2 Frequency-Preserving Lag-SSM (Unwarped Oscillations)

A major limitation of the standard exponential warp is that it "stretches" the basis, turning constant frequencies in the z-domain into "chirps" (variable frequency) in the s-domain. This is suboptimal for signals with sustained oscillatory components (e.g., audio, periodic biomedical signals).

Construction: Define σ_t as a piecewise-linear map that partitions history (the s-domain) into windows of fixed length L. Each window maps linearly to a segment of Z with a decaying length $h_n > 0$ (where $\sum h_n = 1$). For hybrid capability, we augment a Fourier basis on Z (for oscillations) with a Legendre subset (for transients), applying Gram–Schmidt orthogonalization. Full mathematical details appear in Appendix E.

Geometric Insight: This induces a piecewise-constant measure $\omega_t(s) = h_n/L$ on each window. Crucially, this maps the Fourier basis to pure, unwarped sinusoids in the time domain s while preserving orthogonality. The lag operator directly yields the discrete recurrence, bypassing complex ODE derivations despite the discontinuous measure. The result is an SSM implementing a "sliding-window Fourier transform" with infinite decaying memory, optionally hybridized for multi-scale dynamics.

7 DISCUSSION AND FUTURE WORK

This work introduces a novel, geometrically grounded framework for constructing discrete-time SSMs based on inner-product projections and a unique lag operator. This approach offers several advantages over the established multi-stage paradigm:

- **Conceptual Clarity and Unified Memory Control:** Our framework provides a clear, geometric foundation for SSMs. It replaces the complex, multi-stage derivations typical in the HiPPO formulation with a direct projection onto warped bases, where the state transition is intuitively understood as a "re-warping" of the signal's compressed history. Crucially, this geometric view unifies memory control into a single component: the time-warp σ_t. The system's memory profile is dictated by the induced measure $\omega_t = |\sigma_t'|$, which is a direct derivative of the warp. This offers a more straightforward approach than the original HiPPO formulation, which requires a separate "tilting" term for stabilization and measure renormalization.
- **A Modular and Interpretable Design Space:** This unified view of memory leads to a powerful and modular design space. The state dynamics are determined by two independent choices: the canonical basis $\mathbf{\Phi}$ and the time-warping function σ_t. This decoupling is powerful because, as established, the choice of σ_t is a direct and interpretable way to engineer the system's memory via its induced measure. This enables sophisticated designs such as **multi-resolution memory**, in which distinct memory profiles are assigned to different subsets of a basis. For example, low-order basis functions could be paired with a slow-decaying warp to model long-term trends, while high-order functions use a fast-decaying warp to capture fine-grained, recent details. This allows a single SSM to operate on multiple timescales simultaneously. Furthermore, the ability to make σ_t time-dependent provides a natural path toward adaptive, non-stationary dynamics.

Our proof-of-concept validation provides a solid foundation for the framework. Numerical results show precise matrix equivalence with HiPPO-LegS, and a low MSE ($5.76 \cdot 10^{-7}$) between the reconstructed memory states of our model and the baseline. This validates that our framework correctly replicates the core function of the HiPPO recurrence, which is to maintain a compressed representation of history, rather than to achieve perfect reconstruction of the entire past.

In practice, this compressed memory state serves as the foundation for downstream components, such as a prediction or classification layer. Furthermore, the framework extends naturally to multivariate inputs by broadcasting the transition matrix A_t across channels (or inserting a mixing layer on inputs/outputs), exactly as in S4 and Mamba. The framework is also robust to the precise discretization of the input; all standard hold assumptions converge to the same continuous-time generator (Appendix C). By demonstrating that our first-principles framework can produce a memory component that behaves identically to the original, we validate its use as an interchangeable building block. This paves the way for exploring the novel memory designs our framework enables (e.g., the multi-resolution models mentioned above) in future work on downstream tasks. Future research should build on this foundation in several key areas:

- **Analysis of Discretization Schemes:** Our experiments successfully validated the framework using a Zero-Order Hold (ZOH) input matrix. An exciting avenue for future work is to systematically explore other input assumptions, such as First-Order Hold (FOH) (see Appendix C.1), and analyze their resulting stability and scaling properties. This would lead to a deeper understanding of how different signal interpolation methods interact with the system's memory.
- **Benchmarking on Downstream Tasks:** Now that the memory component is validated, future efforts should focus on comprehensive benchmarking. This involves integrating the flexible SSMs derived from our framework into larger models and evaluating their performance on standard downstream tasks, such as long-range sequence classification and time-series forecasting, to quantify the benefits of novel memory designs.

This paper serves as a foundational step, presenting a novel theoretical perspective that clarifies underlying mechanisms in state-space models and paves the way for more flexible and robust sequence modeling.

Acknowledgements

S. Tomonaga and K. Doya were supported by the COI-NEXT grant from the Japan Science and Technology Agency (JST). N. Murata was supported by JSPS KAKENHI Grant Numbers JP23H05492 and JP23K24909.

References

Gu, A. and Dao, T. (2023). Mamba: Linear-time sequence modeling with selective state spaces. *arXiv [cs.LG]*.

Gu, A., Dao, T., Ermon, S., Rudra, A., and Ré, C. (2020). HiPPO: Recurrent memory with optimal polynomial projections. *Neural Information Processing Systems*, abs/2008.07669:1474–1487.

Gu, A., Goel, K., and Re, C. (2021). Efficiently Modeling Long Sequences with Structured State Spaces.

Gu, A., Johnson, I., Timalsina, A., Rudra, A., and Ré, C. (2022). How to Train Your HiPPO: State Space Models with Generalized Orthogonal Basis Projections. *arXiv [cs.LG]*.

Somvanshi, S., Islam, M. M., Mimi, M. S., Polock, S. B. B., Chhetri, G., and Das, S. (2025). From S4 to Mamba: A comprehensive survey on Structured State Space Models. *arXiv [cs.LG]*.

Voelker, A., Kajić, I., and Eliasmith, C. (2019). Legendre Memory Units: Continuous-Time Representation in Recurrent Neural Networks. *Advances in Neural Information Processing Systems*, 32.

Voelker, A. R. and Eliasmith, C. (2018). Improving Spiking Dynamical Networks: Accurate Delays, Higher-Order Synapses, and Time Cells. *Neural computation*, 30(3):569–609.

Checklist

1. For all models and algorithms presented, check if you include:
   (a) A clear description of the mathematical setting, assumptions, algorithm, and/or model.
   [Yes] Detailed in Section 3 and Appendix C.
   (b) An analysis of the properties and complexity (time, space, sample size) of any algorithm.
   [Yes] An analysis of the framework's mathematical properties is provided theoretically in Section 4.1 and numerically in Section 5. A formal computational complexity analysis is outside the scope of this paper.
   (c) (Optional) Anonymized source code, with specification of all dependencies, including external libraries.
   [Yes]

2. For any theoretical claim, check if you include:
   (a) Statements of the full set of assumptions of all theoretical results.
   [Yes] Stated in Section 3 and Section 4.1.
   (b) Complete proofs of all theoretical results.
   [Yes] Details in Appendix B and Appendix C.
   (c) Clear explanations of any assumptions.
   [Yes] Assumptions for the proposed framework are stated in Section 3.

3. For all figures and tables that present empirical results, check if you include:
   (a) The code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL).
   [Yes] We provide supplementary code.
   (b) All the training details (e.g., data splits, hyperparameters, how they were chosen).
   [Yes] Stated in Section 5.1.
   (c) A clear definition of the specific measure or statistics and error bars (e.g., with respect to the random seed after running experiments multiple times).
   [Yes] We report Mean Squared Error (MSE) for signal reconstruction. As our experiment is a deterministic validation of a theoretical derivation, results are consistent across runs, making error bars not applicable for this analysis.
   (d) A description of the computing infrastructure used. (e.g., type of GPUs, internal cluster, or cloud provider).
   [Not Applicable] The numerical validation involves standard matrix operations that are not computationally intensive and can be reproduced on any modern personal computer.

4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets, check if you include:
   (a) Citations of the creator If your work uses existing assets.
   [Yes] Mentioned in Section 5.1
   (b) The license information of the assets, if applicable.
   [Yes] We cite the original papers and provide a footnote in Section 5.1 linking to the public source code repository for the assets used.
   (c) New assets either in the supplemental material or as a URL, if applicable.

[Yes] Code provided in supplemental material.

(d) Information about consent from data providers/curators.
[Not Applicable] We use an artificial dataset. (see Section 5.1.

(e) Discussion of sensible content if applicable, e.g., personally identifiable information or offensive content.
[Not Applicable] We use an artificial dataset from a dynamical model. (see Section 5.1.)

5. If you used crowdsourcing or conducted research with human subjects, check if you include:

(a) The full text of instructions given to participants and screenshots.
[Not Applicable] Did not use crowdsourcing nor research human subjects.

(b) Descriptions of potential participant risks, with links to Institutional Review Board (IRB) approvals if applicable.
[Not Applicable] Did not use crowdsourcing nor research human subjects.

(c) The estimated hourly wage paid to participants and the total amount spent on participant compensation.
[Not Applicable] Did not use crowdsourcing nor research human subjects.

Appendix

A NOTATION SUMMARY FOR DISCRETE-TO-CONTINUOUS MATRICES

Table 4: Summary of key matrices in the discrete-to-continuous limit.

Matrix	Description	Relation
$\boldsymbol{A}_t$, $\boldsymbol{B}_t$	Time-varying discrete transition and input (unit step, eq. (24))	General case; $\boldsymbol{A}_t = \langle \boldsymbol{\Phi}, \boldsymbol{\Phi} \circ \sigma_t \circ \sigma_{t+1}^{-1} \rangle$
$\boldsymbol{A}_\Delta$, $\boldsymbol{B}_\Delta$	Step-size-dependent discrete (arbitrary Δ)	$\boldsymbol{A}_\Delta \approx \boldsymbol{I} + \Delta \boldsymbol{A}_{\text{gen}}$; $\boldsymbol{B}_\Delta \approx \Delta \boldsymbol{B}_{\text{gen}}$
$\boldsymbol{A}_{\text{gen}}$	Continuous state generator	$\boldsymbol{A}_{\text{HiPPO}} = -(\boldsymbol{A}_{\text{gen}} + \boldsymbol{I})^\mathsf{T}$ (tilted/transposed)
$\boldsymbol{B}_{\text{gen}}$	Continuous input generator	Matches HiPPO directly: $(\boldsymbol{B}_{\text{gen}})_n = \phi_n(1) f'(0)$
$\boldsymbol{A}_{\text{HiPPO}}$	Stable HiPPO-LegS matrix (from (Gu et al., 2020))	Derived as special case of $\boldsymbol{A}_{\text{gen}}$

B PROOF OF PROPOSITION 1: CONTINUOUS GENERATOR $\boldsymbol{A}_{\text{gen}}$

This derivation assumes the stationary warping and continuous limit as summarized in Section 4.1.

Proposition 3 (Proposition 1 Restated: Continuous Generator $\boldsymbol{A}_{\text{gen}}$)**.** *For a stationary warping $\sigma_t(s) = f(s-t)$ with inverse $g = f^{-1}$, the generator matrix is*

$$(\boldsymbol{A}_{gen})_{nm} = \left\langle \phi_n, \frac{\phi_m'}{g'} \right\rangle = \int_Z \phi_n(z) \frac{\phi_m'(z)}{g'(z)} \, dz. \tag{34}$$

The derivation follows the backward-lag convention from eq.(21) with $\sigma_t \circ \sigma_{t+\Delta}^{-1}$.

First, find the inverse σ_t^{-1} by solving $\sigma_t(s) = f(s-t)$: $s = t + g(z)$, so $\sigma_t^{-1}(z) = t + g(z)$.

Next, compute $\sigma_t \circ \sigma_{t+\Delta}^{-1}$:

$$\sigma_t \circ \sigma_{t+\Delta}^{-1} = f(\Delta + g(z)). \tag{35}$$

To obtain $\boldsymbol{A}_t \approx \boldsymbol{I} + \Delta \boldsymbol{A}_{\text{gen}}$, evaluate the first-order Taylor expansion of $\phi_m(\sigma_t \circ \sigma_{t+\Delta}^{-1}(z))$ at $\Delta = 0$:

$$\begin{aligned} \text{Generator}(\phi_m) &= \frac{d}{d\Delta} \phi_m\Big(f\big(\Delta + g(z)\big)\Big)\Big|_{\Delta=0} \\ &= \phi_m'\Big(f\big(\Delta + g(z)\big)\Big) f'\big(\Delta + g(z)\big)\Big|_{\Delta=0}. \end{aligned}$$

At $\Delta = 0$, using $f(g(z)) = z$, this is $\phi_m'(z) \cdot f'(g(z))$. Thus,

$$(\boldsymbol{A}_{\text{gen}})_{nm} = \langle \phi_n, \phi_m' \cdot (f' \circ g) \rangle = \left\langle \phi_n, \frac{\phi_m'}{g'} \right\rangle. \tag{36}$$

□

C DERIVATION OF INPUT VECTOR $\boldsymbol{B}$ AND PROOF OF PROPOSITION 2

C.1 The Input Vector $\boldsymbol{B}$ and Different Hold Assumptions

Here, we derive $\boldsymbol{B}_\Delta$ and its limit to $\boldsymbol{B}_{\text{gen}}$ under different hold assumptions, consistent with the continuous ODE $\boldsymbol{c}'(t) = \boldsymbol{A}_{\text{gen}} \boldsymbol{c}(t) + \boldsymbol{B}_{\text{gen}} u(t)$ (see Section 4.1).

In the main text, the new input is modeled generally, leading to the recurrence $\boldsymbol{c}_{t+1} = \boldsymbol{A}_t \boldsymbol{c}_t + \boldsymbol{B}_t u_{t+1}$. The specific form of the discrete input matrix $\boldsymbol{B}_\Delta$ depends on the assumption made about the signal's behavior over the interval $[t, t+\Delta]$. Here, we derive $\boldsymbol{B}_\Delta$ for three common models under a stationary warp $\sigma_t(s) = f(s-t)$, where f is increasing, $f(0) = 1$, and $g = f^{-1}$.

The n-th component of the input matrix $(\boldsymbol{B}_\Delta)_n$ represents the full contribution of the input to the state update, computed as the projection of the input shape, $\mathrm{Input}(s; u_{t+\Delta})$, onto the basis at time $t+\Delta$:

$$\begin{aligned}(\boldsymbol{B}_\Delta)_n &= \langle \psi_{t+\Delta,n}, \mathrm{Input}(s; u_{t+\Delta}) \rangle_{\omega_{t+\Delta}} \\ &= \int_{-\infty}^{t+\Delta} \psi_{t+\Delta,n}(s) \cdot \mathrm{Input}(s; u_{t+\Delta}) \cdot \omega_{t+\Delta}(s)\, ds.\end{aligned} \tag{37}$$

For stationary warping $\sigma_{t+\Delta}(s) = f(s-t-\Delta)$ with the inverse $g = f^{-1}$, substituting $s - t = \Delta + g(z)$ into the input and performing a change of variables, the inner product simplifies to:

$$(\boldsymbol{B}_\Delta)_n = \int_{f(-\Delta)}^{1} \phi_n(z) \cdot \mathrm{Input}(t + \Delta + g(z); u_{t+\Delta})\, dz, \tag{38}$$

assuming f maps the interval appropriately (limits adjusted for the hold).

Note that for Dirac and ZOH, $\mathrm{Input}(s; u_{t+\Delta})$ is defined as a u-independent shape scaled by $u_{t+\Delta}$, so the full contribution is $\boldsymbol{B}_\Delta u_{t+\Delta}$; for FOH, the input shape embeds the linear interpolation of u_t and $u_{t+\Delta}$, which we factor to isolate the matrix component.

C.1.1 Dirac Delta Input (Impulse)

This model assumes the input is an instantaneous impulse at the boundary $s = t + \Delta$, simplifying analysis but approximating discrete semantics in the limit.

$$\mathrm{Input}(s; u_{t+\Delta}) = \delta(s - (t + \Delta)). \tag{39}$$

The projection simplifies to a point evaluation of the warped basis function:

$$\begin{aligned}(\boldsymbol{B}^\delta_\Delta)_n &= \langle \psi_{t+\Delta,n}, \mathrm{Input}(s; u_{t+\Delta}) \rangle_{\omega_{t+\Delta}} \\ &= \int_{-\infty}^{t+\Delta} \psi_{t+\Delta,n}(s)\, \delta(s - (t+\Delta))\, \omega_{t+\Delta}(s)\, ds \\ &= \psi_{t+\Delta,n}(t+\Delta)\, \omega_{t+\Delta}(t+\Delta) \\ &= \phi_n(\sigma_{t+\Delta}(t+\Delta))\, \omega_{t+\Delta}(t+\Delta) \\ &= \phi_n(1)\, |f'(0)|,\end{aligned} \tag{40}$$

assuming f is increasing so $|f'(0)| = f'(0)$. Since this is a boundary evaluation, $(\boldsymbol{B}^\delta_\Delta)_n$ is constant with respect to Δ. The full input contribution is then $(\boldsymbol{B}^\delta_\Delta)_n u_{t+\Delta}$.

C.1.2 Zero-Order Hold (ZOH) Input

This model assumes the input is held constant over the interval, representing a piecewise-constant signal, which more accurately captures discrete-time semantics.

$$\mathrm{Input}(s; u_{t+\Delta}) = \mathbf{1}_{[t,t+\Delta]}(s). \tag{41}$$

Applying the change of variables $z = \sigma_{t+\Delta}(s)$, the integration limits become $[f(-\Delta), 1]$, yielding:

$$(\boldsymbol{B}^{\mathrm{ZOH}}_\Delta)_n = \int_{f(-\Delta)}^{1} \phi_n(z)\, dz. \tag{42}$$

High-order basis functions ($\phi_n(z)$ for large n) oscillate near boundaries, so this integral averages behavior over the interval, scaled by $u_{t+\Delta}$.

C.1.3 First-Order Hold (FOH) Input

This model assumes the input is interpolated linearly from u_t to $u_{t+\Delta}$. The input shape is a linear ramp. To isolate the matrix component (linear in $u_{t+\Delta}$ as per main text), we factor accordingly:

$$\text{Input}(s; u_{t+\Delta}) = \mathbf{1}_{[t,t+\Delta]}(s) \cdot \left[u_t + \frac{u_{t+\Delta} - u_t}{\Delta}(s-t) \right]. \tag{43}$$

After the change of variables where $s - t = \Delta + g(z)$, the projection becomes:

$$\begin{aligned}
(\boldsymbol{B}_\Delta^{\text{FOH}})_n &= \int_{f(-\Delta)}^{1} \phi_n(z) \left[u_t + \frac{u_{t+\Delta} - u_t}{\Delta}(\Delta + g(z)) \right] dz \\
&= \int_{f(-\Delta)}^{1} \phi_n(z) \left[u_{t+\Delta} + \frac{u_{t+\Delta} - u_t}{\Delta} g(z) \right] dz \\
&= u_{t+\Delta} I_1(\Delta) + (u_{t+\Delta} - u_t) \frac{I_g(\Delta)}{\Delta},
\end{aligned} \tag{44}$$

where

$$\begin{aligned}
I_1(\Delta) &= \int_{f(-\Delta)}^{1} \phi_n(z)\, dz, \\
I_g(\Delta) &= \int_{f(-\Delta)}^{1} \phi_n(z) g(z)\, dz.
\end{aligned} \tag{45}$$

The u_t term contributes to the higher-order correction, which vanishes in the limit (see below).

C.2 Proof of Proposition 2: Convergence to the Continuous Generator $\boldsymbol{B}_{\text{gen}}$

While the discrete matrices differ, we now show that the ZOH and FOH models converge to the same continuous generator $\boldsymbol{B}_{\text{gen}}$, consistent with the Dirac case. This demonstrates that for small Δ, the system's dynamics are insensitive to the hold order.

To obtain the generator, assume $u(s)$ is a smooth continuous function, so $u_t = u(t)$ and $u_{t+\Delta} = u(t+\Delta)$.

In the continuous limit, the SSM ODE is $\boldsymbol{c}'(t) = \boldsymbol{A}_{\text{gen}}\boldsymbol{c}(t) + \boldsymbol{B}_{\text{gen}}u(t)$, so the discrete approximation is $\boldsymbol{c}_{t+\Delta} \approx \boldsymbol{c}_t + \Delta[\boldsymbol{A}_{\text{gen}}\boldsymbol{c}_t + \boldsymbol{B}_{\text{gen}}u(t)] + O(\Delta^2)$. Thus, the full input contribution over Δ should be $\boldsymbol{B}_\Delta u_{t+\Delta} \approx \Delta \boldsymbol{B}_{\text{gen}} u(t)$, and

$$\boldsymbol{B}_{\text{gen}} u(t) = \lim_{\Delta \to 0} \frac{\boldsymbol{B}_\Delta u_{t+\Delta}}{\Delta}. \tag{46}$$

Since $\boldsymbol{B}_\Delta$ is the coefficient multiplying $u(t+\Delta)$, we isolate $\boldsymbol{B}_{\text{gen}}$ directly by

$$(\boldsymbol{B}_{\text{gen}})_n = \lim_{\Delta \to 0} \frac{(\boldsymbol{B}_\Delta)_n}{\Delta}. \tag{47}$$

Proposition 4 (Proposition 2 Restated: Continuous Input Generator)**.** *Let the discrete input matrix $\boldsymbol{B}_\Delta$ be the coefficient such that the state update includes the term $\boldsymbol{B}_\Delta u(t+\Delta)$. In the continuous limit, the generator $\boldsymbol{B}_{\text{gen}}$ is given by:*

$$(\boldsymbol{B}_{\text{gen}})_n = \lim_{\Delta \to 0} \frac{(\boldsymbol{B}_\Delta)_n}{\Delta} = \phi_n(1) f'(0). \tag{48}$$

Proof. The proof relies on analyzing the first-order behavior of the ZOH and FOH integrals as $\Delta \to 0$. Recall from Section C.1 (ZOH and FOH definitions) that

$$\begin{aligned}
I_1(\Delta) &= \int_{f(-\Delta)}^{1} \phi_n(z)\, dz, \\
I_g(\Delta) &= \int_{f(-\Delta)}^{1} \phi_n(z) g(z)\, dz.
\end{aligned}$$

For **ZOH** (see eq. (42) in Section C.1.2), the matrix is $(\boldsymbol{B}_\Delta^{\text{ZOH}})_n = I_1(\Delta)$, and the full contribution is $I_1(\Delta) u_{t+\Delta}$.

First, we find the first-order approximation of these integrals. Using the Leibniz integral rule on $I_1(\Delta)$:

$$\frac{d}{d\Delta} I_1(\Delta) = -\phi_n(f(-\Delta)) \cdot f'(-\Delta) \cdot (-1). \tag{49}$$

Evaluating at $\Delta = 0$ (and using $f(0) = 1, g(1) = 0$):

$$\frac{d}{d\Delta} I_1(\Delta)\bigg|_{\Delta=0} = \phi_n(1) f'(0). \tag{50}$$

This implies the first-order Taylor expansion is $I_1(\Delta) = \Delta \cdot \phi_n(1) f'(0) + O(\Delta^2)$. [3]

Similarly for $I_g(\Delta)$:

$$\frac{d}{d\Delta} I_g(\Delta)\bigg|_{\Delta=0} = \phi_n(1) g(1) f'(0) = \phi_n(1) \cdot 0 \cdot f'(0) = 0. \tag{51}$$

This implies $I_g(\Delta) = O(\Delta^2)$, so it vanishes faster than first order.

Now, we apply these limits. For **ZOH**:

$$\lim_{\Delta \to 0} \frac{(\boldsymbol{B}_\Delta^{\text{ZOH}})_n u_{t+\Delta}}{\Delta} = \lim_{\Delta \to 0} \frac{I_1(\Delta) u_{t+\Delta}}{\Delta} \approx \lim_{\Delta \to 0} \frac{\Delta \cdot \phi_n(1) f'(0) u(t)}{\Delta} = \phi_n(1) f'(0) u(t), \tag{52}$$

using $u_{t+\Delta} \to u(t)$ by continuity. Thus $(\boldsymbol{B}_{\text{gen}})_n = \phi_n(1) f'(0)$.

For **FOH** (see Section C.1.3 for the FOH derivation), since u is continuous, $u_{t+\Delta} \to u(t)$ and $u_{t+\Delta} - u_t \to 0$ as $\Delta \to 0$. The full projection is:

$$(\boldsymbol{B}_\Delta^{\text{FOH}})_n = u_{t+\Delta} I_1(\Delta) + (u_{t+\Delta} - u_t) \frac{I_g(\Delta)}{\Delta}. \tag{53}$$

Dividing by Δ and taking the limit:

$$\begin{aligned} \lim_{\Delta \to 0} \frac{(\boldsymbol{B}_\Delta^{\text{FOH}})_n}{\Delta} &= \lim_{\Delta \to 0} \left[u_{t+\Delta} \frac{I_1(\Delta)}{\Delta} + (u_{t+\Delta} - u_t) \frac{I_g(\Delta)}{\Delta^2} \right] \\ &= u(t)[\phi_n(1) f'(0)] + \left(\lim_{\Delta \to 0} (u_{t+\Delta} - u_t) \right) \cdot \left(\lim_{\Delta \to 0} \frac{I_g(\Delta)}{\Delta^2} \right) \\ &= u(t)[\phi_n(1) f'(0)] + 0 \cdot \text{const.} \\ &= \phi_n(1) f'(0) u(t), \end{aligned} \tag{54}$$

Thus $(\boldsymbol{B}_{\text{gen}})_n = \phi_n(1) f'(0)$, matching ZOH and the Dirac boundary case.

The consistency arises because the u_t term in FOH contributes only to higher-order terms (e.g., via $I_g(\Delta)$), which vanish in the $\Delta \to 0$ limit, leaving the $u_{t+\Delta} \approx u(t)$ term dominant. □

C.3 Conclusion for HiPPO-LegS

For the HiPPO-LegS special case ($f(x) = e^x$, $f'(0) = 1$, and $\phi_n(z)$ being normalized Legendre polynomials on $[0, 1]$), all three input models are consistent in the continuous limit, yielding the well-known generator:

$$(\boldsymbol{B}_{\text{gen}})_n = \phi_n(1) \cdot 1 = \sqrt{2n+1}, \tag{55}$$

where the full contribution is $(\boldsymbol{B}_{\text{gen}})_n u(t) = \sqrt{2n+1} u(t)$.

D GEOMETRIC ILLUSTRATION OF THE LAG OPERATOR AND *DOMAIN EXPANSION*

The lag operator $\sigma_t \circ \sigma_{t+1}^{-1}$ geometrically encodes the *domain expansion* of the structured projection basis when time advances from t to $t+1$. Figure 5 visualizes this process for a generic warping function σ_t.

[3] We note that while the Leibniz rule provides a direct path to the boundary derivative, we frame the analysis via first-order Taylor expansion here to emphasize the intuitive linear limit consistent with the main text's generator derivation.

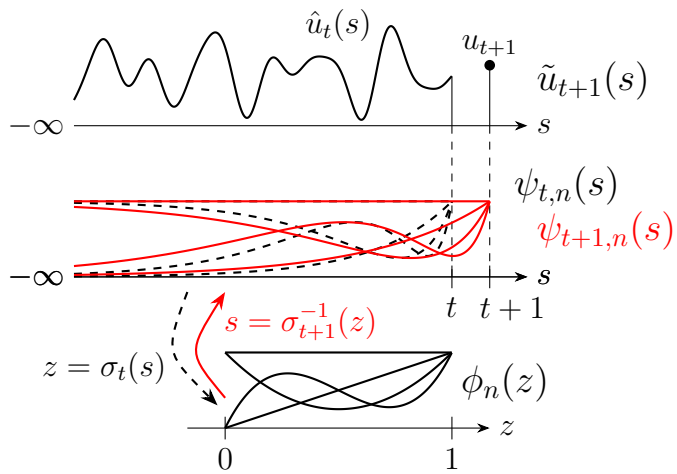

Figure 5: Illustration of the domain-expansion process realized by the lag operator $\sigma_t \circ \sigma_{t+1}^{-1}$. **Top:** The prior approximation $\hat{u}_t(s)$ (black) on $T_t = (-\infty, t]$ is extended by the new input sample u_{t+1} (black dot) to form the interim signal $\tilde{u}_{t+1}(s)$ on $T_{t+1} = (-\infty, t+1]$. **Middle:** The old warped basis functions $\psi_{t,n}(s)$ (dashed black) are re-warped into the new basis $\psi_{t+1,n}(s)$ (solid red) to accommodate the expanded domain. **Bottom:** All basis functions are images of the *same* canonical orthonormal basis $\phi_n(z)$ on $Z = [0, 1]$ under the respective warping maps σ_t and σ_{t+1}. The arrows highlight the geometric action of the lag operator, which is exactly the composition that yields the state-transition matrix A_t via a single inner product (eq. (21)).

This visualization illustrates the geometric action of the lag operator, specifically the domain expansion and compression of the basis functions over a time step. The state transition matrix $\boldsymbol{A}_\Delta$ is simply the algebraic projection of this geometric shift, demonstrating how the system's dynamics are constructed directly from time-warping principles rather than continuous-time ODEs.

E DETAILS ON THE FREQUENCY-PRESERVING LAG-SSM

As briefly sketched in Section 6.2, a major limitation of the standard exponential warp is that it stretches signals in time. A basis function that oscillates at a fixed frequency in the canonical domain Z becomes a chirp (variable frequency) in the time domain s. This makes it inefficient for signals with sustained oscillatory components.

Our framework addresses this by designing a time-warp σ_t that preserves frequency in the s-domain while still providing infinite decaying memory. Figure 6 visualizes the piecewise-linear frequency-preserving time warp of a fixed-frequency oscillatory basis function in the time domain.

E.1 Construction of the Piecewise-Linear Warp

The Frequency-Preserving Lag-SSM is built as a *hybrid* model. It combines two families of bases and associated warps: one subset designed to preserve frequency of oscillation in the s-domain ($\in T_t$), and another subset (e.g., Legendre polynomials with the standard exponential warp) to capture transients and decaying memory. We elaborate on this construction below.

E.2 Geometric Consequence: Unwarped Fourier Basis

The motivation is that even when the canonical basis Φ on Z contains oscillatory components, the standard exponential warp stretches them into chirps (variable frequency) in the s-domain. The piecewise-linear warp introduced below resolves this issue while preserving both orthogonality and infinite decaying memory, as well as the periodicity of the basis functions in the s-domain.

Recall the two domains: the physical time domain $s \in T_t$ and the canonical domain $z \in Z$. The inner product

on Z uses the Lebesgue measure $d\mu(z) = dz$:

$$\langle U, V \rangle = \int_Z U(z)V(z)\, d\mu(z). \tag{56}$$

First, for the s-domain measure ω_t on $(-\infty, t] = T_t$, we consider a step function with infinitely many (decaying) n steps ($n = 1, 2, ...$) of length L and the values are constant across the interval $(t - nL,\, t - (n-1)L]$ of height h_n where $h_n > 0$ and $\sum_{n=1}^{\infty} h_n = 1$. For example, $h_n = C\, e^{-n}$ or $\frac{C}{n^2}$ (C is for normalization).

Then we define the time-warp $\sigma_t(s)$ by dividing Z into intervals of length h_n.

More precisely, for any s in the interval $(t - nL, t - (n-1)L]$, we map it to the Z's divided interval H_n, such that

$$H_n = \left(1 - \sum_{i=1}^{n} h_i,\, 1 - \sum_{i=1}^{n-1} h_i\right). \tag{57}$$

By decomposing s to see its lower bound and mid-point via the length and ratio,

$$s = (t - nL) + \left(\frac{s - t + nL}{L}\right) L. \tag{58}$$

We can get the z in the corresponding interval H_n,

$$\sigma_t = z = \left(1 - \sum_{i=1}^{n} h_i\right) + \left(\frac{s - t + nL}{L}\right) h_n. \tag{59}$$

Now σ_t is a decaying piecewise-linear time-warp yielding a constant measure $\omega_t(s) = |\sigma_t'(s)| = h_n$ over the interval H_n. This means that if we define a Fourier basis in the s-domain with period L, i.e. $\{\sin(2\pi k/L), \cos(2\pi k/L); k = 0, 1, 2, \dots\}$, the waveform is piecewise preserved in the z-domain within each interval H_n, ensuring that orthogonality is maintained in the z-space (Figure 6).

E.3 Extension to Hybrid Model: Transient Decaying Memory

To capture the remaining degrees of freedom (e.g., smooth transients and decaying memory), we can augment the Fourier basis with a Legendre subset under the standard exponential warp. The full hybrid basis is then orthogonalized with respect to the oscillatory basis via the Gram–Schmidt process.

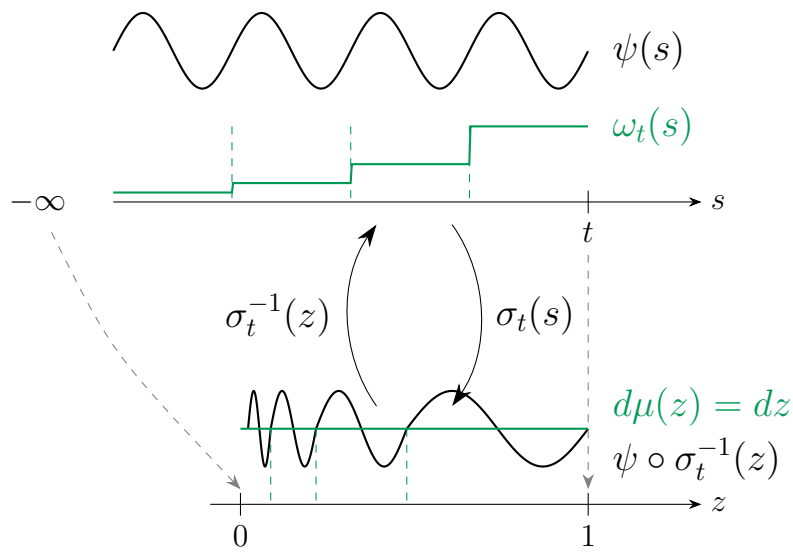

Figure 6: Illustration of the piecewise-linear time warping for the Frequency-Preserving Lag-SSM (Sec. 6.2). The infinite history interval $T_t = (-\infty, t]$ (top, s-axis) is partitioned into fixed-length windows of size L, each weighted by a decaying constant ω_n (green step function $\omega_t(s) = \omega_n/L$ on the n-th window). The invertible map σ_t linearly compresses each window into a segment of the canonical interval $Z = [0, 1]$ whose length is proportional to ω_n. Consequently, the induced measure on Z remains uniform ($d\mu(z) = dz$, green line), so a Fourier basis defined on Z maps back to pure, unwarped sinusoids in the original time domain (visible in the warped signal $u \circ \sigma_t^{-1}(z)$). This yields an SSM that implements a sliding-window Fourier transform with infinite decaying memory, derived directly via the lag operator without any ODE.