The Power of Second Order Methods for Sequence Preconditioning

Annie Marsden * **Elad Hazan** * †

Abstract

Sequence prediction methods for linear dynamical systems (LDSs) with long memory, i.e. marginally stable systems, typically achieve regret that grows polynomially with the hidden dimension of the underlying generative model. While many methods have been developed to address this regime with varying success, we show that simply using the second-order Vovk-Azoury-Warmuth (VAW) algorithm to learn a short autoregressive-with-inputs (ARX) model achieves astoundingly strong results: for bounded and noiseless sequential data from a marginally-stable linear dynamical system with spectra in the complex unit disk except for angular wedge of width δ around the negative real axis, this algorithm achieves dimension-free regret $\mathcal{O}\big(\delta^{-4}\log^2 T\big)$ (with only logarithmic dependence on system norms and the diagonalization condition number). These bounds are state-of-the-art to our knowledge.

The key components for our result come from 1) reformulating "Universal Sequence Preconditioning" (USP) [20] to prove the existence of an optimal setting of autoregressive coefficients, 2) applying VAW, which scales linearly with memory length and only logarithmically with comparator norm, and 3) analyzing Faber polynomials on circular sectors to extend these results to systems with complex spectra.

1 Introduction

A central challenge in sequence prediction is the simultaneous presence of long memory and high hidden dimension of the underlying generative process. The canonical theoretical model to study in this setting is a marginally stable linear dynamical system LDS. Most methods applied to this setting result in a regret bound which grows polynomially with hidden dimension, see for example [14, 17, 4, 12, 23, 24, 18, 2].

Universal Sequence Preconditioning (USP) [20] was the first method, to our knowledge, which simultaneously achieved dimension-free sublinear regret, even in the presence of asymmetric hidden transition matrices with complex eigenvalues. The method they develop is to convolve the original signal with the coefficients of the Chebyshev polynomial, reducing the effective memory to be logarithmic in the horizon T. However, this comes at a cost: the Chebyshev coefficients grow exponentially large with the degree, resulting in large optimal diameter. This creates a mismatch with first-order learning methods, which scale linearly with optimal diameter, yielding suboptimal regret of $\tilde{\mathcal{O}}(T^{11/13})$ [20]. Moreover, while their work did apply to hidden transition matrices with complex eigenvalues, they required the assumption that the complex argument of the eigenvalues vanish at the rate $o(1/\mathrm{poly}\log T)$.

In this work, we take the insight of memory compression from Universal Sequence Preconditioning and use it to design a stronger algorithm which achieves dimension-free and poly-logarithmic cumu-

*Google DeepMind
†Princeton University

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

lative prediction loss for a much broader class of linear dynamical systems. This is an improvement upon all other known methods, illustrated in Table 1.

Dependence on T	**Dependence on d**	**Complex Angle ϕ**	**Methods**
$T^{O(1)}$	$d^{O(1)}$	π	**Direct prediction methods:** Hazan et al. (2018) [14], Cayley-Hamilton and regret minimization (folklore, e.g. [17, 4, 12]) **Recovery methods:** Simchowitz et al. (2019) [23]; Tsiamis & Pappas (2019) [24]; Lee (2022) [18]; Bakshi et al. (2023) [2]
$T^{O(1)}$	$(\log d)^{O(1)}$	$1/\log(T)^{O(1)}$	Marsden & Hazan (2025) [19]
$(\log T)^{O(1)}$	$(\log d)^{O(1)}$	$\pi - \delta$	**Ours**

Table 1: Cumulative prediction error bounds for *marginally stable* linear dynamical systems with *asymmetric* transition matrices. For "Complex Angle ϕ", the results apply when the hidden transition matrix has spectra in the wedge $W_\phi = \{z \in \mathbb{C} : |z| \le 1,\ |\arg(z)| \le \phi\}$ (suppressing dependence on diagonalization conditioning κ and system norms).

Consider a noiseless SISO linear dynamical system with inputs $u_t \in \mathbb{R}$ and outputs $y_t \in \mathbb{R}$ evolving according to:

$$\mathbf{h}_t = \mathbf{A}\mathbf{h}_{t-1} + \mathbf{B}u_t, \qquad y_t = \mathbf{C}\mathbf{h}_t, \qquad \mathbf{h}_0 = \mathbf{0}, \tag{1}$$

where $\mathbf{h}_t \in \mathbb{R}^{d_{\text{hidden}}}$ is the hidden state (with convention $y_s = u_s = 0$ for $s \le 0$). The algorithm we suggest is to predict y_t using the Vovk–Azoury–Warmuth (VAW) forecaster [1, 26] with a short autoregressive-with-inputs (ARX) feature vector $[y_{t-1}, \ldots, y_{t-n}, u_t, \ldots, u_{t-n+1}]$, where n grows logarithmically with the prediction horizon T; see Algorithm 2 for details. By comparison, the original USP algorithm predicts y_t as $-\sum_{i=1}^{n} c_i y_{t-i} + \sum_{j=0}^{n-1} \theta_j u_{t-j}$, where $c_1, \ldots, c_n$ are fixed Chebyshev coefficients and θ_j is learned with a first-order method.

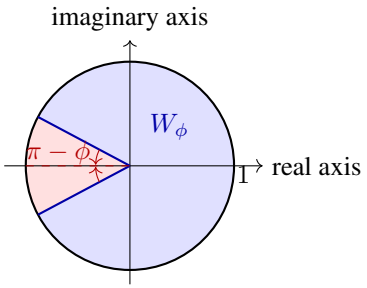


Figure 1: The wedge spectral class $W_\phi = \{z \in \mathbb{C} : |z| \le 1,\ |\arg(z)| \le \phi\}$. In our main result $\phi = \pi - \delta$, so the spectrum may occupy the unit disk except for a symmetric angular gap of width 2δ around the negative real axis.

At a high level, the effectiveness of a preconditioner $c_1, \ldots, c_n$ is governed by two quantities: the effective memory of the transformed representation and the magnitude of the optimal diameter. We can think of USP as a compressed representation of the original signal. The compressed representation requires a history which is linear in the degree of the Chebyshev polynomial, however the coefficient magnitude grows very large. This evokes the Minimum Description Length (MDL) principle [22, 10, 3], which seeks to minimize the combined cost of a model's structural complexity and the bit-length required to represent its parameters. For polynomials, the minimum description length principle suggests that the most efficient representation balances the degree of the polynomial linearly and the coefficients magnitude logarithmically.

Inspired by this principle, we observe that the regret of the VAW forecaster [1, 26] matches this balance: it grows linearly with the effective memory and only logarithmically with the the optimal diameter. In order to apply VAW to this setting, we no longer explicitly "precondition" the signal offline; instead we include $y_{t-1}, \ldots, y_{t-n}$ directly in the feature vector and learn the autoregressive transformation online. USP provides an analytical certificate guaranteeing a low-error short-memory

comparator. By moving to the second order VAW algorithm we improve from a cumulative prediction error of $\mathcal{O}(T^{11/13})$ to $\mathcal{O}\big(\log^2(T)\big)$. Moreover, USP [20] only applies to systems whose hidden transition matrix has spectra close to the real line, the complex argument must be bounded by $1/\mathrm{poly}\log(T)$. Our new approach allows us to consider spectra on the entire complex disk.

In order to show this, we require a new way to guarantee an optimal solution that extends to $W_{\pi-\delta}$, as Chebyshev only applies to $W_{1/\mathrm{poly}\log(T)}$. The solution comes from Faber polynomials adapted to circular sectors. Our analysis gives a deeper understanding of the role of complex spectra in the difficulty of learning a linear dynamical system. More generally, every useful USP preconditioner is a monic polynomial that is small on the spectrum of the hidden transition matrix. Chebyshev polynomials are optimal for nearly real spectra; Faber polynomials are the corresponding extremal objects for more general compact spectral sets, and in this paper we use them for circular sectors.

1.1 Our Results

Our main contributions are as follows:

Polylogarithmic Regret for Sequence Prediction: We propose a second-order finite-history predictor: VAW run on the raw feature vector $(y_{t-1}, \ldots, y_{t-n}, u_t, \ldots, u_{t-n+1})$ with an explicit regularization schedule $\lambda = 1/U_n$ (see Algorithm 2). In Theorem 2 we show that if there exists a degree-n monic polynomial with residual size α_n on the spectrum and coefficient scale G_n, then there exists a comparator for the raw-feature VAW problem, and VAW suffers only $\mathcal{O}(TL_0^2\alpha_n^2 + Y^2 n \log(G_n Y L_0))$ cumulative prediction loss. For comparison, the original first-order method obtains $\mathcal{O}(n^2 G_n\sqrt{T} + \alpha_n T^2)$ cumulative prediction loss (see Theorem D.1 in [20]). Instantiating our theorem with the Chebyshev polynomial gives Corollary 1: for spectra in $[-1, 1]$, the cumulative prediction loss is $\mathcal{O}\big(Y^2 \log^2(TYL_0)\big)$, vastly improving on $\tilde{\mathcal{O}}(T^{11/13})$ from the first-order method.

Extension to Circular-Sector Complex Spectra: We extend Universal Sequence Preconditioning to a strictly broader class of dynamical systems using Faber polynomials for circular sectors [5, 8]. Indeed, let $W_\phi := \{z \in \mathbb{C} : |z| \le 1 \text{ and } \arg(z) \le \phi\}$. Suppose the underlying hidden transition matrix of the data has spectra in $W_{\pi-\delta}$. Lemma 1 shows that for every angular gap $\delta \in (0, \pi)$ there is a real monic polynomial $p_{n,\delta}$ satisfying $\sup_{z\in W_{\pi-\delta}} |p_{n,\delta}(z)| \le 2e^{-n\delta^2/\pi^2}$ and $\max_i |c_i| \le 100^n$, yielding $\mathcal{O}(Y^2\delta^{-4}\log^2(TYL_0))$ cumulative loss across $W_{\pi-\delta}$ (Corollary 2) with memory $n = \Theta(\delta^{-2}\log(TL_0))$. Thus the approximation error remains exponentially small even when the hidden dynamics have large constant complex arguments, provided the spectrum avoids a constant wedge around the negative real axis. Note that this contribution is orthogonal to VAW: one could also apply Lemma 1 to first-order USP to broaden its spectral scope.

1.2 Tightness and the Role of the Angular Gap

Our regret bounds are dimension-free (up to polylogarithmic factors). Dimension-free regret bounds necessitate assuming the system's spectra is in the circular sector (Figure 1). Indeed, if the hidden transition matrix is the shift-permutation matrix (which has spectra in W_π) then the regret of any algorithm must grow linearly with hidden dimension [15]. However, we show that this adversarial example truly relies on having spectra in W_π. Instead, if the spectra is contained in $W_{\pi-\delta}$ for any constant $\delta > 0$, dimension-free regret is attainable.

When the goal is dimension-free prediction, Faber or Chebyshev polynomial preconditioners compress memory exponentially provided the spectrum obeys angular separation from -1. If one allows polynomial dependence on d, the classical Cayley–Hamilton representation uses memory d for arbitrary spectra; meanwhile, our bound $\tilde{\mathcal{O}}(Y^2\delta^{-4})$ already achieves sublinear-in-d error whenever $\delta = \omega(d^{-1/4})$. The remaining gap is quantitative: our wedge guarantee scales as $1/\delta^4$ in loss ($1/\delta^2$ in memory), whereas the permutation lower bound only rules out memory below order $1/\delta$. Closing this polynomial gap remains open.

1.3 Related Work

Our work is most closely related to the literature on prediction in latent linear dynamical systems, rather than recovery. Early spectral-filtering results demonstrated online competitiveness with the

best LDS predictor without explicit system identification, but relied on strong spectral structure like symmetry [16]. [14] later extended spectral filtering to general latent LDSs using a convex relaxation that accommodates phases and no longer requires a symmetric transition matrix; however, in the asymmetric setting their guarantee still has $\mathrm{poly}(T, d)$ dependence (as also obtained via Cayley–Hamilton realization theory [17, 4]).

A second line of work studies *marginally stable* systems more directly. [9] proved no-regret prediction guarantees for marginally stable systems with polynomial system parameter dependence. In the fully observed adversarial setting their bounds remain polynomial in T, while their polylogarithmic-in-T results rely either on stochastic assumptions or on a reduction to finite-memory autoregressive prediction. Related Kalman-filtering works obtain polylogarithmic regret under stochastic noise relative to the Kalman predictor: SLIP assumes the closed-loop matrix $A - KC$ is diagonalizable with real eigenvalues [21], and Tsiamis and Pappas obtain logarithmic regret against the Kalman filter for nonexplosive systems [25]. By contrast, our focus is absolute cumulative squared prediction error for deterministic input–output LDS trajectories.

The direct asymmetric comparison closest in spirit to our setting is Marsden and Hazan [19], which removes hidden-dimension dependence for asymmetric marginally stable LDSs, but retains $\tilde{\mathcal{O}}(T^{11/13})$ horizon dependence and requires $|\arg(z)| \le o(1/\mathrm{poly}\log T)$. Finally, system-identification and recovery methods [23, 24, 18, 2] incur $\mathrm{poly}(d)$ cumulative prediction error, and logarithmic-in-d identification results such as [7] require strict stability. Of particular interest is [11], who curiously obtain a "Pacman shape" identical to our wedge $W_{\pi-\delta}$ as the domain where gradient descent learns LDS parameters, whereas here $W_{\pi-\delta}$ emerges as the domain admitting dimension-free polynomial preconditioning for second-order ARX learning. In short, prior work obtains at most two of marginal stability, asymmetric complex spectra, and $\mathrm{polylog}(T)$ dimension-free error.

In online convex optimization, two prominent second-order algorithms for exp-concave and squared losses are Online Newton Step (ONS) [13] and the Vovk–Azoury–Warmuth (VAW) forecaster [26, 1]. Standard ONS guarantees depend explicitly on the decision-set diameter, which becomes vacuous when applied directly to exponentially large ARX comparators (unless one could establish uniform bounds on the preconditioned signal norm; see Section 4.1). By contrast, VAW requires no prescribed decision-set diameter and depends only logarithmically on the comparator norm $\|\mathbf{x}^*\|_2$, absorbing the exponential blowup of the preconditioner's coefficients. Exploring other second-order methods such as online BFGS or FTRL remains an open direction.

2 Theoretical Results Overview

2.1 Overview of Universal Sequence Preconditioning and Linear ARX Reformulation

For clarity of presentation, we consider the SISO LDS in Eq. (1) with $\mathbf{h}_0 = \mathbf{0}$ and $y_s = u_s = 0$ for $s \le 0$. Unrolling the hidden state $\mathbf{h}_t$ gives $y_t = \sum_{s=0}^{t-1} \mathbf{C}\mathbf{A}^s\mathbf{B}u_{t-s}$. For $\theta_s = \mathbf{C}\mathbf{A}^s\mathbf{B}$, predicting y_t directly from past inputs up to horizon T therefore requires learning T Markov parameters.

The core insight of Universal Sequence Preconditioning [20] is that y_t can be represented, up to negligible error, using only $\mathcal{O}(\log T)$ parameters and history. Let $c_0, c_1, \ldots, c_n$ be the coefficients of any degree-n monic polynomial $p_n(z) = z^n + \sum_{i=1}^{n} c_i z^{n-i}$, with $c_0 = 1$. Algebraic manipulation shows that for all $t \ge 1$ (with the third sum zero for $t \le n$),

$$y_t = \underbrace{-\sum_{i=1}^{n} c_i y_{t-i}}_{\aleph_0} + \underbrace{\sum_{s=0}^{n-1} \left(\sum_{i=0}^{s} c_i \mathbf{C}\mathbf{A}^{s-i}\mathbf{B} \right) u_{t-s}}_{\aleph_1} + \underbrace{\sum_{s=0}^{t-n-1} \mathbf{C} p_n(\mathbf{A}) \mathbf{A}^s \mathbf{B} u_{t-n-s}}_{\aleph_2}. \tag{2}$$

Since the monic degree-n Chebyshev polynomial satisfies $\max_{x\in[-1,1]} |M_n(x)| \le 2^{1-n}$ (extended in [19] to $|\arg(z)| \le 1/(64n^2)$), $\aleph_2$ becomes negligible for $n = \mathcal{O}(\log T)$. Whereas [19] applies fixed Chebyshev coefficients directly to y_t, our key insight and framing is to view Eq. (2) as certifying a $2n$-dimensional linear ARX predictor on raw history: defining

$$\theta_s^p := \sum_{i=0}^{s} c_i \mathbf{C}\mathbf{A}^{s-i}\mathbf{B}, \quad \text{with } c_0 = 1, \tag{3}$$

gives the linear regression relation

$$y_t = -\sum_{i=1}^{n} c_i y_{t-i} + \sum_{s=0}^{n-1} \theta_s^p u_{t-s} + \epsilon_t. \tag{4}$$

For diagonalizable $\mathbf{A} = \mathbf{P}\Lambda\mathbf{P}^{-1}$ with $\kappa = ||\mathbf{P}||_2||\mathbf{P}^{-1}||_2$, $\lambda(\mathbf{A}) \subseteq K \subseteq \{z \in \mathbb{C} : |z| \leq 1\}$, and $|u_t| \leq 1$, any monic polynomial p_n yields $|\epsilon_t| \leq T||\mathbf{C}||_2||\mathbf{B}||_2\kappa \sup_{z\in K} |p_n(z)|$. Taking $K = W_{\pi-\delta}$, sector Faber polynomials achieve exponential decay across constant complex angles.

2.2 Contribution One: The Power of a Second Order Method

Equation (4) suggests learning the full vector $(-c_1, \ldots, -c_n, \theta_0^p, \ldots, \theta_{n-1}^p)$ via online linear regression. Naive first-order methods suffer regret linear in this vector's norm, which grows exponentially with n: $\max_i |c_i|$ grows as $2^{0.3n}$ for Chebyshev preconditioning (Lemma 3.2 of [20]) and is bounded by 100^n for sector Faber preconditioning (Lemma 1). Although θ^p can have a tiny dimension which is only logarithmic in T, its norm does not reflect the reduction in domain. This is where the insight of our work comes in. Inspired by the minimum description length principle, we consider applying a second order learning method to sequence preconditioning. Critically, second-order methods like VAW do not treat the *dimension* and the *norm* of the comparator as equal citizens. We conjecture that exponential coefficient blowup is inherent to any exponentially small monic polynomial:

Conjecture 1. *Let $p_n(z) = z^n + \sum_{j=0}^{n-1} c_j z^j$ denote any monic polynomial, and let $\mathcal{D} \subseteq \mathbb{C}$ be convex and compact with area $D_A > 0$. If $\max_{z\in\mathcal{D}} |p_n(z)| \leq C \cdot 2^{-cn}$ for constants $C, c > 0$ and all $n \geq 1$, then there exist constants $B, b > 0$ such that $\max_{0\leq j\leq n-1} |c_j| \geq B2^{bn}$ for all $n \geq 1$.*

In Appendix A we prove a special case on $[-1, 1]$ at the Chebyshev scale ($c = 1$). Regardless of Conjecture 1, VAW pays only logarithmically in these large coefficients while scaling linearly in $2n$, establishing a clean division of labor between polynomial representation and second-order optimization.

2.3 Contribution Two: Extending to Asymmetric Systems Via Faber Polynomials

Because Theorem 2 is agnostic to the choice of polynomial, extending USP from vanishing angles $|\arg(z)| \leq o(1/\mathrm{poly}\log T)$ to the sector $W_{\pi-\delta}$ in Figure 1 requires only plugging in the circular-sector Faber polynomial bounds of Lemma 1 ($\alpha_n = 2e^{-n\delta^2/\pi^2}$, $G_n = 100^n$). Prior USP analyses required eigenvalue arguments to shrink with T; by contrast, for any constant $\delta > 0$, the Faber preconditioner remains exponentially small across $W_{\pi-\delta}$ while its 100^n coefficient bound enters VAW only logarithmically, yielding the $\pi - \delta$ guarantee in Table 1. Note that the algorithm executed in all cases is the exact same raw-feature VAW forecaster.

This contribution is particularly important for understanding when dimension-free regret is possible. The dimension-dependent lower bound from [15] arises when the system's hidden transition matrix $\mathbf{A}$ has for its spectrum the d roots of unity. Moreover, if regret bounds with dimension dependence are allowed, the classical Cayley–Hamilton representation gives a degree-d autoregressive predictor for arbitrary spectra. Our goal is different: to understand when dimension-free prediction is possible. Our result shows that the complexity of learning an asymmetric system grows in accordance with δ (where the system's spectra lies on $W_{\pi-\delta}$.

3 Main Theorems and Proofs

Theorem 1 states the standard cumulative squared-loss bound of the Vovk–Azoury–Warmuth (VAW) forecaster [1, 26] with explicit regularization parameter $\lambda > 0$.

Theorem 1 (VAW Regret [1, 26])**.** *Assume $|y_t| \leq Y$ and $||\boldsymbol{a}_t||_2 \leq R$ for all $t = 1, \ldots, T$, where $\boldsymbol{a}_t \in \mathbb{R}^m$. For any regularization $\lambda > 0$ and any benchmark $\boldsymbol{x}^* \in \mathbb{R}^m$, VAW predictions $\hat{y}_t = \boldsymbol{a}_t^\top (\lambda I_m + \sum_{s=1}^{t} \boldsymbol{a}_s \boldsymbol{a}_s^\top)^{-1} (\sum_{s=1}^{t-1} y_s \boldsymbol{a}_s)$ satisfy:*

$$\sum_{t=1}^{T} (\hat{y}_t - y_t)^2 \leq \sum_{t=1}^{T} (\boldsymbol{x}^{*\top} \boldsymbol{a}_t - y_t)^2 + \lambda \|\boldsymbol{x}^*\|_2^2 + Y^2 m \log\left(1 + \frac{TR^2}{\lambda m}\right). \tag{5}$$

Algorithm 1 Vovk-Azoury-Warmuth (VAW) Forecaster [1, 26]

Input: regularization $\lambda \geq 1/\|\mathbf{x}^*\|_2^2$
Initialize: $\mathbf{A}_0 = \lambda I_d$, target vector $\mathbf{v}_0 = 0 \in \mathbb{R}^d$
for $t = 1$ to T **do**
    Observe feature vector $\mathbf{a}_t$
    Update curvature matrix: $\mathbf{A}_t = \mathbf{A}_{t-1} + \mathbf{a}_t\mathbf{a}_t^\top$
    Predict $\hat{y}_t = \mathbf{a}_t^\top \mathbf{A}_t^{-1}\mathbf{v}_{t-1}$
    Observe true label y_t and suffer loss $l_t(\hat{b}_t) = \frac{1}{2}(\hat{b}_t - b_t)^2$
    Update target vector: $\mathbf{v}_t = \mathbf{v}_{t-1} + b_t\mathbf{a}_t$
end for

Transforming sequence prediction with short memory into linear prediction with small dimension: In Algorithm 2 we convert sequence prediction into the VAW framework with dimension $m = 2n$. Let $\mathbf{a}_t = [y_{t-1}, \ldots, y_{t-n}, u_t, \ldots, u_{t-n+1}]^\top \in \mathbb{R}^{2n}$ collect the n most recent observations and n most recent inputs (with $y_s = u_s = 0$ for $s \leq 0$). Defining the comparator $\mathbf{x}^* = [-c_1, \ldots, -c_n, \theta_0^p, \ldots, \theta_{n-1}^p]^\top \in \mathbb{R}^{2n}$ from Eq. (3) gives

$$y_t = \mathbf{x}^{*\top}\mathbf{a}_t + \epsilon_t. \tag{6}$$

Algorithm 2 Vovk-Azoury-Warmuth (VAW) Forecaster for Sequence Preconditioning

Input: memory length $n \geq 1$, regularization $\lambda > 0$
Initialize: $\mathbf{S}_0 = \lambda I_{2n}$, target vector $\mathbf{v}_0 = \mathbf{0} \in \mathbb{R}^{2n}$, with $y_s = u_s = 0$ for $s \leq 0$
for $t = 1$ to T **do**
    Observe input u_t and create feature vector $\mathbf{a}_t \leftarrow [y_{t-1}, \ldots, y_{t-n}, u_t, \ldots, u_{t-n+1}]^\top$
    Update curvature matrix: $\mathbf{S}_t = \mathbf{S}_{t-1} + \mathbf{a}_t\mathbf{a}_t^\top$
    Predict $\hat{y}_t = \mathbf{a}_t^\top \mathbf{S}_t^{-1}\mathbf{v}_{t-1}$
    Observe true label y_t, suffer squared loss $(\hat{y}_t - y_t)^2$, and update $\mathbf{v}_t = \mathbf{v}_{t-1} + y_t\mathbf{a}_t$
end for

Theorem 2 (VAW with a polynomial comparator certificate)**.** *Suppose $y_1, \ldots, y_T$ comes from Eq.* (1) *with $\boldsymbol{h}_0 = \boldsymbol{0}$. Assume $\boldsymbol{A} = \boldsymbol{P}\Lambda\boldsymbol{P}^{-1}$ with diagonal Λ, $\max_t |y_t| \leq Y$ for some $Y \geq 1$, and $\max_t |u_t| \leq 1$. Let $K \subseteq \{z \in \mathbb{C} : |z| \leq 1\}$ contain $\lambda(\boldsymbol{A})$, and let $p_n(z) = z^n + \sum_{i=1}^n c_i z^{n-i}$ ($c_0 = 1$) be a monic preconditioning polynomial satisfying*

$$\sup_{z \in K} |p_n(z)| \leq \alpha_n, \qquad \max_{0 \leq i \leq n} |c_i| \leq G_n,$$

where $G_n \geq 1$. Denote $\kappa = ||\boldsymbol{P}||_2||\boldsymbol{P}^{-1}||_2$, let $C_0 \geq ||\boldsymbol{C}||_2, B_0 \geq ||\boldsymbol{B}||_2, \kappa_0 \geq \kappa$ be known upper bounds satisfying $C_0, B_0, \kappa_0 \geq 1$, and define

$$L_0 = 1 + T(1 + \kappa_0 C_0 B_0), \qquad U_n = 2n^3\kappa_0^2 C_0^2 B_0^2 G_n^2.$$

Then Algorithm 2, run with memory length $n \leq T$ and regularization $\lambda = 1/U_n$, satisfies

$$\sum_{t=1}^T (\hat{y}_t - y_t)^2 \leq TL_0^2\alpha_n^2 + 1 + 2Y^2 n \log(1 + TY^2 U_n) = \mathcal{O}\big(TL_0^2\alpha_n^2 + Y^2 n \log(G_n Y L_0)\big).$$

Proof. By Eq. 4, $y_t = -\sum_{i=1}^n c_i y_{t-i} + \sum_{s=0}^{n-1} \theta_s^p u_{t-s} + \epsilon_t$, with $\epsilon_t = 0$ for $t \leq n$ and $\epsilon_t = \sum_{s=0}^{t-n-1} \mathbf{C}p_n(\mathbf{A})\mathbf{A}^s\mathbf{B}u_{t-n-s}$ for $t \geq n+1$. Because $|\lambda_j(\mathbf{A})| \leq 1$,

$$||p_n(\mathbf{A})\mathbf{A}^s||_2 \leq \kappa \max_{z \in \lambda(\mathbf{A})} |p_n(z)z^s| \leq \kappa\alpha_n,$$

so $|\epsilon_t| \leq T||\mathbf{C}||_2\kappa\alpha_n||\mathbf{B}||_2 \leq L_0\alpha_n$. For $\mathbf{x}^* = [-c_1, \ldots, -c_n, \theta_0^p, \ldots, \theta_{n-1}^p]^\top$, $\mathbf{x}^{*\top}\mathbf{a}_t = y_t - \epsilon_t$, so $\sum_{t=1}^T (\mathbf{x}^{*\top}\mathbf{a}_t - y_t)^2 \leq TL_0^2\alpha_n^2$. Since $||\mathbf{A}^j||_2 \leq ||\mathbf{P}||_2||\Lambda^j||_2||\mathbf{P}^{-1}||_2 \leq \kappa$ for all $j \geq 0$, each $0 \leq s < n$ satisfies

$$|\theta_s^p| = \left|\sum_{i=0}^s c_i \mathbf{C}\mathbf{A}^{s-i}\mathbf{B}\right| \leq nG_n\kappa_0 C_0 B_0, \quad \text{so} \quad \|\mathbf{x}^*\|_2^2 \leq nG_n^2 + n^3 G_n^2\kappa_0^2 C_0^2 B_0^2 \leq U_n.$$

Hence $\lambda\|\mathbf{x}^*\|_2^2 \le 1$. The feature vectors in $\mathbb{R}^{2n}$ satisfy $||\mathbf{a}_t||_2 \le \sqrt{n(Y^2+1)} \le \sqrt{2n}Y =: R$. Applying Theorem 1 with $m = 2n$ and $\lambda = 1/U_n$, and noting that $n \le T \le L_0$ and $\kappa_0 C_0 B_0 \le L_0$ imply $\log(1 + TY^2U_n) = \mathcal{O}(\log(G_n Y L_0))$, yields the claim. □

First we show how Theorem 2 substantially improves upon the first-order method from [19] when instantiated with Chebyshev polynomials.

Corollary 1 (Chebyshev instantiation for real spectra)**.** *Under the setting of Theorem 2 with $\lambda(\boldsymbol{A}) \subseteq [-1,1]$, set $n = \lceil \log_2(4TL_0^2) \rceil$ and $\lambda = 1/U_n$ with $G_n = 2^{0.3n}$. Then Algorithm 2 achieves*

$$\sum_{t=1}^{T} (\hat{y}_t - y_t)^2 \le \mathcal{O}\left(Y^2 n \left[n + \log(YL_0)\right]\right) = \mathcal{O}\left(Y^2 \log^2(TYL_0)\right).$$

Proof. If $n > T$, $|y_t| \le Y$ trivially implies $\mathcal{O}(TY^2) = \mathcal{O}(Y^2 n)$ cumulative loss. For $n \le T$, apply Theorem 2 with the degree-n monic Chebyshev polynomial $M_n(z)$, which satisfies $\sup_{z\in[-1,1]} |M_n(z)| \le 2^{1-n} =: \alpha_n$ and $\max_{0\le i\le n} |c_i| \le 2^{0.3n} =: G_n$ (Lemma 3.2 of [19]). Then $TL_0^2\alpha_n^2 = 4TL_0^2 2^{-2n} \le 1$ and $\log G_n = \mathcal{O}(n)$, giving $\mathcal{O}(Y^2 n[n + \log(YL_0)]) = \mathcal{O}(Y^2 \log^2(TYL_0))$. □

Lemma 1 (Faber polynomial bounds for wedge spectra; proof in Appendix B)**.** *For every $\delta \in (0, \pi)$ and $n \ge 1$, there is a real monic polynomial $p_{n,\delta}(z) = z^n + \sum_{i=1}^{n} c_i z^{n-i}$ ($c_0 := 1$) such that*

$$\sup_{z\in W_{\pi-\delta}} |p_{n,\delta}(z)| \le 2e^{-n\delta^2/\pi^2}, \qquad \max_{0\le i\le n} |c_i| \le 100^n.$$

Corollary 2 (Main result: VAW with Faber preconditioning)**.** *Under the setting of Theorem 2 with $\lambda(\boldsymbol{A}) \subseteq W_{\pi-\delta}$ for some $\delta \in (0,\pi)$, set $n = \left\lceil \frac{\pi^2}{\delta^2} \log(4T^2L_0) \right\rceil$ and $\lambda = 1/U_n$ with $G_n = 100^n$. Then Algorithm 2 satisfies*

$$\sum_{t=1}^{T} (\hat{y}_t - y_t)^2 = \mathcal{O}\big(Y^2 n \left[n + \log(YL_0)\right]\big) = \mathcal{O}\left(\frac{Y^2}{\delta^4} \log^2(TYL_0)\right).$$

In particular, for fixed $\delta, Y, \kappa, ||\boldsymbol{C}||_2$, and $||\boldsymbol{B}||_2$, the bound is polylogarithmic in T, and the hidden dimension enters only logarithmically through $\kappa, ||\boldsymbol{C}||_2$, and $||\boldsymbol{B}||_2$.

Proof. If $n > T$, $|y_t| \le Y$ trivially gives $\mathcal{O}(TY^2) = \mathcal{O}(Y^2 n)$ cumulative loss. For $n \le T$, applying Lemma 1 and Theorem 2 with $\alpha_n = 2e^{-n\delta^2/\pi^2}$ and $G_n = 100^n$ gives $\alpha_n \le (2T^2L_0)^{-1}$, so $TL_0^2\alpha_n^2 \le 1/(4T^3) \le 1$. Since $\log(G_n) = \mathcal{O}(n)$ and $n = \mathcal{O}(\delta^{-2}\log(TYL_0))$, the bound follows. □

Remark 1 (Nonzero initial state, known horizon, and parameter estimates)**.** *(i)* ***Nonzero initial state:*** *Corollary 2 assumes $\boldsymbol{h}_0 = \boldsymbol{0}$. For $\boldsymbol{h}_0 \ne \boldsymbol{0}$, Eq.* (4) *acquires for $t \ge n+1$ the additional autonomous residual $\boldsymbol{C}p_n(\boldsymbol{A})\boldsymbol{A}^{t-n}\boldsymbol{h}_0$, bounded in magnitude by $||\boldsymbol{C}||_2\kappa||\boldsymbol{h}_0||_2\alpha_n$; thus replacing L_0 by $L_0 + \kappa_0 C_0||\boldsymbol{h}_0||_2$ controls all steps $t \ge n+1$ with the same exponential decay. (ii)* ***Horizon and norm upper bounds:*** *Setting n assumes a known horizon T (standard in online LDS prediction), but any polynomial upper bound $\widehat{T}$ satisfying $T \le \widehat{T} \le T^q$ for fixed $q \ge 1$ only decreases α_n while multiplying n by at most $O(q)$, preserving the $\mathcal{O}(Y^2\delta^{-4}\log^2(TYL_0))$ guarantee. Likewise, overestimating C_0, B_0, κ_0 affects the bound only logarithmically.*

4 Experiments

Data Generation and Robustness Framing. While Corollary 2 assumes noiseless dynamics, bounded inputs, and cumulative squared loss, our primary experiments are intentionally designed as a *robustness evaluation* testing deviations from the idealized theory—incorporating observation noise, persistent $\mathrm{AR}(1)$ inputs, large unnormalized output scales, and complex spectra. (Full reproducibility details, including the 2×2 real block construction of $\mathbf{A}$, eigenbasis condition numbers κ, seeds, explicit update equations, and harmonized scaling ablations, appear in Appendix C; we omit the ETTh1 benchmark from [19] because it evaluated deep neural networks, whereas VAW is specific to

linear regression.) We evaluate Universal Sequence Preconditioning (USP) on synthetic data generated from an LDS governed by Eq. (1) with additive observation noise $v_t \sim \mathcal{N}(0, \sigma^2)$ ($\sigma = 0.01$). To test the constant-angle regime, we specify the eigenvalues of $\mathbf{A}$ to have magnitude $1-10^{-7}$ and arguments bounded by $\tau = 0.1$ (under a Gaussian basis $\mathbf{P} \sim \mathcal{N}(0,1)^{d_{\text{hidden}} \times d_{\text{hidden}}}$, giving median eigenbasis condition numbers $\kappa = 226.9$ for $d_{\text{hidden}} = 100$ and $\kappa = 41.3$ for $d_{\text{hidden}} = 10$). The matrices $\mathbf{B}$ and $\mathbf{C}$ are sampled from $\mathcal{N}(0,1)$ (unnormalized in Figure 2, and normalized so $\|\mathbf{B}\|_2 = \|\mathbf{C}\|_2 = 1$ in our harmonized sweeps in Appendix C). We evaluate two primary configurations over $T = 5000$:

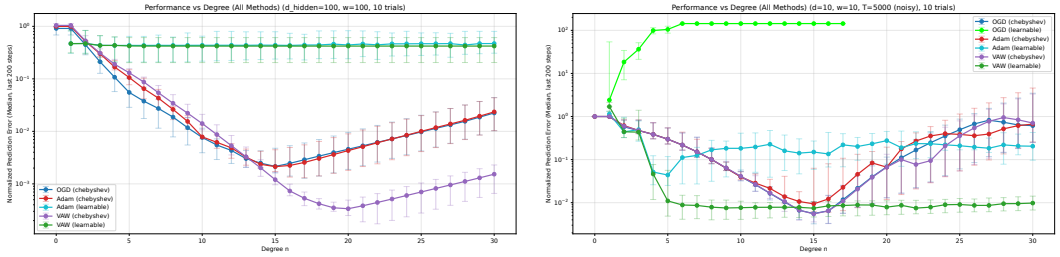
1. **High-dimensional**: $d_{\text{hidden}} = 100$ and baseline input window size $w = 100$.
2. **Low-dimensional**: $d_{\text{hidden}} = 10$ and baseline input window size $w = 10$, where w dictates the input history length for the fixed/unpreconditioned models.

Inputs u_t^{raw} are Rademacher signs in $\{\pm 1\}$ and scaled by $1/\log(5(t+1))$ to simulate a decaying excitation signal; Figure 2a drives the system with strongly correlated $\text{AR}(1)$ inputs $u_t = \rho u_{t-1} + \sqrt{1-\rho^2}\, u_t^{\text{raw}}$ ($\rho = 0.99$), while Figure 2b uses uncorrelated decaying inputs ($\rho = 0$), with all four $(d_{\text{hidden}}, \rho)$ combinations under $\|\mathbf{B}\|_2 = \|\mathbf{C}\|_2 = 1$ reported in Appendix C.

Performance Measurement. We compare Online Gradient Descent (OGD), Adam, and VAW via grid search over learning rates $\eta \in \{10^{-4}, 10^{-3}, 0.01, 0.05, 0.1\}$ (OGD, Adam) and regularization $\lambda \in \{10^{-2}, 10^{-1}, 1, 10, 100, 1000\}$ (VAW). Sweeping polynomial degree n from 0 to 30, we evaluate two polynomial configurations:

1. **Chebyshev**: Fixed coefficients $c_1, \ldots, c_n$ from monic Chebyshev polynomials, targeting the preconditioned signal $\tilde{y}_t = y_t + \sum_{i=1}^{n} c_i y_{t-i}$ using past inputs $[u_t, \ldots, u_{t-w}]^\top \in \mathbb{R}^{w+1}$ as features (note that $n = 0$ gives $\tilde{y}_t = y_t$, providing a direct ablation of VAW, OGD, and Adam on unpreconditioned finite-memory features).
2. **Learnable**: All n autoregressive and $w+1$ input coefficients learned online on raw features $[y_{t-1}, \ldots, y_{t-n}, u_t, \ldots, u_{t-w}]^\top \in \mathbb{R}^{n+w+1}$ for $n \geq 1$ (matching Algorithm 2 when $w = n-1$).

Performance is measured by the steady-state normalized prediction error $\frac{|y_t - \hat{y}_t|}{|y_t^{\text{clean}}| + \epsilon}$ ($\epsilon = 10^{-8}$, with online predictions clipped to $[-10^3, 10^3]$ during evaluation), averaged over the final 200 steps; we report the median and interquartile range (25th–75th percentiles) across 10 independent trials (Figure 2).



(a) Hidden dimension $d = 100$ ($w = 100, \rho = 0.99$). (b) Hidden dimension $d = 10$ ($w = 10, \rho = 0$).

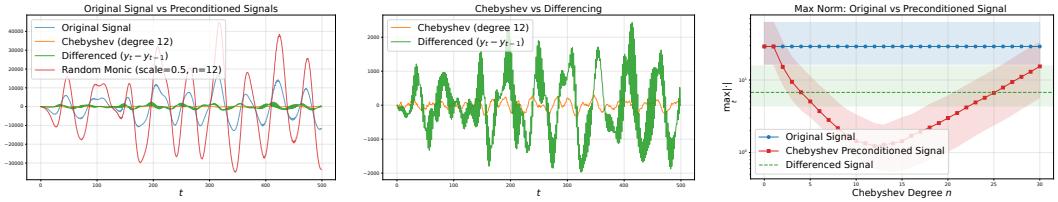
Figure 2: Normalized prediction error versus degree of preconditioning polynomial for (a) $d = 100$ and (b) $d = 10$. In (a), fixed Chebyshev VAW reaches 3.4×10^{-4} median error at $n = 20$ versus 1.04 at unpreconditioned $n = 0$; in (b), learnable ARX VAW avoids high-degree noise amplification, staying below 10^{-2} through $n = 30$.

Results Discussion and Harmonized Scaling. By the Cayley–Hamilton theorem, an exact degree-d_{hidden} predictor exists in the noiseless limit. In Figure 2a, when $d_{\text{hidden}} = 100$ with unnormalized $\mathbf{B}, \mathbf{C}$ (where raw $|y_t| \sim 10^4$), adaptively learning $n + w + 1$ ARX parameters under a fixed grid $\lambda \leq 10^3$ struggles to match fixed Chebyshev preconditioning of degree $n = \mathcal{O}(\log T)$, which compresses the target directly down to 3.4×10^{-4} median error at $n = 20$ (versus 1.04 at unpreconditioned $n = 0$, isolating the gain of USP over VAW alone). Moreover, VAW is substantially more stable than first-order methods (OGD and Adam) to the large preconditioning coefficients: past degree 15, OGD

and Adam degrade, whereas Chebyshev VAW continues improving up to degree 20. Conversely, in Figure 2b ($d_{\text{hidden}} = 10$), fixed Chebyshev methods amplify observation noise $\sigma \sum_{i=0}^{n} |c_i|$ at high degrees (0.71 at $n = 30$), whereas learnable ARX VAW achieves 7.5×10^{-3} at $n = 15$ and remains flat (9.8×10^{-3}) at $n = 30$. In Appendix C (Figures 4 and 5), we verify that once $\|\mathbf{B}\|_2 = \|\mathbf{C}\|_2 = 1$ is normalized so raw outputs match the λ grid, **learnable ARX VAW (Algorithm 2) outperforms all baselines across $n \in [10, 30]$ even at $d_{\text{hidden}} = 100$** (e.g., 0.0072 vs. 0.0225 for Chebyshev VAW and 0.0641 for learnable Adam at $n = 20, \rho = 0.99$).

4.1 Signal Norm and Preconditioning Dynamics

Why does our uniform theory apply VAW to raw ARX features rather than running ONS on the fixed Chebyshev-preconditioned signal $\tilde{y}_t$? Empirically, $\max_t |\tilde{y}_t|$ follows a *U-shape* in the polynomial degree n (Figure 3c): on a noiseless $d_{\text{hidden}} = 100$ system ($\tau = 0.3, \rho = 0.9$), moderate-degree Chebyshev preconditioning dramatically compresses the signal norm relative to raw, differenced ($y_t - y_{t-1}$), and random monic filters (Figures 3a and 3b), dropping the 200-trial median of $\max_t |\tilde{y}_t|$ from 28.96 (raw) and 6.75 (differenced) down to 1.34 at $n = 14$. Beyond this instance-dependent sweet spot, however, exponential Chebyshev coefficient growth dominates, raising the median norm to 15.39 at $n = 30$ (and $42.98 > 28.96$ by $n = 36$). We do not claim VAW is the *only* second-order method that can succeed: if one could establish a uniform theoretical bound on $|\tilde{y}_t|$ at moderate degrees across $W_{\pi-\delta}$, ONS could also yield strong guarantees without learning $c_1, \ldots, c_n$. Absent such a uniform signal-norm theorem, VAW on raw bounded signals ($|y_t| \le Y$) provides our worst-case guarantee because Y stays unamplified while the large comparator norm $\|\mathbf{x}^*\|_2$ enters only logarithmically. Figures 3a and 3b illustrate the raw signal alongside the degree-12 Chebyshev-preconditioned, random monic, and differenced signals, while Figure 3c plots the median and interquartile range of $\max_t |\tilde{y}_t|$ across 200 random trials.



(a) Raw versus preconditioned signals.

(b) Chebyshev preconditioning versus differencing.

(c) Maximum norm versus Chebyshev degree.

Figure 3: Preconditioning empirically reduces the signal norm at moderate degrees before growing exponentially at higher degrees (following a U-shape in (c) across 200 trials). If moderate-degree norm control is provable uniformly, learning the autoregressive coefficients in Algorithm 2 would be unnecessary.

5 Conclusion

Summary We addressed a fundamental tension in sequence prediction for marginally stable linear dynamical systems: the trade-off between effective memory compression and the exponential explosion of parameter magnitudes. By interpreting Universal Sequence Preconditioning (USP) through the lens of the Minimum Description Length (MDL) principle, we established that second-order online learning methods—specifically the Vovk-Azoury-Warmuth (VAW) forecaster—are uniquely equipped to absorb the parameter blowup induced by polynomial preconditioning. Our theoretical framework yields near-optimal guarantees. By shifting the learning burden from the parameter norm to the effective dimension, we exponentially improved the cumulative prediction regret from the $\tilde{\mathcal{O}}(T^{11/13})$ bounds of prior first-order approaches to $\mathcal{O}(\text{polylog}(T))$, completely eliminating polynomial dependencies on both the time horizon and the hidden state dimension. Furthermore, we introduce Faber preconditioning to extend the theoretical viability of USP to a strictly broader class of dynamical systems with constant complex arguments, namely spectra in $W_{\pi-\delta}$ for constant $\delta > 0$. Empirically, our evaluations robustly corroborate these theoretical insights. The Chebyshev-preconditioned VAW algorithm consistently outperforms both first-order methods

and adaptive-polynomial approaches, particularly in adverse regimes characterized by large time horizons, high-dimensional hidden states, and highly correlated input sequences.

Limitations Our theoretically analyzed VAW algorithm requires learning the autoregressive coefficients. This is a slight departure from the spirit of USP since an offline transformation to the target sequence makes it easier to learn whereas in our case we must learn this transformation. However, it is also deeply connected to USP since its proof heavily relies on construction of the benchmark $\mathbf{x}^*$, which is the polynomial preconditioned solution. Moreover, empirical evidence suggests this requirement may be an artifact of our proof technique. Another limitation of this work is the practicality of the VAW algorithm. It requires a much larger memory footprint than the first order methods. The memory footprint grows poly-logarithmically with horizon length and as $\tilde{O}(\delta^{-2})$ in the Faber wedge parameter.

Future Directions Promising directions for future work include extending these second-order preconditioning frameworks to non-linear dynamical systems, perhaps by using the recent breakthrough of [6]. Another important direction is to sharpen the dependence on the angular gap δ, since the current Faber proof gives a δ^{-4} loss dependence while the cyclic-permutation lower bound only forces memory on the order of $1/\delta$. Ultimately, our results demonstrate that when appropriately paired with second-order optimization, universal preconditioning unlocks fundamentally new capabilities for scalable, highly accurate sequence prediction.

References

[1] Katy S Azoury and Manfred K Warmuth. Relative loss bounds for on-line density estimation with the exponential family of distributions. *Machine Learning*, 43(3):211–246, 2001.

[2] Ainesh Bakshi, Allen Liu, Ankur Moitra, and Morris Yau. A new approach to learning linear dynamical systems. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing*, pages 335–348, 2023.

[3] Nicolò Cesa-Bianchi and Gábor Lugosi. *Prediction, learning, and games*. Cambridge university press, 2006.

[4] Chi-Tsong Chen. *Linear System Theory and Design*. Oxford University Press, 3rd edition, 1999.

[5] John P Coleman and Russell A Smith. The faber polynomials for circular sectors. *mathematics of computation*, 49(179):231–241, 1987.

[6] Evan Dogariu, Anand Brahmbhatt, and Elad Hazan. Universal learning of nonlinear dynamics. *arXiv preprint arXiv:2508.11990*, 2025.

[7] Salar Fattahi. Learning partially observed linear dynamical systems from logarithmic number of samples. In *Proceedings of the 3rd Conference on Learning for Dynamics and Control*, volume 144 of *Proceedings of Machine Learning Research*, pages 60–72, 2021.

[8] Karin Gatermann, Christoph Hoffmann, and Gerhard Opfer. Explicit faber polynomials on circular sectors. *Mathematics of computation*, 58(197):241–253, 1992.

[9] Udaya Ghai, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. No-regret prediction in marginally stable systems. In *Proceedings of the 33rd Conference on Learning Theory*, volume 125 of *Proceedings of Machine Learning Research*, pages 1714–1757, 2020.

[10] Peter D Grünwald. *The minimum description length principle*. MIT press, 2007.

[11] Moritz Hardt, Tengyu Ma, and Benjamin Recht. Gradient descent learns linear dynamical systems. *Journal of Machine Learning Research*, 19(29):1–44, 2018.

[12] Elad Hazan. *Introduction to online convex optimization*, volume 2. Now Publishers, Inc., 2016.

[13] Elad Hazan, Amit Agarwal, and Satyen Kale. Logarithmic regret algorithms for online convex optimization. *Machine Learning*, 69(2-3):169–192, 2007.

[14] Elad Hazan, Holden Lee, Karan Singh, Cyril Zhang, and Yi Zhang. Spectral filtering for general linear dynamical systems. In *Advances in Neural Information Processing Systems 31*, pages 4639–4648, 2018.

[15] Elad Hazan and Annie Marsden. Spectral filtering for complex linear dynamical systems. *arXiv preprint arXiv:2601.22400*, 2026.

[16] Elad Hazan, Karan Singh, and Cyril Zhang. Learning linear dynamical systems via spectral filtering. In *Advances in Neural Information Processing Systems 30*, pages 6702–6712, 2017.

[17] Thomas Kailath. *Linear Systems*. Prentice-Hall, 1980.

[18] Holden Lee. Improved rates for prediction and identification of partially observed linear dynamical systems. In *Proceedings of the 33rd International Conference on Algorithmic Learning Theory*, volume 167 of *Proceedings of Machine Learning Research*, pages 668–698, 2022.

[19] Annie Marsden and Elad Hazan. Dimension-free regret for learning asymmetric linear dynamical systems. *arXiv preprint arXiv:2502.06545*, 2025.

[20] Annie Marsden and Elad Hazan. Universal sequence preconditioning. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025.

[21] Paria Rashidinejad, Jiantao Jiao, and Stuart J. Russell. Slip: Learning to predict in unknown dynamical systems with long-term memory. In *Advances in Neural Information Processing Systems 33*, 2020.

[22] Jorma Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

[23] Max Simchowitz, Ross Boczar, and Benjamin Recht. Learning linear dynamical systems with semi-parametric least squares. In *Proceedings of the 32nd Conference on Learning Theory*, volume 99 of *Proceedings of Machine Learning Research*, pages 2714–2802, 2019.

[24] Anastasios Tsiamis and George J. Pappas. Finite sample analysis of stochastic system identification. In *2019 IEEE 58th Conference on Decision and Control (CDC)*, pages 3648–3654. IEEE, 2019.

[25] Anastasios Tsiamis and George J. Pappas. Online learning of the kalman filter with logarithmic regret. *IEEE Transactions on Automatic Control*, 68(5):2774–2789, 2023.

[26] Volodya Vovk. Competitive on-line statistics. *International Statistical Review*, 69(2):213–248, 2001.

NeurIPS Paper Checklist

1. **Claims**

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

   Answer: [Yes]

   Justification: The claims are rigorously proven and experiments are included in support of the claims.

   Guidelines:

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

2. **Limitations**

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

   Answer: [Yes]

   Justification: We discuss limitations in the conclusion regarding artifacts of our proof and the practicality of the proposed algorithm.

   Guidelines:

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

3. **Theory assumptions and proofs**

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

Answer: [Yes]

Justification: Assumptions are clear, theorems and lemmas are referenced.

Guidelines:

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

4. **Experimental result reproducibility**

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

   Answer: [Yes]

   Justification: The experimental section includes the details to replicate the synthetic data and the algorithms used are clear.

   Guidelines:

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

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

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

Answer: [No]

Justification: It is difficult to get approval to release code, however we include all the necessary details to replicate the experiments.

Guidelines:

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

6. **Experimental setting/details**

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

   Answer: [Yes]

   Justification: We specify the algorithms (VAW, OGD, Adam, etc) and the learning rate sweeps we do.

   Guidelines:

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

7. **Experiment statistical significance**

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

   Answer: [Yes]

   Justification: We make clear that our plots are the median with the 25th–75th percentiles represented by the error bars.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
   - The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).

- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

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

   Answer: [No]

   Justification: These are very small synthetic data experiments and can be run with small computational resources.

   Guidelines:

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

9. **Code of ethics**

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

   Answer: [Yes]

   Justification: This paper does not have any obvious impacts on privacy, fairness, safety, discrimination, etc.

   Guidelines:

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

10. **Broader impacts**

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

    Answer: [N/A]

    Justification: This paper doesn't have any clear societal impact.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.

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

11. **Safeguards**

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

    Answer: [N/A]

    Justification: This is a theory paper.

    Guidelines:

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

12. **Licenses for existing assets**

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

    Answer: [N/A]

    Justification: No assets.

    Guidelines:

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

- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

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

    Answer: [N/A]

    Justification: No assets introduced.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: No crowdsourcing or human subjects involved.

    Guidelines:

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

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

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

    Answer: [N/A]

    Justification: No crowdsourcing or human subjects involved.

    Guidelines:

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

16. **Declaration of LLM usage**

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

Answer: [N/A]

Justification: This research does not involve LLMs as any important, original, or non-standard components.

Guidelines:

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

A Coefficient growth of extremal polynomials

We emphasize that this appendix is not used in the proofs of Theorem 2, Corollary 1, or Corollary 2; its purpose is to provide self-contained quantitative motivation for Conjecture 1 with fully explicit constants. The following theorem shows that any monic polynomial that ensures all values in $[-1, 1]$ are exponentially small at the Chebyshev scale must have exponentially large coefficients.

Theorem 3 (Coefficient growth for flat monic polynomials)**.** *Fix $M > 0$ and set $\alpha := \max\{3, 1 + 16M\}$. Let $p(x) = \sum_{j=0}^{n} a_j x^j$ ($a_n = 1$) be a monic polynomial of degree $n \geq 1$. If*

$$\max_{x\in[-1,1]} |p(x)| \leq M 2^{1-n},$$

then for every $n \geq 1$ (i.e. with threshold $n_0(M) = 1$),

$$\max_{0\leq j\leq n} |a_j| \geq \frac{1}{2(n+1)} \left(1 + \frac{1}{\alpha^2 - 1}\right)^n.$$

In particular, at least one coefficient of p is exponentially large in n.

Proof. We split the proof into two parts. The first part is the conceptual reduction, and the second part is the technical no-cancellation estimate.

Part I: Reduction to a no-cancellation statement. Write p in the Chebyshev basis:

$$p(x) = \sum_{k=0}^{n} b_k T_k(x),$$

where T_k is the kth Chebyshev polynomial. Since $T_n(x) = 2^{n-1}x^n +$ lower-degree terms and p is monic, the coefficient of T_n must be $b_n = 2^{1-n}$. Using $T_k(\cos\theta) = \cos(k\theta)$ and orthogonality on $[0, \pi]$, for $0 \leq k \leq n - 1$,

$$\begin{aligned} |b_k| &= \left|\tfrac{2}{\pi} \textstyle\int_0^\pi p(\cos\theta)\cos(k\theta)\, d\theta\right| \\ &\leq \tfrac{2}{\pi} \textstyle\int_0^\pi |p(\cos\theta)|\,|\cos(k\theta)|\, d\theta \\ &\leq 2\max_{x\in[-1,1]} |p(x)| \leq 2M2^{1-n} = 4M2^{-n}. \end{aligned}$$

Thus, after scaling by 2^{n-1}, the polynomial $q(x) := 2^{n-1}p(x)$ has the form

$$q(x) = T_n(x) + \sum_{k=0}^{n-1} c_k T_k(x), \qquad |c_k| \leq 2M.$$

Can the bounded lower-degree Chebyshev terms cancel the coefficient growth of T_n? The answer is no, as shown with explicit constants next.

Lemma 2 (No cancellation, explicit form)**.** *Fix $M > 0$ and $\alpha := \max\{3, 1 + 16M\}$. Suppose*

$$q(x) = T_n(x) + \sum_{k=0}^{n-1} c_k T_k(x) = \sum_{j=0}^{n} d_j x^j, \qquad |c_k| \leq 2M.$$

Then for every $n \geq 1$,

$$\max_{0\leq j\leq n} |d_j| \geq \frac{1}{4(n+1)} \left(2 + \frac{2}{\alpha^2 - 1}\right)^n.$$

Proof. Since $\alpha \geq 3$, setting

$$y := \frac{\alpha - \alpha^{-1}}{2}$$

satisfies $y \geq \frac{4}{3} > 1$. For a polynomial $r(x) = \sum_j r_j x^j$, define $\|r\|_y := \sum_j |r_j| y^j$. Because the nonzero coefficients of T_k alternate in sign and $iy = \frac{(i\alpha)+(i\alpha)^{-1}}{2}$, the Chebyshev identity yields the exact formula

$$\|T_k\|_y = |T_k(iy)| = \frac{\alpha^k + (-1)^k \alpha^{-k}}{2}.$$

Since $\alpha \geq 3$ and $n \geq 1$, $\alpha^{-n} \leq \frac{1}{9}\alpha^n$, giving $\|T_n\|_y \geq \frac{4}{9}\alpha^n > \frac{3}{8}\alpha^n$, while $\|T_k\|_y \leq \alpha^k$ for all $k \geq 0$. Therefore, using $\sum_{k=0}^{n-1} \alpha^k < \frac{\alpha^n}{\alpha-1}$ and $\alpha - 1 \geq 16M$,

$$\|q\|_y \geq \|T_n\|_y - \sum_{k=0}^{n-1} |c_k| \|T_k\|_y \geq \left(\frac{3}{8} - \frac{2M}{\alpha - 1}\right) \alpha^n \geq \frac{1}{4}\alpha^n.$$

On the other hand, setting $D := \max_{0 \leq j \leq n} |d_j|$ and using $y > 1$, $\|q\|_y = \sum_{j=0}^{n} |d_j| y^j \leq D(n+1)y^n$. Hence

$$D \geq \frac{1}{4(n+1)} \left(\frac{\alpha}{y}\right)^n = \frac{1}{4(n+1)} \left(\frac{2\alpha^2}{\alpha^2 - 1}\right)^n = \frac{1}{4(n+1)} \left(2 + \frac{2}{\alpha^2 - 1}\right)^n,$$

proving the lemma. Since $a_j = 2^{1-n} d_j$, $\max_{0 \leq j \leq n} |a_j| = 2^{1-n} D \geq \frac{1}{2(n+1)}(1 + \frac{1}{\alpha^2-1})^n$. □

□

A constant-rate conjecture. The preceding theorem proves exponential coefficient growth at the sharp Chebyshev scale 2^{-n}. Its proof uses this scale essentially: after normalizing by the leading Chebyshev coefficient, the lower-degree Chebyshev coefficients remain uniformly bounded. For a weaker bound such as $\exp(-cn)$ with $c < \log 2$, the same normalization allows the lower-degree Chebyshev coefficients to be exponentially large, so the no-cancellation argument no longer applies. We nevertheless expect the qualitative obstruction to persist.

Conjecture 2 (Coefficient growth at any exponential rate)**.** *For every $c > 0$ there exists $c' > 0$ such that the following holds for all sufficiently large n. If $p_n(x) = x^n + \sum_{j=0}^{n-1} a_j x^j$ is a monic polynomial satisfying $\max_{x \in [-1,1]} |p_n(x)| \leq e^{-cn}$, then $\max_{0 \leq j \leq n} |a_j| \geq e^{c'n}$.*

B Proof of Lemma 1

Recall that Lemma 1 asserts that for every $\delta \in (0, \pi)$ and $n \geq 1$, there is a real monic polynomial

$$p_{n,\delta}(z) = z^n + \sum_{i=1}^{n} c_i z^{n-i}$$

with $c_0 := 1$ such that

$$\sup_{z \in W_{\pi-\delta}} |p_{n,\delta}(z)| \leq 2e^{-n\delta^2/\pi^2}, \qquad \max_{0 \leq i \leq n} |c_i| \leq 100^n.$$

Proof. **Part I: Exponentially small norm on the wedge.** For $\theta \in (0, \pi)$, define

$$W_\theta = \{z \in \mathbb{C} : |z| \leq 1,\ |\arg(z)| \leq \theta\}.$$

We use the following sector-specific Faber estimate for circular sectors, due to Coleman and Smith [5] and stated in the convenient normalization of Gatermann, Hoffmann, and Opfer [8]. Let $F_{n,\theta}$ be the monic degree-n Faber polynomial of W_θ, and let $\rho(\theta)$ be the transfinite diameter of W_θ. Coleman and Smith compute

$$\rho(\theta) = t^{-t}(2-t)^{t-2}, \qquad t = \frac{\theta}{\pi}.$$

In the notation of [8, Eq. (2.6)], the scaled Faber polynomial is $\widetilde{F}_{n,\theta}(z) = \rho(\theta)^{-n} F_{n,\theta}(z)$. Moreover, [8, Eq. (3.10)] gives $\sup_{z \in W_\theta} |\widetilde{F}_{n,\theta}(z)| \leq 2$. Consequently, $\sup_{z \in W_\theta} |F_{n,\theta}(z)| \leq 2\rho(\theta)^n$.

Set $\theta = \pi - \delta$ and $\eta = \frac{\delta}{\pi}$. Then $t = \frac{\theta}{\pi} = 1 - \eta$, and therefore $\rho(\pi - \delta) = (1-\eta)^{-(1-\eta)}(1+\eta)^{-(1+\eta)}$. Thus, for $0 \leq \eta < 1$,

$$\log\left(\frac{1}{\rho(\pi - \delta)}\right) = (1+\eta)\log(1+\eta) + (1-\eta)\log(1-\eta) = \sum_{k=1}^{\infty} \frac{2\eta^{2k}}{(2k)(2k-1)} \geq \eta^2.$$

Hence $\rho(\pi-\delta) \le e^{-\eta^2} = e^{-\delta^2/\pi^2}$, and consequently $\sup_{z\in W_{\pi-\delta}} |F_{n,\pi-\delta}(z)| \le 2e^{-n\delta^2/\pi^2}$. Finally, $W_{\pi-\delta}$ is invariant under complex conjugation. Therefore, replacing $F_{n,\pi-\delta}$ by

$$p_{n,\delta}(z) = \frac{F_{n,\pi-\delta}(z) + \overline{F_{n,\pi-\delta}(\overline{z})}}{2}$$

preserves monicity, gives real coefficients, and satisfies $\sup_{z\in W_{\pi-\delta}} |p_{n,\delta}(z)| \le 2e^{-n\delta^2/\pi^2}$.

Part II: Coefficient bound. Write $p_{n,\delta}(z) = z^n + \sum_{i=1}^n c_i z^{n-i}$ with $c_0 = 1$. Since $[0,1] \subseteq W_{\pi-\delta}$, Part I gives $\sup_{x\in[0,1]} |p_{n,\delta}(x)| \le 2$. Define $q(x) = p_{n,\delta}\big(\frac{x+1}{2}\big)$, so $\sup_{x\in[-1,1]} |q(x)| \le 2$. Writing $q(x) = \sum_{k=0}^n b_k T_k(x)$ in the Chebyshev basis, orthogonality gives

$$b_0 = \frac{1}{\pi}\int_0^\pi q(\cos\theta)\, d\theta, \qquad b_k = \frac{2}{\pi}\int_0^\pi q(\cos\theta)\cos(k\theta)\, d\theta \quad (k \ge 1),$$

so $|b_k| \le 4$ for $0 \le k \le n$. The recurrence $T_{k+1}(x) = 2xT_k(x) - T_{k-1}(x)$ implies $\|T_k\|_1 \le 3^k$, hence $\|q\|_1 \le \sum_{k=0}^n |b_k|\,\|T_k\|_1 \le 4(n+1)3^n \le 24^n$. Finally, since $p_{n,\delta}(z) = q(2z-1)$ and $\|(2z-1)^j\|_1 = 3^j \le 3^n$ for all $j \le n$,

$$\max_{0\le i\le n} |c_i| \le \sum_{i=0}^n |c_i| = \|p_{n,\delta}\|_1 \le 3^n \|q\|_1 \le 72^n \le 100^n.$$

Combining Parts I and II proves the lemma. □

C Experimental Reproducibility and Supplementary Evaluations

Real-Valued Construction of Asymmetric A, Conditioning κ, and Seeds. For a real LDS, nonreal eigenvalues occur in complex-conjugate pairs $re^{\pm i\theta}$ with $r = 1 - 10^{-7}$ and $\theta \sim \mathrm{Unif}[-\tau, \tau]$ (with random sign ± 1 on the real part in Section 4, $\tau = 0.1$ in Figure 2, and $\tau = 0.3$ in Figure 3). Each conjugate pair is embedded along the block-diagonal of a real modal matrix $\mathbf{D} \in \mathbb{R}^{d_{\text{hidden}} \times d_{\text{hidden}}}$ via the 2×2 rotation block

$$\mathbf{J}(r,\theta) = r\begin{bmatrix} \pm\cos\theta & \sin|\theta| \\ -\sin|\theta| & \pm\cos\theta \end{bmatrix},$$

with any remaining odd coordinate assigned $\pm r$. Because each 2×2 rotation block is unitarily diagonalizable over $\mathbb{C}^2$, forming $\mathbf{A} = \mathbf{P}\mathbf{D}\mathbf{P}^{-1}$ from an i.i.d. Gaussian basis $\mathbf{P} \sim \mathcal{N}(0,1)^{d_{\text{hidden}} \times d_{\text{hidden}}}$ yields exact complex eigenbasis condition number $\kappa(\mathbf{V}) = \|\mathbf{V}\|_2\|\mathbf{V}^{-1}\|_2 = \|\mathbf{P}\|_2\|\mathbf{P}^{-1}\|_2$. Across the deterministic random seeds used in our evaluations, the empirical condition numbers are:

- $d_{\text{hidden}} = 10$ (10 trials, seeds 42–51): median $\kappa = 41.3$, IQR $[16.6, 67.1]$, range $[11.1, 299.1]$.
- $d_{\text{hidden}} = 100$ (10 trials, seeds 42–51): median $\kappa = 226.9$, IQR $[110.8, 3064.1]$, range $[94.6, 16741.9]$.
- $d_{\text{hidden}} = 100$ (200 trials, seeds 1000–1199 in Figure 3c): median $\kappa = 331.5$, IQR $[186.8, 688.5]$ (single-trajectory Figures 3a–3b use seed 42, $T = 500$, and random monic coefficients $c_1, \ldots, c_{12} \sim \mathcal{N}(0, 0.5^2)$).

Explicit Update Equations and Hyperparameter Selection Protocol. At each step $t = 1, \ldots, T$ (with zero pre-history $y_{\le 0} = 0, u_{\le 0} = 0$):

- **Learnable (ARX) configuration:** With raw feature vector $\mathbf{a}_t = [y_{t-1}, \ldots, y_{t-n}, u_t, \ldots, u_{t-w}]^\top \in \mathbb{R}^{n+w+1}$ ($w = n-1$ in Algorithm 2) and unclipped iterate prediction $\tilde{\hat{y}}_t = \mathbf{x}_t^\top \mathbf{a}_t$, OGD updates $\mathbf{x}_{t+1} = \mathbf{x}_t - \eta(\tilde{\hat{y}}_t - y_t)\mathbf{a}_t$, Adam applies standard bias-corrected moment updates ($\beta_1 = 0.9, \beta_2 = 0.999, \epsilon_{\text{adam}} = 10^{-8}$) to $\mathbf{g}_t = (\tilde{\hat{y}}_t - y_t)\mathbf{a}_t$, and VAW updates $\mathbf{S}_t = \lambda\mathbf{I} + \sum_{s=1}^t \mathbf{a}_s\mathbf{a}_s^\top$, $\mathbf{v}_{t-1} = \sum_{s=1}^{t-1} y_s\mathbf{a}_s$, predicting $\hat{y}_t^{\text{VAW}} = \mathbf{a}_t^\top \mathbf{S}_t^{-1}\mathbf{v}_{t-1}$.

- **Fixed Chebyshev configuration:** Using monic degree-n Chebyshev coefficients $c_1, \ldots, c_n$, the preconditioned target $\tilde{y}_t = y_t + \sum_{i=1}^{n} c_i y_{t-i}$ is predicted from the input window $\mathbf{u}_t = [u_t, \ldots, u_{t-w}]^\top \in \mathbb{R}^{w+1}$; since $\hat{y}_t = \hat{\tilde{y}}_t - \sum_{i=1}^{n} c_i y_{t-i}$, the prediction error satisfies $\hat{y}_t - y_t = \hat{\tilde{y}}_t - \tilde{y}_t$ identically (with $n = 0$ giving $\tilde{y}_t = y_t$, the unpreconditioned baseline).

Reported predictions are clipped to $[-10^3, 10^3]$ at evaluation time, and for each trial, method, and setting, the best finite hyperparameter on the shared grid ($\eta \in \{10^{-4}, 10^{-3}, 0.01, 0.05, 0.1\}$; $\lambda \in \{10^{-2}, 10^{-1}, 1, 10, 100, 1000\}$) is selected.

Harmonized $\|\mathbf{B}\|_2 = \|\mathbf{C}\|_2 = 1$ Evaluations Across Persistent $\mathrm{AR}(1)$ and Decaying Inputs. In Figure 2a, unnormalized $\mathbf{B}, \mathbf{C} \sim \mathcal{N}(0, 1)$ with $\rho = 0.99$ produced raw signal amplitudes $|y_t| \sim 10^4$, causing raw ARX Gram entries $y_{t-i}^2 \sim 10^8$ to overwhelm the fixed $\lambda \leq 10^3$ grid during burn-in. Figures 4 and 5 evaluate harmonized unit-norm matrices $\|\mathbf{B}\|_2 = \|\mathbf{C}\|_2 = 1$ ($\sigma = 0.01, T = 5000$) across both $d_{\text{hidden}} \in \{100, 10\}$ for strongly correlated $\mathrm{AR}(1)$ inputs ($\rho = 0.99$) and uncorrelated decaying inputs ($\rho = 0$). Across all four regimes, **learnable ARX VAW (Algorithm 2) achieves the lowest steady-state error for all** $n \in [10, 30]$: at $n = 20$, VAW (learnable) achieves median normalized error 0.0072 ($d = 100, \rho = 0.99$), 0.0052 ($d = 10, \rho = 0.99$), 0.0257 ($d = 100, \rho = 0$), and 0.0082 ($d = 10, \rho = 0$), compared to 0.0225, 0.0102, 0.0748, and 0.0843 for VAW (chebyshev) (which suffers from high-degree observation-noise amplification $\sigma \sum_{i=0}^{n} |c_i|$) and 0.0641, 0.0874, 0.2164, and 0.1414 for Adam (learnable).

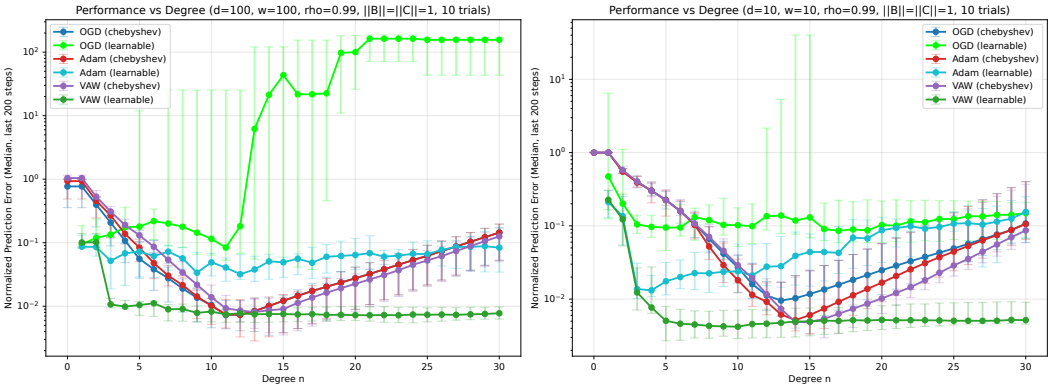


Figure 4: Harmonized $\|\mathbf{B}\|_2 = \|\mathbf{C}\|_2 = 1$ evaluation under persistent $\mathrm{AR}(1)$ inputs ($\rho = 0.99, \sigma = 0.01, T = 5000, 10$ trials) for $d_{\text{hidden}} = 100$ (left) and $d_{\text{hidden}} = 10$ (right).

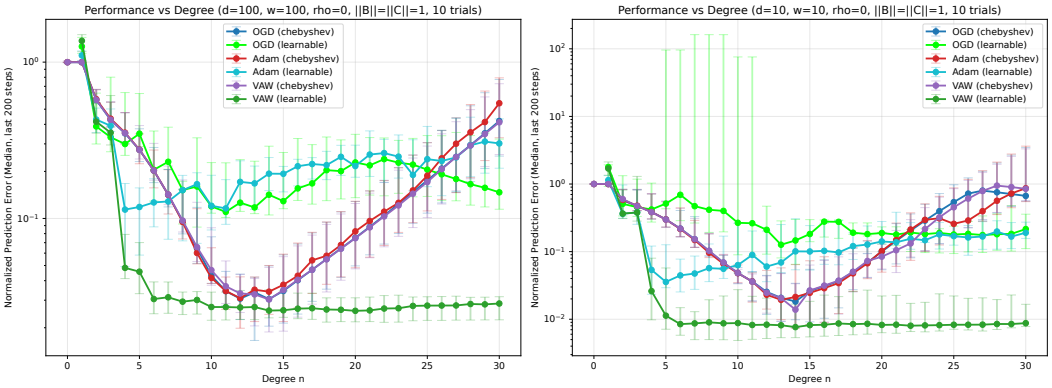


Figure 5: Harmonized $\|\mathbf{B}\|_2 = \|\mathbf{C}\|_2 = 1$ evaluation under uncorrelated decaying inputs ($\rho = 0, \sigma = 0.01, T = 5000, 10$ trials) for $d_{\text{hidden}} = 100$ (left) and $d_{\text{hidden}} = 10$ (right).