S^3 : Learnable Spline-Wavelets for State Space Models

Daniel Cai[*]
Brown University
daniel_cai@brown.edu

Randall Balestriero
Brown University

Abstract

A foundational principle of State Space Models (SSMs) involves efficient function compression: compressing input signals into a low-dimensional coefficient vector via projection onto orthogonal bases. Recent work has extended this framework from orthogonal bases towards general frames, which allows for diversity and flexibility in this step of SSM construction. However, this framework yields no principled way of choosing the right basis, and manual selection is challenging when the data regime is unknown. To this end, we propose learnable spline-wavelets for SSMs (S^3), addressing this concern by making the frame itself learnable. The resulting construction matches or exceeds the best fixed wavelet frames on various synthetic test suites, and we further show that the spline wavelet learns data-dependent characteristics.

1. Introduction

In recent years, SSMs have re-emerged as a competitive alternative to attention-based architectures on long-sequence tasks. Fundamentally, they are based on an older principle of online function approximation via projection onto a fixed basis (HiPPO) [6]. The HiPPO construction derives, from a chosen basis and measure, a linear ODE whose state at time t is represented by the optimal coefficient vector of the input signal.

The original HiPPO construction is closed-form and only applicable for orthogonal bases, with specific identities proposed for both Legendre and Laguerre polynomials. SSMs for frame-agnostic representation (SaFARi) [1] remove this restriction: given any densely-sampled frame, we may compute state-transition matrices numerically from the frame and its dual. This generalization yields SSMs based on wavelet frames, which offer superior flexibility and performance on signals with sharp transitions and localized energy.

However, the SaFARi framework also requires a static frame, which must therefore be chosen based on prior knowledge of data characteristics. This choice naturally lends an inductive bias in the SSM, and existing work applying SaFARi typically settles on the Daubechies family [4], with no principled justification beyond convention.

We propose a class of *learnable wavelets* for SSMs, that do not require manual selection. S^3 lives under the SaFARi framework, but is fully differentiable. Our contributions are two-fold:

- A fully differentiable construction of spline-wavelets for SSMs, allowing end-to-end backpropagation through frame and dual construction and SSM rollout.
- Empirical evidence that S^3 matches or exceeds the best fixed wavelet performance over various synthetic and natural tasks, and learns data characteristics even from random initialization.

2. Background

2.1. HiPPO and Online Function Approximation

The goal of online function compression is to maintain a coefficient vector $c(t) \in \mathbb{R}^N$ at each time t, representing the optimal N-term approximation of the input signal $u(\tau)$ for $\tau \in [0, t]$ in some orthogonal basis under a measure μ. HiPPO does so by deriving a linear ODE for $c(t)$:

$$\dot{c}(t) = -Ac(t) + Bu(t),$$

with A and B static transition matrices determined by the choice of Laguerre/Legendre polynomial basis, and scaled/translated measure. Specifically, the first *scaled* measure μ_{sc} dilates the basis as t grows, so that the window $[0, t]$ is always covered. The *translated* measure μ_{tr} instead slides a fixed-width window. Then, discretization of the ODE is traditionally achieved via the Generalized Bilinear Transform (GBT) yielding a linear recurrence $c_{n+1} = \bar{A}c_n + \bar{B}u_n$ that runs online at $O(N)$ per step for the structured transition matrices of HiPPO.

The original HiPPO methodology relies on closed-form identities for Legendre and Laguerre polynomials to derive the state transition matrices A, B. Notably, the framework is heavily restricted to orthogonal bases, and moreover there are very few well-behaved bases that admit such identities.

[*]Paper approved as a Master's report for this author.

2.2. SaFARi: Frame-Agnostic Construction

SaFARi removes this basis requirement. Instead, given a frame $\{\phi_i\}_{i=1}^N$ sampled at L grid points on $[0, 1]$, take $F \in \mathbb{R}^{N\times L}$ with $F_{ij} = \phi_i(t_j)$ and $dF_{ij} = \phi_i'(t_j)$. Then, compute the canonical dual via pseudoinverse, $D = (F^+)^T$. In our implementation, the transition matrices follow:

$$\begin{aligned}
\text{scaled:} \quad & A = I + dF \cdot \operatorname{diag}(t/L) \cdot D^T, \\
& B = F[:, -1], \\
\text{translated:} \quad & A = L \cdot F[:, 0]\, D[:, 0]^T + dF \cdot D^T, \\
& B = F[:, -1].
\end{aligned}$$

Unlike HiPPO, this construction is entirely agnostic to the choice of frame. WaLRUS [2] instantiates the SSM with Daubechies-based orthonormal frame and reports strong reconstruction performance on standard benchmarks. However, like HiPPO, the wavelet frame itself is still fixed throughout training, and must be manually tuned as a hyperparameter.

3. S^3: Learnable Spline-Wavelets for SSMs

To this end, we propose a family of differentiable spline-wavelets, that yield a learnable frame through training. Specifically, since the SaFARi pipeline is differentiable in its frame and derivative (F and dF), we aim to generate them from parameter θ in a fully differentiable manner, thus allowing the frame to adapt to the signal characteristics.

Cubic Hermite Spline-Wavelets. To generate F, we leverage cubic Hermite spline wavelets, which yield a parameterization of θ that is flexible, admits a closed-form derivative, and decouples parameter count from discretization length L. Following [3], we place n uniformly spaced knots $\{k_0 = 0, k_1, \dots, k_n = 1\}$ on the unit interval, and assign each knot a learnable value p_i and tangent m_i. The collection $\theta := \{(p_i, m_i)\}_{i=0}^n$ uniquely determines, on each subinterval $[k_i, k_{i+1})$, a cubic polynomial f_i via Hermite interpolation of (p_i, m_i) and (p_{i+1}, m_{i+1}). The mother wavelet is then

$$\psi_\theta(t) = f_i(t), \quad t \in [k_i, k_{i+1}), \quad i = 0, 1, \dots, n-1.$$

We enforce three constraints on ψ_θ before each forward pass to maintain wavelet admissibility [9]: zero values at the boundary knots ($p_0 = p_n = 0$), zero mean across the interior ($\sum_{i=1}^{n-1} p_i = 0$), and unit energy ($\int_0^1 \psi_\theta^2 = 1$).

This construction yields two useful properties.

- Both $\psi_\theta(t)$ and $\psi_\theta'(t)$ have analytic expressions in the standard Hermite basis, so F and dF can be evaluated without finite differences.
- The parameter count of θ scales with n rather than L, allowing for high-resolution discretization of our frame while keeping the optimization low-dimensional.

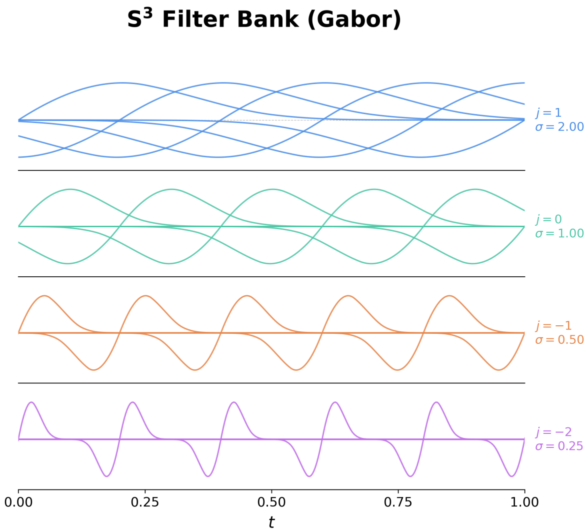


Figure 1. Multi-resolution wavelet filterbank generated from the mother wavelet ψ_θ via scale-shift discretization.

Subsampling. From the admissible mother wavelet ψ_θ, we construct a multi-resolution dictionary by the standard scale-shift discretization [9]:

$$\{\psi_{j,n}\}_{(j,n)\in\mathbb{Z}^2} = \left\{ \frac{1}{\sqrt{a^j}}\, \psi_\theta\left(\frac{t - n\mu_0 a^j}{a^j} \right) \right\}.$$

We sweep j over a contiguous range $[j_{\min}, j_{\max}]$ and n over a fixed number of shifts per scale, giving a dictionary of total size N.

Moment Regularization. To improve the regularity of the wavelet, we also enforce vanishing moments via a soft penalty. For each desired moment order $k = 0, \dots, m-1$, we add a term $\lambda_m (\sum_i t_i^k \psi_\theta(t_i))^2$ to the loss.

Algorithm 1 Differentiable Forward Pass

Require: Signal $u \in \mathbb{R}^T$, parameters θ, threshold rcond
Ensure: Reconstruction $\hat{u} \in \mathbb{R}^T$
Evaluate $F, dF \in \mathbb{R}^{N\times L}$ from θ
Compute $D = \operatorname{pinv}(F, \mathrm{rcond})^T$
Construct A, B per the chosen measure
for $t = 0, 1, \dots, T-1$ **do**
  $c_{t+1} \leftarrow \bar{A}_t\, c_t + \bar{B}_t\, u_t$
end for
Reconstruct $\hat{u} = D^T c_T \cdot L$ and resample to length T

The full forward pass is summarized in Algorithm 1, and we observe that every operation is differentiable. In particular, PyTorch's [10] `linalg.pinv` backpropagates through the SVD, and the GBT is a sequence of matrix products in A and B. The reconstruction loss $\|\hat{u} - u\|_2^2$

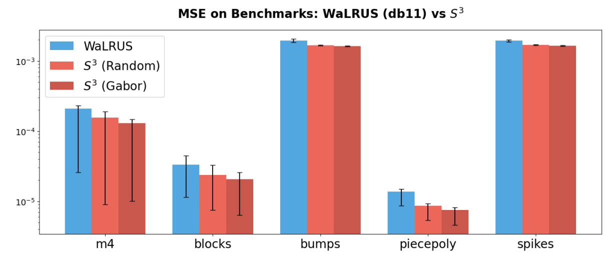


Figure 2. Test MSE on synthetic and natural evaluation suites. S^3 with Gabor initialization outperforms WaLRUS-db11 for controlled filter counts. Random initialization remains competitive but exhibits higher variance across seeds.

is therefore a function of θ, achieving end-to-end differentiability from our reconstruction to our frame parameters.

It is worth noting that determining whether the resulting wavelet dictionary strictly follows the frame inequalities depends on numerical constraints for ψ_θ, a, and μ_0 that are impossible to tractably characterize and verify. Empirically, we find that taking μ_0 small enough and constraining ψ_θ to have at least one vanishing moment (i.e, subsampling sufficiently thoroughly and enforcing regularity) yields stable training.

4. Experiments

4.1. Datasets

We follow the evaluation protocol of WaLRUS and SaFARi.

M4 Monthly. This natural dataset consists of 300 natural sequences from the M4 forecasting competition [8], resampled to length $L = 500$ and unit-normalized. All series are derived from real-world economic signals.

Synthetics. Our synthetic test suite is derived from four signal classes [5] — blocks, bumps, piecewise polynomials (piecepoly), and spikes — generated through the WaLRUS reference implementation. Each class tackles a different feature regime, and we use 500 samples per class at $L = 500$.

4.2. Baselines

We compare S^3 against two baselines, controlled for same filterbank length N: WaLRUS with Daubechies-11 (based on the configuration of [2]), as well as the *best fixed* Daubechies wavelet *per dataset*, selected via a sweep of db1 through db16. The latter is a proxy of comparison for S^3 against the optimally-selected wavelet.

4.3. Configuration

Our parameter selection is described below for reproducibility:

Dataset	Best db	S^3 (Gabor)
M4	db6 (3.96×10^{-5})	1.30×10^{-5}
Blocks	db11 (1.9×10^{-5})	6.05×10^{-6}
Bumps	db8 (3.90×10^{-3})	1.62×10^{-3}
Piecepoly	db8 (1.03×10^{-5})	7.55×10^{-6}
Spikes	db8 (2.47×10^{-3})	1.64×10^{-3}

Table 1. Best test MSE (mean over 5 seeds) over synthetic and natural signals. S^3 with Gabor initialization exceeds the performance of the best-performing Daubechies wavelet frame across all datasets.

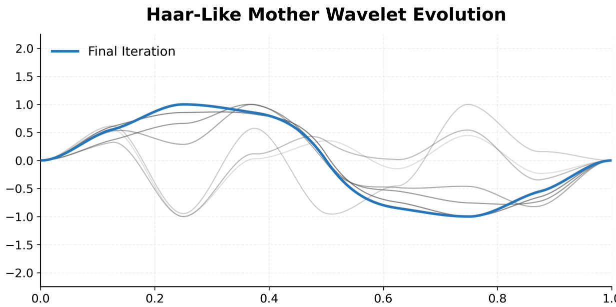
- Spline knots: $n \in \{5, 8\}$
- Scale range: $(j_{\min}, j_{\max}) \in \{(-2, 2), (-3, 1)\}$
- Shifts per scale: 80
- Optimizer: Adam [7] with $lr = 5 \times 10^{-2}, 10^{-3}$ for control points and tangents, respectively.
- Vanishing-moment weight: $\lambda_m = 10^{-2}$
- Training: 100 epochs, mean test MSE reported across 5 seeds

5. Results

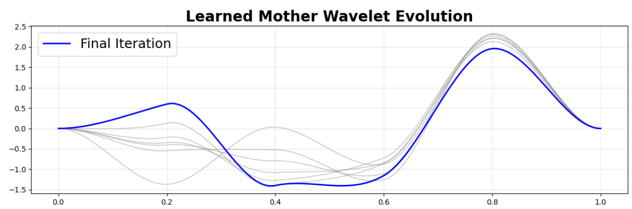
Learned spline-wavelets outperform the best fixed wavelet. Over all synthetic and natural sequences, S^3 from both random and Gabor initialization consistently outperforms both baselines (Daubechies and optimal fixed-wavelet) (Table 1). We note that the biggest gains occur on synthetic blocks and piecewise polynomials, where Daubechies tend to oversmooth. Notably, on M4, where signal characteristics vary more substantially, S^3 still learns a wavelet that performs on-par even from random initialization.

Learned wavelets reflect signal structure. Figure 3 displays the learned mother wavelet evolution for block-signals and M4. According to classical theory, Haar wavelets are typically preferred the former signal-type, due to their localized analysis of abrupt changes. Surprisingly, through training S^3 converges to a wavelet that resembles the step-like profile of Haar. On M4, the learned wavelet is asymmetric and slightly oscillatory, similar to the structure of the baseline Daubechies wavelet. In both examples, we proceeded from random initialization, yet our optimization objective yielded different wavelets that closely reflected the data characteristics.

Learning is sensitive to matrix conditioning. We note that for random initialization, we observed substantial run-to-run variance in the final test loss. Some seeds converged to a low-loss landscape within a few epochs, while others stalled several times higher with no convergence. Gabor initialization was far more stable. We suggest a rea-



(a) Blocks



(b) M4

Figure 3. Evolution of the learned mother spline-wavelet during training, with random init. On blocks (top), the wavelet converges to a step-like profile resembling Haar. On M4 (bottom), it converges to an asymmetric, regular structure resembling Daubechies.

sonable explanation for this phenomenon, tracing back to the pseudoinverse-based dual: $D = \mathrm{pinv}(F, \mathrm{rcond})^T$ truncates singular values of F below $\mathrm{rcond} \cdot \sigma_{\max}$. If F has singular values clustered near this threshold, small parameter changes can cause many singular values to cross the cutoff, yielding ill-conditioned gradients for the affected filters. Under random initialization, this may be particularly pronounced, leading to the optimization issues observed.

6. Discussion

Across our experiments, S^3 consistently wins on both natural and synthetic 1D signals. In particular, S^3 significantly outperforms the baselines when the feature regime is poorly matched to Daubechies frame, and performs on-par when the best fixed wavelet is already well-suited for the task.

While S^3 demonstrates strong performance on 1D reconstruction tasks, we note several limitations:

- **Computational cost.** Since S^3 is not diagonalizable by construction, it results in a $O(T)$ recurrence. Moreover, the transition matrices are a function of θ and thus change throughout training; per-step cost is $O(N^2)$ from the GBT inverse, and frame construction costs $O(NL + N^2 L)$ from F and the pseudoinverse. This complexity is negligible for shorter sequences, but fails to scale to the deep-SSM regime.
- **Initialization sensitivity.** As mentioned previously, S^3 is sensitive to matrix conditioning, with substantial run-to-run variance in final test performance. Given this dependence on initialization, we suggest that practitioners leveraging S^3 default to Gabor initialization.

7. Conclusion

Under the SaFARi framework, choosing the appropriate frame is a computationally expensive and challenging task on diverse data modes. To alleviate this limitation, we present S^3, a learnable extension in which the mother wavelet is spline-parameterized and thus trainable end-to-end by construction. On both synthetic and natural 1D signals, S^3 matches or exceeds the performance of traditional fixed wavelets. Moreover, the learned wavelet quickly adapts to the data characteristics. Given the promising results on 1D signals but aforementioned limitations in applications to deep SSMs, we invite further research, particularly into constructions that are frame-theoretic and computationally tractable.

References

[1] Hossein Babaei, Mel White, Sina Alemohammad, and Richard G. Baraniuk. Safari: State-space models for frame-agnostic representation, 2025. 1

[2] Hossein Babaei, Mel White, Sina Alemohammad, and Richard G. Baraniuk. Walrus: Wavelets for long-range representation using ssms, 2025. 2, 3

[3] Randall Balestriero, Romain Cosentino, Hervé Glotin, and Richard Baraniuk. Spline filters for end-to-end deep learning. In *International conference on machine learning*, pages 373–382. PMLR, 2018. 2

[4] Ingrid Daubechies. *Ten Lectures on Wavelets*. SIAM, 1992. 1

[5] David L. Donoho and Iain M. Johnstone. Adapting to unknown smoothness via wavelet shrinkage. *Journal of the American Statistical Association*, 90(432):1200–1224, 1995. 3

[6] Albert Gu, Tri Dao, Stefano Ermon, Atri Rudra, and Christopher Re. Hippo: Recurrent memory with optimal polynomial projections, 2020. 1

[7] Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014. 3

[8] Spyros Makridakis, Evangelos Spiliotis, and Vassilios Assimakopoulos. The m4 competition: 100,000 time series and 61 forecasting methods. *International Journal of Forecasting*, 36(1):54–74, 2020. 3

[9] Stéphane Mallat. *A wavelet tour of signal processing (2. ed.)*. Academic Press, 1999. 2

[10] Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, 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 *Advances in Neural Information Processing Systems 32*, 2019. 2