An Uncertainty Principle for Linear Recurrent Neural Networks

Alexandre François ALEXANDRE.FRANCOIS@INRIA.FR
INRIA, Ecole Normale Superieure, PSL Research University, France

Antonio Orvieto ANTONIO@TUE.ELLIS.EU
Max Planck Institute for Intelligent Systems, ELLIS Institute Tübingen, Germany

Francis Bach FRANCIS.BACH@INRIA.FR
INRIA, Ecole Normale Superieure, PSL Research University, France

Editors: Nika Haghtalab and Ankur Moitra

Abstract

We consider linear recurrent neural networks, which have become a key building block of sequence modeling due to their ability for stable and effective long-range modeling. In this paper, we aim at characterizing this ability on the simple but core copy task, whose goal is to build a linear filter of order S that approximates the filter that looks K time steps in the past (which we refer to as the shift-K filter), where K is larger than S. Using classical signal models and quadratic cost, we fully characterize the problem by providing lower bounds of approximation, as well as explicit filters that achieve this lower bound up to constants. The optimal performance highlights an uncertainty principle for this task: the optimal filter has to average values around the K-th time step in the past with a range (width) that is proportional to K/S.

Keywords: Recurrent neural networks, sequence models, state-space models, rational approximations.

1. Introduction

Since their early development (Rumelhart et al., 1986; Elman, 1990), recurrent neural networks (RNNs) have advanced machine learning for sequential data, with milestones such as echo-state networks (Jaeger, 2001) and LSTMs (Hochreiter and Schmidhuber, 1997). However, two problems severely limit the application of classical RNNs in modern times: (1) GPU hardware optimized for large matrix operations struggles with efficient sequential processing, and (2) RNNs are notoriously difficult to train due to vanishing and exploding gradients (Bengio et al., 1994; Pascanu et al., 2013). As a result, transformers (Vaswani et al., 2017) have emerged as the dominant solution for sequence processing, offering desirable scalability properties and less challenging optimization. However, the attention mechanism powering transformers relies on computing pairwise interactions between inputs at each timestamp, resulting in a squared inference and memory complexity $O(L^2)$ in the sequence length L. Instead, classical RNNs require one pass through the data to recurrently update their hidden state, bringing their complexity down to $O(L)$. This property is particularly desirable in the long-context setting (e.g., analysis of long documents or genomics).

Indeed, in the interest of efficiency, we have recently witnessed a *resurgence of new RNNs* in state-of-the-art industry-size applications such as language modeling (Gu and Dao, 2024; Peng et al., 2024; Qin et al., 2024; De et al., 2024; Yang et al., 2024b). Sparked from the S4 model (Gu et al., 2022b), these new recurrences offer $O(L)$ complexity as classical RNNs, yet are parallelizable

on modern hardware like attention. At the core of their efficiency is a simplified recurrence that is *linear* in the hidden state:

$$x_n = A_n x_{n-1} + B_n u_n, \tag{1}$$

where u_n is the input data at timestamp n, x_n is the hidden state (which is a linear combination of inputs $u_1, u_2, \ldots, u_n$), and A_n, B_n are input-controlled transition matrices with a special parametrization (Orvieto et al., 2023). Compared to previous RNNs, A_n and B_n have *no dependency on the hidden state*—a feature which reduces expressivity (Merrill et al., 2024; Cirone et al., 2024) but unlocks GPU-efficient processing (Martin and Cundy, 2018; Smith et al., 2023).

New linear RNNs offer improved inference complexity and competitive performance on language modeling tasks (Dao and Gu, 2024; Waleffe et al., 2024), as well as state-of-the-art results on several other domains including vision (Liu et al., 2024; Li et al., 2025; Liang et al., 2024; Xing et al., 2024), audio generation (Goel et al., 2022), online learning (Zucchet et al., 2023), reinforcement learning (Lu et al., 2023) and genome analysis, where the $O(L)$ complexity can tackle long DNA sequences (Nguyen et al., 2024).

Despite the practical advantages of new linear recurrent mechanisms, we are at a very early evaluation stage in regards to assessing and understanding the capabilities and optimization properties of such systems when compared to (1) transformers and (2) non-linear (classic) RNNs. While several works are devoted to establishing a direct connection between transformers and linear RNNs (Katharopoulos et al., 2020; Schlag et al., 2021; Ali et al., 2024; Sieber et al., 2024), others point to fundamental and drastic differences in regards to expressivity and basic capabilities. Further, despite Orvieto et al. (2024); Wang and Xue (2024); Cirone et al. (2024) provide infinite-width theoretical guarantees for the expressivity of deep architectures based on linear RNNs, other works focusing on specific reasoning tasks of general interest in language modeling tell a different story: Arora et al. (2023) identified in the problem of selective *copying* (i.e., of recalling a specific value from the past, when presented the relative key) a fundamental discrepancy between attention and RNNs: building up a memory of past inputs, as opposed to attention's direct edges between tokens, can fundamentally limit finite-width performance of linear RNN based models. This finding inspired a formal investigation by Jelassi et al. (2024), who proved that perfectly retrieving (i.e., with zero error) inputs from distant past requires the RNN width to increase linearly with the sequence length. This in contrast to attention-based models, that can build associative mechanisms to solve such tasks with 2 layers Olsson et al. (cf. 2022). Similarly, Arora et al. (2024) and Bhattamishra et al. (2024) identify a tradeoff between model's state size and recall capability, highlighting key differences between attention-based models and efficient recurrent alternatives.

Inspired both by the practical relevance of new linear RNNs and by the need of further theoretical investigations of their basic properties, in this work we mathematically investigate arguably the most basic long-range task: recalling inputs seen K timestamps before the current processing step. Such task has a close relation to the *copy task* by Jelassi et al. (2024), while being simpler and with a clear challenge: successful replay as K increases. As Jelassi et al. (2024), we are specifically interested in characterizing optimal performance as a function of the recall range K and the memory size S—the dimension of the hidden state x. Yet, while Jelassi et al. (2024) work in the finite-vocabulary input setting standard in language modeling, assuming no particular structure in the recurrence, we take instead a signal processing approach, which allows us to characterize in detail the tradeoff between long-memory requirements (large K) and optimal recall resolution under reduced memory size ($S < K$). Similarly, both Arora et al. (2024) and Bhattamishra et al. (2024) consider inputs coming from a finite dictionary and only perfect recall. We instead consider

a richer scenario inspired by natural data (smooth, real-valued, and potentially correlated inputs) where approximate recall is tolerable and one is interested in precisely quantifying the tradeoff between precision and long-range capabilities. Since the task is independent from the input value to recall, we restrict our attention to the case where in Eq. (1), A_n and B_n are input-independent and hence fixed matrices: A, B. Further, as common in modern RNNs, we consider without loss in generality[1] the diagonal case $A = \mathrm{diag}(a)$. For one-dimensional input sequences and a final sum operation, if the RNN is initialized with zero memory, the scalar output sequence $(y_n)_{n\geq 0}$ can be computed through a *convolution*:

$$x_n = \mathrm{diag}(a)x_{n-1} + u_n b, \quad y_n = 1^\top x_n \quad \Longrightarrow \quad y_n = (c*u)_n = \sum_{k=0}^{n} c_k u_{n-k} \quad \text{with} \quad c_k = \sum_{s=1}^{S} a_s^k b_s. \tag{2}$$

The purpose of Eq. (2) is to introduce the causal filter defined by the RNN. The task consists in finding potentially complex vectors a, b such that $y_n \approx u_{n-K}$ for all n. This is equivalent to requiring the sequence $(c_k)_{k\geq 0}$ to approximate the *shift-K* filter $d = \delta_K$ (which is a sequence which is zero everywhere except at position K, that is, $d_k = 1_{k=K}$).

In order to assess the approximation of $d = \delta_K$ by c in the form of Eq. (2), we consider the idealized situation of infinite-length random stationary signals $(u_n)_{n\in\mathbb{Z}}$, and consider the expected loss function at time $n = 0$, $\mathbb{E}\big[|(c*u)_0 - (d*u)_0|^2\big]$, where the expectation $\mathbb{E}$ is taken with respect to the distribution of the random sequence (u_n). We thus look at the action of filters c of the form in Eq. (2) on input signals (u_n) defined on $\mathbb{Z}$. By stationarity of (u_n) and the law of large number, this is equivalent to the mean-square-error over the entire sequence:

$$\mathbb{E}\big[|(c*u)_0 - (d*u)_0|^2\big] = \lim_{N\to+\infty} \frac{1}{N+1} \sum_{n=0}^{N} |(c*u)_n - (d*u)_n|^2. \tag{3}$$

We study this loss function for u being the white noise (problem becomes $\min_{a,b} \|c(a,b) - d\|_2^2$), and for simple auto-correlations $\mathbb{E}[u_k \bar{u}_{k'}] = \rho^{|k-k'|}$ for $\rho \in [0,1)$ ($\rho = 0$ corresponding to white noise).

Contributions. We make the following contributions:

1. We provide a lower bound on the best possible value for the shift-K loss above (optimized with respect to a and b) using tools from the approximation of rational functions (Baratchart et al., 2016) and Cauchy matrices (Yang, 2003). For white noise, we obtain the lower bound $1 - \frac{S}{K}$, showing that a large copy lag K leads to an increase in error. This is made more precise with more general ρ's, with the lower bound $\big(1 - \frac{3S}{K}\frac{1}{1-\rho}\big)_+$, showing that a small error can be obtained for autocorrelated input signals.

2. We find a closed-form solution to the shift-K problem close to our lower bound (with matching behavior up to constants, and thus nearly optimal). Our solution allows us to instantiate an

1. This equivalence is often used in linear systems theory (Hespanha, 2018). Let us start from $x_n = Ax_{n-1} + Bu_n$. Over the space of $S \times S$ non-diagonal real matrices A, the subset of those non-diagonalizable in the complex domain has measure zero (Bhatia, 2013). Thus, with arbitrarily small perturbations, $A = Q\mathrm{diag}(a)Q^{-1}$. This implies $Q^{-1}x_n = \mathrm{diag}(a)(Q^{-1}x_{n-1}) + (Q^{-1}B)u_n$. Renaming $x_n \leftarrow Q^{-1}x_n$ and $B \leftarrow Q^{-1}B$ yields a diagonal complex-valued recurrence.

uncertainty principle, providing a clear intuition on resolution/memory tradeoffs (see Fig. 1). This interpretation is specific to the non-discrete nature of our setting and our interest in approximate recall: at a fixed state size S, the optimal filter linear RNNs can produce sets a tradeoff between precision and recall range K. Our uncertainty principle can be seen as follows: let S be fixed, defining the "nature" of our setting (fundamental constant, akin to $\hbar$ in the physical domain), along with the definition of our task. One can write $(1-\text{loss}) \times K \simeq S$. This form is similar to the one in physics: $\Delta x \times \Delta p \geq \hbar/2$, where Δx and Δp represent the uncertainties in position and momentum, and $\hbar$ sets the scale of quantum effects. In addition, our closed-form solution for a in Eq. (2) allows us to motivate from a task-specific memorization perspective the successful S4D-Lin initialization by Gu et al. (2022a) — the simplest linear RNN initialization allowing to solve the most challenging tasks in the long-range arena (Tay et al., 2020).

The loss of our near-optimal solution illustrates the trade-off between recall range (K), memory size (S) and recall precision (i.e., the concentration of the filter c in Eq. (2) around the spike δ_K). Surprisingly, our finding can be formulated as an uncertainty principle:

> *Learning a filter c centered around a large K for a fixed state-size S is relatively easy, yet increasing time-horizon K comes at the expense of resolution (width[2] of the filter).*

As illustrated in Fig. 1, perfect recall is eventually achieved at $S = K$. For $K > S$, the width of the filter around the correct location is proportional to K/S.

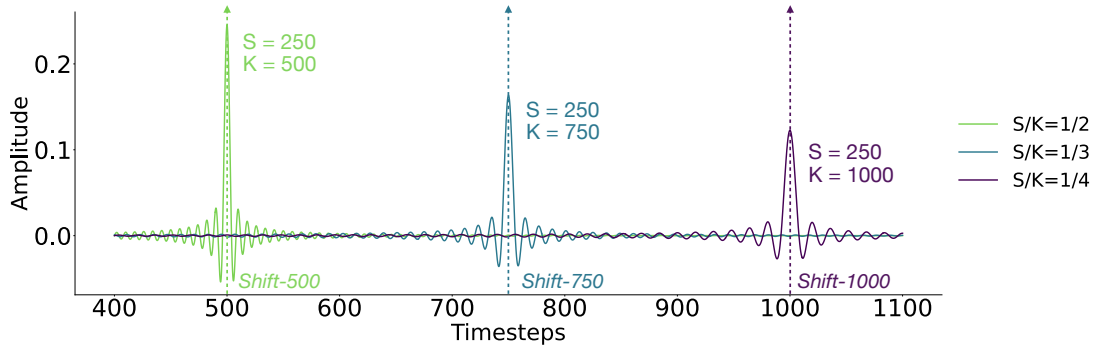

Figure 1: *Learning to shift-K with linear recurrences exhibits an uncertainty principle. For fixed $S = 250$, different values of K induce different performances: the smaller the ratio S/K, the lower the peak of the filter and the larger the width. For a fixed memory size S, increasing the time horizon is feasible, but comes at the expense of resolution. For $K > S$, the width of the filter around the correct location is K/S.*

2. Notation and Main Results

In this paper, we operate in the complex domain as usually done in the literature on SSMs (Gu et al., 2022a; Orvieto et al., 2023). The reason for this choice in the literature is motivated by good

2. For a filter designed to approximate the shift-K, we call width the width at the halfway height of the peak centered in K. The narrower the width, the better the approximation of the shift-K. The width can be interpreted as the resolution of the filter.

performance, increased expressivity guarantees (Orvieto et al., 2024; Ran-Milo et al., 2024), and most of all by the equivalence between dense linear RNNs and diagonal complex-valued RNNs[1].

Time domain. Starting from Eq. (2), given a generic real-valued input $u = (u_n)_{n\in\mathbb{Z}}$, the output $y = (y_n)_{n\in\mathbb{Z}}$ of a linear RNN with parameters $(a, b) \in \mathbb{C}^S \times \mathbb{C}^S$ can be written as $y_n = (c * u)_n := \sum_{k=0}^{\infty} c_k u_{n-k}$, where the convolution kernel $c = (c_k)_{k\in\mathbb{N}}$ is defined by (a, b) as:

$$c_k = \sum_{s=1}^{S} a_s^k b_s, \tag{4}$$

for all $k \in \mathbb{N}$. Let $d = (d_k)_{k\in\mathbb{N}}$ be a second convolution kernel processing the input—the one we would like to approximate with our RNN (that is, $d_k = 1_{k=K}$). One can compute the expected squared norm between outputs of $d * u$ and $c * u$ at only a single $n \in \mathbb{Z}$; as shown in Eq. (3), this is also the mean-squared error over the entire sequence:

$$\mathbb{E}\Big[|(d*u)_n - (c*u)_n|^2\Big] = \mathbb{E}\Big[\Big|\sum_{k=0}^{\infty}(c_k - d_k)u_{n-k}\Big|^2\Big] = \sum_{k,k'=0}^{\infty}(c_k - d_k)(\bar{c}_{k'} - \bar{d}_{k'})\mathbb{E}[u_{n-k}u_{n-k'}].$$

Using stationarity of the signal u, $\mathbb{E}[u_{n-k}u_{n-k'}] = \gamma(k - k')$ only depends on $k - k'$, and, we get our objective function, to be optimized with respect to the RNN parameters (a, b):

$$\mathcal{L}_{\text{time}}(c, d) = \sum_{k,k'=0}^{\infty}(c_k - d_k)(\bar{c}_{k'} - \bar{d}_{k'})\gamma(k - k'), \tag{5}$$

where $\gamma(k - k')$ is the auto-correlation function that captures average temporal dependencies, weighting the contribution of errors based on time step correlations (Brockwell and Davis, 2002). When there is no ambiguity on the filters $(c_k)_{k\in\mathbb{N}}, (d_k)_{k\in\mathbb{N}}$, we will refer to the loss $\mathcal{L}_{\text{time}}(c, d)$ as $\mathcal{L}_{\text{time}}$. We adopt the common choice $\gamma(k) = \rho^{|k|}$ with $\rho \in [0, 1)$, also used recently by Zucchet and Orvieto (2024), where $\rho = 0$ corresponds to uncorrelated white noise, where $\mathcal{L}_{\text{time}}(c, d) = \sum_{k=0}^{\infty} |c_k - d_k|^2$, and $\rho \to 1$ reflects strong temporal dependencies.

Frequency domain. In this work, we aim at approximating the action of the shift-K filter $d = \delta_K := (1_{k=K})_{k\in\mathbb{N}}$. We find it convenient to process the loss above in frequency domain. The discrete-time Fourier transforms and Parseval's theorem allow to write the loss in Eq. (5) as

$$\mathcal{L}_{\text{freq}}(C, D) = \frac{1}{2\pi}\int_{-\pi}^{\pi} |C(e^{i\omega}) - D(e^{i\omega})|^2 \Gamma(e^{i\omega}) d\omega, \tag{6}$$

where $C(e^{i\omega}) = \sum_{s=1}^{S} \frac{b_s}{1-a_s e^{-i\omega}}$ is a rational function of $e^{-i\omega}$, $D(e^{i\omega}) = e^{-iK\omega}$ (Discrete Fourier Transform (DFT) of a shifted Dirac impulse) and $\Gamma(e^{i\omega}) = \frac{1-\rho^2}{|1-\rho e^{-i\omega}|^2}$, thus turning the problem to that of rational approximations on the unit circle (Baratchart et al., 2016). See Appendix B.4 and D.1 for more details. When there is no ambiguity on the Fourier transforms C and D, we will refer to the loss $\mathcal{L}_{\text{freq}}(C, D)$ as $\mathcal{L}_{\text{freq}}$.

Overview. In Section 4, we provide a lower bound on $\mathcal{L}_{\text{time}}$ suggesting that having a small state size does not necessarily imply a short memory capacity; however, the bound also shows that this comes at the cost of a degraded resolution. This provides a first connection with our uncertainty principle in the context of learning shifts with linear models and reveals a fundamental tradeoff between the time horizon of our copy task and the performance of the filter, given a fixed size of the model. Furthermore, we highlight the significant role of data autocorrelation, demonstrating that while linear models struggle to retain white noise, their performance improves substantially when dealing with autocorrelated data, which may better reflect real-world scenarios. In Section 5, we establish our uncertainty principle by deriving a closely matching upper bound. To do this, we consider the loss $\mathcal{L}_{\text{freq}}$ to carefully design a new filter that performs similar to the lower bound, up to a constant factor. This representation, providing meaningful results in practice, gives insights on the behavior of linear RNNs as they implement longer memory.

To summarize, our insights stem from two main results. The first one is a lower bound on the best possible error greater than $1-\frac{S}{K}$ for $\rho=0$ and $\left(1-\frac{3S}{K}\frac{1}{1-\rho}\right)_+$ for $\rho\in[0,1)$ (see Theorems 2 and 3). The second is an upper bound (construction of an explicit filter) that matches the lower-bound up to a constant factor (thus establishing[3] our uncertainty principle), as informally described below in Theorem 1 and illustrated in Fig. D.2.

Theorem 1 (upper bound, informal) *Let S be odd and $T=\frac{S-1}{2}$. The filter defined by Eq. (4) with $a_s=\exp\left(-\frac{\alpha}{K}\right)\exp\left(i\frac{\pi s}{K}\right)$ and $b_s\propto(-1)^s$ for $s\in[\![-T,T]\!]$[4], achieves an approximation error comparable to the lower bound up to a constant factor in the context of white noise data. This is because our filter accurately approximates a shift-K in frequency domain over the window $\left[-\frac{\pi T}{K},\frac{\pi T}{K}\right]$, vanishing outside this range.*

The connection between Theorem 1 and HiPPO initialization (Gu et al., 2023) is presented in Sec. 5.3

3. Related Works

Attention and RNNs. In order to reduce the $O(L^2)$ complexity burden in transformers, techniques such as patching (Dosovitskiy et al., 2021; Pagnoni et al., 2024), gradient checkpointing (Chen et al., 2016), and FlashAttention (Dao et al., 2022; Dao, 2023) become crucial when training and deploying models at scale. Despite this limitation, transformers successfully power most state-of-the-art architectures we use today: beyond large language models (Brown et al., 2020; Team et al., 2024), attention found widespread application in vision (Dosovitskiy et al., 2021), graphs (Ma et al., 2023) and DNA (Dalla-Torre et al., 2024). Nevertheless, the quadratic complexity of attention has remained a pressing limitation, prompting numerous efforts over the years to develop efficient approximations (Wang et al., 2020; Choromanski et al., 2020; Chen et al., 2021; Lee-Thorp et al., 2022) that inevitably bring attention closer to RNNs (Schlag et al., 2021; Katharopoulos et al., 2020). Indeed, more recently, we have witnessed a resurgence of RNNs in state-of-the-art industry-size applications such as language modeling. Sparked by S4 (Gu et al., 2020, 2022b), which surpassed attention on long-range reasoning tasks (Tay et al., 2020), we have seen a drastic increase in

3. Our uncertainty principle, as formulated in the introduction, is first suggested by our lower bound but only formally implied by (and immediately follows from) the combination with our matching upper bound.
4. While in our introduction, for clarity, we considered $s\in[\![1,S]\!]$, hereafter, for ease of notation and to emphasize symmetry in our filter, we reparametrize $s\in[\![-T,T]\!]$ where $T=\frac{S-1}{2}$. The dimension of our hidden state in Eq. (2) remains S.

the usage of RNNs in deep architectures, albeit in a linear form that guarantees both $O(L)$ memory/inference complexity and fast computation on GPUs (Martin and Cundy, 2018) while matching or surpassing transformers on downstream tasks: a prime example are state-space models (SSMs) such as Mamba (Gu and Dao, 2024), along with architectures based on RNNs (De et al., 2024; Peng et al., 2024; Yang et al., 2024a).

Special initialization of SSMs. While in natural language SSMs are relatively robust to initialization, on challenging long-range reasoning or memorization tasks careful initialization is crucial (Gu et al., 2023; Orvieto et al., 2023; Trockman et al., 2024). The used schemes stem from the HiPPO theory by Gu et al. (2020): a special initialization can provably construct features related to the coefficients of optimal polynomial approximations of the input signal. Despite this intriguing connection, already S4 (Gu et al., 2022b), the first SSM, deviates quite significantly from the HiPPO prescription. Latest initialization such as S4D-Lin (Gu et al., 2022a) or the one by Orvieto et al. (2023) are only vaguely related to the HiPPO and present a single non-trivial property: recurrent eigenvalues (a in Eq. (2)) are complex-valued, with coupled phase and magnitude. Our theory provides a formal justification of this choice from a memorization perspective.

Theoretical guarantees for (non)-linear RNNs. Expressivity of standard nonlinear RNNs has been extensively studied from a Turing completeness perspective (Siegelmann and Sontag, 1992; Korsky, 2019). Taking instead the signal processing angle, Hanson and Raginsky (2020) proved that wide enough non-linear RNNs can approximate up to vanishing precision non-linear time-homogeneous systems of differential equations driven by input paths. The argument used here is based on Barron's theorem (Barron, 1993) for approximation of continuous functions with neural networks with one hidden layer. Regarding instead linear RNNs such as Eq. (1), results are more recent and have been mostly driven by deep learning developments. Li et al. (2022) showed that linear time-invariant RNNs (A_n and B_n independent of n, as in this paper) can approximate arbitrary convolution filters as the hidden state size S grows to infinity. Further, Hanson and Raginsky (2019) proved that stacking exponentially (in the sequence length) many temporal convolution filters, chained together with ReLU activations, leads to approximation of arbitrary non-linear filters. Recent works (Orvieto et al., 2024; Wang and Xue, 2024) prove the universality of linear recurrences (one layer) when equipped with a fixed (timestamp independent) point-wise MLP acting across the recurrence output, with intriguing connections to Volterra series (Boyd and Chua, 1985). Finally, expressivity of latest models such as Mamba has been studied by (Cirone et al., 2024). Further, language-specific capabilities of new SSM and RNNs have been studied by Merrill et al. (2024) (state tracking) and Jelassi et al. (2024) (copying).

4. Lower Bound

We aim to establish a lower bound on the approximation error $\mathcal{L}_{\text{time}}(c, d)$ where c has the RNN form in Eq. (4). By deriving this lower bound, we provide a theoretical benchmark for evaluating the effectiveness of linear time-invariant filters in our shift-K task. Importantly, we demonstrate that the derived lower bound depends on the ratio $\frac{S}{K}$, where S represents the hidden dimension and K is the horizon of our copy task.

To gain deeper insights into the performance of these filters, we analyze the approximation error in two scenarios: the case of white noise ($\rho = 0$), and the case of autocorrelated data ($\rho > 0$).

4.1. White Noise

With white noise input, $\mathcal{L}_{\text{time}}(c,d)$ has a simpler squared ℓ_2-norm formulation:

$$\mathcal{L}_{\text{time}}(c,d) = \sum_{k=0}^{+\infty} |c_k - d_k|^2 = 1 + \sum_{k=0}^{+\infty} |c_k|^2 - 2\,\text{Re}\Big(\sum_{k=0}^{+\infty} c_k d_k\Big), \tag{7}$$

where we recall that c has the form in Eq. (4) and that the shift-K filter $d_k = 1_{k=K}$ has norm one. The following theorem shows a lower bound.

Theorem 2 (Lower bound of the approximation error—white noise) *Let S and K be two positive integers. The approximation error $\mathcal{L}_{time}(c,d)$ of the shift-K filter d by a filter c of the form in Eq. (4) is lower bounded by $1 - \frac{S}{K+1}$.*

Proof (Sketch, see full proof in Appendix C.1). Given the form of c as $c_k = \sum_{s=1}^S b_s a_s^k$, the loss in Eq. (7) has an explicit expression by summing geometric series over k, leading to:

$$\mathcal{L}_{\text{time}}(c,d) = 1 + \sum_{s,s'=1}^{S} \frac{b_s \bar{b}_{s'}}{1 - a_s \bar{a}_{s'}} - 2\,\text{Re}\Big(\sum_{s=1}^{S} b_s a_s^K\Big). \tag{8}$$

We thus want to maximize with respect to a_s, b_s, $s \in [\![1, S]\!]$, the following quantity:

$$2\text{Re}\Big(\sum_{s=1}^{S} b_s a_s^K\Big) - \sum_{s,s'=1}^{S} \frac{b_s \bar{b}_{s'}}{1 - a_s \bar{a}_{s'}},$$

which is equal to

$$\langle \bar{b}, a^K \rangle + \langle a^K, \bar{b} \rangle - \langle \bar{b}, C\bar{b} \rangle, \tag{9}$$

where C is an $S \times S$ matrix with entries $C_{ss'} = \dfrac{1}{1 - a_s \bar{a}_{s'}}$, and $\langle \cdot, \cdot \rangle$ is the standard Hermitian product. This is a quadratic form in b, and thus we can maximize with respect to b in closed form, leading to the performance criterion $\mathcal{L}_{\text{time}} = 1 - F_K$, with

$$F_K = \langle a^K, C^{-1} a^K \rangle = \sum_{s,s'=1}^{S} \bar{a}_s^K (C^{-1})_{ss'} a_{s'}^K. \tag{10}$$

This is a function of the a_s's only since we have maximized out the b_s's. This function is rational but has a complicated expression. In order to bound it, we notice that the matrix C has some "displacement structure" similar to Cauchy matrices (Yang, 2003), that is,

$$C - \text{Diag}(a) C\, \text{Diag}(\bar{a}) = 1_S 1_S^\top,$$

which leads to, after some manipulations, to a "closed form" expression for the inverse C^{-1}:

$$(C^{-1})_{ss'} \Big(\frac{1}{\bar{a}_s a_{s'}} - 1\Big) = u_s \bar{v}_{s'},$$

with $u = C^{-1}1_S$ and $v = \operatorname{Diag}(\bar{a})^{-1}C^{-1}\operatorname{Diag}(a)^{-1}1_S \propto u$. Moreover, the vector u happens to have a simple characterization through rational functions as

$$\sum_{s=1}^{S} \frac{u_s}{1 - z\bar{a}_s} = 1 - \prod_{s=1}^{S} \bar{a}_s \prod_{s=1}^{S} \frac{a_s - z}{1 - z\bar{a}_s}.$$

This allows to characterize the Fourier series of F_K and get an explicit bound using properties from rational approximations on the unit circle (Baratchart et al., 2016). See details in Appendix C.1. ■

The lower bound established in Theorem 2 demonstrates that the approximation error remains close to 1 when the ratio S/K is small. This highlights the inherent difficulty of approximating shift-K filter using linear recurrences in this regime. Nevertheless, by increasing the dimension of the parameters S, with fixed K, we can hope to achieve a better loss. This shows a fundamental tradeoff in the linear model's ability to solve the copy task, in the context of white noise, connected to our uncertainty principle. Allowing auto-correlated signals gives a finer picture, this is explored in the next subsection.

4.2. Autoregressive Autocorrelation

In this context, we consider a non-zero correlation factor defined as $\gamma(k) = \rho^{|k|}$ to account for the temporal structure in the data. This approach with $\rho > 0$ simulates situations with real-life data, whose autocorrelation is often modeled this way (Brockwell and Davis, 2002). The loss function writes in this case:

$$\mathcal{L}_{\text{time}}(c, d) = \sum_{k,k'=0}^{+\infty} (c_k - d_k)(\bar{c}_{k'} - \bar{d}_{k'})\rho^{|k-k'|}, \tag{11}$$

where c_k is defined as in Eq. (2), and (d_k) is given by $d_k = 1_{k=K}$. The following theorem extends Theorem 2 to all ρ's. We use the notation $(y)_+ = \max(y, 0)$.

Theorem 3 (Lower bound of the approximation error—auto-correlated noise) *Let S and K be two integers. The approximation error $\mathcal{L}_{time}(c, d)$ of the shift-K filter d by a filter c of the form in Eq. (4) is lower bounded by, for the autoregressive autocorrelation, $\left(1 - \frac{S}{K}\frac{3}{1-\rho}\right)_+$.*

Proof (Sketch, see full proof in Appendix C.2). Let (c_k) be a linear-time filter such that $c_k = \sum_{s=1}^{S} b_s a_s^k$, and let (d_k) be defined as $d_k = 1_{k=K}$. Let $w_s = b_s a_s/(a_s - \rho)$, for $s \in [\![1, S]\!]$. We can compute $\mathcal{L}_{\text{time}}(c, d)$ in closed form by explicit summation leading to

$$\mathcal{L}_{\text{time}}(c, d) = 1 - 2(1 - \rho^2)\operatorname{Re}\left(\sum_{s=1}^{S+1} \frac{w_s a_s^K}{1 - a_s\rho}\right) + (1 - \rho^2)\sum_{s,s'=1}^{S+1} \frac{w_s \bar{w}_{s'}}{1 - a_s\bar{a}_{s'}},$$

where we have used the specificity of the auto-correlation function to create a new pole, defined as $a_{S+1} = \rho$ and the constraint $\sum_{s=1}^{S+1} w_s a_s^{-1} = 0$ holds for some vector w to be optimized. The minimum with respect to w with the constraint is greater than the unconstrained minimizer, equal to (using the fact that this is a quadratic problem): $H_K = 1 - (1-\rho^2)\sum_{s,s'=1}^{S+1} \frac{\bar{a}_s^K}{1 - \bar{a}_s\rho}\frac{a_{s'}^K}{1 - a_{s'}\rho}(C^{-1})_{ss'}$,

where $C_{ss'} = \frac{1}{1-a_s\bar{a}_{s'}}$ is a matrix of a similar form as in the proof of Theorem 2. The proof then follows similarly by using explicit expressions of matrix inverses. ■

Therefore, in the autocorrelated case, $\mathcal{L}_{\text{time}}$ exhibits a lower bound that depends on the ratio $\frac{S}{K}$. The error diminishes as ρ approaches 1, indicating that memorization may become more effective in the limit of strong autocorrelation. This behavior also suggests that memorization performance is intrinsically linked to the spectral characteristics of the data. Specifically, linear filters are incapable of precisely solving the copy task for time horizons K larger than S in the presence of poorly correlated data, as in white noise. However, reducing the spectral domain, by imposing autocorrelation in the data, concentrates the signal's energy within specific frequency bands and significantly improves performance. Next, we further investigate this behavior by designing an explicit filter.

5. Upper Bound

In this section, we complement the previous results, which showed that the lower bound of $\mathcal{L}_{\text{time}}(c, d)$ for the copy task using linear systems depends on the ratio $\frac{S}{K}$. We present a closed-form parameterization of the filter that achieves a similar performance differing only by a constant factor. This parameterization serves as an upper bound on the achievable approximation accuracy of linear RNNs on the shift-K task. In particular, we provide explicit expressions for the learnable parameters a_s and b_s, accompanied by a theoretical analysis of their performance. This formulation establishes a theoretical upper limit for the smallest attainable error and highlights the behavior of a "good" filter. Since this upper limit also depends on the ratio $\frac{S}{K}$, we can infer conclusions about the optimal behavior of the filter, establishing our uncertainty principle and particularly its relation to the spectral width of the data. To present our intuitions and results, we convert the time-domain loss $\mathcal{L}_{\text{time}}$ in Eq. (5) into its frequency-domain counterpart $\mathcal{L}_{\text{freq}}$ in Eq. (6).

5.1. Parameterization of the Filter

Here, we introduce a new filter inspired by the frequency representation of the problem, which achieves good results on the copy task. $\mathcal{L}_{\text{freq}}(C, D)$ in Eq. (6) suggests that a good filter (c_k), denoted by $C(e^{i\omega})$ in the frequency domain, should approximate as best as possible the complex exponential $e^{-iK\omega}$ for $\omega \in [-\pi, \pi]$. Additionally, Sec. 4 demonstrated that it is not possible to achieve an error smaller than $1 - \frac{S}{K}$ when solving the copy task using linear models on white noise. Based on these results, we consider a greedy approach, where each individual term $\frac{b_s}{1-a_s e^{-i\omega}}$ of $C(e^{i\omega})$ captures a single oscillation of the complex exponential. This should result in an error that depends on $\frac{S}{K}$, and motivates the following representation.

Parameterization of the a_s. The parameters a_s govern the filter's ability to refer to earlier time steps, making them the most critical components of the recurrence. To provide finer control around a part of the complex unit circle, we employ an exponential parameterization, for S odd:

$$a_s = \exp\left(-\frac{\alpha}{K}\right)\exp\left(i\frac{\pi s}{K}\right), \quad s \in [\![-T, T]\!], \text{ with } S = 2T + 1, \tag{12}$$

where $0 < \alpha \ll K$ so that $|a_s| < 1$ (for stability of the system). The a_s's have a constant modulus defined by the parameter α, while their phases are uniformly distributed around the unit circle, separated by an angular distance of $\frac{\pi}{K}$.

Remark: This representation in Eq. (12) ensures that the majority of the weight in each individual term $\frac{b_s}{1-a_s e^{-i\omega}}$ is concentrated around the frequency $\frac{\pi s}{K}$, effectively capturing a single oscillation of the complex exponential. Our goal is to fit S oscillations of $e^{-iK\omega}$, which would result in a loss proportional to $\frac{S}{K}$.

Parameterization of the b_s. We can obtain the b_s's by an approximate minimization as follows:

Lemma 4 *Let the parameters a_s of the filter be defined as in Eq. (12), where α is a positive real number. The asymptotic optimal parameters (when $K \to +\infty$) b_s that minimize $\mathcal{L}_{freq}$ are given by:*

$$b_s = \frac{e^{-\alpha}(e^{2\alpha} - e^{-2\alpha})}{2K}(-1)^s, \quad s \in [\![-T, T]\!]. \tag{13}$$

Proof As highlighted in Eq. (9), the approximation error, defined in terms of a_s and b_s, is quadratic and convex with respect to b_s. Hence, the optimal solution is given by:

$$b = C^{-1}\bar{a}^K, \tag{14}$$

where the matrix C is defined as $C_{ss'} = \frac{1}{1-a_s\bar{a}_{s'}}$. Using asymptotic expansions for large K, the eigenvector of C corresponds to $z = ((-1)^s)_s \in \mathbb{R}^S$ associated to the eigenvalue $\frac{2}{e^{2\alpha}-e^{-2\alpha}}$, yielding the result, thanks to the asymptotic expansion of (a_s). See full proof in Appendix D.2. ■

Note that (a_s) and (b_s) form complex conjugate pairs; this allows to obtain a real filter.

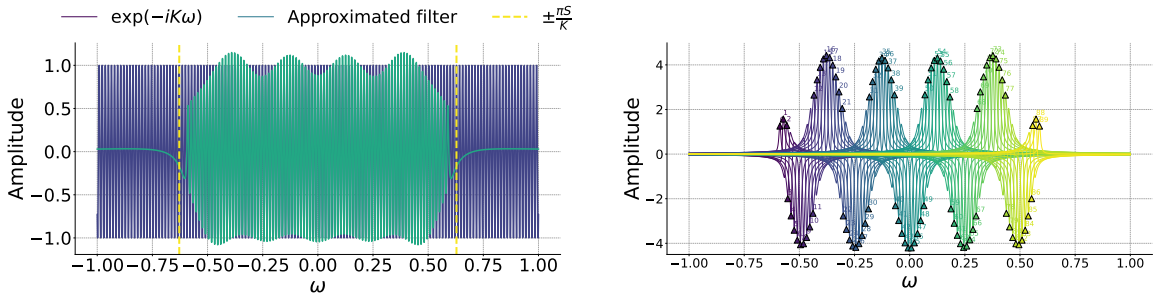

Figure 2: *Poor performance of the filter for white noise data is due to its approximation of the complex exponential over a limited frequency window of size $\frac{\pi S}{K}$. Left: The target filter $\exp(-iK\omega)$ (blue) for $K = 450$, and the approximated filter using linear recurrences (green) for $S = 90$. The approximation is reasonably accurate within the frequency window of size $\frac{\pi S}{K}$, indicated by the dashed yellow lines. Outside this window, the filter is zero, demonstrating the inability of filters based on linear recurrences to perfectly memorize long-range data with broad spectra. Right: Contributions from all individual terms $\frac{b_s}{1-a_s e^{-i\omega}}$ for $s \in [\![-45, 45]\!]$. Each individual term captures one oscillation of the complex exponential, making their contributions highly localized. This design reflects the structure of the filter's parameters.*

Lemma 5 *Let K and S be two large integers such that $S \ll K$. Consider the parameters $(a_s)_{s\in[\![-T,T]\!]}$ from Eq. (12) and $(b_s)_{s\in[\![-T,T]\!]}$ from Eq. (13). Then the spectral representation of the filter is given by*

$$C(e^{i\omega}) = \sum_{s=-T}^{T} \frac{b_s}{1 - a_s e^{-i\omega}} = \sum_{s=-T}^{T} \frac{(-1)^s e^{-\alpha}(e^{2\alpha} - e^{-2\alpha})}{2K\left(1 - e^{-\frac{\alpha}{K}} e^{i(\frac{\pi s}{K} - \omega)}\right)}.$$

In the time domain, this filter is approximately equivalent to a shifted sine cardinal, see Fig. D.1, highlighting its inherent smoothness and symmetry. The positions of its parameters on the complex plane are strongly influenced by the ratio S/K, which corresponds to the horizon of the copy task relative to the order of the linear recurrence. This dependency captures the trade-off between long-term memory and the granularity of the recurrence structure.

Theorem 6 (Upper bound of the error) *Consider (c_k) the filter defined in Lemma 5, and $(d_k) = (1_{k=K})$ the shift-K filter. Then, for $S, K \to +\infty$ with $S/K \to 0$, we have*

$$\mathcal{L}_{\textit{time}}(c, d) \sim 1 - \frac{e^{-2\alpha}(e^{2\alpha} - e^{-2\alpha})}{2} \times \frac{S}{K}. \tag{15}$$

Remark: Note that the relation of $\mathcal{L}_{\text{time}}$ to α in Theorem 6 incites to take α very large. Nevertheless, this would cause the value of the norm of b_s in Eq. (13) to explode. In practice, we always chose $\alpha = 1$, a choice which also leads to a closer match with HiPPO initialization (see Section 5.3). We further note that our bound on the b_s in Lemma 4 is strictly related to the discussion around benefits of complex parametrization by Ran-Milo et al. (2024): as a_s' magnitude increases at equal phase, approximating arbitrary filters requires exploding coefficients (cf. their Theorem 2).

When $S \ll K$, the term $\frac{S}{K}$ becomes very small, causing the error in Eq. (15) to approach 1. We recover the result of Section 4, obtaining a loss that is similar up to a constant factor. This approximation error for our filter serves as an upper bound for the approximation of shift-K filter by linear recurrences.

This behavior can be attributed to the inherent properties of the filter, as illustrated in Fig. 2. The filter approximates reasonably well all the oscillations of $e^{-iK\omega}$ over the frequency window $[\frac{-\pi T}{K}, \frac{\pi T}{K}]$ and vanishes outside this window. Each individual term of the partial fraction decomposition is responsible for capturing a peak of the complex exponential. Therefore, data exhibiting large frequency spectrum like white noise cannot be memorized properly, explaining the poor performance of the filter on our copy task. This is how we designed it, to catch up with the lower bound. This is made precise in the following theorem.

Theorem 7 *For α real and positive and $\Omega = \frac{K\omega}{\pi}$,*

$$C(e^{i\omega}) = \sum_{s=-T}^{T} \frac{(-1)^u e^{-\alpha}(e^{2\alpha} - e^{-2\alpha})}{2K\big(1 - e^{-\frac{\alpha}{K}} e^{i(\frac{\pi u}{K} - \frac{\pi\Omega}{K})}\big)} \underset{\substack{S\to+\infty \\ S/K\to 0}}{\sim} \begin{cases} \frac{e^{-\alpha}(e^{2\alpha}-e^{-2\alpha})}{2} \times \frac{i(-1)^{T+1}\times 2\lfloor\Omega\rfloor}{2\pi(\lfloor\Omega\rfloor - T)(\lfloor\Omega\rfloor + T)} & \text{if } |\Omega| > T, \\ \frac{e^{-\alpha}(e^{2\alpha}-e^{-2\alpha})}{e^{\alpha}e^{i\pi\Omega} - e^{-\alpha}e^{-i\pi\Omega}} & \text{if } |\Omega| < T. \end{cases}$$

In particular, we obtain that $C(e^{i\omega})$ tends to 0 if Ω is out of the window $[-T, T]$, while inside the window, we obtain some oscillations around 1 whose magnitude depends on Ω (see Fig. 2).

When both S and K tend to infinity, K being significantly larger than S, the filter converges to a rectangular window function on the frequency interval $[-\frac{\pi T}{K}, \frac{\pi T}{K}]$, taking the value oscillating around 1 within this interval and 0 outside it. See an illustration in Fig. D.2. This limitation highlights why the filter performs poorly on white noise, as the uniform spectral density of white noise extends far beyond this narrow frequency window. Conversely, as the frequency window narrows (determined by the autocorrelation $\Gamma(e^{i\omega})$), the filter becomes better aligned with the target response, leading to improved performance.

5.2. Performance in the Autocorrelated case

The autocorrelation factor $\Gamma(e^{i\omega})$ has a narrowing effect in the frequency domain, reducing the bandwidth of frequencies over which $\mathcal{L}_{\text{freq}}(C, D)$ is evaluated. See Appendix B.4 for more details. Since our filter is specifically designed to accurately approximate the oscillations of the complex exponential over a frequency window of size $\frac{\pi S}{K}$, it follows logically that the loss decreases as the autocorrelation factor ρ approaches 1.

We can compute the loss of the idealized filter in the frequency domain $1_{|\omega| \leqslant \frac{2\pi S}{K}} e^{-iK\omega}$:

$$1 - \frac{1}{2\pi} \int_{-\frac{\pi S}{2K}}^{\frac{\pi S}{2K}} \Gamma(e^{i\omega}) = 1 - \frac{2}{\pi} \arctan\left(\frac{1+\rho}{1-\rho} \tan \frac{\pi S}{K}\right) \sim 1 - \frac{2}{\pi} \frac{1+\rho}{1-\rho} \frac{\pi S}{K},$$

when S/K goes to zero, which is corresponding to the lower bound in Theorem 3.

5.3. Connection with HiPPO Initialization

HiPPO theory (Gu et al., 2020) was crucial for the development of modern recurrent models. The main result of this theory is that linear continuous-time ODEs (linear RNNs, when discretized) can perform online compression of smooth input signals by storing projection onto an S-dimensional (S is the dimension of x in Eq. (1)) polynomial basis. Starting from a *dense* HiPPO-inspired A matrix, Gupta et al. (2022) first proposed to initialize the A matrix in Eq. (1) as the diagonal part of its "diagonal plus low rank" approximation. Gu et al. (2022a) additionally simplified this expression conjecturing (see their Conjecture 5) a simplified closed-form solution that works well in practice: $a_s = \exp\left(-\frac{\Delta}{2}\right) \exp\left(i\pi s \Delta\right)$ (S4D-Lin). The parameter Δ here is a learnable coefficient resulting from discretization of the approximate HiPPO system. There is no theory indicating how to initialize this coefficient, though further studies (Gu et al., 2023) suggest initializing near $1/K$ (K being the sequence length) yields good results. Our theory gives grounding to this initialization practice, as well as to the S4D-Lin approximation, using a different viewpoint: our closed-form approximation for the filter δ_K in Eq. (12) is $a_s = \exp\left(-\frac{\alpha}{K}\right) \exp\left(i\frac{\pi s}{K}\right)$, with $s \in [\![-T, T]\!]$ and $S = 2T + 1$. According to Lemma 4, for numerical stability α should be a small scalar. For $\alpha = 1/2$, we get $\exp\left(-\frac{1}{2K}\right) \exp\left(i\frac{\pi s}{K}\right)$, i.e., exactly S4DLin with $\Delta = 1/K$. We believe this connection to be a piece of evidence motivating correlation between magnitude and phase in modern variants of S4.

6. Experiments

We conclude our analysis of the copy task problem with numerical experiments illustrating the potential benefits of initialization of the parameters of linear models using representations from Eq. (12) and Eq. (13). We consider the following task: Given a dataset of autoregressive sequences $U = (u_1, u_2, \ldots, u_N)$ of length N, generated as: $u_n = \rho u_{n-1} + \varepsilon_n$, $\varepsilon_n \sim \mathcal{N}(0, \sigma^2)$, $u_1 \sim \mathcal{U}(0, 1)$, where $\rho \in [0, 1)$ is the correlation factor and $\sigma^2 = 1 - \rho^2$, the task is for the model to restore the output $Y = u_{t^*}$ for a fixed index t^* in the sequence. This boils down to learning a shift of $K^* = N - t^*$ with a finite number of samples. We use an input-independent linear model as in Eq. (2), where the vector a is initialized with Eq. (12), and vector b with Eq. (13). In an initial set of experiments, we demonstrate the advantages of initializing with linearly-spaced phases for tasks with a large horizon, compared to random initialization with phases sampled across the entire disk. Subsequently, we assess the robustness of gridded initialization to variations in K_{init}, highlighting

its flexibility—a crucial property for real-world applications where the task horizon is typically uncertain. See results in Fig. 3, where (left plot) we see that our filter yields increasing benefits as ρ grows compared to initialization with random phases, and (right plot) the optimal performance is obtained with the correct $K_{\text{init}} = K^*$ for initialization, yet the method remains robust even when initialized far from the optimal value. In all experiments, we took $S = 128$.

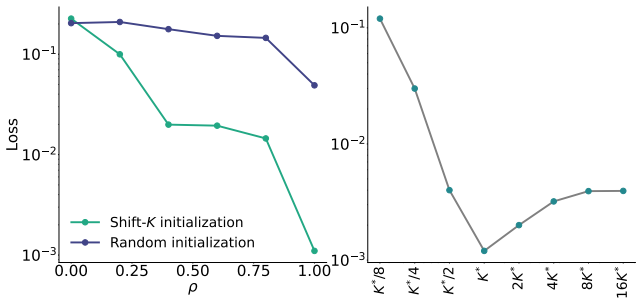

Figure 3: Initialization with regularly spaced phases enhances robustness and outperforms random initialization near the unit disk. Left: For $N = 1500$ and $t^* = 200$, initialization using our filter defined in Eq. (12) and Eq. (13). Right: For $N = 2250$ and $t^* = 250$, the task consists of learning a shift-K filter with $K^* = 2000$. Here, $\rho = 0.7$.

7. Conclusion

We demonstrated that the performance of linear models on a simplified copy task, when applied to stationary input sequences, depends on the ratio $\frac{S}{K}$, where S denotes the size of the state space, and K represents the lag of the copy task. This analysis revealed a form of uncertainty principle governing the resolution of our copy task with linear recurrences. To explain this trade-off between memory capacity and filter performance, we introduced a new filter that achieves the same performance on our copy task up to constants. This representation offers fresh insights into the filter's behavior, particularly in the spectral domain. As highlighted by Orvieto et al. (2023) and further elaborated by Gu et al. (2022b), the initialization of the recurrence matrix's entries plays a crucial role in achieving high performance. Specifically, these studies constrain both the magnitudes and phases of the diagonal entries to depend on $\frac{1}{\Delta}$, where Δ has an order of magnitude similar to the sequence length. In this paper, we aim to provide an explanation for the efficacy of this specific initialization: it arises from the linear model's endeavor to retain certain elements of the sequence, thereby approximating the shifted Dirac function.

Acknowledgments

This work has received support from the French government, managed by the National Research Agency, under the France 2030 program with the reference "PR[AI]RIE-PSAI" (ANR-23-IACL-0008). Antonio Orvieto is supported by the Hector Foundation.

References

Ameen Ali, Itamar Zimerman, and Lior Wolf. The hidden attention of Mamba models. *arXiv preprint arXiv:2403.01590*, 2024.

Simran Arora, Sabri Eyuboglu, Aman Timalsina, Isys Johnson, Michael Poli, James Zou, Atri Rudra, and Christopher Re. Zoology: Measuring and improving recall in efficient language models. In *International Conference on Learning Representations*, 2023.

Simran Arora, Sabri Eyuboglu, Michael Zhang, Aman Timalsina, Silas Alberti, Dylan Zinsley, James Zou, Atri Rudra, and Christopher Ré. Simple linear attention language models balance the recall-throughput tradeoff. In *International Conference on Machine Learning*, 2024.

Laurent Baratchart, Sylvain Chevillard, and Tao Qian. Minimax principle and lower bounds in H2-rational approximation. *Journal of Approximation Theory*, 206:17–47, 2016.

Andrew R. Barron. Universal approximation bounds for superpositions of a sigmoidal function. *IEEE Transactions on Information Theory*, 39(3):930–945, 1993.

Yoshua Bengio, Patrice Simard, and Paolo Frasconi. Learning long-term dependencies with gradient descent is difficult. *IEEE Transactions on Neural Networks*, 5(2):157–166, 1994.

Rajendra Bhatia. *Matrix Analysis*, volume 169. Springer Science & Business Media, 2013.

Satwik Bhattamishra, Michael Hahn, Phil Blunsom, and Varun Kanade. Separations in the representational capabilities of transformers and recurrent architectures. In *Advances in Neural Information Processing Systems*, 2024.

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

Peter J. Brockwell and Richard A. Davis. *Introduction to Time Series and Forecasting*. Springer, 2002.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 2020.

D. Calvetti and L. Reichel. On the solution of Cauchy systems of equations. *Electronic Transactions on Numerical Analysis*, 4:125–137, 1996.

Tianqi Chen, Bing Xu, Chiyuan Zhang, and Carlos Guestrin. Training deep nets with sublinear memory cost. *arXiv preprint arXiv:1604.06174*, 2016.

Yifan Chen, Qi Zeng, Heng Ji, and Yun Yang. Skyformer: Remodel self-attention with Gaussian kernel and Nystrom method. *Advances in Neural Information Processing Systems*, 2021.

Krzysztof Marcin Choromanski, Valerii Likhosherstov, David Dohan, Xingyou Song, Andreea Gane, Tamas Sarlos, Peter Hawkins, Jared Quincy Davis, Afroz Mohiuddin, Lukasz Kaiser, et al. Rethinking attention with performers. In *International Conference on Learning Representations*, 2020.

Nicola Muca Cirone, Antonio Orvieto, Benjamin Walker, Cristopher Salvi, and Terry Lyons. Theoretical foundations of deep selective state-space models. In *Advances in Neural Information Processing Systems*, 2024.

Hugo Dalla-Torre, Liam Gonzalez, Javier Mendoza-Revilla, Nicolas Lopez Carranza, Adam Henryk Grzywaczewski, Francesco Oteri, Christian Dallago, Evan Trop, Bernardo P. de Almeida, Hassan Sirelkhatim, et al. Nucleotide transformer: building and evaluating robust foundation models for human genomics. *Nature Methods*, 21(11):1–11, 2024.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning. *arXiv preprint arXiv:2307.08691*, 2023.

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

Tri Dao, Dan Fu, Stefano Ermon, Atri Rudra, and Christopher Ré. Flashattention: Fast and memory-efficient exact attention with io-awareness. *Advances in Neural Information Processing Systems*, 2022.

Soham De, Samuel L. Smith, Anushan Fernando, Aleksandar Botev, George Cristian-Muraru, Albert Gu, Ruba Haroun, Leonard Berrada, Yutian Chen, Srivatsan Srinivasan, et al. Griffin: Mixing gated linear recurrences with local attention for efficient language models. *arXiv preprint arXiv:2402.19427*, 2024.

Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. An image is worth 16x16 words: Transformers for image recognition at scale. In *International Conference on Learning Representations*, 2021.

Jeffrey L. Elman. Finding structure in time. *Cognitive Science*, 14(2):179–211, 1990.

Karan Goel, Albert Gu, Chris Donahue, and Christopher Ré. It's raw! audio generation with state-space models. *International Conference on Machine Learning*, 2022.

Robert M. Gray. Toeplitz and circulant matrices: A review. *Foundations and Trends in Communications and Information Theory*, 2(3):155–239, 2006.

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

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

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

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

Albert Gu, Isys Johnson, Aman Timalsina, Atri Rudra, and Christopher Re. How to train your HIPPO: State space models with generalized orthogonal basis projections. In *International Conference on Learning Representations*, 2023.

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

Joshua Hanson and Maxim Raginsky. Universal approximation of input-output maps by temporal convolutional nets. *Advances in Neural Information Processing Systems*, 2019.

Joshua Hanson and Maxim Raginsky. Universal simulation of stable dynamical systems by recurrent neural nets. In *Learning for Dynamics and Control*, 2020.

Joao P. Hespanha. *Linear Systems Theory*. Princeton University Press, 2018.

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

Herbert Jaeger. The "echo state" approach to analysing and training recurrent neural networks-with an erratum note. *German National Research Center for Information Technology GMD Technical Report*, 2001.

Samy Jelassi, David Brandfonbrener, and Sham M. Kakade. Repeat after me: Transformers are better than state space models at copying. In *International Conference on Machine Learning*, 2024.

Angelos Katharopoulos, Apoorv Vyas, Nikolaos Pappas, and François Fleuret. Transformers are RNNs: Fast autoregressive transformers with linear attention. In *International Conference on Machine Learning*, 2020.

Samuel A. Korsky. *On the Computational Power of RNNs*. PhD thesis, Massachusetts Institute of Technology, 2019.

James Lee-Thorp, Joshua Ainslie, Ilya Eckstein, and Santiago Ontanon. Fnet: Mixing tokens with Fourier transforms. In *North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 2022.

Kunchang Li, Xinhao Li, Yi Wang, Yinan He, Yali Wang, Limin Wang, and Yu Qiao. Videomamba: State space model for efficient video understanding. In *European Conference on Computer Vision*, 2025.

Zhong Li, Jiequn Han, E Weinan, and Qianxiao Li. Approximation and optimization theory for linear continuous-time recurrent neural networks. *Journal of Machine Learning Research*, 23 (42):1–85, 2022.

Dingkang Liang, Xin Zhou, Wei Xu, Xingkui Zhu, Zhikang Zou, Xiaoqing Ye, Xiao Tan, and Xiang Bai. PointMamba: A simple state space model for point cloud analysis. In *Advances in Neural Information Processing Systems*, 2024.

Yue Liu, Yunjie Tian, Yuzhong Zhao, Hongtian Yu, Lingxi Xie, Yaowei Wang, Qixiang Ye, Jianbin Jiao, and Yunfan Liu. VMamba: Visual state space model. In *Advances in Neural Information Processing Systems*, 2024.

Chris Lu, Yannick Schroecker, Albert Gu, Emilio Parisotto, Jakob Foerster, Satinder Singh, and Feryal Behbahani. Structured state space models for in-context reinforcement learning. In *Advances in Neural Information Processing Systems*, 2023.

Liheng Ma, Chen Lin, Derek Lim, Adriana Romero-Soriano, Puneet K Dokania, Mark Coates, Philip Torr, and Ser-Nam Lim. Graph inductive biases in transformers without message passing. In *International Conference on Machine Learning*, 2023.

Eric Martin and Chris Cundy. Parallelizing linear recurrent neural nets over sequence length. In *International Conference on Learning Representations*, 2018.

William Merrill, Jackson Petty, and Ashish Sabharwal. The illusion of state in state-space models. In *International Conference on Machine Learning*, 2024.

Eric Nguyen, Michael Poli, Matthew G. Durrant, Brian Kang, Dhruva Katrekar, David B. Li, Liam J. Bartie, Armin W. Thomas, Samuel H. King, Garyk Brixi, et al. Sequence modeling and design from molecular to genome scale with evo. *Science*, 386(6723), 2024.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, et al. In-context learning and induction heads. *arXiv preprint arXiv:2209.11895*, 2022.

Alan V. Oppenheim, Alan S. Willsky, and S. Hamid Nawab. *Signals & Systems (2nd ed.)*. Prentice-Hall, 1996.

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

Antonio Orvieto, Soham De, Caglar Gulcehre, Razvan Pascanu, and Samuel L. Smith. Universality of linear recurrences followed by non-linear projections: Finite-width guarantees and benefits of complex eigenvalues. In *International Conference on Machine Learning*, 2024.

Artidoro Pagnoni, Ram Pasunuru, Pedro Rodriguez, John Nguyen, Benjamin Muller, Margaret Li, Chunting Zhou, Lili Yu, Jason Weston, Luke Zettlemoyer, et al. Byte latent transformer: Patches scale better than tokens. *arXiv preprint arXiv:2412.09871*, 2024.

Razvan Pascanu, Tomas Mikolov, and Yoshua Bengio. On the difficulty of training recurrent neural networks. In *International Conference on Machine Learning*, 2013.

Bo Peng, Daniel Goldstein, Quentin Anthony, Alon Albalak, Eric Alcaide, Stella Biderman, Eugene Cheah, Xingjian Du, Teddy Ferdinan, Haowen Hou, et al. Eagle and Finch: RWKV with matrix-valued states and dynamic recurrence. *arXiv preprint arXiv:2404.05892*, 2024.

Zhen Qin, Songlin Yang, Weixuan Sun, Xuyang Shen, Dong Li, Weigao Sun, and Yiran Zhong. HGRN2: Gated linear RNNs with state expansion. *arXiv preprint arXiv:2404.07904*, 2024.

Yuval Ran-Milo, Eden Lumbroso, Edo Cohen-Karlik, Raja Giryes, Amir Globerson, and Nadav Cohen. Provable benefits of complex parameterizations for structured state space models. *arXiv preprint arXiv:2410.14067*, 2024.

David E. Rumelhart, Paul Smolensky, James L. McClelland, and G. Hinton. Sequential thought processes in pdp models. *Parallel Distributed Processing: Explorations in the Microstructures of Cognition*, 2:3–57, 1986.

Imanol Schlag, Kazuki Irie, and Jürgen Schmidhuber. Linear transformers are secretly fast weight programmers. In *International Conference on Machine Learning*, 2021.

Valery Serov. *Fourier series, Fourier Transform and their Applications to Mathematical Physics*, volume 197. Springer, 2017.

Jerome Sieber, Carmen Amo Alonso, Alexandre Didier, Melanie Zeilinger, and Antonio Orvieto. Understanding the differences in foundation models: Attention, state space models, and recurrent neural networks. In *Advances in Neural Information Processing Systems*, 2024.

Hava T. Siegelmann and Eduardo D. Sontag. On the computational power of neural nets. In *Proceedings of the fifth Annual Workshop on Computational Learning Theory*, 1992.

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

Yi Tay, Mostafa Dehghani, Samira Abnar, Yikang Shen, Dara Bahri, Philip Pham, Jinfeng Rao, Liu Yang, Sebastian Ruder, and Donald Metzler. Long range arena: A benchmark for efficient transformers. In *International Conference on Learning Representations*, 2020.

Gemini Team, Petko Georgiev, Ving Ian Lei, Ryan Burnell, Libin Bai, Anmol Gulati, Garrett Tanzer, Damien Vincent, Zhufeng Pan, Shibo Wang, et al. Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. *arXiv preprint arXiv:2403.05530*, 2024.

Asher Trockman, Hrayr Harutyunyan, J Zico Kolter, Sanjiv Kumar, and Srinadh Bhojanapalli. Mimetic initialization helps state space models learn to recall. *arXiv preprint arXiv:2410.11135*, 2024.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

Roger Waleffe, Wonmin Byeon, Duncan Riach, Brandon Norick, Vijay Korthikanti, Tri Dao, Albert Gu, Ali Hatamizadeh, Sudhakar Singh, Deepak Narayanan, et al. An empirical study of Mamba-based language models. *arXiv preprint arXiv:2406.07887*, 2024.

Shida Wang and Beichen Xue. State-space models with layer-wise nonlinearity are universal approximators with exponential decaying memory. In *Advances in Neural Information Processing Systems*, 2024.

Sinong Wang, Belinda Z. Li, Madian Khabsa, Han Fang, and Hao Ma. Linformer: Self-attention with linear complexity. *arXiv preprint arXiv:2006.04768*, 2020.

Zhaohu Xing, Tian Ye, Yijun Yang, Guang Liu, and Lei Zhu. Segmamba: Long-range sequential modeling Mamba for 3D medical image segmentation. In *International Conference on Medical Image Computing and Computer-Assisted Intervention*, 2024.

Songlin Yang, Bailin Wang, Yikang Shen, Rameswar Panda, and Yoon Kim. Gated linear attention transformers with hardware-efficient training. In *International Conference on Machine Learning*, 2024a.

Songlin Yang, Bailin Wang, Yu Zhang, Yikang Shen, and Yoon Kim. Parallelizing linear transformers with the delta rule over sequence length. *arXiv preprint arXiv:2406.06484*, 2024b.

Zheng-Hong Yang. Generalized confluent Cauchy–Vandermonde matrices: displacement structures, inversion formulas and tangential interpolations. *Journal of Computational and Applied Mathematics*, 154(2):355–371, 2003.

Nicolas Zucchet and Antonio Orvieto. Recurrent neural networks: vanishing and exploding gradients are not the end of the story. In *Advances in Neural Information Processing Systems*, 2024.

Nicolas Zucchet, Robert Meier, Simon Schug, Asier Mujika, and João Sacramento. Online learning of long-range dependencies. In *Advances in Neural Information Processing Systems*, 2023.

In this Appendix, we provide a detailed proof for all our theoretical results. We start in Appendix A with an equivalence of various representations of linear RNNs, then in Appendix B with a review of fundamentals of signal processing.

Appendix Contents.

A. Recurrent Neural Networks and Diagonal forms

Linear recurrent networks such as SSMs, in their simplest form, are causal models acting on a d dimensional input sequence with L elements $U \in \mathbb{R}^{d\times L}$, producing an output sequence $Y \in \mathbb{R}^{d\times L}$ through a filtering process parametrized by variables $A \in \mathbb{R}^{N\times N}$, $B \in \mathbb{R}^{N\times d}, P \in \mathbb{R}^{d\times N}$. Let $U_n \in \mathbb{R}^d$ denote the n-th timestamp data contained in U, a linear RNN processes the inputs as follows (Gu et al., 2022a; Orvieto et al., 2023)

$$X_n = AX_{n-1} + BU_n, \qquad Y_n = PX_n. \tag{16}$$

Proposition 8 (Linear RNNs and convolution form) *Let $A \in \mathbb{R}^{S\times S}$ such that A is diagonal, $B \in \mathbb{R}^{S\times 1}, P \in \mathbb{R}^{1\times S}$, and $u = (u_n)_{n\in\mathbb{Z}}$ be a univariate input signal. The output signal $(y_n)_{n\in\mathbb{Z}}$ can write*

$$y_n = \sum_{k=0}^{\infty} c_k u_{n-k}$$

with $c_k = \sum_{s=1}^{S} a_s^k b_s$.

Proof We have $A = \begin{pmatrix} a_1 & \dots & \\ & \ddots & \\ & \dots & a_S \end{pmatrix}, B = \begin{pmatrix} b_1 \\ \vdots \\ b_S \end{pmatrix}, P = \begin{pmatrix} p_1 & \dots & p_S \end{pmatrix}.$

$$\begin{aligned} X_n &= AX_{n-1} + Bu_n \\ &= A(AX_{n-2} + Bu_{n-1}) + Bu_n = \dots = \sum_{k=0}^{n} A^k B u_{n-k} \\ &= \sum_{k=0}^{n} \begin{pmatrix} a_1^k & \dots & \\ & \ddots & \\ & \dots & a_S^k \end{pmatrix} \begin{pmatrix} b_1 \\ \vdots \\ b_S \end{pmatrix} u_{n-k} = \sum_{k=0}^{n} \begin{pmatrix} a_1^k b_1 \\ \vdots \\ a_S^k b_s \end{pmatrix} u_{n-k}. \end{aligned}$$

Finally,

$$\begin{aligned} y_n = \begin{pmatrix} p_1 & \dots & p_S \end{pmatrix} X_n = \sum_{k=0}^{n} \begin{pmatrix} p_1 & \dots & p_S \end{pmatrix} \begin{pmatrix} a_1^k b_1 \\ \vdots \\ a_N^k b_N \end{pmatrix} u_{n-k} &= \sum_{s=1}^{S}\sum_{k=0}^{n} p_s a_s^k b_s u_{n-k} \\ &= \sum_{k=0}^{n} u_{n-k} \sum_{s=1}^{S} a_s^k b_s p_s = \sum_{k=0}^{n} u_{n-k} c_k, \end{aligned}$$

with $c_k = \sum_{s=1}^{S} a_s^k b_s p_s$. In this paper, we consider without loss of generality $\begin{pmatrix} p_1 \dots p_s \end{pmatrix} = \begin{pmatrix} 1 \dots 1 \end{pmatrix}$. ■

B. Some fundamentals of signal processing

In this section, we will recall some fundamentals definitions and results in signal processing. We will only look at discrete-time signals. Throughout this section, we denote $(x_n)_{n\in\mathbb{Z}}$ or x_n a discrete time signal, and x_k the value taken by the signal at time k. For example, let us denote (e_n) the impulse signal such that

$$e_n = \begin{cases} 1, n = 0 \\ 0, n \neq 0. \end{cases} \tag{17}$$

This signal is useful because the response of a system to a impulse signal gives a lot of insights. In particular it fully describes a linear time-invariant system. For more on signal processing, we refer the reader to Oppenheim et al. (1996).

B.1. Linear Time-invariant systems

A system is said to be *time-invariant* if its response to a certain input signal does not depend on time. It is said to be *linear* if its output response to a linear combinations of inputs is the same linear combinations of the output responses of the individual inputs. A system is said to be *causal* if the output at a present time depends on the input up the present time only.

There exist several ways to represent the input-output behavior of LTI system. We will only look at the impulse response representation (convolution).

Proposition 9 (Convolution) *Let h_n be the impulse response of an LTI system H (i.e., the output of system H subject to input e_n), and x_n be an input signal. In this case, the output signal of the system y_n writes*

$$y_n = \sum_{k=-\infty}^{+\infty} x_k h_{n-k}. \tag{18}$$

Causal systems. The output y_n of a causal system depends only on past or present values of the input. This forces $h_k = 0$ for $k < 0$ and the convolution sum is rewritten

$$y_n = \sum_{k=0}^{+\infty} h_k x_{n-k}.$$

Stable systems. A system is stable if the output is guaranteed to be bounded for every bounded input.

B.2. Discrete-Time Fourier Transform

In this section, we denote x_n a complex-valued discrete-time signal.

Definition 10 *The discrete-time Fourier transform of signal x_n is given by*

$$X(\omega) = \sum_{n=-\infty}^{+\infty} x_n e^{-i\omega n}.$$

This function takes values in the frequency space. The inverse discrete-time Fourier transform is given by

$$x_n = \frac{1}{2\pi} \int_0^{2\pi} X(\omega) e^{i\omega n} d\omega.$$

The Discrete-Time Fourier transform presents some notable properties that we recall in Table B.1.

Property	Relation
Time Shifting	$x_{n-k} \overset{DTFT}{\longleftrightarrow} e^{-i\omega k} X(\omega)$
Convolution in Time	$x_n * y_n \overset{DTFT}{\longleftrightarrow} X(\omega)Y(\omega)$
Frequency Differentiation	$j\frac{d}{d\omega}X(\omega) \overset{DTFT}{\longleftrightarrow} -nx_n$
Differencing in Time	$x_n - x_{n-1} \overset{DTFT}{\longleftrightarrow} \left(1 - e^{-i\omega}\right) X(\omega)$

Table B.1: Properties of the Discrete-Time Fourier Transform (DTFT). For each property, assume $x_n \overset{DTFT}{\longleftrightarrow} X(\omega)$ and $y_n \overset{DTFT}{\longleftrightarrow} Y(\omega)$.

We recall Parseval's theorem that establishes a fundamental equivalence between the inner product of two signals in the time domain and their corresponding representation in the frequency domain.

Theorem 11 (Parseval) *For two complex-valued discrete-time signals (x_n) and (y_n) with discrete-time Fourier transforms $X(e^{i\omega})$ and $Y(e^{i\omega})$, Parseval's theorem yields:*

$$\sum_{n=-\infty}^{+\infty} x_n \overline{y_n} = \frac{1}{2\pi}\int_0^{2\pi} X(e^{i\omega})\overline{Y(e^{i\omega})}d\omega. \tag{19}$$

In particular, Parseval's theorem yields an energy conservation result:

$$\sum_{n=-\infty}^{+\infty} |x_n|^2 = \frac{1}{2\pi}\int_0^{2\pi} |X(e^{i\omega})|^2 d\omega.$$

The following proposition will be useful in our lower bound proof in Appendix C.1.

Proposition 12 *Let w_n be a causal discrete-time complex-valued signal with Fourier transform $W(\omega)$. We have the following equality:*

$$\sum_{L=0}^{+\infty} L|w_L|^2 = \frac{i}{2\pi}\int_0^{2\pi} \frac{dW(\omega)}{d\omega}\overline{W}(\omega)d\omega.$$

Proof By definition of the DTFT, $W(\omega) = \sum_{L=0}^{+\infty} w_L e^{-i\omega L}$. Therefore,

$$\begin{aligned}\sum_{L=0}^{+\infty} L|w_L|^2 = \sum_{L=0}^{+\infty} Lw_L\bar{w}_L &= \frac{1}{2\pi}\sum_{L=0}^{+\infty}\sum_{L'=0}^{+\infty} Lw_L\bar{w}_{L'}\int_0^{2\pi} e^{-i\omega(L-L')}d\omega \\ &= \frac{i}{2\pi}\int_0^{2\pi}\sum_{L=0}^{+\infty} -iL\omega_L e^{-iL\omega}\sum_{L'=0}^{+\infty}\bar{w}_{L'}e^{iL'\omega}d\omega.\end{aligned}$$

Provided that the sequence $(Lw_L)_{L\geq 0}$ is summable, $\frac{dW(\omega)}{d\omega} = \sum_{L=0}^{+\infty} -iLw_L e^{-i\omega L}$, which proves the result. ■

B.3. Fourier series

We recall basics of Fourier Series. For more about Fourier series and their applications, we refer the reader to Serov (2017).

Definition 13 (Fourier series) *Let $f : \mathbb{R} \to \mathbb{R}$ be a piecewise continuous and 2π-periodic function. The Fourier series of f is the series of functions*

$$S(f) = \sum_{n=-\infty}^{+\infty} c_n(f) e^{int},$$

where $c_n(f)$ are the Fourier coefficients of f, such that

$$c_n(f) = \frac{1}{2\pi} \int_{-\pi}^{\pi} f(t) e^{-int} dt.$$

The partial sums of these series write

$$S_n(f)(t) = \sum_{k=-n}^{n} c_k(f) e^{ikt}$$

Theorem 14 (Dirichlet) *Let f be piecewise $\mathcal{C}^1$ and 2π-periodic. Therefore, for every $x \in \mathbb{R}$, $S_n(f)(x)$ converges to*

$$\frac{f(x+0) + f(x-0)}{2},$$

where $f(x+0)$ (resp. $f(x-0)$) denotes the right-hand (resp. left-hand) limit of f at x.

Remark: If the function f is not 2π-periodic, its graph on the interval $[0, 2\pi]$ can be extended periodically over $\mathbb{R}$. In this case, Dirichlet's theorem is applicable at potential discontinuities at 0 and 2π.

B.4. A natural pair for autocorrelation

A natural parametrization is to represent autocorrelation with $\gamma(k) = \rho^{|k|}$ with $|\rho| < 1$, as done in the main paper. This models exponentially decreasing autocorrelation between data. The natural associated time-frequency pair to represent is

$$(\gamma(k), \Gamma(e^{i\omega})) = (\rho^{|k|}, \frac{1-\rho^2}{|1-\rho e^{-i\omega}|^2}).$$

Indeed, as $|\rho| < 1$, the sequence $(\rho^{|k|} e^{ik\omega})_{k\in\mathbb{Z}}$ is summable, γ admits a Fourier transform that we denote Γ. For $\omega \in \mathbb{R}$.

$$\begin{aligned}\Gamma(e^{i\omega}) &= \sum_{k=-\infty}^{+\infty} \rho^{|k|} e^{-i\omega k} = \sum_{k=1}^{+\infty} \rho^k e^{i\omega k} + \sum_{k=0}^{+\infty} \rho^k e^{-i\omega k} \\ &= \frac{1}{1-\rho e^{i\omega k}} - 1 + \frac{1}{1-\rho e^{-i\omega k}} = \frac{1-\rho^2}{|1-\rho e^{-i\omega}|^2}.\end{aligned}$$

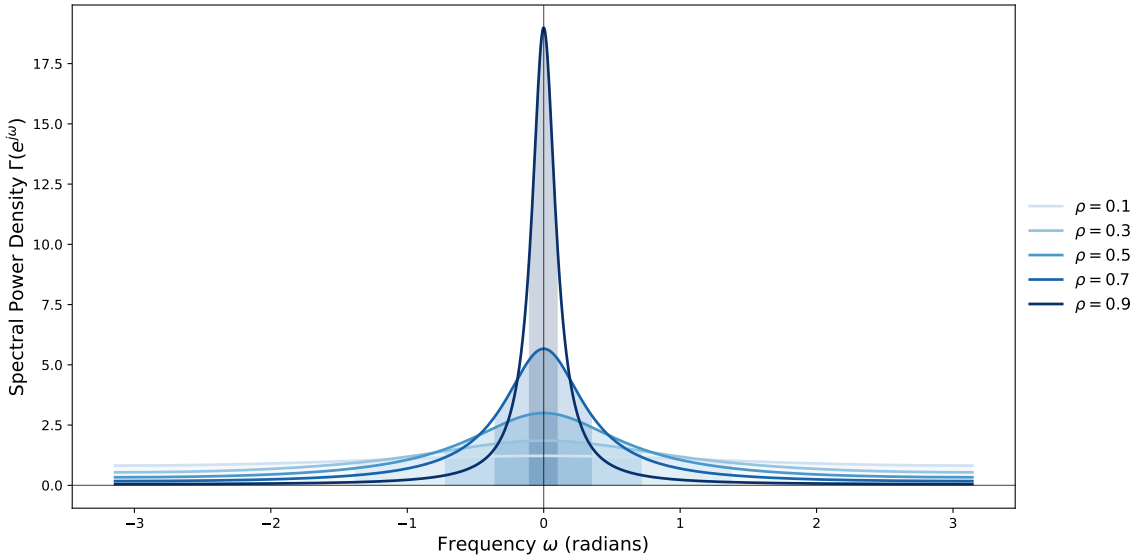

Figure B.1: *The autocorrelation factor ρ determines the width of the spectral power density $\Gamma(e^{i\omega})$. The larger ρ, the narrower the spectral power density. This means that increasing ρ in $\mathcal{L}_{freq}(c,d)$ narrows the bandwidth over which we evaluate the difference $|C(e^{i\omega}) - D(e^{i\omega})|^2$, leading to improved performance.*

C. Lower bound

In this section, we provide the proofs of the two lower bounds.

C.1. White noise case (Theorem 2)

We start by the representation of our loss function as a quadratic form.

Proposition 15 *In the white noise case, the correlation factor ρ is null. The loss $\mathcal{L}_{time}(c,d)$ writes*

$$\mathcal{L}_{time}(c,d) = 1 + \sum_{k=0}^{+\infty} |c_k|^2 - 2\mathrm{Re}\big(\sum_{k=0}^{+\infty} c_k d_k\big),$$

where $c_k = \sum_{s=1}^{S} a_s^k b_s$. Therefore, the loss writes

$$\mathcal{L}_{time}(c,d) = 1 + \sum_{s,s'}^{S} \frac{b_s \bar{b}_{s'}}{1 - a_s \bar{a}_{s'}} - 2\mathrm{Re}\big(\sum_{s=1}^{S} b_s a_s^K\big).$$

Proof On the one hand,

$$\begin{aligned}\sum_{k=0}^{+\infty} |c_k|^2 &= \sum_{k=0}^{+\infty} \big|\sum_{s=1}^{S} a_s^k b_s\big| = \sum_{k=0}^{+\infty} \sum_{s=1}^{S} \sum_{s'=1}^{S} a_s^k \bar{a}_{s'}^k b_s b_{s'} \\ &= \sum_{s=1}^{S} \sum_{s'=1}^{S} b_s b_{s'} \sum_{k=0}^{+\infty} a_s^k \bar{a}_{s'}^k = \sum_{s=1}^{S} \sum_{s'=1}^{S} b_s b_{s'} \frac{1}{1 - a_s \bar{a}_{s'}}.\end{aligned}$$

On the other hand,

$$\mathrm{Re}\big(\sum_{k=0}^{+\infty} c_k d_k\big) = c_K d_K = \sum_{s=1}^{S} b_s a_s^K.$$

Hence the result. ■

We can now maximize it in closed form.

Proposition 16 (Performance criterion) *Minimizing the loss in Proposition 15 boils down to maximizing the following performance criterion*

$$F_K = \sum_{s,s'=1}^{S} \bar{a}_s^K (C^{-1})_{ss'} a_{s'}^K,$$

where $C_{ss'} = \frac{1}{1-a_s \bar{a}_{s'}}$.

Proof The loss $\mathcal{L}_{\text{time}}$ writes

$$1 + \langle \bar{b}, C\bar{b} \rangle - \langle \bar{b}, a^K \rangle - \langle a^K, \bar{b} \rangle.$$

We thus want to maximize with respect to a_s and b_s the quantity

$$\langle \bar{b}, a^K \rangle + \langle a^K, \bar{b} \rangle - \langle \bar{b}, C\bar{b} \rangle.$$

This is convex and quadratic with respect to b, and the minimizer $\bar{b}^*$ is $C^{-1}a^K$[5], leading to the performance criterion

$$F_K = \langle a^K, C^{-1} a^K \rangle = \sum_{s,s'=1}^{S} \bar{a}_s^K (C^{-1})_{ss'} a_{s'}^K.$$

■

We can now move to the proof of Theorem 2, by first analyzing properties of the matrix C.

Linear algebra preview. We use the similarities with Cauchy matrices and their so-called displacement structure (Yang, 2003; Calvetti and Reichel, 1996).

Starting from

$$C - \mathrm{Diag}(a) C \, \mathrm{Diag}(\bar{a}) = 1_S 1_S^\top,$$

we get by post multiplying by $\mathrm{Diag}(\bar{a})^{-1}$,

$$C \, \mathrm{Diag}(\bar{a})^{-1} - \mathrm{Diag}(a) C = 1_S 1_S^\top \, \mathrm{Diag}(\bar{a})^{-1}$$

and thus, by pre and post multiplying by C^{-1}:

$$\mathrm{Diag}(\bar{a})^{-1} C^{-1} - C^{-1} \, \mathrm{Diag}(a) = C^{-1} 1_S 1_S^\top \, \mathrm{Diag}(\bar{a})^{-1} C^{-1},$$

5. The loss with respect to b is clearly holomorphic and therefore twice differentiable. Denoting $\mathcal{L}(b) = \langle \bar{b}, a^K \rangle + \langle a^K, \bar{b} \rangle - \langle \bar{b}, C\bar{b} \rangle$, we have $\frac{\partial \mathcal{L}(b)}{\partial b \partial b} = -C$. As $C_{ss'} = \frac{1}{1-a_s \bar{a}_{s'}}$ with $|a_s| < 1$, classical results yield that C is positive semi-definite.

leading to

$$\text{Diag}(\bar{a})^{-1}C^{-1}\ \text{Diag}(a)^{-1} - C^{-1} = C^{-1}1_S 1_S^\top\ \text{Diag}(\bar{a})^{-1}C^{-1}\ \text{Diag}(a)^{-1} = uv^*,$$

with $u = C^{-1}1_S$ and $v = \ \text{Diag}(\bar{a})^{-1}C^{-1}\ \text{Diag}(a)^{-1}1_S$. This leads to a closed form expression for the inverse:

$$(C^{-1})_{ss'}\big(\frac{1}{\bar{a}_s a_{s'}} - 1\big) = u_s \bar{v}_{s'}.$$

We get

$$v - u = \big[\ \text{Diag}(\bar{a})^{-1}C^{-1}\ \text{Diag}(a)^{-1} - C^{-1}\big]1_S = uv^*1_S = u1_S^\top\ \text{Diag}(\bar{a})^{-1}C^{-1}\ \text{Diag}(a)^{-1}1_S,$$

which leads to $v = u(1 + 1_S^\top\ \text{Diag}(\bar{a})^{-1}C^{-1}\ \text{Diag}(a)^{-1}1_S)$. Moreover we can write

$$\begin{aligned} 1_S^\top\ \text{Diag}(\bar{a})^{-1}C^{-1}\ \text{Diag}(a)^{-1}1_S &= 1_S^\top(C^{-1} + uv^*)1_S = 1_S^\top(C^{-1} + C^{-1}1_S v^*)1_S \\ &= 1_S^\top C^{-1}1_S \cdot (1 + v^*1_S) \\ &= 1_S^\top C^{-1}1_S \cdot (1 + 1_S^\top\ \text{Diag}(\bar{a})^{-1}C^{-1}\ \text{Diag}(a)^{-1}1_S), \end{aligned}$$

which leads to $1_S^\top\ \text{Diag}(\bar{a})^{-1}C^{-1}\ \text{Diag}(a)^{-1}1_S = \frac{1_S^\top C^{-1}1_S}{1 - 1_S^\top C^{-1}1_S}$, and thus $1 + 1_S^\top\ \text{Diag}(\bar{a})^{-1}C^{-1}\ \text{Diag}(a)^{-1}1_S = \frac{1}{1 - 1_S^\top C^{-1}1_S} = \frac{1}{1 - u^\top 1_S}$. Moreover, we have for any $z \in \mathbb{C}$, if all a_s are distinct:

$$\sum_{s'=1}^S \frac{u_{s'}}{1 - z\bar{a}_{s'}} = 1 - \prod_{s'=1}^S \bar{a}_{s'} \prod_{s'=1}^S \frac{a_{s'} - z}{1 - z\bar{a}_{s'}}$$

(the two rational functions have the same degrees, the same poles and are equal for $z = a_1, \dots, a_S$), which leads to for $z = 0$,

$$\sum_{s'=1}^S u_{s'} = 1_S^\top C^{-1}1_S = \sum_{s'=1}^S u_{s'} = 1 - \prod_{s'=1}^S |a_{s'}|^2,$$

and thus $1 - 1_S^\top C^{-1}1 = \prod_{s'=1}^S |a_{s'}|^2$.

We have, if $|z| = 1$,

$$\Big| \prod_{s'=1}^S \frac{a_{s'} - z}{1 - z\bar{a}_{s'}} \Big| = 1,$$

which will be used in the bound (such expressions are typically referred to as Blaschke products (Baratchart et al., 2016), and are know to have unit magnitude).

Proof of the lower bound (by upper bounding F_K). We have, using our linear algebra preview,

$$F_K = \langle a^K, C^{-1}a^K \rangle = \sum_{s,s'=1}^S \bar{a}_s^K (C^{-1})_{ss'} a_{s'}^K = \sum_{s,s'=1}^S (\bar{a}_s a_{s'})^{K+1} \frac{u_s \bar{v}_{s'}}{1 - \bar{a}_s a_{s'}}.$$

We get, using our linear algebra results,

$$F_K - F_{K+1} = \sum_{s,s'=1}^{S} (\bar{a}_s a_{s'})^{K+1}(1-\bar{a}_s a_{s'})\frac{u_s \bar{v}_{s'}}{1-\bar{a}_s a_{s'}} = \frac{1}{\prod_{s'=1}^{S}|a_{s'}|^2}\Big|\sum_{s=1}^{S}\bar{a}_s^{K+1}u_s\Big|^2.$$

This leads to

$$F_K \;=\; \sum_{L=K}^{+\infty}(F_L - F_{L+1}) = \sum_{L=K+1}^{+\infty}\Big|\sum_{s=1}^{S}\bar{a}_s^L u_s\Big|^2 \frac{1}{\prod_{s'=1}^{S}|a_{s'}|^2}.$$

We have:

$$\sum_{L=K+1}^{+\infty}\Big|\sum_{s=1}^{S}\bar{a}_s^L u_s\Big|^2 \;\leqslant\; \frac{1}{K+1}\sum_{L=0}^{+\infty} L\Big|\sum_{s'=1}^{S}\bar{a}_{s'}^L u_{s'}\Big|^2 \text{ since } 1_{L\geqslant K+1} \leqslant \frac{L}{K+1}.$$

We consider the sequence $w_L = \sum_{s=1}^{S}\bar{a}_s^L u_s$, with Fourier series

$$W(\omega) = \sum_{L=0}^{+\infty} w_L e^{-i\omega L} = \sum_{s=1}^{S}\frac{u_s}{1-\bar{a}_s e^{-i\omega}} = 1 - \prod_{s'=1}^{S}\bar{a}_{s'}\prod_{s'=1}^{S}\frac{a_{s'}-e^{-i\omega}}{1-e^{-i\omega}\bar{a}_{s'}}.$$

We then use Proposition 12 to write:

$$\sum_{L=0}^{+\infty} L|w_L|^2 = \frac{i}{2\pi}\int_0^{2\pi} W'(\omega)\overline{W(\omega)}d\omega,$$

leading to

$$\begin{aligned}
&\sum_{L=K+1}^{+\infty}\Big|\sum_{s=1}^{S}\bar{a}_s^L u_s\Big|^2 \\
\leqslant\;& \frac{1}{K+1}\frac{i}{2\pi}\int_0^{2\pi}\frac{d}{d\omega}\Big[-\prod_{s=1}^{S}\bar{a}_s\prod_{s=1}^{S}\frac{a_s-e^{i\omega}}{1-e^{i\omega}\bar{a}_s}\Big]\overline{\Big(1-\prod_{s'=1}^{S}\bar{a}_{s'}\prod_{s'=1}^{S}\frac{a_{s'}-e^{i\omega}}{1-e^{i\omega}\bar{a}_{s'}}\Big)}d\omega \\
=\;& \frac{1}{K+1}\frac{i}{2\pi}\int_0^{2\pi}\frac{d}{d\omega}\Big[\prod_{s=1}^{S}\bar{a}_s\prod_{s=1}^{S}\frac{a_s-e^{i\omega}}{1-e^{i\omega}\bar{a}_s}\Big]\overline{\Big(\prod_{s'=1}^{S}\bar{a}_{s'}\prod_{s'=1}^{S}\frac{a_{s'}-e^{i\omega}}{1-e^{i\omega}\bar{a}_{s'}}\Big)}d\omega.
\end{aligned}$$

We now have, by taking derivatives of the product:

$$\begin{aligned}
\frac{d}{d\omega}\Big[\prod_{s=1}^{S}\frac{a_s-e^{i\omega}}{1-e^{i\omega}\bar{a}_s}\Big] &= \prod_{s=1}^{S}\frac{a_s-e^{i\omega}}{1-e^{i\omega}\bar{a}_s}\sum_{s=1}^{S}\frac{1-e^{i\omega}\bar{a}_s}{a_s-e^{i\omega}}\frac{d}{d\omega}\Big[\frac{a_s-e^{i\omega}}{1-e^{i\omega}\bar{a}_s}\Big] \\
&= \prod_{s=1}^{S}\frac{a_s-e^{i\omega}}{1-e^{i\omega}\bar{a}_s}\sum_{s=1}^{S}\frac{1-e^{i\omega}\bar{a}_s}{a_s-e^{i\omega}}\frac{d}{d\omega}\Big[\frac{1}{\bar{a}_s}+\frac{a_s-\frac{1}{\bar{a}_s}}{1-e^{i\omega}\bar{a}_s}\Big] \\
&= \prod_{s=1}^{S}\frac{a_s-e^{i\omega}}{1-e^{i\omega}\bar{a}_s}\sum_{s=1}^{S}\frac{1-e^{i\omega}\bar{a}_s}{a_s-e^{i\omega}}\Big[(1-|a_s|^2)\frac{-ie^{i\omega}}{(1-e^{i\omega}\bar{a}_s)^2}\Big] \\
&= \prod_{s=1}^{S}\frac{a_s-e^{i\omega}}{1-e^{i\omega}\bar{a}_s}\sum_{s=1}^{S}(1-|a_s|^2)\frac{-i}{|e^{-i\omega}-\bar{a}_s|^2}.
\end{aligned}$$

This leads to, using the unit magnitude of $\frac{a_s-e^{i\omega}}{1-e^{i\omega}\bar{a}_s}$,

$$F_K \leqslant \frac{1}{K+1}\frac{1}{2\pi}\sum_{s=1}^{S}(1-|a_s|^2)\int_0^{2\pi}\frac{1}{|a_s-e^{i\omega}|^2}d\omega = \frac{S}{K+1},$$

using an explicit integration $\frac{1}{2\pi}\int_0^{2\pi}\frac{1}{|a_s-e^{i\omega}|^2}d\omega = \frac{1}{1-|a_s|^2}$.

The approximation error $\mathcal{L}_{\text{time}}(c,d)$ is thus $1-F_K$, which leads to the desired result.

C.2. Autocorrelated case (Theorem 3)

We follow the same proof technique as for Theorem 2, and compute first an explicit expression of the loss, this time, by introducing a new a_s, equal to ρ, with the introduction of new weights $w_s = b_s a_s/(a_s-\rho)$ for $s \in \{1,\dots,S\}$, the weight w_{S+1} being determined by the linear constraint.

Lemma 17 *In the autocorrelated case ($\rho \neq 0$), $\mathcal{L}_{time}(c,d)$ as in Eq. (11) writes*

$$1-2(1-\rho^2)\text{Re}\Big(\sum_{s=1}^{S+1}\frac{w_s a_s^k}{1-a_s\rho}\Big)+(1-\rho^2)\sum_{s,s'}^{S+1}\frac{w_s\bar{w}_{s'}}{1-a_s\bar{a}_{s'}}, \tag{20}$$

where $a_{S+1}=\rho$ and the constraint $\sum_{s=1}^{S+1}w_s a_s^{-1}=0$ holds.

Proof We aim to minimize

$$\sum_{k,k'}(c_k-d_k)(c_{k'}-d_{k'})\gamma(k-k'),$$

where $\gamma(k-k')=\rho^{|k-k'|}$. Denoting $C(e^{i\omega})$, $D(e^{i\omega})$ and $\Gamma(e^{i\omega})$ the Fourier transforms of (c_n), (d_n) and (γ_n) respectively, Parseval's theorem yields

$$\sum_{k,k'}(c_k-d_k)(c_{k'}-d_{k'})\gamma(k-k') = \frac{1}{2\pi}\int_{-\pi}^{\pi}\big|C(e^{i\omega})-D(e^{i\omega})\big|^2\Gamma(e^{i\omega})d\omega.$$

We have $D(e^{i\omega}) = e^{-iK\omega}$ (Fourier transform of a shifted Dirac at timestep K), and

$$C(e^{i\omega}) = \sum_{k=0}^{+\infty} \sum_{s=1}^{S} b_s a_s^k e^{-i\omega k} = \sum_{s=1}^{S} \frac{b_s}{1 - a_s e^{-i\omega}},$$
$$\Gamma(e^{i\omega}) = \sum_{k=-\infty}^{+\infty} \gamma(k) e^{-i\omega k} = \frac{1}{1 - \rho e^{-i\omega}} \frac{1 - \rho^2}{1 - \rho e^{i\omega}}.$$

The criterion becomes (with an error of 1 if $C = 0$):

$$\begin{aligned}
& \frac{1}{2\pi} \int_0^{2\pi} |D(e^{i\omega}) - C(e^{i\omega})|^2 \Gamma(e^{i\omega}) d\omega \\
= \; & \frac{1 - \rho^2}{2\pi} \int_0^{2\pi} \left| D(e^{i\omega}) \frac{1}{1 - \rho e^{-i\omega}} - C(e^{i\omega}) \frac{1}{1 - \rho e^{-i\omega}} \right|^2 d\omega \\
= \; & 1 - \frac{1 - \rho^2}{2\pi} 2 \, \mathrm{Re}\Big(\int_0^{2\pi} \overline{D(e^{i\omega}) \frac{1}{1 - \rho e^{-i\omega}}} C(e^{i\omega}) \frac{1}{1 - \rho e^{-i\omega}} \Big) d\omega \\
& \qquad + \frac{1 - \rho^2}{2\pi} \int_0^{2\pi} \left| C(e^{i\omega}) \frac{1}{1 - \rho e^{-i\omega}} \right|^2 d\omega.
\end{aligned}$$

We have

$$\frac{1}{1 - a_s e^{-i\omega}} \frac{1}{1 - \rho e^{-i\omega}} = \frac{1}{a_s - \rho} \Big(\frac{a_s}{1 - a_s e^{-i\omega}} - \frac{\rho}{1 - \rho e^{-i\omega}} \Big),$$

and thus

$$\begin{aligned}
C(e^{i\omega}) \frac{1}{1 - \rho e^{-i\omega}} &= \sum_{s=1}^{S} \frac{b_s}{a_s - \rho} \Big(\frac{a_s}{1 - a_s e^{-i\omega}} - \frac{\rho}{1 - \rho e^{-i\omega}} \Big) \\
&= \sum_{s=1}^{S+1} \frac{w_s}{1 - a_s e^{-i\omega}},
\end{aligned}$$

with $w_s = b_s a_s / (a_s - \rho)$, $a_{S+1} = \rho$, and the constraint $\sum_{s=1}^{S+1} w_s a_s^{-1} = 0$. The criterion becomes

$$1 - (1 - \rho^2) \sum_{s=1}^{S+1} 2\mathrm{Re}\Big(\frac{w_s a_s^K}{1 - a_s \rho} \Big) + (1 - \rho^2) \sum_{s,s'=1}^{S+1} \frac{\bar{w}_s w_{s'}}{1 - a_s \bar{a}'_s},$$

after straightforward computations. ∎

Proof of Theorem 3. The minimum with respect to w in Eq. (20) with the constraint is greater than the unconstrained minimizer, equal to

$$H_K = 1 - (1 - \rho^2) \sum_{s,s'=1}^{S+1} \frac{\bar{a}_s^K}{1 - \bar{a}_s \rho} \frac{a_{s'}^K}{1 - a_{s'} \rho} (C^{-1})_{ss'},$$

where we recall that $C_{ss'} = \frac{1}{1 - a_s \bar{a}_{s'}}$.

Using linear algebra properties from above with $S+1$ zeros and poles, we get

$$\begin{aligned} H_K &= 1-(1-\rho^2)\sum_{s,s'=1}^{S+1}\frac{1}{1-\bar{a}_s\rho}\frac{1}{1-a_{s'}\rho}\frac{(\bar{a}_s a_{s'})^{K+1}u_s\bar{v}_{s'}}{1-\bar{a}_s a_{s'}} \\ &= 1-(1-\rho^2)\frac{1}{\prod_{s=1}^{S+1}|a_s|^2}\sum_{s,s'=1}^{S+1}\frac{1}{1-\bar{a}_s\rho}\frac{1}{1-a_{s'}\rho}\frac{(\bar{a}_s a_{s'})^{K+1}u_s\bar{u}_{s'}}{1-\bar{a}_s a_{s'}}, \end{aligned}$$

where we recall that $u = C^{-1}1_S$ and $v = \text{Diag}(\bar{a})^{-1}C^{-1}\text{Diag}(a)^{-1}1_S$.

We have

$$\begin{aligned} H_{K+1}-H_K &= \frac{1-\rho^2}{\prod_{s=1}^{S+1}|a_s|^2}\sum_{s,s'=1}^{S+1}\frac{1}{1-\bar{a}_s\rho}\frac{1}{1-a_{s'}\rho}(\bar{a}_s a_{s'})^{K+1}u_s\bar{u}_{s'} \\ &= \frac{1-\rho^2}{\prod_{s=1}^{S+1}|a_s|^2}\Big|\sum_{s=1}^{S+1}\frac{1}{1-\bar{a}_s\rho}\bar{a}_s^{K+1}u_s\Big|^2, \end{aligned}$$

leading to

$$\begin{aligned} H_K &= \sum_{L=K}^{+\infty}(H_L-H_{L+1})+1 \\ &= 1-\frac{1-\rho^2}{\prod_{s=1}^{S+1}|a_s|^2}\sum_{L=K}^{+\infty}\Big|\sum_{s=1}^{S+1}\frac{1}{1-\bar{a}_s\rho}\bar{a}_s^L\bar{a}_s u_s\Big|^2 \\ &\geqslant 1-\frac{1}{K}\frac{1-\rho^2}{\prod_{s=1}^{S+1}|a_s|^2}\sum_{L=0}^{+\infty}L\Big|\sum_{s=1}^{S+1}\frac{1}{1-\bar{a}_s\rho}\bar{a}_s^L\bar{a}_s u_s\Big|^2, \end{aligned}$$

using $1_{L\geqslant K}\leqslant\frac{L}{K}$.

The sequence $w_L=\sum_{s=1}^{S+1}\frac{1}{1-\bar{a}_s\rho}\bar{a}_s^L\bar{a}_s u_s$, has Fourier series

$$\begin{aligned} W(\omega) &= \sum_{L=0}^{+\infty}w_L e^{-i\omega L}=\sum_{L=0}^{+\infty}e^{-i\omega L}\sum_{s=1}^{S+1}\frac{1}{1-\bar{a}_s\rho}\bar{a}_s^L\bar{a}_s u_s \\ &= \sum_{s=1}^{S+1}\frac{1}{1-\bar{a}_s\rho}\frac{\bar{a}_s u_s}{1-\bar{a}_s e^{-i\omega}}=\sum_{s=1}^{S+1}u_s\Big(\frac{1}{1-\bar{a}_s\rho}-\frac{1}{1-\bar{a}_s e^{-i\omega}}\Big)\frac{1}{\rho-e^{-i\omega}} \\ &= \frac{1}{\rho-e^{-i\omega}}\Big(\prod_{s=1}^{S+1}\bar{a}_s\Big)\Big(\prod_{s=1}^{S+1}\frac{a_s-e^{-i\omega}}{1-e^{-i\omega}\bar{a}_s}-\prod_{s=1}^{S+1}\frac{a_s-\rho}{1-\rho\bar{a}_s}\Big) \\ &= \frac{1}{\rho-e^{-i\omega}}\Big(\prod_{s=1}^{S+1}\bar{a}_s\Big)\prod_{s=1}^{S+1}\frac{a_s-e^{-i\omega}}{1-e^{-i\omega}\bar{a}_s}, \end{aligned}$$

because of the link between u, C and rational functions.

We have:

$$\begin{aligned}
1-H_K &\leqslant \frac{1-\rho^2}{K}\frac{1}{\prod_{s=1}^{S+1}|a_s|^2}\sum_{L=0}^{+\infty}L|w_L|^2 \\
&= \frac{1-\rho^2}{K}\frac{1}{\prod_{s=1}^{S+1}|a_s|^2}\frac{i}{2\pi}\int_0^{2\pi}W'(\omega)\overline{W(\omega)}d\omega \text{ using properties of Fourier Series,} \\
&= \frac{1-\rho^2}{K}\frac{i}{2\pi}\int_0^{2\pi}\frac{d}{d\omega}\Big(\frac{1}{\rho-e^{-i\omega}}\prod_{s=1}^{S+1}\frac{a_s-e^{-i\omega}}{1-e^{-i\omega}\bar{a}_s}\Big)\overline{\frac{1}{\rho-e^{-i\omega}}\prod_{s=1}^{S+1}\frac{a_s-e^{-i\omega}}{1-e^{-i\omega}\bar{a}_s}}d\omega \\
&= \frac{1-\rho^2}{K}\frac{i}{2\pi}\int_0^{2\pi}\frac{-ie^{-i\omega}}{(\rho-e^{-i\omega})^2}\prod_{s=1}^{S+1}\frac{a_s-e^{-i\omega}}{1-e^{-i\omega}\bar{a}_s}\overline{\frac{1}{\rho-e^{-i\omega}}\prod_{s=1}^{S+1}\frac{a_s-e^{-i\omega}}{1-e^{-i\omega}\bar{a}_s}}d\omega \\
&\quad +\frac{1-\rho^2}{K}\frac{i}{2\pi}\int_0^{2\pi}\frac{1}{\rho-e^{-i\omega}}\frac{d}{d\omega}\Big(\prod_{s=1}^{S+1}\frac{a_s-e^{-i\omega}}{1-e^{-i\omega}\bar{a}_s}\Big)\overline{\frac{1}{\rho-e^{-i\omega}}\prod_{s=1}^{S+1}\frac{a_s-e^{-i\omega}}{1-e^{-i\omega}\bar{a}_s}}d\omega.
\end{aligned}$$

Using the following identities,

$$\begin{aligned}
\frac{a_s-e^{-i\omega}}{1-e^{-i\omega}\bar{a}_s} &= \frac{1}{\bar{a}_s}+\frac{a_s-1/\bar{a}_s}{1-e^{-i\omega}\bar{a}_s}, \\
\frac{d}{d\omega}\Big(\frac{a_s-e^{-i\omega}}{1-e^{-i\omega}\bar{a}_s}\Big) &= \frac{a_s-1/\bar{a}_s}{(1-e^{-i\omega}\bar{a}_s)^2}\bar{a}_s(-ie^{-i\omega}) = ie^{-i\omega}\frac{1-|a_s|^2}{(1-e^{-i\omega}\bar{a}_s)^2}, \\
\Big|\frac{a_s-e^{-i\omega}}{1-e^{-i\omega}\bar{a}_s}\Big| &= 1,
\end{aligned}$$

we get

$$\begin{aligned}
1-H_K &\leqslant \frac{1-\rho^2}{K}\frac{1}{2\pi}\int_0^{2\pi}\frac{e^{-i\omega}}{(\rho-e^{-i\omega})^2(\rho-e^{i\omega})}d\omega \\
&\qquad +\frac{1-\rho^2}{K}\sum_{s=1}^{S+1}(1-|a_s|^2)\frac{1}{2\pi}\int_0^{2\pi}\frac{1}{|\rho-e^{-i\omega}|^2}\frac{1}{|a_s-e^{-i\omega}|^2}d\omega \\
&= \frac{1-\rho^2}{K}\frac{1}{2\pi}\int_0^{2\pi}\frac{e^{-i\omega}}{(\rho-e^{-i\omega})^2(\rho-e^{i\omega})}d\omega \\
&\quad +\frac{(1-\rho^2)^2}{K}\frac{1}{2\pi}\int_0^{2\pi}\frac{1}{|\rho-e^{-i\omega}|^4}d\omega+\frac{1-\rho^2}{K}\sum_{s=1}^{S}(1-|a_s|^2)\frac{1}{2\pi}\int_0^{2\pi}\frac{1}{|\rho-e^{-i\omega}|^2}\frac{1}{|a_s-e^{-i\omega}|^2}d\omega \\
&= -\frac{1}{K}\frac{1}{1-\rho^2}+\frac{1}{K}\frac{1+\rho^2}{1-\rho^2}+\frac{1-\rho^2}{K}\sum_{s=1}^{S}(1-|a_s|^2)\frac{1}{2\pi}\int_0^{2\pi}\frac{1}{|\rho-e^{-i\omega}|^2}\frac{1}{|a_s-e^{-i\omega}|^2}d\omega
\end{aligned}$$

by exact integration. Then, using $\frac{1}{|\rho-e^{-i\omega}|^2}\leqslant\frac{1}{(1-\rho)^2}$, and $\frac{1}{2\pi}\int_0^{2\pi}\frac{1}{|a_s-e^{-i\omega}|^2}d\omega=\frac{1}{1-|a_s|^2}$, we get

$$1-H_K \leqslant \frac{1}{K}\frac{\rho^2}{1-\rho^2}+\frac{1+\rho}{1-\rho}\frac{S}{K}\leqslant\frac{1}{K}\frac{\rho}{1-\rho}+\frac{2}{1-\rho}\frac{S}{K}=\frac{1}{K}\frac{1}{1-\rho}(\rho+2S).$$

Thus, we get an approximation error greater than $\Big(1-\frac{1}{K}\frac{3S}{1-\rho}\Big)_+$. (since it is always nonnegative).

D. Upper bound

Here, we prove the results of Section 5 of the paper. We begin by proving the expression of $\mathcal{L}_{\text{freq}}$, to then justify the parametrization of the optimal b_s in Eq. (12). We finally compute the asymptotic loss in Eq. (15) and Theorem 7.

D.1. Loss in frequency domain

Here, we prove the expression of the counterpart of $\mathcal{L}_{\text{time}}$ in the frequency domain, $\mathcal{L}_{\text{freq}}$. More explicitly, we give a proof of Eq. (6).

Proof Denote (z_k) the discrete-time filter such that

$$z_k = \sum_{k'=0}^{+\infty} (c_{k'} - d_{k'})\gamma(k-k').$$

Therefore, (z_k) is a convolution between $(c_k - d_k)$ and γ_k,

$$\sum_{k,k'=0}^{+\infty} (c_k - d_k)(c_{k'} - d_{k'})\gamma(k-k') = \sum_{k=0}^{+\infty}(c_k - d_k)z_k.$$

According to Parseval's theorem and denoting $C(e^{i\omega}), D(e^{i\omega})$ and $\Gamma(e^{i\omega})$ the respective Fourier transforms, we have:

$$\begin{aligned}
\sum_{k=0}^{+\infty}(c_k - d_k)z_k &= \frac{1}{2\pi}\int_0^{2\pi} Z(\omega)\overline{(C(e^{i\omega}) - D(e^{i\omega}))}d\omega \\
&= \frac{1}{2\pi}\int_0^{2\pi} \Gamma(e^{i\omega})(C(e^{i\omega}) - D(e^{i\omega}))\overline{(C(e^{i\omega}) - D(e^{i\omega}))}d\omega \\
&= \frac{1}{2\pi}\int_0^{2\pi} \left|C(e^{i\omega}) - D(e^{i\omega})\right|^2\Gamma(e^{i\omega})d\omega,
\end{aligned}$$

by the convolution property of the DTFT. Finally, (d_k) being the shifted impulse filter, its Fourier transform is $D(e^{i\omega}) = e^{-iK\omega}$. The Fourier transform $C(e^{i\omega})$ of (c_k) is given by

$$C(e^{i\omega}) = \sum_{s=1}^{S} b_s \sum_{k=-\infty}^{\infty} \left(a_s e^{-i\omega}\right)^k = \sum_{s=1}^{S} \frac{b_s}{1 - a_s e^{-i\omega}}.$$

■

D.2. Parametrization of the optimal b_s

For the sake of conciseness, we sometimes denote a and b for the vectors (a_s) and (b_s) respectively. Before proving Lemma 4, we show three general lemmas on Fourier series and remainder of series. We will use them later in the proof of Lemma 4.

Lemma 18 *Let $\alpha \in \mathbb{C}$ and $S \in \mathbb{N}$. Consider the infinite Toeplitz matrix T defined by*

$$T(s,s') \;=\; \frac{1}{2\alpha - \mathrm{i}(s-s')\pi}.$$

Then, $\frac{2}{e^{2\alpha}-e^{-2\alpha}}$ is an asymptotic eigenvalue of T, associated with the eigenvector $z = ((-1)^s)_{s\in\mathbb{Z}}$.

Proof We compute the action of T on z:

$$(Tz)_s = \sum_{s'} T(s,s')(-1)^{s'} = \sum_{s'} \frac{e^{\mathrm{i}\pi s'}}{2\alpha - \mathrm{i}\pi s + \mathrm{i}\pi s'}.$$

This expression resembles a Fourier series evaluated at $\omega = \pi$. For $k \in \mathbb{Z}$, consider:

$$\frac{1}{2\pi}\int_0^{2\pi} e^{-\frac{2\alpha}{\pi}(\omega-\pi)} e^{-\mathrm{i}k\omega}\, d\omega = \frac{1}{2}\cdot\frac{e^{2\alpha}-e^{-2\alpha}}{2\alpha + \mathrm{i}k\pi}.$$

Thus:

$$\frac{1}{2\alpha + \mathrm{i}k\pi} = \frac{1}{2\pi}\cdot\frac{2}{e^{2\alpha}-e^{-2\alpha}}\int_0^{2\pi} e^{-\frac{2\alpha}{\pi}(\omega-\pi)} e^{-\mathrm{i}k\omega}\, d\omega. \tag{21}$$

We recognize this as the Fourier coefficient of the function $f_\alpha(\omega) = \frac{2}{e^{2\alpha}-e^{-2\alpha}} e^{-\frac{2\alpha}{\pi}(\omega-\pi)}$. Therefore, according to Dirichlet's theorem, for $s \in \mathbb{Z}$:

$$f_{2\alpha-\mathrm{i}\pi s}(\pi) = \sum_{s'} \frac{e^{\mathrm{i}\pi s'}}{2\alpha - \mathrm{i}\pi s + \mathrm{i}\pi s'} = \sum_{s'} \frac{(-1)^{s'}}{2\alpha - \mathrm{i}\pi(s-s')}.$$

Simplifying, we find: $\dfrac{2}{e^{2\alpha-\mathrm{i}\pi s} - e^{-2\alpha+\mathrm{i}\pi s}} = \displaystyle\sum_{s'} \frac{(-1)^{s'}}{2\alpha - \mathrm{i}\pi(s-s')}$. Finally, we obtain $(Tz)_s = \dfrac{2}{e^{2\alpha}-e^{-2\alpha}} z_s$, and thus: $Tz = \dfrac{2}{e^{2\alpha}-e^{-2\alpha}} z$, proving that $\frac{2}{e^{2\alpha}-e^{-2\alpha}}$ is an eigenvalue associated with $z = ((-1)^s)_{s\in\mathbb{Z}}$.

■

Now, we prove two general results on remainders of alternating series.

Lemma 19 *We have:*

$$R_N = \sum_{n=N}^{+\infty} \frac{(-1)^n}{n} = \frac{(-1)^N}{2N} + \frac{1}{2}\sum_{n=N}^{+\infty} \frac{(-1)^n}{n(n+1)}.$$

Proof On the one side, by grouping two consecutive terms,

$$R_N + R_{N+1} = \sum_{n=N}^{+\infty} \frac{(-1)^n}{n} + \sum_{n=N+1}^{+\infty} \frac{(-1)^n}{n} = \sum_{n=N}^{+\infty} \left\{\frac{(-1)^n}{n} - \frac{(-1)^n}{n+1}\right\} = \sum_{n=N}^{+\infty} \frac{(-1)^n}{n(n+1)},$$

and on the other side,

$$R_{N+1} = R_N - \frac{(-1)^N}{N}.$$

Therefore, $2R_N = \dfrac{(-1)^N}{N} + \displaystyle\sum_{n=N}^{+\infty} \frac{(-1)^n}{n(n+1)}$. ■

Lemma 20 *For $\alpha \in \mathbb{C}$ such that $Re(\alpha) > 0$, we have*

$$\sum_{n=N}^{+\infty} \frac{(-1)^n}{\alpha + i\pi n} = \frac{i}{\pi} \times \frac{(-1)^N}{2N} + o(\frac{1}{N}).$$

Proof We denote $R_N = \sum_{n=N}^{+\infty} \frac{(-1)^n}{\alpha - i\pi n}$ and $S_N = \sum_{n=N}^{+\infty} \frac{(-1)^n}{-i\pi n}$. Consequently,

$$R_N - S_N = \sum_{n=N}^{+\infty} (-1)^n \big(\frac{1}{\alpha - i\pi n} + \frac{1}{i\pi n}\big) = \alpha \sum_{n=N}^{+\infty} \frac{(-1)^n}{\alpha i\pi n + \pi^2 n^2},$$

leading to

$$\begin{aligned} R_N &= S_N + \alpha \sum_{n=N}^{+\infty} \frac{(-1)^n}{\alpha i\pi n + \pi^2 n^2} = -\frac{i}{\pi} \sum_{n=N}^{+\infty} \frac{(-1)^n}{n} + \alpha \sum_{n=N}^{+\infty} \frac{(-1)^n}{\alpha i\pi n + \pi^2 n^2} \\ &= -\frac{i}{\pi} \times \frac{(-1)^N}{2N} - \frac{i}{\pi} \sum_{n=N}^{+\infty} \frac{(-1)^n}{n(n+1)} + \frac{\alpha}{\pi} \sum_{n=N}^{+\infty} \frac{(-1)^n}{\alpha i n + \pi n^2}, \end{aligned}$$

where the last line stems from Lemma 19. The result then follows by classical results on the remainder of alternating series.[6] ■

Proof of Lemma 4 We adopt the parametrization in Eq. (12) for the poles (a_s) and aim to determine the optimal parameters (b_s) under this configuration. Since $\mathcal{L}_{\text{time}}(c, d)$ is convex with respect to (b_s), setting the gradient to zero yields:

$$b = C^{-1} \bar{a}^K,$$

where $C_{ss'} = \frac{1}{1 - a_s \bar{a}_{s'}}$ for s, s' in $[\![-T, T]\!]$ where we recall that $S = 2T + 1$. Let's denote the matrix M such that $M(s, s') = \frac{K}{2\alpha - i\pi(s - s')} + \frac{1}{2}$ for s, s' in $[\![-T, T]\!]$.

First, we remark that, for $s \in [\![-T, T]\!]$,

$$(a^K)_s = (-1)^s e^{-\alpha} \in \mathbb{R}.$$

We derive the asymptotic expansion for the optimal b, using a coordinate-wise approach. We recall that we place ourselves in the case $1 \ll S \ll K$. Let us denote $z = ((-1)^s)_{s \in [\![-T,T]\!]}$. For $s \in [\![-T, T]\!]$,

6. Alternating series' criterion: Let (a_n) be a positive and decreasing sequence such that $a_n \to 0$. Then, the series $\sum_n (-1)^n a_n$ converges. Denoting $R_n = \sum_{k=n+1}^{+\infty} (-1)^k a_k$, we have $|R_n| \leq a_{n+1}$.

$$
\begin{aligned}
(Cz)_s =& (Mz)_s + (Cz)_s - (Mz)_s \\
&= \sum_{s'=-T}^{T} \frac{(-1)^{s'} K}{2\alpha - i\pi(s-s')} \\
&\quad + \sum_{s'=-T}^{T} \frac{(-1)^{s'}}{2} + \sum_{s'=-T}^{T} (-1)^{s'} \Big[\frac{1}{1 - e^{-\frac{2\alpha}{K}} e^{\frac{i\pi(s-s')}{K}}} - \frac{K}{2\alpha - i\pi(s-s')} - \frac{1}{2} \Big] \\
&= \sum_{s'=-T}^{T} \frac{(-1)^{s'} K}{2\alpha - i\pi(s-s')} + \frac{1}{2} + \sum_{s'=-T}^{T} (-1)^{s'} \Big[\frac{2\alpha - i\pi(s-s') - K(1 - e^{-\frac{2\alpha}{K}} e^{\frac{i\pi(s-s')}{K}})}{(1 - e^{-\frac{2\alpha}{K}} e^{i\pi(s-s')})(2\alpha - i\pi(s-s'))} - \frac{1}{2} \Big] \\
&= \sum_{s'=-T}^{T} \frac{(-1)^{s'} K}{2\alpha - i\pi(s-s')} + \frac{1}{2} \\
&\quad + \frac{1}{K} \sum_{s'=-T}^{T} (-1)^{s'} \Big[\frac{\frac{4\alpha^3}{3} - \pi^2 \alpha (s-s')^2 + i \left(\pi^3 \frac{(s-s')^3}{6} - 2\pi\alpha^2 (s-s') \right) + o(1)}{(2\alpha - i\pi(s-s') + o(1))(2\alpha - i\pi(s-s'))} \Big]. \qquad (22)
\end{aligned}
$$

But,

$$
\left[\frac{-\pi^2 \alpha (s-s')^2 + i \left(\pi^3 \frac{(s-s')^3}{6} - 2\pi\alpha^2 (s-s') \right) + o(1)}{(2\alpha - i\pi(s-s') + o(1))(2\alpha - i\pi(s-s'))} \right] \underset{+\infty}{\sim} i\pi \frac{s-s'}{6},
$$

So there exists a sequence $(\epsilon_{s'})$ for $s' \in \mathbb{Z}$ such that $\epsilon_{s'} \to 0$ and

$$
\begin{aligned}
&\sum_{s'=-T}^{T} (-1)^{s'} \Big[\frac{\frac{4\alpha^3}{3} - \pi^2 \alpha (s-s')^2 + i \left(\pi^3 \frac{(s-s')^3}{6} - 2\pi\alpha^2 (s-s') \right) + o(1)}{(2\alpha - i\pi(s-s') + o(1))(2\alpha - i\pi(s-s'))} \\
&= \sum_{s'=-T}^{T} (-1)^{s'} \frac{\alpha}{3} + \sum_{s'=-T}^{T} (-1)^{s'} \frac{i\pi(s-s')}{6} (1 + \epsilon_{s'}) \\
&= (-1)^T [\frac{\alpha}{3} - \frac{i\pi s}{6}] + \sum_{s'=-T}^{T} (-1)^{s'} \frac{i\pi(s-s')\epsilon_{s'}}{6} = O(1) \qquad (23)
\end{aligned}
$$

Note that we had to keep the constant term $\frac{4\alpha^3}{3}$ which corresponds to $s = s'$.

Plugging this into eq (22), this leads to

$$
(Cz)_s = \sum_{s'=-T}^{T} \frac{(-1)^{s'} K}{2\alpha - i\pi(s-s')} + \frac{1}{2} + O(\frac{1}{K}) = (-1)^s \frac{2K}{e^{2\alpha} - e^{-2\alpha}} + O(\frac{K}{T}),
$$

where we used Lemma 20. We can deduce from this coordinate-wise equation that

$$
Cz = \frac{2K}{e^{2\alpha} - e^{-2\alpha}} z + O(\frac{K}{T}).
$$

We finally use the bounded nature of C's condition number[7] to apply C^{-1} and to show:

$$C^{-1}\bar{a}^K \sim \frac{e^{2\alpha}-e^{-2\alpha}}{2K}z.$$

This is valid when $T \to +\infty, T/K \to 0$. ■

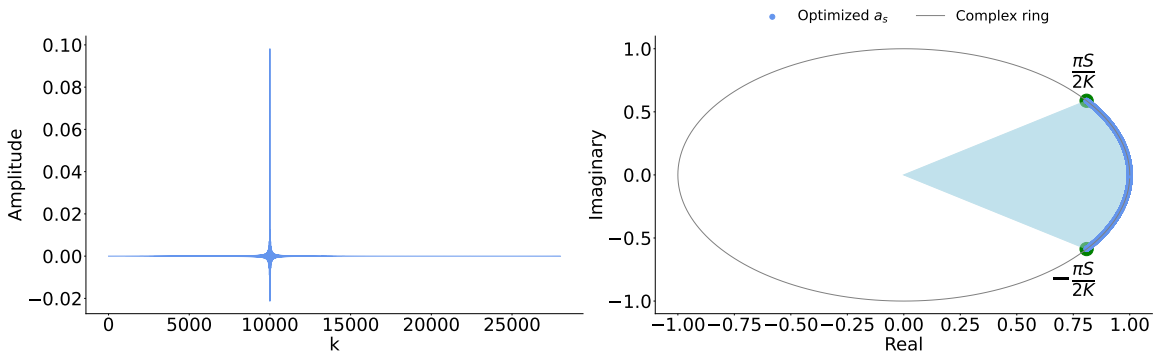

Figure D.1: *Left: Real values of filter in Eq. (5) in the time-domain. Right: Positions of the a_s on the unit disk. In this case, $S = 100, K = 10000$. The x-axis on the left images represents the timesteps. The representation in Eq. (5) concentrates the poles a_s on a slice of the unit disk, whose size depends on the ratio $\frac{S}{K}$. The a_s operate in pairs of complex conjugates, ensuring that the final filter remains real. Each a_s approximates a single oscillation of the complex exponential, with the oscillations spaced by a distance of $\frac{\pi}{K}$. Therefore when K increases (for fixed S), the slice size decreases, imposing smaller phase shifts to capture long-range dependencies in the data (see Orvieto et al. (2023)). This parametrization allows to build filters than can look far back in time.*

D.3. Upper bound of the loss

In this section, we will prove Theorem 6 through an asymptotic expansion. We will use general results on the remainder of alternate series in Lemmas 19 and 20. We also start with a lemma very similar to Lemma 18. We use all of them later in the proof of Theorem 6.

Lemma 21 *Let $\alpha \in \mathbb{C}$ such that $\mathrm{Re}(\alpha) > 0$. Then,* $\displaystyle\sum_{s=-\infty}^{+\infty} \frac{(-1)^s}{2\alpha - i\pi s} = \frac{2}{e^{2\alpha}-e^{-2\alpha}}.$

Proof We consider $\sum_{s=-\infty}^{+\infty} \frac{(-1)^s}{2\alpha - i\pi s}$. This looks like a Fourier series evaluated in $\omega = \pi$. We will look for a function f_α such that $c_s(f_\alpha) = \frac{1}{2\alpha - i\pi s}$. Denoting f_α such that $f_\alpha(\omega) = \frac{2e^{\frac{-2\alpha}{\pi}(\omega-\pi)}}{e^{2\alpha}-e^{-2\alpha}}$, we have,

$$c_s(f_\alpha) = \frac{1}{2\pi}\int_0^{2\pi} \frac{2e^{\frac{2\alpha}{\pi}(\omega-\pi)}}{e^{2\alpha}-e^{-2\alpha}} e^{-is\omega} d\omega = \frac{1}{\pi}\frac{1}{(e^{2\alpha}-e^{-2\alpha})(\frac{2\alpha}{\pi}-is)}(e^{2\alpha}-e^{-2\alpha}) = \frac{1}{2\alpha - i\pi s}.$$

7. This is due to the link between eigenvalues of Toeplitz matrices and the Fourier series of the first row Gray (2006), and the relationship $C_{ss'} = \frac{1}{1-\exp(-2\alpha/K)\exp(i\pi(s-s')/K)} \sim \frac{K}{2\alpha - i\pi(s-s')}$ together with Eq. (21).

Therefore, using Dirichlet's theorem, f_α is the appropriate function and

$$f_\alpha(\pi) = \sum_s \frac{(-1)^s}{2\alpha - i\pi s} \Leftrightarrow \sum_{s=-\infty}^{+\infty} \frac{(-1)^s}{2\alpha - i\pi s} = \frac{2}{e^{2\alpha} - e^{-2\alpha}}.$$

■

Proof of Theorem 6 We recall that we have the following asymptotic representations for our filter:

$$a_s = e^{-\frac{\alpha}{K}} e^{i\frac{s\pi}{K}}, s \in [\![-T, T]\!]$$
$$b_s = \frac{e^{-\alpha}(e^{2\alpha} - e^{-2\alpha})}{2K}(-1)^s, s \in [\![-T, T]\!],$$

and that $\mathcal{L}_{\text{time}}(c, d)$ can therefore write:

$$\begin{aligned}
\mathcal{L}_{\text{time}}(c,d) &= \sum_{s,s'=-T}^{T} \frac{b_s \bar{b}_{s'}}{1 - a_s \bar{a}_{s'}} - 2\sum_{s=-T}^{T} b_s a_s^K + 1 \\
&= \frac{e^{-2\alpha}(e^{2\alpha} - e^{-2\alpha})^2}{4K^2} \sum_{s,s'=-T}^{T} \frac{(-1)^{s+s'}}{1 - e^{-\frac{2\alpha}{K}} e^{(s-s')i\frac{\pi}{K}}} \\
&\quad - 2\sum_{s=-T}^{T} \frac{(-1)^s e^{-2\alpha}(e^{2\alpha} - e^{-2\alpha})(-1)^s}{2K} + 1 \\
&= \frac{e^{-2\alpha}(e^{2\alpha} - e^{-2\alpha})^2}{4K^2} \sum_{s,s'=-T}^{T} \frac{(-1)^{s+s'}}{1 - e^{-\frac{2\alpha}{K}} e^{(s-s')i\frac{\pi}{K}}} - \frac{e^{-2\alpha}(e^{2\alpha} - e^{-2\alpha})S}{K} + 1.
\end{aligned}$$

First, we prove that

$$\sum_{s,s'=-T}^{T} \frac{(-1)^{s+s'}}{1 - e^{\frac{-2\alpha}{K}} e^{(s-s')i\frac{\pi}{K}}} = \sum_{s=-T}^{T} (-1)^s K \sum_{s'=-T}^{T} \frac{e^{i\pi s'}}{2\alpha - i\pi(s - s')} + O(1), \tag{24}$$

when $T \to +\infty, T/K \to 0$. Let us compute:

$$\begin{aligned}
&\sum_{s,s'=-T}^{T} \frac{(-1)^{s+s'}}{1 - e^{\frac{-2\alpha}{K}} e^{(s-s')i\frac{\pi}{K}}} - \sum_{s=-T}^{T} (-1)^s K \sum_{s'=-T}^{T} \frac{e^{i\pi s'}}{2\alpha - i\pi(s - s')} \\
&= \sum_{s,s'=-T}^{T} (-1)^{s+s'} \Big[\frac{1}{1 - e^{-\frac{2\alpha}{K}} e^{i\frac{\pi}{K}(s-s')}} - \frac{K}{2\alpha - i\pi(s - s')}\Big] \\
&= \sum_{s,s'=-T}^{T} (-1)^{s+s'} \Big[\frac{2\alpha - i\pi(s - s') - K[1 - e^{-\frac{2\alpha}{K}} e^{i\frac{\pi}{K}(s-s')}]}{(1 - e^{-\frac{2\alpha}{K}} e^{i\frac{\pi}{K}(s-s')})(2\alpha - i\pi(s - s'))}\Big].
\end{aligned}$$

Let us finish the computation:

$$= \sum_{s,s'=-T}^{T} (-1)^{s+s'} \frac{2\alpha - i\pi(s-s') - [2\alpha - i\pi(s-s') - \frac{4\alpha^2}{2K} + \frac{\pi^2}{2K}(s-s')^2 + \frac{4i\pi\alpha}{K}(s-s') + o(\frac{1}{K})]}{(1 - e^{-\frac{2\alpha}{K}} e^{i\frac{\pi}{K}(s-s')})(2\alpha - i\pi(s-s'))}$$

$$= \sum_{s,s'=-T}^{T} \frac{(-1)^{s+s'}}{K} \times \frac{2\alpha^2 - \frac{\pi^2}{2}(s-s')^2 - 2i\pi\alpha(s-s') + o(1)}{(1 - e^{-\frac{2\alpha}{K}} e^{i\frac{\pi}{K}(s-s')})(2\alpha - i\pi(s-s'))}$$

$$= \sum_{n=-2T}^{2T} \frac{(-1)^n}{K} \times \frac{2\alpha^2 - \frac{\pi^2 n^2}{2} - 2i\pi\alpha n + o(1)}{(1 - e^{-\frac{2\alpha}{K}} e^{i\frac{\pi}{K}n})(2\alpha - i\pi n)}(2T + 1 - |n|) \text{ with the change of variable } n = s - s'$$

$$= \sum_{n=-2T}^{2T} \frac{(-1)^n}{K}(2T + 1 - |n|) f(n) \text{ with } f(n) = \frac{2\alpha^2 - \frac{\pi^2 n^2}{2} - 2i\pi\alpha n + o(1)}{(1 - e^{-\frac{2\alpha}{K}} e^{i\frac{\pi}{K}n})(2\alpha - i\pi n)}.$$

But notice that

$$f(n) = \frac{2\alpha^2 - \frac{\pi^2 n^2}{2} - 2i\pi\alpha n + o(1)}{(1 - e^{-\frac{2\alpha}{K}} e^{i\frac{\pi}{K}n})(2\alpha - i\pi n)} = \frac{K(2\alpha^2 - \frac{\pi^2 n^2}{2} - 2i\pi\alpha n + o(1))}{(2\alpha - i\pi n + o(1))(2\alpha - i\pi n)},$$

therefore

$$\sum_{n=-2T}^{2T} \frac{(-1)^n}{K}(2T + 1 - |n|) f(n) = \sum_{n=-2T}^{2T} (-1)^n (2T + 1 - |n|) \frac{2\alpha^2 - \frac{\pi^2 n^2}{2} - 2i\pi\alpha n + o(1)}{(2\alpha - i\pi n + o(1))(2\alpha - i\pi n)}$$

$$\sim \sum_{n=-2T}^{2T} (-1)^n (2T + 1 - |n|) \frac{-\frac{\pi^2 n^2}{2}}{-\pi^2 n^2} \sim \frac{1}{2},$$

using the same reasoning as for equation (23). This shows

$$\sum_{n=-2T}^{2T} \frac{(-1)^n}{K}(2T + 1 - |n|) f(n) = O(1). \tag{25}$$

We can therefore conclude:

$$\mathcal{L}_{\text{time}}(c,d) = 1 - \frac{e^{-2\alpha}(e^{2\alpha} - e^{-2\alpha})S}{K} + \frac{e^{-2\alpha}(e^{2\alpha} - e^{-2\alpha})^2}{4K^2} \times \sum_{s=-T}^{T} (-1)^s K \sum_{s'=-T}^{T} \frac{e^{i\pi s'}}{2\alpha - i\pi(s-s')} + O(\frac{1}{K^2})$$

$$= 1 - \frac{e^{-2\alpha}(e^{2\alpha} - e^{-2\alpha})S}{K} + \frac{e^{-2\alpha}(e^{2\alpha} - e^{-2\alpha})^2}{4K} \Big[\sum_{s=-T}^{T} (-1)^s \Big(\sum_{s'=-\infty}^{+\infty} \frac{(-1)^{s'}}{(2\alpha - i\pi s) + i\pi s'} + O(\frac{1}{T}) \Big) \Big] + O(\frac{1}{K^2})$$

$$= 1 - \frac{e^{-2\alpha}(e^{2\alpha} - e^{-2\alpha})S}{K} + \frac{e^{-2\alpha}(e^{2\alpha} - e^{-2\alpha})^2}{4K} \Big[\sum_{s=-T}^{T} (-1)^s \frac{2 \times (-1)^s}{e^{2\alpha} - e^{-2\alpha}} + O(\frac{1}{T}) \Big] + O(\frac{1}{K^2})$$

$$= 1 - \frac{e^{-2\alpha}(e^{2\alpha} - e^{-2\alpha})S}{K} + \frac{e^{-2\alpha}(e^{2\alpha} - e^{-2\alpha})S}{2K} + O(\frac{1}{TK})$$

$$= 1 - \frac{e^{-2\alpha}(e^{2\alpha} - e^{-2\alpha})S}{2K} + O(\frac{1}{TK}),$$

where we used Lemmas 20 and 21. ■

D.4. Proof of Theorem 7

In this section, we give the proof for Theorem 7 that describes the asymptotic behavior of the transfer function $C(e^{i\omega})$. First, we refer the reader to Lemmas 19 and 20 where we derive results on the remainder of some series. Then in Lemmas 22 and 23, we derive an asymptotic new form for the transfer function. We combine all these lemmas to prove Theorem 7.

Lemma 22 *For α real and positive and Ω real,*

$$\frac{1}{2}\sum_{s=-T}^{T}\frac{e^{-\alpha}(e^{2\alpha}-e^{-2\alpha})(-1)^s}{K(1-e^{-\alpha/K}e^{i\pi(\frac{s}{K}-\frac{\Omega}{K})})}=\frac{1}{2}\sum_{s=-T}^{T}\frac{e^{-\alpha}(e^{2\alpha}-e^{-2\alpha})(-1)^s}{\alpha-i\pi(s-\Omega)}+O(\frac{1}{K}),$$

as $T\to+\infty, T/K\to 0$.

Proof

$$\begin{aligned}
&\sum_{s=-T}^{T}\frac{(-1)^s}{K(1-e^{-\alpha/K}e^{i\pi(\frac{s}{K}-\frac{\Omega}{K})})}-\sum_{s=-T}^{T}\frac{(-1)^s}{\alpha-i\pi(s-\Omega)}\\
&=\sum_{s=-T}^{T}(-1)^s\Big[\frac{1}{K(1-e^{-\alpha/K}e^{i\frac{\pi}{K}(s-\Omega)})}-\frac{1}{\alpha-i\pi(s-\Omega)}\Big]\\
&=\sum_{s=-T}^{T}(-1)^s\frac{\alpha-i\pi(s-\Omega)-K(1-e^{-\alpha/K}e^{\frac{i\pi}{K}(s-\Omega)})}{K\big(1-e^{-\alpha/K}e^{i\frac{\pi}{K}(s-\Omega)}\big)\big(\alpha-i\pi(s-\Omega)\big)}\\
&=\sum_{s=-T}^{T}(-1)^s\frac{\alpha-i\pi(s-\Omega)-K\big(\frac{\alpha}{K}-\frac{i\pi}{K}(s-\Omega)-\frac{\alpha^2}{2K^2}+\frac{\pi^2}{2K^2}(s-\Omega)^2+\frac{i\pi\alpha}{K^2}(s-\Omega)+o(\frac{1}{K^2})\big)}{K\big(1-e^{-\alpha/K}e^{i\frac{\pi}{K}(s-\Omega)}\big)\big(\alpha-i\pi(s-\Omega)\big)}\\
&=\sum_{s=-T}^{T}(-1)^s\frac{\frac{\alpha^2}{2K}-\frac{\pi^2}{2K}(s-\Omega)^2-\frac{i\pi\alpha}{K}(s-\Omega)+o(\frac{1}{K})}{K\big(1-e^{-\alpha/K}e^{i\frac{\pi}{K}(s-\Omega)}\big)\big(\alpha-i\pi(s-\Omega)\big)}\\
&=\frac{1}{K}\sum_{s=-T}^{T}(-1)^s\frac{\frac{\alpha^2}{2}-\frac{\pi^2}{2}(s-\Omega)^2-i\pi\alpha(s-\Omega)+o(1)}{[\alpha-i\pi(s-\Omega)+o(1)][\alpha-i\pi(s-\Omega)]}=O(\frac{1}{K}).
\end{aligned}$$

We used a similar argument as the one in eq. (23) and eq. (25). The constant $\frac{e^{-\alpha}(e^{2\alpha}-e^{-2\alpha})}{2}$ does not impact our computation. ∎

Lemma 23 *For α real and positive,*

$$\frac{1}{2}\sum_{s=-T}^{T}\frac{(-1)^s e^{-\alpha}(e^{2\alpha}-e^{-2\alpha})}{\alpha-i\pi(s-\Omega)}\underset{\substack{S\to+\infty\\ S/K\to 0}}{\sim}\begin{cases}\frac{e^{-\alpha}(e^{2\alpha}-e^{-2\alpha})}{2}\times\frac{i(-1)^{T+1}\times 2\lfloor\Omega\rfloor}{2\pi(\lfloor\Omega\rfloor-T)(\lfloor\Omega\rfloor+T)} & \text{if } |\Omega|>T,\\ \frac{e^{-\alpha}(e^{2\alpha}-e^{-2\alpha})}{e^{\alpha}e^{i\pi\Omega}-e^{-\alpha}e^{-i\pi\Omega}} & \text{if } |\Omega|<T.\end{cases}$$

Proof *We decompose $\Omega=\lfloor\Omega\rfloor+\beta=n+\beta$, and look at two different cases.*

First case: $|\Omega| < T$. *We have:*

$$\frac{1}{2}\sum_{s=-T}^{T}(-1)^s\frac{e^{-\alpha}(e^{2\alpha}-e^{-2\alpha})}{\alpha-i\pi(s-n-\beta)}$$
$$=\frac{(-1)^n}{2}\sum_{s=-(T+n)}^{T-n}\frac{e^{-\alpha}(e^{2\alpha}-e^{-2\alpha})}{\alpha-i\pi(s-\beta)}=\frac{(-1)^n}{2}\sum_{s=-(T+n)}^{T-n}\frac{e^{-\alpha}(e^{2\alpha}-e^{-2\alpha})}{\tilde{\alpha}-i\pi s}\ \textit{where}\ \tilde{\alpha}=\alpha+i\pi\beta$$
$$=\frac{(-1)^n}{2}\sum_{s=-\infty}^{+\infty}\frac{e^{-\alpha}(e^{2\alpha}-e^{-2\alpha})}{\tilde{\alpha}-i\pi s}$$
$$+\Big[\frac{(-1)^n}{2}\sum_{s=-(T+n)}^{T-n}\frac{e^{-\alpha}(e^{2\alpha}-e^{-2\alpha})}{\tilde{\alpha}-i\pi s}-\frac{(-1)^n}{2}\sum_{s=-\infty}^{+\infty}\frac{e^{-\alpha}(e^{2\alpha}-e^{-2\alpha})}{\tilde{\alpha}-i\pi s}\Big]$$
$$=\frac{(-1)^n}{2}\sum_{s=-\infty}^{+\infty}\frac{e^{-\alpha}(e^{2\alpha}-e^{-2\alpha})}{\tilde{\alpha}-i\pi s}-\frac{i}{\pi}\times\frac{(-1)^{T-n+1}}{2(T-n+1)}+\frac{i}{\pi}\times\frac{(-1)^{T+n+1}}{2(T+n+1)}+o\Big(\frac{1}{T-n}\Big)\ \textit{(Lemma 20)}$$
$$=(-1)^n\frac{e^{-\alpha}(e^{2\alpha}-e^{-2\alpha})}{e^{\alpha}e^{i\pi\beta}-e^{-\alpha}e^{-i\pi\beta}}+O\Big(\frac{1}{T}\Big)=\frac{e^{-\alpha}(e^{2\alpha}-e^{-2\alpha})}{e^{\alpha}e^{i\pi\Omega}-e^{-\alpha}e^{-i\pi\Omega}}+O\Big(\frac{1}{T}\Big).$$

Second case: $|\Omega| > T$. *We consider again:*

$$\frac{(-1)^n}{2}\sum_{s=-(T+n)}^{T-n}\frac{e^{-\alpha}(e^{2\alpha}-e^{-2\alpha})}{\tilde{\alpha}-i\pi s}=\frac{(-1)^n e^{-\alpha}(e^{2\alpha}-e^{-2\alpha})}{2}\Big[\sum_{s=-\infty}^{T-n}\frac{1}{\tilde{\alpha}-i\pi s}-\sum_{s=-\infty}^{-T-n-1}\frac{1}{\tilde{\alpha}-i\pi s}\Big]$$
$$=\frac{(-1)^n e^{-\alpha}(e^{2\alpha}-e^{-2\alpha})}{2}\Big[\sum_{s=n-T}^{+\infty}\frac{1}{\tilde{\alpha}+i\pi s}-\sum_{s=T+n+1}^{+\infty}\frac{1}{\tilde{\alpha}+i\pi s}\Big]$$
$$=\frac{e^{-\alpha}(e^{2\alpha}-e^{-2\alpha})}{2}\times\frac{i(-1)^{T+1}\times 2n}{2\pi(n-T)(T+n)}+o\Big(\frac{1}{T}\Big)\quad\textit{(Lemma 20)}\,.$$

■

Proof of Theorem 7 For $\Omega\in\mathbb{R}$,

$$C(e^{i\omega})=\sum_{s=-T}^{T}\frac{e^{-\alpha}(e^{2\alpha}-e^{-2\alpha})}{2K}\frac{(-1)^s}{1-e^{-\alpha/K}e^{i\frac{\pi}{K}(s-\Omega)}}$$
$$=\sum_{s=-T}^{T}\frac{e^{-\alpha}(e^{2\alpha}-e^{-2\alpha})}{2}\times\frac{(-1)^s}{\alpha-i\pi(s-\Omega)}$$
$$+\Big[\sum_{s=-T}^{T}\frac{e^{-\alpha}(e^{2\alpha}-e^{-2\alpha})}{2K}\times\frac{(-1)^s}{1-e^{-\alpha/K}e^{i\frac{\pi}{K}(s-\Omega)}}-\sum_{s=-T}^{T}\frac{e^{-\alpha}(e^{2\alpha}-e^{-2\alpha})}{2}\frac{(-1)^s}{\alpha-i\pi(s-\Omega)}\Big].$$

We then use Lemma 22 and Lemma 23 to conclude.

■

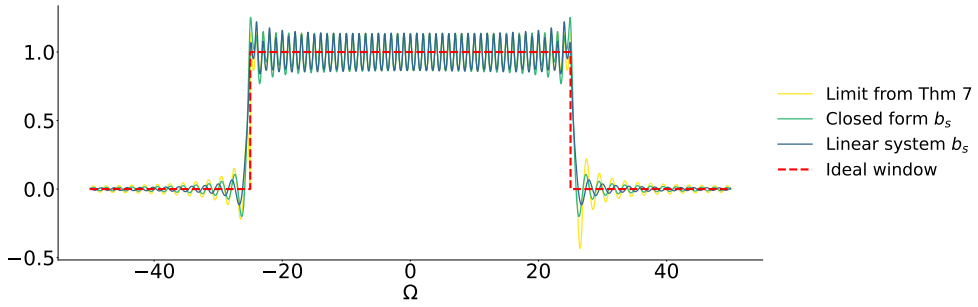

Figure D.2: *Behavior of* $\frac{C(e^{i\omega})}{D(e^{i\omega})}$*, where* $C(e^{i\omega})$ *is the Fourier transform of our near-optimal filter in Theorem 1 and* $D(e^{i\omega}) = e^{-iK\omega}$ *is the Fourier transform of our Shift-*K *filter. Perfect match between filters implies the ratio is* 1 *for all* ω*. If instead this equality holds in a window, then the filter would effectively act as a Shift-*K *for inputs with frequency content* $\Gamma(e^{-i\omega})$ *in the same window. For* $S = 51, K = 500$*, we denote* $T = \frac{S-1}{2}$ *and plot the ratio* $\frac{C(e^{i\omega})}{D(e^{i\omega})}$ *with respect to* $\Omega = \frac{K\omega}{\pi}$ *(to dilate the space). The asymptotic ratio* $\frac{C(e^{i\omega})}{D(e^{i\omega})}$ *(yellow) from Theorem 7, the same ratio for linear models with* (b_s) *given by Eq. (13) (green), and for* (b_s) *given by linear system inversion Eq. (14) (blue) are compared. The model effectively approximates the shift-*K *operation, within the frequency window* $[-\frac{\pi T}{K}, \frac{\pi T}{K}]$*, while vanishing outside this window, leading to a time resolution (inverse of filter width) of* S/K*. This behavior underscores the uncertainty principle associated with the filter: for small* S/K *ratios and uncorrelated data, the approximation holds over a narrow frequency range. As autocorrelation increases, the approximation domain shrinks, enhancing accuracy. In red, we show the perfect window (value of 1 on* $[-\frac{\pi T}{K}, \frac{\pi T}{K}]$ *and* 0 *outside).*

E. Experiments

In this section, we present a series of experiments designed to validate our theoretical findings in a practical setting. Specifically, we assess whether our conclusions hold when transitioning from an idealized infinite-data framework to real-world scenarios with a limited number of samples.

Let us first introduce the linear recurrent neural network (RNN) used in our study. It is defined by the following recurrence relations:

$$\begin{aligned} h_0 &= 0, \\ x_{t+1} &= Ax_t + Bu_{t+1}, \\ y_t &= Cx_t, \end{aligned}$$

where $x_t \in \mathbb{R}^{d_{\text{hidden}}}$ represents the hidden state, $u_t \in \mathbb{R}$ is the input, and $y_t \in \mathbb{R}$ is the output. The network parameters consist of $A \in \mathbb{C}^{d_{\text{hidden}} \times d_{\text{hidden}}}$, $B \in \mathbb{C}^{d_{\text{hidden}} \times 1}$, and $C \in \mathbb{C}^{d_{\text{hidden}}}$.

Without loss of generality, we adopt a diagonal representation for the matrix A. The choice of its initial eigenvalues depends on the specific experiment: we either use a random initialization or employ the structured initialization given by Eq. (12).

In the simple experiments conducted below, the objective is to learn a single filter. Consequently, there is no need to decompose the matrix A into multiple diagonal blocks. The matrix C is initialized as:

$$C_{\text{init}} = \begin{pmatrix} 1 & \dots & 1 \end{pmatrix} \in \mathbb{R}^{d_{\text{hidden}} \times 1}.$$

and the entries of B are initialized given by Eq. (13).

The synthetic dataset consists of autoregressive sequences $X = (u_1, u_2, \dots, u_N)$ of length N, generated as:

$$u_n = \rho u_{n-1} + \epsilon_k, \quad \epsilon_k \sim \mathcal{N}(0, 1-\rho^2), \quad u_1 \sim U(0,1). \tag{26}$$

The objective is to learn a mapping with linear recurrences $f : X \to Y$, where the target is given by:

$$Y = u_{t^*} \tag{27}$$

This corresponds to learning a shift of $N - t^*$ with finite samples.

E.1. Random initialization vs. Shift-K initialization

In this first set of experiments, we analyze the impact of initializing the complex diagonal entries a_s of the linear RNN using phases that are uniformly distributed over a segment of the unit disk, with a constant radius close to 1, as described in the parametrization in Eq. (12). Additionally, the parameters b_s are initialized following the parametrization given in Eq. (13). We call this initialization the shift-K initialization. We compare this approach to a standard random initialization to evaluate potential benefits in terms of performance and stability.

	Random init.	**Shift-K init.**
Batch size	[20, 50, 100]	[20, 50, 100]
Number Samples	130000	130000
Sequence length	1500	1500
Position of t^*	200	200
Hidden neurons	128	128
Input / output dimension	1	1
Learning rates	[0.01, 0.005, 0.001, 0.0001]	[0.01, 0.005, 0.001, 0.0001]
Weight decay	10^{-5}	10^{-5}
ρ	{0, 0.2, 0.4, 0.6, 0.8, 1}	{0, 0.2, 0.4, 0.6, 0.8, 1}
a_s param.	$a_u = e^{-\alpha/K_{\text{init}}} e^{i\epsilon_u \pi}, \varepsilon \sim \mathcal{U}(-1,1)$	$a_u = e^{-\alpha/K_{\text{init}}} e^{iu\frac{\pi}{K_{\text{init}}}}$
b_s param.	$b_u = \frac{e^{-\alpha}(e^{2\alpha}-e^{-2\alpha})}{2K_{\text{init}}} \times (-1)^u$	$b_u = \frac{e^{-\alpha}(e^{2\alpha}-e^{-2\alpha})}{2K_{\text{init}}} \times (-1)^u$
α	1	1
K_{init}	1300	1300
Number epochs	60	60

Table E.1: Experimental details for Figure 3 (left). We use [...] to denote hyperparameters that were scanned over with grid search and {...} to denote the variable of interest for the figure. We chose the same representation for b_s in both cases because we observed small impact of this parameter on the final results.

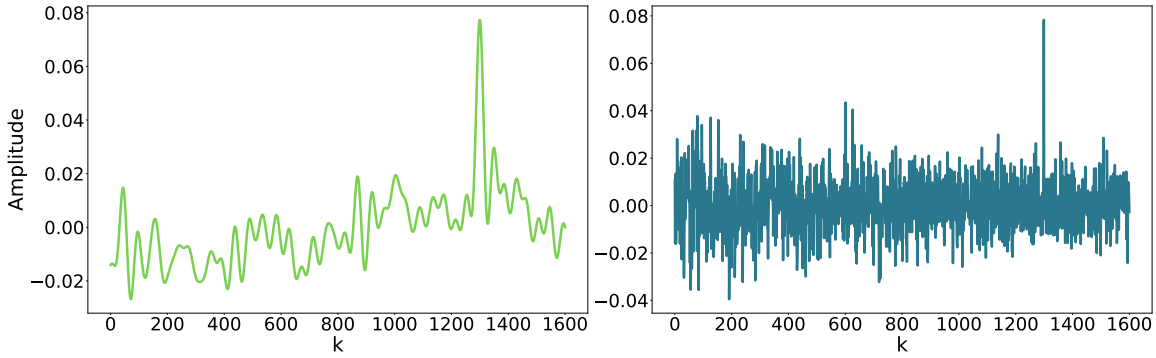

Figure E.1: *Comparison of Filters Obtained with Different Initialization Methods. Left: Filter obtained using our proposed shift-K initialization method, which exhibits a more structured and interpretable pattern. Right: Filter obtained with random initialization, which appears significantly noisier, indicating less effective memory propagation.*

E.2. Robustness of Shift-K initialization

In this second set of experiments, we investigate the robustness of our initialization scheme with respect to inaccuracies in the choice of K_{init} when initializing a_s as in Eq. (12). In practical applications, the actual shift of the sequence is often unknown, making it impossible to initialize with the exact optimal value of K. A robust initialization method should exhibit resilience to such uncertainties, allowing for performance stability within a reasonable range of K_{init} values.

Shift-K init.
Batch size: [20, 50, 100]
Number of Samples: 150000
Sequence length: 2250
Position of t^*: 250
Hidden neurons: 128
Input / output dimension: 1
Learning rates: [0.01, 0.005, 0.001, 0.0001]
Weight decay: 10^{-5}
ρ: 0.7
a_s param.: $a_u = e^{-\alpha/K_{\text{init}}} e^{iu\frac{\pi}{K_{\text{init}}}}$
b_s param.: $b_u = \frac{e^{-\alpha}(e^{2\alpha}-e^{-2\alpha})}{2K_{\text{init}}} \times (-1)^u$
α: 1
K_{init}: $\{250, 500, 1000, 2000, 4000, 8000, 16000, 32000\}$
Number of epochs: 60

Table E.2: Experimental details for Figure 3 (right).